

**Identification and Analysis of Key Candidate Genes and Pathways in Lung
Adenocarcinoma by Integrated Bioinformatics**

Prateek Gupta

Leya Joykutty

American Heritage School

Leya Joykutty

Abstract

Lung adenocarcinoma (LUAD) accounts for forty percent of all lung cancer cases in the United States. Treatment currently focuses on surgery and chemotherapy but survival rates remain lower than fifteen percent. The purpose of this investigation is to use an integrated bioinformatics approach to identify key mutated genes in LUAD that are differentially expressed, environmentally responsive, functionally important, and affect survival rates in patients. First, 23,483 genes from nine cancer databases and published studies were identified that are either differentially expressed, mutated, or environmentally responsive. From these, there were 107 significant genes common to all 9 sources used for further analysis. The Database for Annotation, Visualization and Integrated Discovery showed that many of the 107 genes were significantly involved in common processes and pathways, such as DNA replication, DNA repair, ATP binding, and pathways in cancer. Among the 107 genes, the Comparative Toxicogenomics Database found 38,532 gene-environment interactions. A protein-protein interaction network was mapped with the Search Tool for the Retrieval of Interacting Genes/Proteins. This revealed 1116 protein-to-protein interactions among the 107 genes, 25 of which had more than 30 interactions each. Finally, these 25 genes were analyzed with OncoLnc, which resulted in three genes, Checkpoint Kinase 1, Cyclin E1 and Exonuclease 1 (EXO1), that had a significant correlation between increased differential expression in LUAD and worsened patient survival. Among these, EXO1 has been the least studied in LUAD. This study can help focus research on genes likely significant in LUAD and use those genes as targets in gene therapies.

Background

There were an estimated 2.1 million lung cancer cases and 1.8 million deaths globally in 2018, representing almost 1 in 5 (18.4%) of all cancer-related deaths (Bray et al., 2018). In 2020, there were an estimated 228,820 new cases and 135,720 deaths due to lung cancer in the United States (American Cancer Society, 2020). Lung cancer is largely divided into two sub groups, non-small cell lung carcinoma (NSCLC), accounting for 80% of cases, and small cell lung carcinoma, accounting for 20% of cases (Zheng, 2016). Lung adenocarcinoma, a subset of NSCLC, is the most common form of lung cancer, representing about 40% of cases in the U.S. (Myers & Wallen, 2020), and it begins in the glandular cells of the lung (Zheng, 2016). Treatment focuses on surgical resection and definitive radiotherapy, and newer developments like the use of bevacizumab in chemotherapy or external beam therapy have made treatment much more effective. Despite these advancements, lung adenocarcinoma remains deadly with survival rates ranging from 12-15% (Myers & Wallen, 2020).

A key aspect of lung adenocarcinoma that makes it different from other types of lung cancers is that it is prevalent among smokers AND non-smokers, with 10-40% of those diagnosed reported having no smoking history. In fact, the spectrum of mutated genes for smokers vs. non-smokers is largely distinct (Hu & Chen, 2015). While the reason for this is unknown, evidence suggests genetic mutations playing a major influence (Greulich, 2010). Thus, it is important to study gene-environment interactions, which analyze how exposure to different chemicals affect the expression, activity, and other behaviors of a gene. This plays a key role in genomic differentiation in cancer (Liu et al., 2008). Understanding gene environment interactions can allow for the avoidance of such factors, leading to lower incidence for many diseases (Manuck & McCaffery, 2014).

In addition to understanding environmental interactions, disruptions in the genetic makeup of cells itself which leads to cancer is crucial in finding targets for cancer treatments. These mutations can affect protein function, so exploring protein-protein interactions and the functions of these interconnected proteins provides additional insight into understanding cancer processes (Rao et al., 2014). Once significant genes are identified, gene-therapy can be used as a method of treatment for cancer that makes modifications to the human genome to prevent the development of tumors (Gonçalves & Paiva, 2017). Studies have shown that this method shows promise when paired with chemo or radiation therapy (Lara-Guerra & Roth, 2016).

The purpose of this investigation is to: (1) Combine data from multiple cancer-specific databases and literature on genetic mutations to generate a list of genes that are significantly differentially expressed, significantly mutated, and environmentally responsive; (2) Determine protein-protein interactions of these significant genes and identify clusters of important genes; (3) Analyze shared biological functions and pathways for these significant genes; and (4) Assess the correlation between gene expression and survival for significant genes to create a list of genes important for further investigation in lung adenocarcinoma.

It is hypothesized that if data from multiple databases and literature on genetic mutations, expressions, and gene-environment responses in lung adenocarcinoma are combined, then significant genes can be identified. It is also hypothesized that if databases and tools are used to analyze protein-protein interactions, functions, and correlation with patient survival for these genes, then a few key genes can be identified for further research in lung adenocarcinoma.

