5-4-2022

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

Babatunde Bello

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

LIST OF TABLES

LIST OF FIGURES

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 BACKGROUNDS

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 dat**a.** 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].

Table 2.1 GATK table Sample

```
#fileformat=VCFv4.2
##FORMAT=<ID=GT,Number=1,Type=Integer,Description="Genotype">
##FORMAT=<ID=GP,Number=G,Type=Float,Description="Genotype Probabilities">
##FORMAT=<ID=PL,Number=G,Type=Float,Description="Phred-scaled Genotype Likelihoods">

#CHROM  POS     ID        REF   ALT    QUAL  FILTER  INFO  FORMAT      SAMP001     SAMP002
20      1291018 rs11449   G     A      .     PASS    .     GT          0/0         0/1
20      2300608 rs84825   A     AAGG   .     PASS    .     GT:GP       0/1:.       0/1:0.03,0.97,0
20      2301308 rs84823   TAAC  G      .     PASS    .     GT:PL       ./.:.       1/1:10,5,0
```

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

#ROW_NO.	INPUT	PROTEIN_ID	LENGTH	STRAND	CODON_CHANGE	POS	RESIDUE_REF	RESIDUE_ALT	TYPE
185	1,91745,C,T	ENSP00000414022	696	-1	AT[G/A] CCG	1	M	I	Single AA Change
536	1,1197697,A,G	ENSP00000423880	105	-1	CTC CA[T/C] CAT	82	H	H	Synonymous
605	1,1267057,A,T	ENSP00000344411	852	1	ATG AA[A/T] ATG	77	K	N	Single AA Change
633	1,1290173,C,T	ENSP00000399229	450	-1	TGT [G/A]GC CTG	280	G	S	Single AA Change
1230	1,1720498,G,A	ENSP00000367869	340	-1	GAC [C/T]GG GCA	304	R	W	Single AA Change
2982	1,3809539,A,G	ENSP00000355306	243	-1	TCT [T/C]CA GCG	13	S	P	Single AA Change
2983	1,3809544,C,T	ENSP00000355306	243	-1	GTG C[G/A]G TCT	11	R	Q	Single AA Change
4137	1,5947486,T,A	ENSP00000367398	1426	-1	TCC C[A/T]C GAG	782	H	L	Single AA Change
5504	1,7724628,T,C	ENSP00000306522	1673	1	CAC C[T/C]C ATG	674	L	P	Single AA Change
5509	1,7737674,T,A	ENSP00000306522	1673	1	GGT C[T/A]T GTG	932	L	H	Single AA Change

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.

