Functional Enrichment Analysis of Transcriptomics Data of Breast Cancer RNA-seq

Babatunde Bello

Follow this and additional works at: https://scholarworks.gsu.edu/cs_theses

Recommended Citation
doi: https://doi.org/10.57709/28959410
Functional Enrichment Analysis of Transcriptomics Data of Breast Cancer RNA-seq

by

Babatunde L Bello

Under the Direction of Murray Patterson, PhD

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2022
ABSTRACT

The biochemical mechanism driving cancer metastasis and primary cancer invasion of new sites are still unclear as the process can be complex. The mutations in somatic cells often include mutation drivers and some passenger mutations. In this study, we have analyzed RNA-Seq datasets from primary breast cancer and metastatic lung cancer for differentially expressed gene lists to gain insight into transcriptomic profiles of the two conditions. The gene lists are analyzed for pathway and functional enrichment annotations. It is interesting to note that the top enriched pathways are major ones involving some connected cancer-related signaling processes. The enriched gene sets from this analysis includes ones connected to cancer proliferation, progression, and metastatic invasions. The pathways and genes show some overlapping networks and connections that may be key to finding potential mutation driver genes.

INDEX WORDS: Functional, Annotation, Enrichment, Transcriptomics, Cancer
Functional Enrichment Analysis of Transcriptomics Data of Breast Cancer RNA-seq

by

Babatunde L. Bello

Committee Chair: Murray Patterson

Committee: Alex Zelikovsky

Pavel Skums

Electronic Version Approved:

Office of Graduate Services
College of Arts and Sciences
Georgia State University
May 2022
DEDICATION

To moimi and papami, my darling siblings and friends turned families.
ACKNOWLEDGEMENTS

I would like to thank everyone who played a key role during my state and experience here at Georgia State University - the Chemistry and Computer science departments. I want to thank my advisors Dr. Murray Patterson and Dr. Alex Zelikovsky, for their guidance and training throughout my research work in the department. I also want to acknowledge Dr. Ritu Aneja and her collaboration group for making the data available for this work and support. I also want to appreciate my committee member, Dr. Pavel Skums, for his help, academic advice, and comments on this thesis work. I want to appreciate Dr. Ion Mandoiu for his advice on guides on this work as well. I must appreciate my research colleagues Fil, Jaspreet, Sarwan, Prakash, Sai, Armel, Bikram, Sara, Ria, Akshay, and others not mentioned. I must say be thanks to my family and friends - AuntyB, Uncle Segun, Moimi, Papami, ‘daddy Pastor’, Temitope, ‘bro Gbenga’, Olawale, Otunba Taye, Tomi, Feran, Folu, Edwin, Anu, KB Debayo, Abbey, Jide, KennyBrown Tolu and some ‘friends-not-named’ - some for moral support, some for their prayers, some for patience and understanding, some for sharing the ups and downs of the flow together! And, thanks to God for His Grace and capacity to see this to this end.
### TABLE OF CONTENTS

**ACKNOWLEDGEMENTS** ........................................................................................................... V

**LIST OF TABLES** ....................................................................................................................... VIII

**LIST OF FIGURES** .................................................................................................................... IX

1 **INTRODUCTION** ........................................................................................................................................ 1

1.1 Transcriptomic Profile Pathway Analysis of RNA Sequence ................................................ 1

1.2 RNA-Seq Technique Analysis and Applications ................................................................. 2

1.3 Cancer RNA-Seq Data Analysis and Functional Annotation ........................................... 3

1.4 Aims of Analysis .................................................................................................................... 4

2 **CONCEPTS AND BACKGROUNDS** ...................................................................................... 5

2.1 Differential Gene Expression ............................................................................................................... 5

2.2 IsoEM2 and IsoDE2 Package for Differential Expression Analysis .................................... 5

2.3 Pathway Analysis for DE Gene Lists ........................................................................................... 6

2.4 Functional Enrichment Analysis and Annotation of DE Gene List ................................. 6

2.4.1 *Overrepresentation Enrichment Analysis (ORA)* ......................................................... 7

2.4.2 *Gene Set Enrichment Analysis (GSEA)* ............................................................................. 7

2.5 Other Tools and Concept for Analysis ....................................................................................... 8

2.5.1 *GATK-HaplotypeCaller* .................................................................................................... 8

2.5.2 *Provean Annotation Tool* ................................................................................................... 9