Methods and Data Analysis

Differentially Expressed Genes and Genetic Mutations from Databases and Literature

First, literature search for studies that used microarray or RNA-sequencing to determine differential expression of genes in lung adenocarcinoma tissue vs normal tissue yielded two selected studies: (Liang et al., 2020; Lim et al., 2018). Additionally, The Lung Cancer Explorer (LCE) database was used to collect additional datasets on differential expression of genes in lung adenocarcinoma (Cai et al., 2019). From both studies and the database, the gene symbols, differential expression values (to determine under vs over-expression), and adjusted p-values (by the Benjamini-Hochberg method) were recorded. Genes in this study were considered significantly differentially expressed if they had an adjusted p-value $< (0.05 / \# \text{ of genes})$, based on applying a Bon Ferroni correction. The intersection of the three lists of significant genes made a singular list of over- and under-expressed genes in lung adenocarcinoma.

Second, literature search for studies that used exome or whole genome sequencing to describe genetic mutations in lung adenocarcinoma patients yielded 3 studies: (Campbell et al., 2016; Collisson et al., 2014; Imielinski et al., 2012). Additionally, The Catalogue of Somatic Mutations in Cancer (COSMIC) database was also used to collect genetic mutation data in lung adenocarcinoma. Gene symbols, type of mutation, and adjusted p-values were recorded, and the same significance threshold was applied as described for differential expression above. The intersection of the 4 lists of significant mutated genes was found to make up a singular list of genes with mutations in lung adenocarcinoma.

Gene-Environment Interaction from Databases

The Comparative Toxicogenomics Database (CTD) was used to get lists of chemical-gene interactions in lung adenocarcinoma and the number of published references supporting

each interaction. Interactions with fewer than 10 references were discarded. The Environmental Genome Project (EGP) was used to get a list of environmentally responsive genes. The intersection of these lists made a list of environmentally responsive genes in adenocarcinoma.

Determining Curated List of Significant Genes (Referred to as curated genes)

The intersection/overlap of all three significant gene lists from above were found to create a list of statistically significantly mutated, differentially expressed and environmentally responsive genes in lung adenocarcinoma. Using this list, a list of significant under expressed genes and a list of significant over expressed were found using the expression data collected (a positive differential expression value indicated an over expression while a negative represented an under expression).

Analyzing Gene-Environment Interactions for Curated Significant Genes

The curated list of genes was inputted into the chemical associations tool on CTD to find a list of environmental effects on each of the genes. The gene symbol, chemical name, interaction effect and number of published references supporting the interaction were recorded. An interaction was deemed significant if it had 10 or more supporting published references.

Analyzing the Gene Ontology and Pathways for Curated Genes

The Database for Annotation, Visualization, and Integrated Discovery (DAVID) was used to find the biological processes, cellular components, molecular functions, and signaling pathways that were shared among the curated genes. The background and annotation were set to Homo Sapiens. The curated list of under expressed genes and the curated list of over expressed genes was inputted into the DAVID tool in the same way as the original curated list of genes. The tool provides a modified Fisher's Exact test for significance of gene-enrichment.

Analyzing Protein to Protein Interactions for Curated Significant Genes

The original list of curated genes was entered into the Search Tool for the Retrieval of Interacting Genes/Proteins v11.0b (STRING) to find the protein-to-protein interactions, which is based on 7 types of evidence (fusion, neighborhood, co-occurrence, experimental, text-mining, database, and co-expression). This generated a protein-to-protein interaction map of the curated list of genes. The output file was then exported into the Molecular Complex Detection (MCODE) plugin in Cytoscape v3.8.2 for visualization and clustering. Key “hub” genes with 30 or more protein to protein interactions were identified. Using the MCODE plugin based on graph theory methods, genes were clustered together into two subsets of densely connected regions (Modules 1 and 2) in the large protein network. DAVID functional analysis was conducted as described above for Module 1 and Module 2.

Analyzing the Effect of Gene Expression on Patient Survival for “Hub” Genes

The hub genes, genes that had 30 or more protein-to protein interactions, isolated from Cytoscape were entered into the tool OncoLnc to collect data on the relationship between Lung adenocarcinoma patient survival and gene expression. This revealed a cox coefficient for gene expression (OncoLnc calculated this using a multivariate Cox regression with age, sex, and gene expression as the input variables) where a positive coefficient corresponds with a worse survival rate with higher levels of gene expression while a negative coefficient corresponds with a better prognosis with higher levels of gene expression. It also provided a False Discovery Rate (FDR) corrected p-value. The Kaplan-Meier survival curves for genes that had an FDR of below 0.05 were graphed inside of OncoLnc. Patients were subclassified into the low-expression group and high-expression if median expression was in bottom 25 percentile or top 25 percentile.