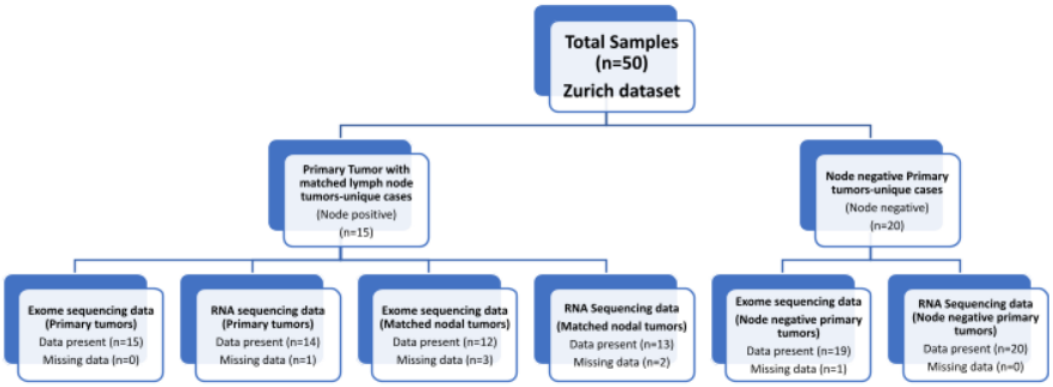


Figure 3.1 Chart for Sample Used for Analysis

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

Gene ID	Confident log2 FC	Single run log2 FC	c1 average TPM	c2 average TPM
BOK	-1.000056373	-1.121601485	3.52572024	1.620362406
ADGRB1	-1.000072119	-1.038886449	5.768848672	2.807715914
RAP1GAP	-1.000939253	-1.054216619	9.234011102	4.446717791
CPT1A	-1.001839863	-1.039018566	47.99653336	23.35791729
PAQR8	-1.002507521	-1.181136348	2.82536454	1.245997068
L1TD1	-1.002628765	-1.075824093	5.503625245	2.610919801
MTHFSD	-1.003470938	-1.020660636	10.4810142	5.165993349
SEC31B	-1.003580168	-1.041933653	35.30563921	17.14710454
TOX	-1.003639355	-1.02251562	11.98202296	5.898237756
IRF2BPL	-1.00379213	-1.068639691	10.45906799	4.986552374
RAPGEF3	-1.004521117	-1.064442563	6.571225527	3.142079956
KRT16P2	-1.0051287	-1.056831102	5.673500462	2.727176416
FOXF2	-1.005515556	-1.124286107	5.753987913	2.639520888
TMEM18	-1.00580043	-1.07014388	4.002426879	1.906241846
IGFLR1	-1.00633949	-1.023772093	16.27760234	8.005792432
METTL7A	-1.006863713	-1.047177318	55.86409329	27.0334187
CYP21A1P	-1.007001833	-1.190073173	3.810067708	1.669878878
TTC30A	-1.007288846	-1.153569532	3.126987129	1.405617199
FAM86HP	-1.007363604	-1.15641456	4.847660432	2.174786963

Table 4.2 The DE Gene Lists for Metastasis Result Sample

Gene ID	Confident log2 FC	Single run log2 FC	c1 average TPM	c2 average TPM
KCTD5	1.000091946	1.055624634	12.22443429	25.40992888
ATXN7L1	1.00010367	1.045167036	17.84757057	36.83034068
NMD3	1.000178251	1.029564774	37.90343476	77.37638693
PDZK1IP1	1.000442786	1.031231772	13.225601	27.03006639
AP4S1	1.000647004	1.037592219	1.165380872	2.392292434
ZSCAN16	1.00100557	1.063800679	19.7576361	41.30198229
PLD2	1.001007536	1.025507324	24.0110508	48.87869834
RRM2B	1.001169731	1.073004123	7.23671841	15.22467745
JAZF1	1.001960695	1.101686632	3.336018564	7.159277244
AGPAT2	1.002562444	1.015086788	2.826155487	5.711729389
UNC13A	1.002645983	1.103186558	2.036939916	4.37593098
PATL1	1.003121778	1.020772356	32.15419025	65.24100989
PNMA8A	1.003191662	1.012750992	1.945594575	3.925733132
VAPA	1.003641796	1.030581676	27.08995937	55.34066215
IMPAD1	1.003681018	1.032397404	12.8665735	26.31755209
ITPR2	1.00380077	1.004842012	120.7545408	242.3210025
ZNF638	1.004244229	1.015691273	95.27249446	192.6287364
PLA2G4A	1.00439818	1.021391449	14.03375028	28.48676991
PCMTD1	1.00443181	1.01942672	88.63657948	179.6763883

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 4.3 The KOBAS Annotation Result Sample

ID	ZURICH_01 P-Value	ZURICH_02 P-Value	ZURICH_04 P-Value	ZURICH_05 P-Value	P-
hsa01100	0.000107597	1.89E-10	0.408371	4.87E-07	
hsa04512	2.64E-05	8.12E-14	5.78E-07	3.57E-12	
hsa04974	0.021168293	1.13E-12	1.28E-10	8.33E-08	
hsa04151	0.000275952	1.93E-16	7.59E-06	5.56E-12	
hsa05200	2.51E-05	2.97E-15	4.02E-05	1.13E-09	
hsa04510	0.002089862	8.10E-17	2.82E-06	3.04E-12	
hsa05165	0.00237879	6.27E-11	7.34E-09	3.40E-11	
hsa04060	2.64E-07	6.71E-07	7.46E-09	0.0130555	
hsa05410	0.001251796	2.06E-13	0.0169704	8.73E-07	