3 **DATASETS AND METHODS** .................................................................................................... 10
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Datasets for Analysis ..................................................................</td>
<td>10</td>
</tr>
<tr>
<td>3.2</td>
<td>IsoEM2 and IsoDE2 ........................................................................</td>
<td>10</td>
</tr>
<tr>
<td>3.3</td>
<td>KOBAS Web-based Pathway Annotation ............................................</td>
<td>11</td>
</tr>
<tr>
<td>3.4</td>
<td>ClusterProfiler Package ..................................................................</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>RESULTS AND DISCUSSION ..................................................................</td>
<td>12</td>
</tr>
<tr>
<td>4.1</td>
<td>Results for IsoDE Analysis and Annotation ....................................</td>
<td>12</td>
</tr>
<tr>
<td>4.2</td>
<td>Provean Analysis Result Sample ..................................................</td>
<td>14</td>
</tr>
<tr>
<td>4.3</td>
<td>Enrichment Analysis for Gene List Analysis Results .......................</td>
<td>14</td>
</tr>
<tr>
<td>4.3.1</td>
<td><em>Overrepresentation Enrichment Analysis (ORA) Results</em> ...................</td>
<td>14</td>
</tr>
<tr>
<td>4.3.2</td>
<td><em>Gene Enrichment Analysis (GSEA) Results</em> ...................................</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>CONCLUSION ......................................................................................</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>REFERENCES .....................................................................................</td>
<td>23</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 2.1 GATK Table Sample ................................................................. 9
Table 2.2 A Sample of Provean Analysis Output .............................................. 9
Table 4.1 The DE Gene Lists for Primary Result Sample ................................ 12
Table 4.2 The DE Gene Lists for Metastasis Result Sample .............................. 13
Table 4.3 The KOBAS Annotation Result Sample .......................................... 13
Table 4.4 The Provean Annotation Output Sample ............................................. 14
LIST OF FIGURES

Figure 3.1 Chart for Samples Used for Analysis ................................................................. 10
Figure 4.1 Dot Plot Chart for ORA for NO_02 Samples ...................................................... 15
Figure 4.2 UpSet Plot for ORA for NO_02 Samples ............................................................ 17
Figure 4.3 UpSet Plot for ORA for NO_07 samples ........................................................... 17
Figure 4.4 Enrichment Map for NO_02 Samples ............................................................... 18
Figure 4.5 Gene Concept Newtork (C-netplot) for NO_02 Samples ................................. 20
Figure 4.6 Enrichment Map for NO_07 Samples ............................................................... 20
Figure 4.7 Gene Concept Network (C-netplot) for NO_07 Samples ................................. 21
1  INTRODUCTION

1.1  Transcriptomic Profile Pathway Analysis of RNA Sequence

Transcriptomic profiling is a way of studying gene expression by analyzing the entire set of RNA transcripts in a cell population at a given time or condition. RNA sequence (RNA-Seq) analysis has become one major approach for accessing and quantifying changes in cellular transcripts by identifying genes and pathways that are either upregulated or downregulated in cells. The cellular transcriptome is a term used to describe a complete set of RNA transcripts in a cell or cell population [1].

Genes in the genomes of cells are transcribed into mRNA molecules that are either translated into proteins or functions are tRNA or rRNA. This way, the encoded genetic information - Adenine (A), Guanine (G), Thymine (T) and Cytosine (C.) - DNA sequences in genes are copied into corresponding mRNA transcripts (with C replaced with U in transcripts). The different forms of RNA molecules - mRNA, tRNA, rRNA, etc. – that are transcribed from genes can be quantified to estimate the expression levels of gene transcripts in a cell or tissue at a particular condition. This way cells dynamically express and translate specific parts of the DNA blueprint to respond to need and changes, be it internal or external [2]. This way, transcriptomic analysis of RNA sequences from RNA extracts from experimental samples provides a means of tapping into the flow of information in the central dogma of the molecular biology model [3]. One primary objective of a transcriptomic analysis of RNA-seq has been found helpful as a guide to interpret the functional elements of the genome and diseases [1].
1.2 RNA-Seq Technique Analysis and Applications

RNA-Seq technique is a significant method for gene expression profiles from biological samples. The approach uses high-throughput sequencing technologies to generate raw RNA sequences from biological samples to provide insights into the cellular transcriptome. It enables a means to characterize the RNA contents and profile of biological materials at the time of collection. However, given the current limitation of RNA-seq tech, the transcript information is not sequenced in its entirety but rather in a short piece of reads of base-pair sections [4]. The RNA-Seq data analysis includes data preprocessing, checking for quality, mapping of reads to the corresponding genes on a reference genome (in cases where reference genomes are available), differential gene expression computation (between different samples), and many statistical analyses and other downstream analyses and interpretations [2, 4]. RNA-seq technologies have aided in understanding the complexity of transcriptions, gene expression, and regulation of biological processes. It has aided in differential gene expression of disease conditions, sRNA profiling and alternative splicing activities, variant detections [5]. For instance, RNA-seq analysis has been helpful in comparative studies of differential gene expression across different drug treatments and disease conditions.

In the studies, RNA-seq has been applied to the functional analysis of RNA-seq data from primary breast cancer and metastatic lung cancer sites for differential gene expressions and annotation of genes to pathways and enriched gene sets that may be further explored for mutational changes in the exome data of patient sources.
1.3 Cancer RNA-Seq Data Analysis and Functional Annotation

Breast cancers have been second on the list of cancer-related deaths among women in America [6]. Most of these deaths are attributable to breast cancer metastasis. The incidence rate of cancer worldwide is about 1.6 million cases per year [7]. U.S. breast cancer statistics have it that about 1 out of 8 women will develop invasive breast cancer in the United States [8].