RESULTS

Differentially Expressed Genes and Mutations from Databases & Literature

Overall, there were a total of 23,483 unique genes found from the 7 lung adenocarcinoma databases/studies that were either mutated or differentially expressed in lung adenocarcinoma. The distribution of significantly differentially expressed genes from each data set is shown in Table 1, and significantly mutated genes in Table 2.

Number of Genes	Lung Cancer Explorer	Lim 2018	Liang 2020
Over-Expressed	3387	3453	3428
Under-Expressed	2221	2809	3491
Total	5608	6262	6919

Table 1. *Number of Significantly Over- and Under-Expressed Genes in Lung Adenocarcinoma.*

There are 1,637 common significantly differentially expressed genes among these 3 data sets.

	Campbell 2016	COSMIC	Imielinski 2012	Collison 2014
Number of Mutated Genes	421	16383	243	757

Table 2. *Number of Significantly Mutated Genes in Lung Adenocarcinoma.* There are 300 common significantly mutated genes among these 4 data sets.

Gene-Environment Interaction from Databases

From CTD and EGP, the databases about environmental interactions, there were 21,441 genes with gene-chemical interaction. After discarding those with fewer than 10 references to support the interaction in CTD, there were 583 common environmentally responsive genes.

Curated List of Significant Genes (i.e., curated genes)

The intersection of all 9 data sets results in 107 genes that were significantly differentially expressed, mutated, and environmentally responsive (Table 3). This set of 107 significant genes, hereafter referred to as “curated genes”, was used for further analysis.

	Gene Symbol
Over-Expressed Curated Genes (64)	ABCC1, BLM, BRAF, BRCA1, CAPN1, CASP8, CCNE1, CCNT2, CDC6, CDH1, CDK4, CDKN2A, CHEK1, CHEK2, CP, DCLRE1A, DCLRE1B, DCLRE1C, E2F3, EGF, EGFR, ERBB2, ERCC4, EXO1, FANCA, FANCD2, FDXR, FGF13, GPX1, HK2, HSPA6, LIG1, MCM2, MCM3, MCM4, MCM6, MDM2, MDM4, MMP1, MMP2, MMP9, MN1, MSH2, MSH6, MYC, NF1, NRAS, OXR1, PLA2G4A, PMS1, POLD1, PRKDC, RECQL4, RUNX2, SLC4A2, TAF1, TDG, TDP1, TFDP1, TIMELESS, TP53, TRPM2, WRN, XDH
Under-Expressed Curated Genes (43)	ADH1B, APP, ATM, ATRX, BIRC3, CCND1, CCND3, CDC14A, CDC42BPB, CYP4V2, DAPK1, EDNRB, ERCC5, ESR1, FGFR1, FGFR2, FGFR3, FGFR4, FMO2, FMO3, GAB1, HSPA2, IGF1, IGF1R, IGF2R, ITGAL, MCL1, MSR1, MTHFR, NFKB1, NR3C1, POLI, PRKD1, PTEN, PTGS2, RB1, REV3L, SOD2, TAF1C, TGFB1, TGFB2, TGFB3, TJP1

Table 3. Curated List of 107 Differentially Expressed, Mutated and Environmentally Responsive Genes in Lung Adenocarcinoma.

Gene-Environment Interactions for Curated Significant Genes

The curated genes were then inputted into the Comparative Toxicogenomic Database (CTD) to find the gene environment interactions for each of the 107 curated genes. 38,532 gene-environment interactions were found, and a sample of the top 5 is provided in Table 4.

Gene Symbol	Chemical Name	Interaction	References
ESR1	Estradiol	Increases activity	179
ESR1	bisphenol A	Affects binding & increases activity	76
MYC	(+)-JQ1 compound	Decreases expression	72
CCND1	Estradiol	Increases expression	52
ESR1	Estradiol	Decreases expression	50

Table 4. Sample of Gene-Environment Interactions found in Mutated, Differentially Expressed, and Environmentally Affected Genes in Lung Adenocarcinoma.

Gene Ontology and Pathways for Curated Genes

The curated gene list was inputted into the Database for Annotation, Visualization, and Integrated Discovery (DAVID), for functions and pathways these are collectively enriched in.