Max	Avg	Combo	2nd Max	2nd Combo
0.78273546	0.12506867	3.15E-140	0.34154767	1.28E-141
2.29E-05	0.00686442	4.77E-66	0.02889752	4.38E-66
1.13E-08	0.02229613	1.35E-64	0.03390477	4.63E-65
0.00187965	0.0078209	6.76E-61	1.12E-05	4.83E-61
0.0185655	0.01117109	1.50E-59	3.43E-06	1.33E-59
4.54E-05	0.04363922	7.18E-58	0.09586434	1.13E-58
0.00026804	0.05146542	6.88E-51	0.30608126	1.58E-51
6.79E-08	0.05034126	4.17E-48	1.89E-20	1.07E-48
2.39E-06	0.06689774	1.02E-41	0.25553207	1.53E-42

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

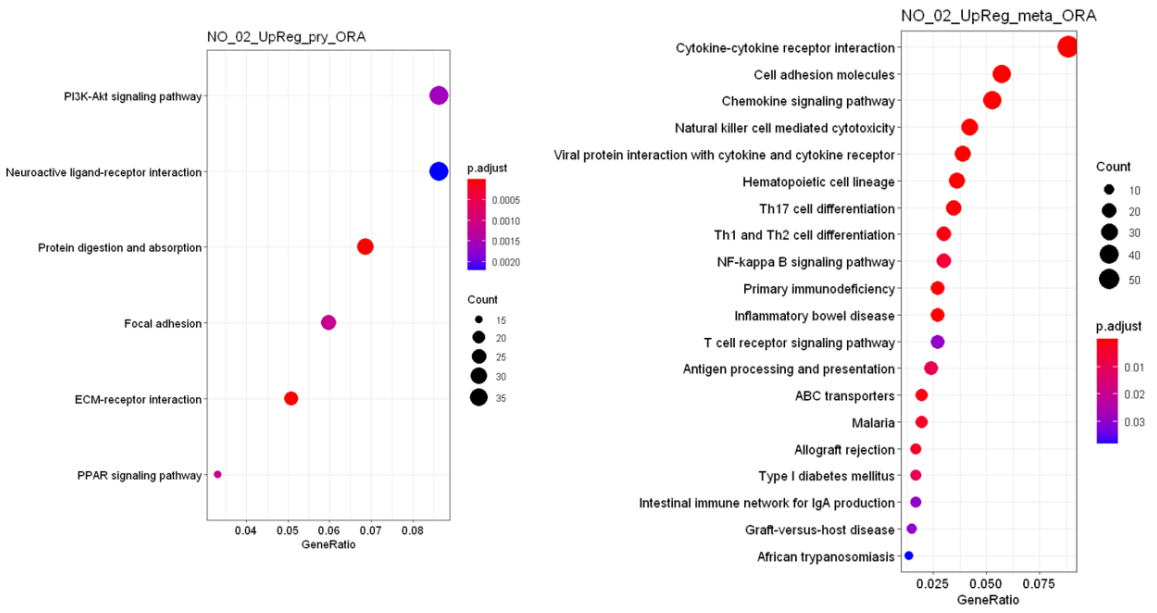
#ROW_NO.	INPUT	PROTEIN_ID	LENGTH	STRAND	CODON_CHANGE	POS	RESIDUE_REF	RESIDUE_ALT	TYPE
185	1,91745,C,T	ENSP00000414022	696	-1	AT[G/A] CCG	1	M	I	Single AA Change
536	1,1197697,A,G	ENSP00000423880	105	-1	CTC CA[T/C] CAT	82	H	H	Synonymous
605	1,1267057,A,T	ENSP00000344411	852	1	ATG AA[A/T] ATG	77	K	N	Single AA Change
633	1,1290173,C,T	ENSP00000399229	450	-1	TGT [G/A]GC CTG	280	G	S	Single AA Change
1230	1,1720498,G,A	ENSP00000367869	340	-1	GAC [C/T]GG GCA	304	R	W	Single AA Change
2982	1,3809539,A,G	ENSP00000355306	243	-1	TCT [T/C]CA GCG	13	S	P	Single AA Change
2983	1,3809544,C,T	ENSP00000355306	243	-1	GTG C[G/A]G TCT	11	R	Q	Single AA Change
4137	1,5947486,T,A	ENSP00000367398	1426	-1	TCC C[A/T]C GAG	782	H	L	Single AA Change
5504	1,7724628,T,C	ENSP00000306522	1673	1	CAC C[T/C]C ATG	674	L	P	Single AA Change
5509	1,7737674,T,A	ENSP00000306522	1673	1	GGT C[T/A]T GTG	932	L	H	Single AA Change

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].