Cancer is a result of accumulations of mutations in somatic cells that impair normal cell-growth regulations but give in selective abnormal proliferation to cancer cells. Metastasis – the spread of tumor cells from their original sites to the rest or some specific areas in the body – usually develops when cancer cells become genetically unstable and adapt to new tissue microenvironments other than their primary site [9]. These mutations are called drivers or driver mutations [10, 11]. Some accumulated somatic driver mutations include single-nucleotide variants (SNVs), copy-number variants (CNVs), insertions, and deletions (indels). The characterization of these mutations has been central to understanding cancer invasions and metastasis and keys to diagnoses and therapeutics developments. Besides driver mutations, there are passenger mutations that do not have any causative and contributive effects on metastatic progressions. However only a fraction of these driver mutations has been noted to reside in cancer genes. Fundamental difficulty is analyzing and interrogating the cancer genome data to identify driver mutations differently from passenger mutations from datasets [12][13].
1.4 Aims of Analysis

In this study, we have conducted a comparative analysis of RNA-seq datasets of primary breast cancer and metastatic lung cancer to identify differentially expressed gene lists and identify lists of potential genes and associated enriched pathways. Here, we have used functional enrichment analysis and annotation of these gene sets as steps for finding useful biological insights underlying how mutational alterations in these genes lead to primary breast cancer invading new sites generally - and specifically lung tissue sites, given basic metadata we have about the sample sets for this work.
2 CONCEPTS AND BACKGROUND

2.1 Differential Gene Expression

The cellular expression of genes is highly regulated and based on need. The relative abundance of different RNA transcripts can be used to estimate and determine the expression level of genes in a given biological condition. It is possible to determine which genes are differentially expressed significantly in the pool of gene transcripts for comparative studies of two or more conditions [14, 15]. This approach has been widely applied for the comparative study between primary cancer and secondary cancer sites [16]. There are many available computational and statistical packages and software to differentially quantify gene expression from raw RNA-Seq reads for different conditions such as HISAT, edgeR, DESeq2, ISoEM2 and IsoDE2 [15, 17–19].

2.2 ISOEM2 and ISODE2 Package for Differential Expression Analysis

In this study, we used ISOEM2 and ISODE2 [18], which are bootstrapping-based estimations for confidence intervals of differential gene (DE) expression levels from RNA-seq data. ISOEM2 uses a probabilistic model that is based on Expectation- Maximization to estimate fragments per kilobase million (FPKM) and RNA transcripts per million (TPM) level for genes using bootstrapping. ISODE2 computes the DGE analysis from FPKM/TPM values from ISOEM2 estimates. ISOEM2 reportedly gives one of the highest accuracies compared with some other RNA-seq quantification tools in the studies [19, 20]. Like many DE tools, the output for differential gene expression includes mean values of gene expression, fold-change, Log2 of Fold Change, TPM, and FPKM with a cutoff of 0.05 for the P-value for each generated gene list output.
2.3 **Pathway Annotation for DE Gene Lists**

Pathway or pathway enrichment analysis is a crucial step in interpreting high-throughput data for genomic data. The analysis of RNA-Seq typically yields a long list of differentially expressed genes which can be challenging to interrogate for biological insights. Even after gene lists are filtered, at fold change > |2| and or p-value < 0.05, the significant genes often remain in a range greater than 1000s. Pathway enrichment analysis is often used to extract meaning from such lists of genes to identify biological pathways, physical interaction and protein functions [21, 22]. This way a long list of genes are grouped into smaller gene sets of related functions and regulation based on publicly available knowledge of databases such Gene Ontology (GO) terms [23] or Kyoto Encyclopedia of Genes and Genomes (KEGG) [24].

2.4 **Functional enrichment Analysis and Annotation of DE Gene Lists**

Functional enrichment analysis (also called gene set analysis, GSA) is a statistical method used to interpret and make sense of the huge lists of DE genes returned from RNA-Seq analysis. Enrichment analysis is used basically to check if a group of genes shares some functional properties more or less frequently than it can be expected or attributable to chances. DE gene lists return may be over 1000 even after applying the P-value cut-off for significant levels. Thus, enrichment analyses are conducted to sub-group genes together in a few biological functions based on statistical significance and function relationship [25]. The aim is to find biological annotations that are over-represented in a list of genes associated with an experimental condition studied with respect to reference backgrounds [26]. Genes with similar gene expression often share related metabolic and cellular function or common functions.
2.4.1 Overrepresentation Enrichment Analysis (ORA)

ORA is a technique used to check if a set of biochemical processes is over-represented or enriched in the gene lists generated from the experimental process (i.e., experimentally derived genes). ORA identifies gene sets from pathways that are present more than would be expected by chance. The p-value is calculated using equation(i) below. N stands for the number of genes in the background, M for number of genes in a distribution, n for the size of the gene list of interest and k is the number of the gene within that list which are annotated to the gene set. The background distribution is all the genes that have been annotated, and p-values is adjusted for multiple comparisons [27].

\[
p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}
\]

equation(i)

2.4.2 Gene Set Enrichment Analysis (GSEA)

GSEA is an analytical method for interpreting gene expression data at the level of gene sets - meaning a group of genes that share common biological functions, chromosomal location, or regulations. The gene sets are known based on available biological knowledge for biological pathways and co-expression from experimental procedures [26]. At the conception of the idea of GSEA, Mootha et al. in their work, thought that changes in the expression of genes that is associated with diseases may be manifest in the levels of biochemical pathways or co-regulated genes, rather than it is for individual genes[28]. The technique aggregates genes per statistics across genes in a gene set to make it easily detect cases where all genes in a predefined set change in a small and coordinated way. This technique works by ranking genes based on their phenotypes and then checking if members
of a predefined set of genes (S) are distributed randomly in a ranked gene list (L) or primarily found at the top or bottom of ranks. The method also computes an Enrichment Score (ES), Significant Level or p-value of the ES, and adjustment Multiple Hypothesis Testing for the analysis. ES indicates the degree to which a set of genes S is over-represented at the top of a ranked gene list, L. A null distribution is generated using a permutation test of gene labels in the gene list L. The p-value of the observed ES is then calculated relative to null distribution, and q-values are calculated for FDR control [25].