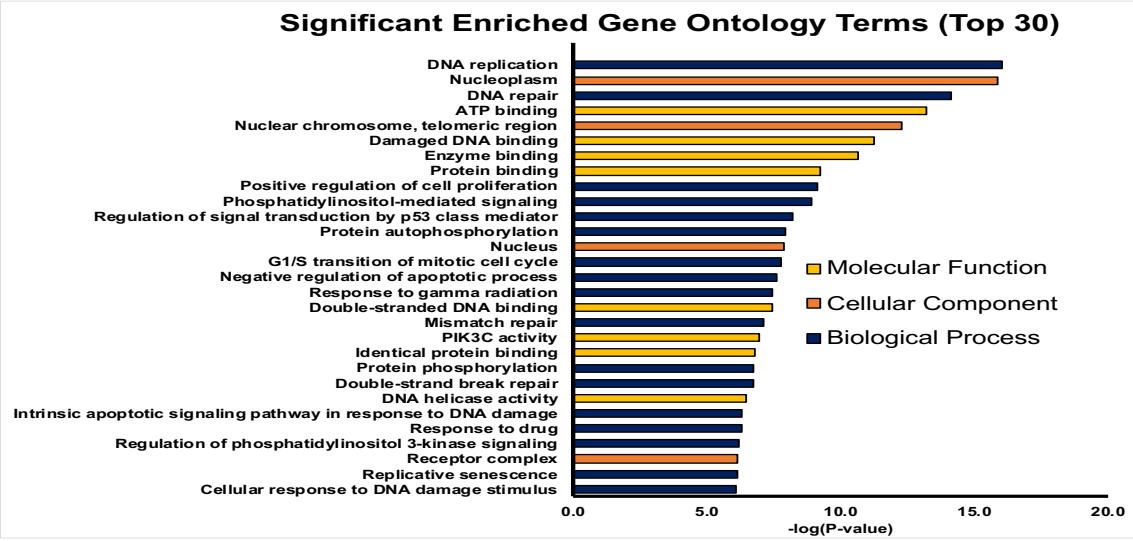


Figure 1. *Significant Enriched Gene Ontology Terms (Top 30).* This bar graph shows the top 30 most significantly shared/enriched gene ontology terms (biological processes, cellular components and molecular functions) for the set of 107 curated genes.

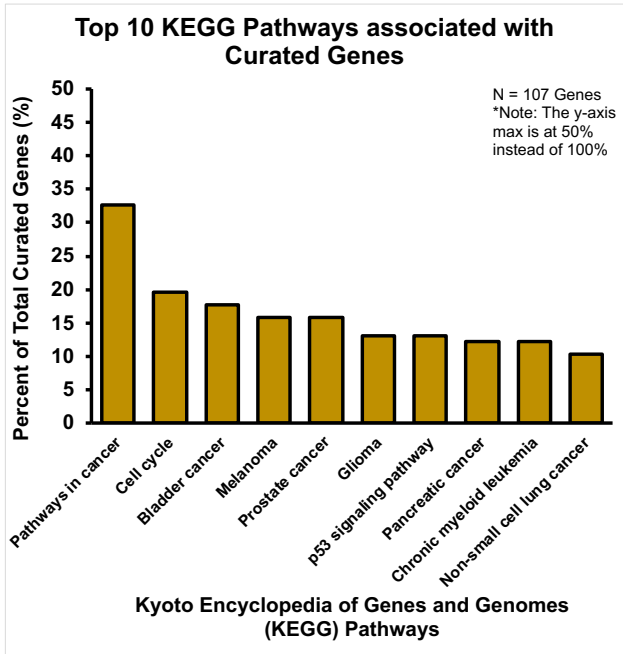


Figure 2. *Top 10 Pathways Enriched for the Curated Genes.*

After analyzing the attributes that all 107 genes affected, the list of over-expressed genes and under expressed genes were separately analyzed with DAVID to see the enrichment of biological processes, cellular components, molecular functions, and

pathways in over-expressed (Table 5) vs under-expressed (Table 6) genes.

<i>Type of Gene Ontology</i>	<i>Term</i>	<i>% of Genes</i>	<i>P-value</i>
Molecular Functions	damaged DNA binding	15.625	1.35E-12
Molecular Functions	ATP binding	39.0625	1.61E-10
Molecular Functions	double-stranded DNA binding	12.5	1.79E-08
Cellular Processes	nucleoplasm	56.25	8.86E-14
Cellular Processes	nuclear chromosome, telomeric region	18.75	5.91E-13
Cellular Processes	nucleolus	23.4375	8.22E-07
Biological Processes	DNA replication	23.4375	4.42E-16
Biological Processes	DNA repair	23.4375	1.53E-13
Biological Processes	regulation of signal transduction by p53 class mediator	15.625	8.24E-10
KEGG Pathways	Bladder cancer	23.4375	1.21E-20
KEGG Pathways	Cell cycle	25	1.03E-14
KEGG Pathways	Pathways in cancer	31.25	3.09E-11

Table 5. *Gene Ontology of Overexpressed Curated Genes.* This table represents the three most significant processes for each of the four functional categories in DAVID.