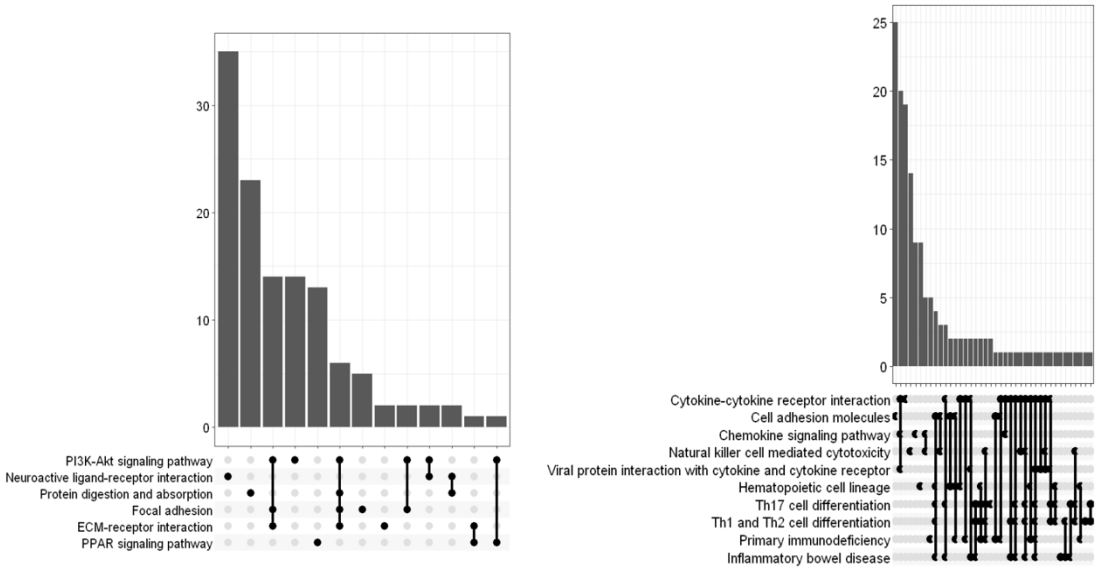


(a) (b)

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.

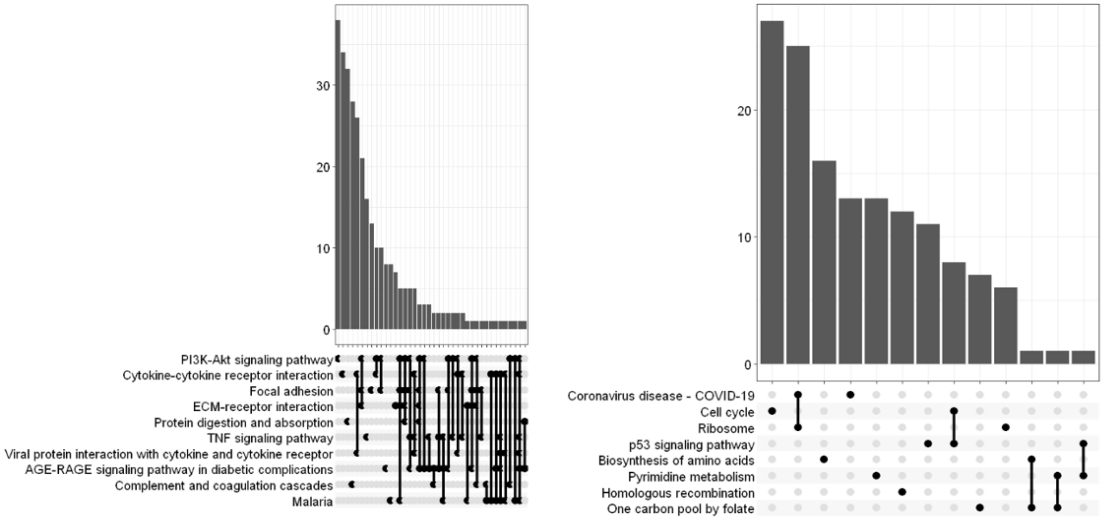
PPAR signaling pathway, identified to be enrich in primary breast cancer, is a key participant in regulation of cellular processes associated with carcinogenesis, cell differentiation, proliferation, survival and apoptosis affecting family of nuclear hormone receptor, peroxisome proliferator-activated 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.