2.5 Other Tools and Concepts for Analysis

2.5.1 GATK-HaplotypeCaller

GATK-HaplotypeCaller[29] is a terminal-based tool to call SNPs and indels simultaneously via a local de-novo assembly of haplotypes in the active region of DNA-seq data. The input for the program is BAM files from exome - metastasis and primary datasets – from which variant calls are made compared with the human reference genome. The GATK Haplotype Caller returns a VCF file with inferred SNVs, CNVs, and indels and substituted or altered nucleotides, fig.2. The red part shows identified alterations – for ID rs84825. It indicates ‘Some insertion of 'AAGG' modifies 'A' from the human reference genome in this sample of patients [29].
2.5.2 Proveans Annotation Tool

PROVEANS [30] is a bioinformatic software tool to predict whether amino acid substitutions or indels will impact the biological function of proteins and genes. The device is used to filter the VCF variants – output from GATK caller-identified variants – to identify nonsynonymous and indel variants that may have potential functional imports cellular activities [30].

Table 2.2 A Sample of Provean Analysis Output
3 DATA SETS AND METHODS

3.1 Datasets for Analysis

The datasets used for this analysis include RNA sequences collected from breast cancer patients in Norway through collaborator Ritu Aneja’s lab and groups. The data include two sets of RNA-Seq for primary breasts and metastasis (lungs) sites, Figure 3.1. Differential analysis was done on the two sets of pairs such that primary cancer sites are compared with metastatic (sites in the case being lungs) in the following analysis.

![Figure 3.1 Chart for Sample Used for Analysis](image)

3.2 IsoEM2 and IsoDE2

IsoDE2 is used to estimate gene differential expression for the three different conditions based on the output generated from IsoEM2 components of the package. The tool is run on the Galaxy platform [31]. The final output data from the tool are the differential gene ID and the associated fold changes for the metastasis and primary tumors. The IsoDE2 device returns a differential expressed gene and associated fold-change value for comparing gene sets from primary RNA-seq and metastatic RNA-seq data for three conditions considered in the analysis. The ratio of folds greater than 1 indicates an increased gene expression level. The log\_FC is reported alongside and calculated in equation (i) below.
\[ \log_{b}(x/y) = \log_{b}(x) - \log_{b}(y) \]

equation(i)

3.3 KOBAS Web-based Pathway Annotation

KOBAS [32] – the web-server tool annotates the identified gene sets - from IsoDE2 analysis of RNA-seq reads – as input by mapping the genes with known pathways and diseases in the KEGG database, Gene Ontology. The outputs from the tool include the pathway ID and the cluster of genes set for it and p-value calculations for the pathway. The statistical significance value of P-value of 0.05 is used to screen the returned output from KOBAS gene-enrichment annotation. KOBAS results for all patients are collected for analysis, where means, average, and corrected p-values are calculated using the equation in equation(ii) below. The k value = P_1*P_2*…. P_n for p-value for identical pathways across each patient for each pathway ID.

\[ k \sum_{i=0}^{n-1} \frac{(-\ln(k))^i}{i!} \]

equation(ii)

3.4 ClusterProfiler Package

ClusterProfiler package is an R package for performing statistical analysis and visualization of functional profiles genes and gene clusters for thousands of species. It supports a tidy interface for manipulation and visualization of gene enrichment results easy interpretation of gene list, gene clusters and pathway annotations. The ClusterProfiler library is equipped to efficiently conduct over-representation analysis (ORA) and gene set enrichment analysis (GSEA) using database information of KEGG and GO [25].
4 RESULTS AND DISCUSSION

4.1 Results for IsoDE2 Analysis and Annotation

Some of the results obtained are reported here. IsoDE2 results for condition one that select upregulated genes for primary breast cancer. Table 4.1. and Table 4.2 shows snapshots of sample output for a cutoff of p-value 0.05. A total of 1274 and 1847 genes were estimated to be upregulated in primary breast cancers and metastatic lung cancer respectively for NO_02 sample.