<i>Type of Gene Ontology</i>	<i>Function</i>	<i>% of Genes</i>	<i>P-value</i>
Molecular Functions	1-phosphatidylinositol-3-kinase activity	13.9534884	6.69E-08
Molecular Functions	fibroblast growth factor-activated receptor activity	9.30232558	1.43E-07
Molecular Functions	fibroblast growth factor binding	11.627907	2.83E-07
Cellular Processes	receptor complex	13.9534884	1.05E-05
Cellular Processes	cell surface	18.6046512	2.16E-04
Cellular Processes	nucleoplasm	39.5348837	2.64E-04
Biological Processes	phosphatidylinositol-mediated signaling	18.6046512	7.35E-09
Biological Processes	phosphatidylinositol-3-phosphate biosynthetic process	13.9534884	1.35E-07
Biological Processes	protein autophosphorylation	18.6046512	2.09E-07
KEGG Pathways	Pathways in cancer	34.8837209	5.73E-09
KEGG Pathways	Prostate cancer	18.6046512	3.32E-07
KEGG Pathways	FoxO signaling pathway	18.6046512	5.75E-06

Table 6. *Gene Ontology of Under Expressed Curated Genes.* This table represents the three most significant processes for each of 4 functional categories in DAVID.

Protein to Protein Interactions for Curated Significant Genes

Using STRING and MCODE in Cytoscape, the protein-to-protein interactions (PPI) were analyzed for the curated set of 107 genes. STRING outputted 103 proteins with interactions (and 4 that did not –CYP4V2, ADH1B, SLC4A2, FMO3). There were a total of 1116 protein-to-protein interactions, resulting in an average of 20.9 interactions per protein.

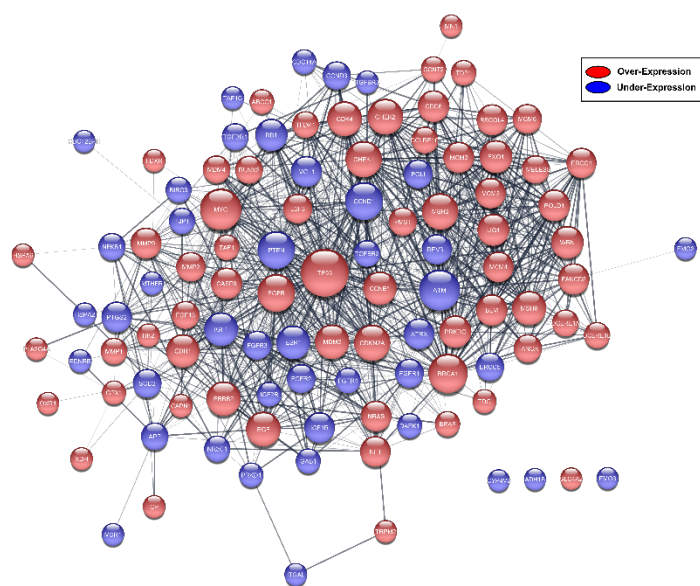


Figure 3. *Network of Protein-to-Protein Interactions for Curated Genes.*

Each circle represents a protein; larger size corresponds to more interactions.

Each line represents a protein-to-protein interaction; thicker lines correspond to more evidence for the interaction. Proteins > 30 interactions were called key “hub” genes. Using the

MCODE plugin, the genes clustered into module 1 and module 2.

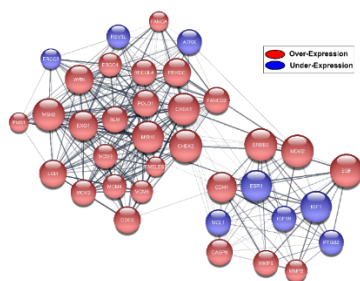


Figure 4. *Network of Protein-to-Protein Interactions for Module 1 Genes.* Module 1 had 36 proteins, 310 protein-to-protein interactions, and average of 17.2 interactions/protein.

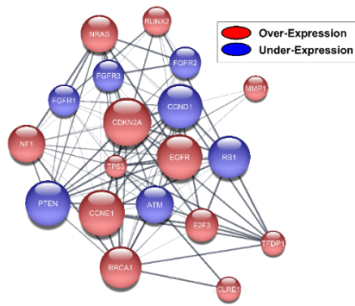


Figure 5. Network of Protein-to-Protein Interactions for Module 2 Genes. Module 2 cluster had 19 proteins, 112 protein-to-protein interactions, and average of 11.8 interactions.

Module 1 and Module 2 genes were then analyzed with DAVID for functional enrichment (Table 7 and 8).

<i>Description</i>	<i>Percent</i>	<i>P-Value</i>
DNA replication (Biological Process)	41.7	3.48E-20
nucleoplasm (Cellular Component)	69.4	1.78E-12
ATP binding (Molecular Function)	50.0	1.12E-09
Cell cycle (KEGG Pathway)	25.0	3.73E-08
Pathways in cancer (KEGG Pathway)	33.3	3.75E-07
DNA replication (KEGG Pathway)	16.7	3.81E-07

Table 7. Gene Ontology for Module 1 Genes from DAVID. This table represents the most significant process for the 4 functional categories in Module 1.