(a) (b)

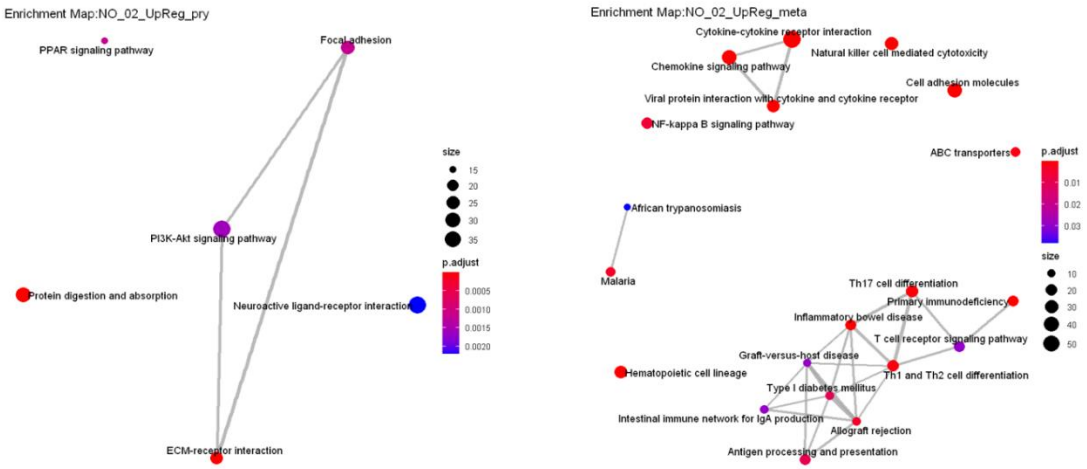
Figure 4.2 UpSet Plot for ORA for NO_02 Samples (a) primary breast cancer identifies six pathways: PI3K-Akt signaling pathway, focal adhesion, ECM-receptor signaling, PPAR pathway shows shared connection (b) metastatic lungs overrepresented pathway includes cytokine-cytokine receptor related pathway, Th17, Th1 and Th2 cell differentiation pathway with multiple connecting dots which key play in cancer and breast cancer progressions.



(a) (b)

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

signaling, TNF pathway show shared connection as cancer associated pathways (b) metastatic lungs overrepresented pathway includes cell cycle, p53 signaling, one carbon pool by folate, pyrimidine metabolism connecting dots which key play in cancer and breast cancer progressions.



(a) (b)

Figure 4.4 Enrichment Map for NO_02 Samples (a) primary breast cancer (b) Metastasis lung cancer showing the connection between enriched pathways involving pathways for cancer progression and metastasis