*Table 4.1 The DE Gene Lists for Primary Result Sample*

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Confident log2 FC</th>
<th>Single run log2 FC</th>
<th>c1 average TPM</th>
<th>c2 average TPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOK</td>
<td>-1.000056373</td>
<td>-1.121601485</td>
<td>3.52572024</td>
<td>1.620362406</td>
</tr>
<tr>
<td>ADGRB1</td>
<td>-1.000072119</td>
<td>-1.038886449</td>
<td>5.768848672</td>
<td>2.807715914</td>
</tr>
<tr>
<td>RAP1GAP</td>
<td>-1.000939253</td>
<td>-1.054216619</td>
<td>9.234011102</td>
<td>4.446717791</td>
</tr>
<tr>
<td>CPT1A</td>
<td>-1.001839863</td>
<td>-1.039018566</td>
<td>47.995653356</td>
<td>23.35791729</td>
</tr>
<tr>
<td>PAQR8</td>
<td>-1.002507521</td>
<td>-1.181136348</td>
<td>2.82536454</td>
<td>1.245997068</td>
</tr>
<tr>
<td>L1TD1</td>
<td>-1.002628765</td>
<td>-1.075824093</td>
<td>5.503625245</td>
<td>2.610919801</td>
</tr>
<tr>
<td>MTHFD1</td>
<td>-1.003470938</td>
<td>-1.020660636</td>
<td>10.4810142</td>
<td>5.165993349</td>
</tr>
<tr>
<td>SEC31B</td>
<td>-1.003580168</td>
<td>-1.041933653</td>
<td>35.30563921</td>
<td>17.14710454</td>
</tr>
<tr>
<td>TOX</td>
<td>-1.003693955</td>
<td>-1.02251562</td>
<td>11.98202296</td>
<td>5.898237756</td>
</tr>
<tr>
<td>IRF2BL1</td>
<td>-1.00379213</td>
<td>-1.068639691</td>
<td>10.45906799</td>
<td>4.986552374</td>
</tr>
<tr>
<td>RAPGEF3</td>
<td>-1.004521117</td>
<td>-1.064442563</td>
<td>6.571225527</td>
<td>3.142079956</td>
</tr>
<tr>
<td>KRT16P2</td>
<td>-1.0051287</td>
<td>-1.056831102</td>
<td>5.673500462</td>
<td>2.727176416</td>
</tr>
<tr>
<td>FOXF2</td>
<td>-1.005515556</td>
<td>-1.124286107</td>
<td>5.753987913</td>
<td>2.639520888</td>
</tr>
<tr>
<td>TMEM18</td>
<td>-1.00580043</td>
<td>-1.07014388</td>
<td>4.002426879</td>
<td>1.906241846</td>
</tr>
<tr>
<td>IGFLR1</td>
<td>-1.00633949</td>
<td>-1.023772093</td>
<td>16.27760234</td>
<td>8.005792432</td>
</tr>
<tr>
<td>METTL7A</td>
<td>-1.006863713</td>
<td>-1.047177318</td>
<td>35.8649329</td>
<td>27.0334187</td>
</tr>
<tr>
<td>CYP21A1P</td>
<td>-1.007001833</td>
<td>-1.190073173</td>
<td>3.810067708</td>
<td>1.669878878</td>
</tr>
<tr>
<td>TTC30A</td>
<td>-1.007288846</td>
<td>-1.153569332</td>
<td>3.126987129</td>
<td>1.405617199</td>
</tr>
<tr>
<td>FAM86HP</td>
<td>-1.007363604</td>
<td>-1.15641456</td>
<td>4.847660432</td>
<td>2.17478963</td>
</tr>
</tbody>
</table>
Combined analysis of annotated pathways from KOBAS tool, Table 4. The statistical significance cutoff on the P-values was set at 0.05 for all the red-painted values. The pathways are literature mined for information on the pathways. The results are further sorted based on the pathway database used for the annotation.

**Table 4.2 The DE Gene Lists for Metastasis Result Sample**

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Confident log2 FC</th>
<th>Single run log2 FC</th>
<th>c1 average TPM</th>
<th>c2 average TPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>KCTD5</td>
<td>1.000019146</td>
<td>1.055624634</td>
<td>12.22443429</td>
<td>25.40992888</td>
</tr>
<tr>
<td>ATXN7L1</td>
<td>1.00010367</td>
<td>1.045167036</td>
<td>17.84757057</td>
<td>36.83034068</td>
</tr>
<tr>
<td>NMD3</td>
<td>1.00178251</td>
<td>1.029546747</td>
<td>37.90343476</td>
<td>77.37386931</td>
</tr>
<tr>
<td>PDZK1IP1</td>
<td>1.00042786</td>
<td>1.031231772</td>
<td>13.225601</td>
<td>27.03006639</td>
</tr>
<tr>
<td>AP4S1</td>
<td>1.000647004</td>
<td>1.037592219</td>
<td>1.16538872</td>
<td>2.392292434</td>
</tr>
<tr>
<td>ZSCAN16</td>
<td>1.00100557</td>
<td>1.063800679</td>
<td>19.7576361</td>
<td>41.30198229</td>
</tr>
<tr>
<td>PLD2</td>
<td>1.001007536</td>
<td>1.025507324</td>
<td>24.0110508</td>
<td>48.87869834</td>
</tr>
<tr>
<td>RRM2B</td>
<td>1.001169731</td>
<td>1.073004123</td>
<td>7.23617841</td>
<td>15.22467745</td>
</tr>
<tr>
<td>JAZF1</td>
<td>1.001960695</td>
<td>1.101686632</td>
<td>3.33601856</td>
<td>7.15927244</td>
</tr>
<tr>
<td>AGPAT2</td>
<td>1.002562444</td>
<td>1.015086788</td>
<td>2.826155487</td>
<td>5.711729839</td>
</tr>
<tr>
<td>UNC13A</td>
<td>1.002645983</td>
<td>1.103186558</td>
<td>2.036993916</td>
<td>4.37593098</td>
</tr>
<tr>
<td>PAT1</td>
<td>1.003121178</td>
<td>1.020772356</td>
<td>32.15419025</td>
<td>65.24100989</td>
</tr>
<tr>
<td>PNMA8A</td>
<td>1.003191662</td>
<td>1.012750992</td>
<td>1.94559457</td>
<td>3.925733132</td>
</tr>
<tr>
<td>VAPA</td>
<td>1.003641796</td>
<td>1.030581676</td>
<td>27.08939537</td>
<td>55.34066215</td>
</tr>
<tr>
<td>IMPAD1</td>
<td>1.003681084</td>
<td>1.023297404</td>
<td>12.8656735</td>
<td>26.31755209</td>
</tr>
<tr>
<td>ITPR2</td>
<td>1.00380077</td>
<td>1.004842012</td>
<td>120.7545408</td>
<td>242.3210025</td>
</tr>
<tr>
<td>ZNF638</td>
<td>1.004244229</td>
<td>1.015691273</td>
<td>95.27249446</td>
<td>192.6287364</td>
</tr>
<tr>
<td>PLA2G4A</td>
<td>1.00439818</td>
<td>1.021391449</td>
<td>14.03375028</td>
<td>28.48676991</td>
</tr>
<tr>
<td>PCMTD1</td>
<td>1.00443181</td>
<td>1.01942672</td>
<td>88.63657948</td>
<td>179.6763883</td>
</tr>
</tbody>
</table>