<i>Description</i>	<i>Percent</i>	<i>P-Value</i>
Ras protein signal transduction (Biological Process)	26.3	8.11E-07
nucleoplasm (Cellular Component)	68.4	1.19E-06
fibroblast growth factor-activated receptor activity (Molecular Function)	15.8	1.07E-05
Bladder cancer (KEGG Pathway)	47.4	3.24E-14
Prostate cancer (KEGG Pathway)	52.631579	2.67E-13
Melanoma (KEGG Pathway)	47.368421	3.46E-12

Table 8. Gene Ontology for Module 2 Genes from DAVID. This table represents the most significant process for the 4 functional categories in Module 2.

Effect of Gene Expression on Patient Survival for “Hub” Genes

Using data produced from STRING and the MCODE plugin, 25 hub genes were determined (>30 protein interactions). Using OncoLnc, which contains gene expression and patient survival data, the prognostic value of each gene was analyzed (Table 9).

Gene	mRNA Expression (Tumor vs Normal)	# of protein-protein interactions (STRING)	Cox coefficient (OncoLnc)	FDR-adjusted p- value (OncoLnc)
TP53	Over	82	-0.0682	0.6092
MYC	Over	61	0.0999	0.4102
ATM	Under	60	-0.0906	0.4822
BRCA1	Over	54	0.1771	0.1120
CCND1	Under	50	0.0899	0.4935
EGFR	Over	49	0.0999	0.4356
PTEN	Under	49	-0.0359	0.8033
CHEK1	Over	45	0.3362	0.0045**
MSH2	Over	45	0.0643	0.6268
CDKN2A	Over	45	0.0564	0.6763
CHEK2	Over	43	0.0109	0.9383
CDK4	Over	40	0.0503	0.7061
EGF	Over	40	-0.0269	0.8546
ERBB2	Over	39	-0.0733	0.5911
MDM2	Over	39	0.0322	0.8292
MSH6	Over	38	0.1301	0.2765
IGF1	Under	37	-0.0474	0.7205
ESR1	Under	37	-0.0292	0.8497
CCNE1	Over	36	0.2136	0.0496**
CDH1	Over	35	0.0009	0.9933
EXO1	Over	34	0.3242	0.0045**
WRN	Over	33	0.0126	0.9383
PRKDC	Over	32	0.1860	0.1037
RB1	Under	32	-0.0683	0.6092
LIG1	Over	31	0.1267	0.2880

Table 9. Protein Interactions and Survival Analysis of the 25 Hub Genes (>30 protein

interactions). A positive Cox coefficient indicates that a gene with higher expression correlates with worse survival; the opposite is true for negative coefficient. FDR-adjusted p-value gives a significance of the gene in a Cox regression model accounting for gene expression, age, and sex.

3 of the 25 genes had a statistically significant correlation between their mRNA expression level and survival status and are all over-expressed in lung adenocarcinoma. Higher expression levels of checkpoint kinase 1 (CHEK1), exonuclease 1 (EXO1), and cyclin E1

(CCNE1) were correlated with poor prognosis of patients with lung adenocarcinoma. Using OncoLNC, the Kaplan-Meier Survival curves were graphed for EXO1, CHEK1, and CCNE1.