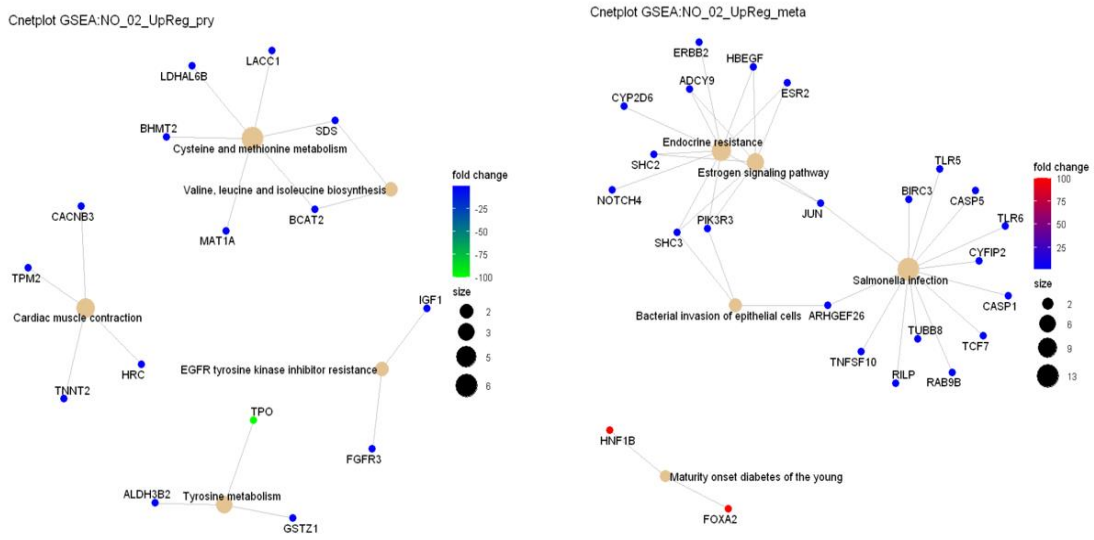
TNF (Tumor necrosis factor) signaling pathway is annotated to be enriched in NO_O7 primary sample, which may be another point of interest for investigation. Increasing findings have implicated TNF - for example, TNF-α is known for modulating cancer-related inflammation and promoting tumor progressions - 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

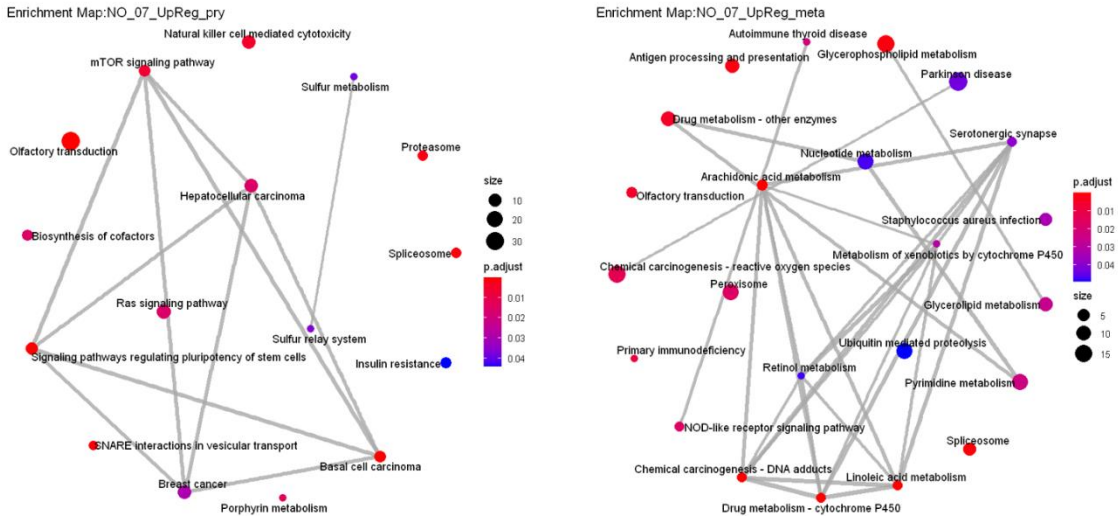
The enriched gene sets in Figure 4.5 and 4.6 from the GSEA - particularly the one involving interconnection between two or more cancer-related pathways - are critical for exploration and mutation analysis - alongside the critical genes identified over-represented in pathway enrichment analysis (ORA).

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.