**Table 4.3 The KOBAS Annotation Result Sample**

<table>
<thead>
<tr>
<th>ID</th>
<th>ZURICH_01 P-Value</th>
<th>ZURICH_02 P-Value</th>
<th>ZURICH_04 P-Value</th>
<th>ZURICH_05 P-Value</th>
<th>Max</th>
<th>Avg</th>
<th>Combo</th>
<th>2nd Max</th>
<th>2nd Combo</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa01100</td>
<td>0.000107597</td>
<td>1.89E-10</td>
<td>0.408371</td>
<td>4.87E-07</td>
<td>0.78273546</td>
<td>0.12506867</td>
<td>3.15E-14</td>
<td>0.34154767</td>
<td>1.28E-14</td>
</tr>
<tr>
<td>hsa04512</td>
<td>2.64E-05</td>
<td>8.12E-14</td>
<td>5.78E-07</td>
<td>3.57E-12</td>
<td>2.29E-05</td>
<td>0.00686442</td>
<td>4.77E-66</td>
<td>0.02889752</td>
<td>4.38E-66</td>
</tr>
<tr>
<td>hsa04974</td>
<td>0.021168293</td>
<td>1.13E-12</td>
<td>1.28E-10</td>
<td>8.33E-08</td>
<td>1.13E-08</td>
<td>0.02229613</td>
<td>1.35E-64</td>
<td>0.03390477</td>
<td>4.63E-65</td>
</tr>
<tr>
<td>hsa04151</td>
<td>0.000275952</td>
<td>1.93E-16</td>
<td>7.59E-06</td>
<td>5.56E-12</td>
<td>0.00187965</td>
<td>0.0078209</td>
<td>6.76E-61</td>
<td>1.12E-05</td>
<td>4.83E-61</td>
</tr>
<tr>
<td>hsa05200</td>
<td>2.51E-05</td>
<td>2.97E-15</td>
<td>4.02E-05</td>
<td>1.13E-09</td>
<td>0.0185655</td>
<td>0.0117109</td>
<td>1.50E-59</td>
<td>3.43E-06</td>
<td>1.33E-59</td>
</tr>
<tr>
<td>hsa04510</td>
<td>0.002089862</td>
<td>8.10E-17</td>
<td>2.82E-06</td>
<td>3.04E-12</td>
<td>4.54E-05</td>
<td>0.04363922</td>
<td>7.18E-58</td>
<td>0.09586434</td>
<td>1.13E-58</td>
</tr>
<tr>
<td>hsa05165</td>
<td>0.00237879</td>
<td>6.27E-11</td>
<td>7.34E-09</td>
<td>3.40E-11</td>
<td>6.0026804</td>
<td>0.05456452</td>
<td>6.88E-51</td>
<td>0.30608126</td>
<td>1.58E-51</td>
</tr>
<tr>
<td>hsa04060</td>
<td>2.64E-07</td>
<td>6.71E-07</td>
<td>7.46E-09</td>
<td>0.0130555</td>
<td>6.79E-08</td>
<td>0.05034126</td>
<td>4.17E-48</td>
<td>1.89E-20</td>
<td>1.07E-48</td>
</tr>
<tr>
<td>hsa05410</td>
<td>0.001251796</td>
<td>2.06E-13</td>
<td>0.0169704</td>
<td>8.73E-07</td>
<td>2.39E-06</td>
<td>0.06689774</td>
<td>1.02E-41</td>
<td>0.25553207</td>
<td>1.53E-42</td>
</tr>
</tbody>
</table>
4.2 Provean Annotation Result Sample

This sample output, table 4.4 from PROVEAN tool. The predicted functional effects of mutation include damaging and neutral and the associated mutation, genes, and protein. This is an ongoing analysis. Further co-analysis and integration of RNA-seq and DNA-seq data will identify and characterize drivers – SNVs, CNVs, Idels, and SNPs associated with the different metastatic sites and pathways.