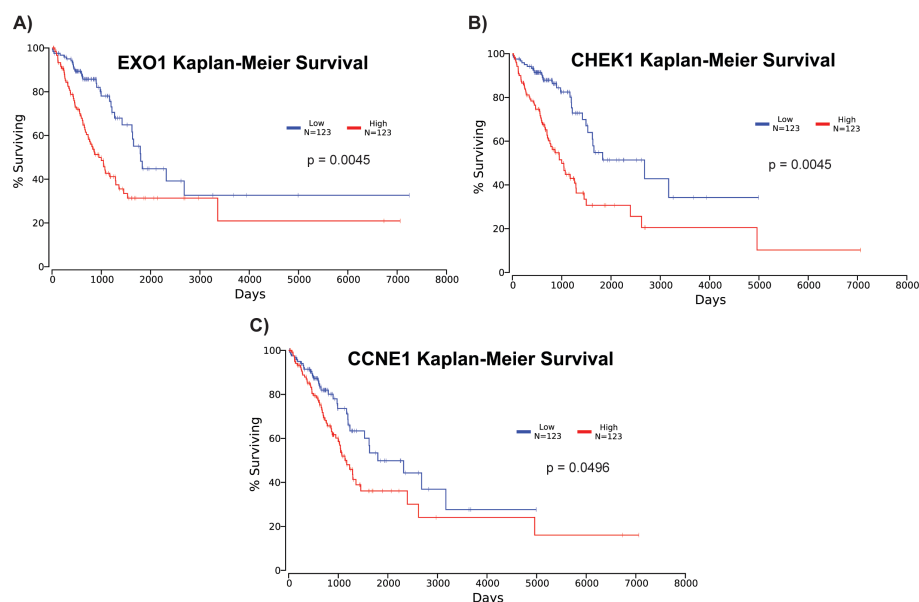


Figure 6. *Kaplan-Meier Survival Curves for Significant Hub Genes.* A. EXO1 B. CHEK1 C. CCNE1. These show that over-expression of these genes results in lower survival.

Conclusion

107 significantly mutated, differentially expressed and environmentally responsive genes in lung adenocarcinoma using databases and publication were identified. Of these 107 curated genes, 64 were overexpressed and 43 were under expressed (Table 3 & Figure 1). Key genes were identified by integrating functional analysis, protein-protein interactions, and survival analysis, which include the 25 hub genes. 3 were significantly associated with patient survival.

Gene-environment interactions of all 107 curated genes were identified with Comparative Toxicogenomic Database (CTD) (Table 4). The interaction between Estrogen Receptor 1 (ESR1) and Estradiol, an estrogen hormone, had the most supporting references (179). ESR1 is often seen mutated in many breast cancers. Since ESR1 is under expressed in lung adenocarcinoma,

estradiol could be used to increase ESR1 activity in patients. A similar analysis for any of the gene-environment interactions can determine a positive or negative effect for each chemical.

The functional analysis in DAVID (Figures 1-2) show that the two most common biological process the curated genes affected was DNA repair and the regulation of cell proliferation both (17.76% of genes). Since DNA repair and regulation of cell proliferation are important regulated processes in cancer, the curated genes may be significant to cancer development. Tumor must interrupt DNA repair to generate harmful mutations, and tumors bypass regulation of cell cycle, such as the G1 checkpoint, to continue proliferating. The most common cellular component enriched in the curated genes was the nucleus (56.07% of genes), which is important because tumors interfere with DNA replication and transcription. The most common molecular function enriched in the curated genes was protein binding (80.37% of genes), which is important because interference in protein binding results in dysregulation of cell signals. The most common KEGG Pathway affected was pathways in cancer. Over-expressed genes were enriched in DNA regulation processes, such as damaged DNA binding, DNA replication, mismatch repair and the nucleus (Table 5). Under-expressed genes were enriched in protein binding and apoptosis (Table 6), so tumors can evade cell death.

25 hub genes from STRING and 2 modules (clusters) from MCODE were identified. To check if the mathematical separation of the modules was functionally significant, enrichment for gene ontology and KEGG pathway in DAVID were performed. A gene was considered a hub gene if it had > 30 protein-protein interactions, so it was highly interconnected with other genes.

The survival rates of patients with high expressions versus low expressions for each of the 25 hub genes was analyzed in OncoLnc (Table 9). Exonuclease 1 (EXO1), Checkpoint Kinase 1 (CHEK1), Cyclin E1 (CCNE1) all had adjusted p-values < 0.05, meaning that there was

a significant difference in prognosis of patients with high vs low expression of the gene. CHEK1 and CCNE1 have been commonly associated with other tumors but EXO1 is not as commonly known to have any effect (Wu et al., 2019; Zhao et al., 2019). EXO1 encodes for a protein involved in DNA repair pathways, specifically on mismatch repairs and recombination (Myler et al., 2016). In lung adenocarcinoma, the mutated EXO1 gene was overexpressed indicating that it is most likely an oncogene (Table 9). This is further proven by the Kaplan-Meier survival curve that shows higher expression of the gene results in worse patient survival (Figure 6). EXO1 is also highly involved with other proteins, having 34 protein-to-protein interactions (Table 9), indicating mutations of this gene can affect many other proteins in cancer processes (Table 9).

Applications

The first application is that the curated list of genes as well as the module 1 and 2 lists of genes can be targeted for gene therapy. They have been shown to be statistically significantly mutated and differentially expressed in lung adenocarcinoma, so targeted gene therapy either through vesicle gene replacement or gene knockdown could be applied to treat the tumor. The second application is that an analysis of gene ontologies and processes can be used by determine when gene therapy would be best suited. For example, gene therapy that targeted genes that allow for proliferation would be most beneficial in the early stage to prevent the tumor from growing. The third application is the use of gene-environment interactions to find potential treatment. If chemicals that increase activity in genes that are under expressed, those can be used to develop medication to prevent the under expression necessary for the tumor to survive.