(a) (b)

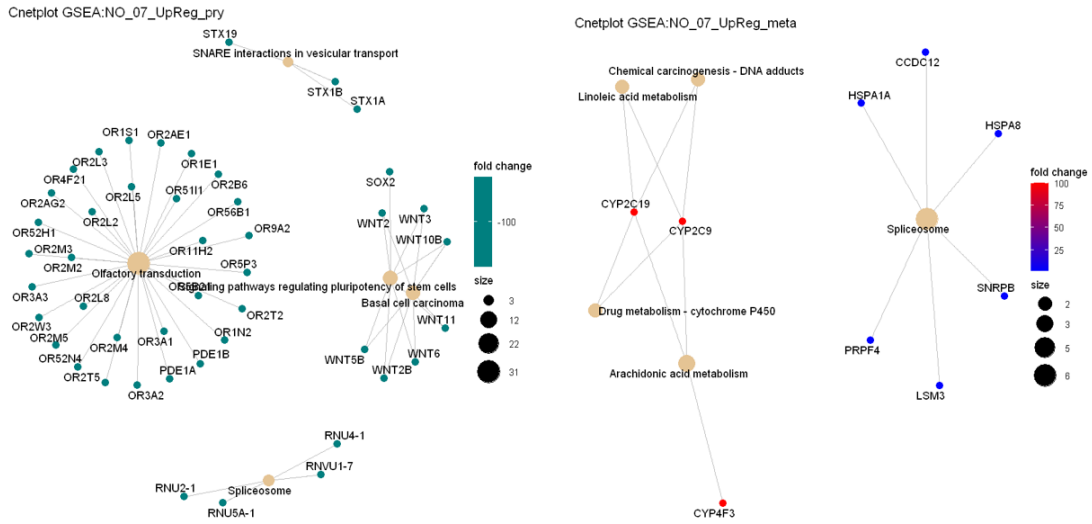
Figure 4.5 Gene Concept Network (C-netplot) for NO_02 Samples (a) primary and (b) metastasis samples show genes in selected enriched pathways for gene overlaps with other pathways. Gene linkage across pathways is shown for c-neplot for the map for upregulated pathways in primary samples. In metastasis, it shows Endocrine resistance and estrogen signaling pathway gene overlapping.



(a) (b)

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



(a) (b)

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

1. Wolf JBW (2013) Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Mol E.coli Resour 13:559–572

2. Kukurba KR, Montgomery SB (2015) RNA Sequencing and Analysis. Cold Spring Harb Protoc 2015:951–969

3. Crick F (1970) Central dogma of molecular biology. Nature 227:561–563

4. Chen Y, Lun ATL, Smyth GK (2016) From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Research 5:1438

5. Applications of RNA-Seq. https://www.cd-genomics.com/resourse-Applications-of-RNA-Seq.html#:~:text=An%20important%20application%20of%20RNA-seq%20is%20the%20comparison,isoforms%20and%20precise%20assessment%20of%20their%20expression%20levels. Accessed 24 Apr 2022

6. Tarver T (2012) Cancer Facts & Figures 2012. American Cancer Society (ACS). Journal of Consumer Health On the Internet 16:366–367

7. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A (2015) Global cancer statistics, 2012. CA Cancer J Clin 65:87–108

8. Website. https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html.

9. Gupta GP, Massagué J (2006) Cancer Metastasis: Building a Framework. Cell 127:679–695

10. Nik-Zainal S, Davies H, Staaf J, et al (2019) Author Correction: Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 566:E1

11. Martincorena I, Campbell PJ (2015) Somatic mutation in cancer and normal cells. Science 349:1483–1489

12. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW (2013) Cancer genome landscapes. Science 339:1546–1558

13. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ (2017) Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 171:1029–1041.e21

14. Oshlack A, Robinson MD, Young MD (2010) From RNA-seq reads to differential

expression results. Genome Biology 11:220

15. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols 11:1650–1667

16. Website. Chen, L., Chu, C., Lu, J., Kong, X., Huang, T., & Cai, Y. D. (2015). Gene Ontology and KEGG Pathway Enrichment Analysis of a Drug Target-Based Classification System. PloS one, 10(5), e0126492. https://doi.org/10.1371/journal.pone.0126492.

17. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140

18. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550

19. Mandric I, Temate-Tiagueu Y, Shcheglova T, Al Seesi S, Zelikovsky A, Măndoiu II (2017) Fast bootstrapping-based estimation of confidence intervals of expression levels and differential expression from RNA-Seq data. Bioinformatics 33:3302–3304

20. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M (2015) Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol 16:150

21. Paczkowska M, Barenboim J, Sintupisut N, et al (2020) Integrative pathway enrichment analysis of multivariate omics data. Nat Commun 11:735

22. (2021) Pathway Analysis: ANOVA vs. Enrichment Analysis. In: Partek Inc. https://www.partek.com/pathway-analysis-anova-vs-enrichment-analysis/. Accessed 25 Apr 2022

23. Ashburner M, Ball CA, Blake JA, et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

24. Kanehisa M (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28:27–30

25. Wu T, Hu E, Xu S, et al (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation 2:100141

26. Subramanian A, Tamayo P, Mootha VK, et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102:15545–15550

27. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20:3710–3715

28. Mootha VK, Lindgren CM, Eriksson K-F, et al (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34:267–273

29. Auwera GA, Carneiro MO, Hartl C, et al (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics. https://doi.org/10.1002/0471250953.bi1110s43

30. Choi Y, Chan AP (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31:2745–2747

31. Community TG, The Galaxy Community, Afgan E, et al (2022) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Research. https://doi.org/10.1093/nar/gkac247

32. Bu D, Luo H, Huo P, et al (2021) KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. Nucleic Acids Research 49:W317–W325

33. Keppler-Noreuil KM, Rios JJ, Parker VER, Semple RK, Lindhurst MJ, Sapp JC, Alomari A, Ezaki M, Dobyns W, Biesecker LG (2015) *PIK3CA* -related overgrowth spectrum (PROS): Diagnostic and testing eligibility criteria, differential diagnosis, and evaluation. American Journal of Medical Genetics Part A 167:287–295

34. Keppler-Noreuil KM, Sapp JC, Lindhurst MJ, et al (2014) Clinical delineation and natural history of the PIK3CA-related overgrowth spectrum. Am J Med Genet A 164A:1713–1733

35. He Y, Sun MM, Zhang GG, Yang J, Chen KS, Xu WW, Li B (2021) Targeting PI3K/Akt signal transduction for cancer therapy. Signal Transduct Target Ther 6:425

36. Website. Reddy, A. T., Lakshmi, S. P., & Reddy, R. C. (2016). PPARγ as a Novel Therapeutic Target in Lung Cancer. PPAR research, 2016, 8972570. https://doi.org/10.1155/2016/8972570.

37. Berger J, Moller DE (2002) The mechanisms of action of PPARs. Annu Rev Med 53:409–435

38. Jiang Y, Zou L, Zhang C, et al (2009) PPARγ and Wnt/β-Catenin pathway in human breast cancer: expression pattern, molecular interaction and clinical/prognostic correlations. Journal of Cancer Research and Clinical Oncology 135:1551–1559

39. Website. Aggarwal, B. B. Nuclear factor-kappaB: the enemy within. Cancer Cell. 6, 203–208, https://doi.org/10.1016/j.ccr.2004.09.003 (2004).

40. Darnay BG, Aggarwal BB (1997) Early events in TNF signaling: a story of associations and dissociations. J Leukoc Biol 61:559–566

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

42. Wu Y, Zhou BP (2010) TNF-α/NF-κB/Snail pathway in cancer cell migration and invasion. British Journal of Cancer 102:639–644

43. Liu W, Lu X, Shi P, Yang G, Zhou Z, Li W, Mao X, Jiang D, Chen C (2020) TNF-α increases breast cancer stem-like cells through up-regulating TAZ expression via the non-canonical NF-κB pathway. Sci Rep 10:1804

44. Kim R, Kaneko M, Arihiro K, Emi M, Tanabe K, Murakami S, Osaki A, Inai K (2006) Extranuclear expression of hormone receptors in primary breast cancer. Ann Oncol 17:1213–1220

45. Yu H (2000) Role of the Insulin-Like Growth Factor Family in Cancer Development and Progression. Journal of the National Cancer Institute 92:1472–1489