Table 4.4 The Provean Annotation Output Sample

<table>
<thead>
<tr>
<th>#ROW_NO.</th>
<th>INPUT</th>
<th>PROTEIN_ID</th>
<th>LENDRAN</th>
<th>THRD</th>
<th>CODON_CHANGE</th>
<th>POS</th>
<th>RESIDUE_REF</th>
<th>RESIDUE_ALT</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>185</td>
<td>1,91745,C,T</td>
<td>ENSP00000414022</td>
<td>696</td>
<td>-1</td>
<td>AT[G/A]CCG</td>
<td>1</td>
<td>M</td>
<td>I</td>
<td>Single AA Change</td>
</tr>
<tr>
<td>536</td>
<td>1,1197697,A,G</td>
<td>ENSP00000423880</td>
<td>105</td>
<td>-1</td>
<td>CTC CA[T/C]CAT</td>
<td>82</td>
<td>H</td>
<td>H</td>
<td>Synonymous</td>
</tr>
<tr>
<td>6051</td>
<td>1267057,A,T</td>
<td>ENSP0000344411</td>
<td>852</td>
<td>1</td>
<td>ATG AA[A/T]ATG</td>
<td>77</td>
<td>K</td>
<td>N</td>
<td>Single AA Change</td>
</tr>
<tr>
<td>633</td>
<td>1,1290173,C,T</td>
<td>ENSP00000399229</td>
<td>450</td>
<td>-1</td>
<td>TGT [G/A]GCCTG</td>
<td>280</td>
<td>G</td>
<td>S</td>
<td>Single AA Change</td>
</tr>
<tr>
<td>12301</td>
<td>1720498,G,A</td>
<td>ENSP00000367869</td>
<td>340</td>
<td>-1</td>
<td>GAC [C/T]GGGCA</td>
<td>304</td>
<td>R</td>
<td>W</td>
<td>Single AA Change</td>
</tr>
<tr>
<td>29821</td>
<td>3809539,A,G</td>
<td>ENSP00000355306</td>
<td>243</td>
<td>-1</td>
<td>TCT [T/C]CA GCG</td>
<td>13</td>
<td>S</td>
<td>P</td>
<td>Single AA Change</td>
</tr>
<tr>
<td>2983</td>
<td>1,3809544,C,T</td>
<td>ENSP00000355306</td>
<td>243</td>
<td>-1</td>
<td>GTG [G/A]GTCTT</td>
<td>11</td>
<td>R</td>
<td>Q</td>
<td>Single AA Change</td>
</tr>
<tr>
<td>4137</td>
<td>1,5947486,T,A</td>
<td>ENSP00000367398</td>
<td>1426</td>
<td>-1</td>
<td>TCC [A/T]CGAG</td>
<td>782</td>
<td>H</td>
<td>L</td>
<td>Single AA Change</td>
</tr>
<tr>
<td>5504</td>
<td>1,7724628,T,C</td>
<td>ENSP00000306522</td>
<td>1673</td>
<td>1</td>
<td>CAC [T/C]CAATG</td>
<td>674</td>
<td>L</td>
<td>P</td>
<td>Single AA Change</td>
</tr>
<tr>
<td>5509</td>
<td>1,7737674,T,A</td>
<td>ENSP00000306522</td>
<td>1673</td>
<td>1</td>
<td>GGT [T/A]GTG</td>
<td>932</td>
<td>L</td>
<td>H</td>
<td>Single AA Change</td>
</tr>
</tbody>
</table>

4.3 Enrichment Analysis for Gene Lists Analysis Results

4.3.1 Overrepresentation Enrichment Analysis (ORA) Results

Usually not all pathways identified enriched in ORA may have relevance for further, however, as seen Fig 4.1 and 4.2, 4.4, some of the upregulated pathways identified in the ORA analysis of sample NO_O2 and NO_07 are known key pathways for breast cancer progression. PI3K-Akt (phosphatidylinositol 3-kinase) signaling pathway, for instance, has been known to be aberrantly activated by various mechanisms related to mutation in PIK3CA of various
cancer types. Mutation in PIK3CA genes will result in unchecked cell growth and proliferation and as such has been a target for drug development and therapy [33–35].

Figure 4.1 Dot Plot Chart for ORA for NO_02 Samples (a) Primary Cancer upregulated pathways- seven pathway identified to be significantly over represents including PI3K-Akt signaling pathway, PPAR signaling pathway and ECM-receptor interactions (b) Metastatic Cancer 20 over-represented pathway including some known pathway for breast and lung cancer progression.
PPAR signaling pathway, identified to be enrich in primary breast cancer, is a key participant in regulation of cellular processes associated with carcinogenesis, cell differentiation, proliferation, survival and apoptosis affecting family of nuclear hormone receptor, peroxisome proliferator-activated receptor (PPARs). There is accumulating evidence showing that PPARy (a subtype of PPAR - agonists tend to prevent cancer cells from acquiring potential for the migratory and metastatic invasion into new sites. The complexity of the PPAR family in lung cancer has been well documented and identified to be significant for potential therapy [36–38]. Thus, PPAR signaling pathway identified to be among the over-represented pathway in primary breast cancer for NO_02 primary breast cancer may be a pointer to further investigate these pathways, other connecting pathways and associated genes and their relevance to metastatic lung cancer for more insights.

T helpers - Th1, Th12, Th17 - cell differentiation have been identified in many human tumor micro-environments even though their roles of some are unclear. Recent study identifies Th17 as a prognostic composite biomarker for T cell non-inflamed triple negative breast cancer (TNBC). As such with Th cell differentiation - being identified as upregulated in NO_02 sample Figure 4.3 of enriched pathway in metastatic lung cancer, it may be good exploration to examine further associated genes in Figure 4.4 for key participation as drivers in this sample.
Figure 4.2 UpSet Plot for ORA for NO_02 Samples (a) primary breast cancer identifies six pathways: PI3K-Akt signaling pathway, focal adhesion, ECM-receptor signaling, PPAR pathway shows shared connection (b) metastatic lungs overrepresented pathway includes cytokine-cytokine receptor related pathway, Th17, Th1 and Th2 cell differentiation pathway with multiple connecting dots which key play in cancer and breast cancer progressions.

Figure 4.3 Upsets Plot for ORA for NO_07 Samples (a) Primary breast cancer identifies 10 pathways of with PI3K-Akt signaling pathway, Focal adhesion, ECM-receptor
TNF (Tumor necrosis factor) signaling pathway is annotated to be enriched in NO_07 primary sample, which may be another point of interest for investigation. Increasing findings have implicated TNF - for example, TNF-α is known for modulating cancer-related inflammation and promoting tumor progressions - which are vital to cancer invasion and metastasis progression. Studies have identified TNF-α, in conjunction with the NF-kB pathway, to activate JNK, MAPKs, AKT, and other non-canonical NF-kB pathways. In fact, TNF-α is recorded to up-regulate more than 400 inflammatory genes involved in cell-adhesion molecules, inflammatory cytokines, and chemokines [39–43]. In fact, TNF-α is recorded to up-regulate more than 400 inflammatory genes involved in cell-adhesion
molecules enriched in pathways for crosstalk between tumor cell, stromal and immune cells with primary and secondary cancer sites, Figures 4.1 and 4.4.

4.3.2 Gene Enrichment Analysis (GSEA) Results

The enriched gene sets in Figure 4.5 and 4.6 from the GSEA - particularly the one involving interconnection between two or more cancer-related pathways - are critical for exploration and mutation analysis - alongside the critical genes identified over-represented in pathway enrichment analysis (ORA).

ErbB2 gene identified enriched in NO_12 Figure 4.5. The over-expression of ErbB2 has been identified to contribute to some breast cancer growth by activating PI3K/AKT signaling pathways crosstalk with ER-α [44]. The ErbB2 is found enriched for the estrogen signaling pathway and estrogen resistance pathway, Figure 4.5. Overlapping genes may be further studied for more insight. IGF1 (insulin-like growth factor 1) gene is identified to be enriched for EGFR HIF-1 signaling pathway, tyrosine kinase inhibitor resistance and Hypertrophic cardiomyopathy, Figure 4.5. IGF-1 is essential for the growth and survival of cancer cells and has a connection with the initiation of downstream signal transduction pathways involving Ras/Raf/ERK and PI3k/AKT/mTOR[45], Figure 4.4 – 4.5.
Figure 4.5 Gene Concept Network (C-netplot) for NO_02 Samples (a) primary and (b) metastasis samples show genes in selected enriched pathways for gene overlaps with other pathways. Gene linkage across pathways is shown for c-neplot for the map for upregulated pathways in primary samples. In metastasis, it shows Endocrine resistance and estrogen signaling pathway gene overlapping.

Figure 4.6 Enrichment Map for sample NO_07 (a) primary and (b) metastasis show genes involved in the enriched pathways for gene overlaps between pathways. Gene linkage
across pathways as shown for c-neplot for map for upregulated pathways in primary samples. Both primary and metastasis show multiple gene overlap.

Figure 4.7 Gene Concept Network (C-netplot) for NO_07 Samples (a) primary and (b) metastasis show genes involved in the enriched pathways for gene overlaps between pathways. Gene linkage across pathways is shown for c-neplot for map for upregulated pathways in primary samples. Both primary and metastasis show multiple gene overlap.
5 CONCLUSION

In this thesis, we have successfully analyzed RNA-Seq data for primary breast cancer and metastasis lung cancer for differentially expressed gene lists for each of pairs of samples. The annotated enriched pathways and gene from this analysis that can be explored further for more biological insight and mutational analysis. The critical cancer-related pathways identified in this analysis include PI3K-Akt signaling pathway, PPAR signaling pathway, T helpers - Th1, Th12, Th17 - cell differentiation and ECM-receptor interactions, Ras signaling pathways, TNF signaling pathway, estrogen signaling pathway, and estrogen resistance pathway. All of which have some associations and implications in breast cancer progression, proliferation, and invasion of secondary sites. The enriched genes identified include ErbB2, IGF-1, FOXA2 for there is increasing evidence they have associated with cancer proliferation, invasion, and metastasis. Future work will include comprehensive annotation of the identified enriched genes for mutational changes and studies how these changes may be contributing to their biochemical participation in driving metastatic invasion and other related activities.
REFERENCES


expression results. Genome Biology 11:220


41. Aggarwal BB, Takada Y Pro-apoptotic and Anti-apoptotic Effects of Tumor Necrosis Factor in Tumor Cells. Cancer Treatment and Research 103–127