Future Research

Future research would primarily be composed of conducting experiments in vitro on the effects changing each gene deemed significant. A first test that could be conducted would be to

perform a gene knockdown or knockout on one of the overexpressed or mutated gene to see what effects it has on proliferation, cell migrations, apoptosis, and other cellular processes important for cancer. A second test that could be conducted would be to use extracellular vesicles to deliver replacement genes for the mutated genes in the lung adenocarcinoma cells and then analyzing the effect on cell viability and proliferation and metastasis of the tumor. A third test would be on the effect of different chemicals on a culture could be conducted. This would be used to see if using a chemical to reverse the effects of the oncogene could work, for example using a chemical that reduces expression on an over expressed gene. All of these tests would eventually lead to a development of new treatment for lung adenocarcinoma. Lastly, an analysis using the same methods can be conducted for other genetic diseases to allow for understandings of the genetic makeup and developments of treatments for different diseases.

Limitations

The first major limitation of this study was that everything is based on computational associations and correlations so further experiments are required to make definitive claims about causation. The second major limitation was the inability to test each of the gene's effects in vitro so more data on the true effects of each gene's involvement. Some unintended effects or unforeseen consequences may arise as a result of tampering in the genetic makeup of the cell. The third major limitation is that although DAVID provided info about what processes and pathways each gene was involved in; it doesn't explain how it affects each pathway and whether the effect was different in lung adenocarcinoma versus normal glandular cells.

Bibliography

American Cancer Society. (2020). Key Statistics for Lung Cancer. *American Cancer Society*.
Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global

cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*.

<https://doi.org/10.3322/caac.21492>

Cai, L., Lin, S. Y., Girard, L., Zhou, Y., Yang, L., Ci, B., Zhou, Q., Luo, D., Yao, B., Tang, H., Allen, J., Huffman, K., Gazdar, A., Heymach, J., Wistuba, I., Xiao, G., Minna, J., & Xie, Y. (2019). LCE: an open web portal to explore gene expression and clinical associations in lung cancer. *Oncogene*. <https://doi.org/10.1038/s41388-018-0588-2>

Campbell, J. D., Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., Shukla, S. A., Guo, G., Brooks, A. N., Murray, B. A., Imielinski, M., Hu, X., Ling, S., Akbani, R., Rosenberg, M., Cibulskis, C., Ramachandran, A., Collisson, E. A., Kwiatkowski, D. J., ... Meyerson, M. (2016). Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics*. <https://doi.org/10.1038/ng.3564>

Collisson, E. A., Campbell, J. D., Brooks, A. N., Berger, A. H., Lee, W., Chmielecki, J., Beer, D. G., Cope, L., Creighton, C. J., Danilova, L., Ding, L., Getz, G., Hammerman, P. S., Hayes, D. N., Hernandez, B., Herman, J. G., Heymach, J. V., Jurisica, I., Kucherlapati, R., ... Cheney, R. (2014). Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature*, *511*(7511), 543–550. <https://doi.org/10.1038/nature13385>

Gonçalves, G. A. R., & Paiva, R. de M. A. (2017). Gene therapy: advances, challenges and perspectives. In *Einstein (Sao Paulo, Brazil)* (Vol. 15, Issue 3, pp. 369–375). Instituto de Ensino e Pesquisa Albert Einstein. <https://doi.org/10.1590/S1679-45082017RB4024>

Greulich, H. (2010). The genomics of lung adenocarcinoma: Opportunities for targeted therapies.

Genes and Cancer, 1(12), 1200–1210. <https://doi.org/10.1177/1947601911407324>

Hu, Y., & Chen, G. (2015). Pathogenic mechanisms of lung adenocarcinoma in smokers and non-smokers determined by gene expression interrogation. *Oncology Letters*.
<https://doi.org/10.3892/ol.2015.3462>

Imielinski, M., Berger, A. H., Hammerman, P. S., Hernandez, B., Pugh, T. J., Hodis, E., Cho, J., Suh, J., Capelletti, M., Sivachenko, A., Sougnez, C., Auclair, D., Lawrence, M. S., Stojanov, P., Cibulskis, K., Choi, K., De Waal, L., Sharifnia, T., Brooks, A., ... Meyerson, M. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6), 1107–1120. <https://doi.org/10.1016/j.cell.2012.08.029>

Lara-Guerra, H., & Roth, J. A. (2016). Gene therapy for lung cancer. *Critical Reviews in Oncogenesis*, 21(1–2), 115–124. <https://doi.org/10.1615/CritRevOncog.2016016084>

Liang, J., Li, H., Han, J., Jiang, J., Wang, J., Li, Y., Feng, Z., Zhao, R., Sun, Z., Lv, B., & Tian, H. (2020). Mex3a interacts with LAMA2 to promote lung adenocarcinoma metastasis via PI3K/AKT pathway. *Cell Death and Disease*, 11(8). <https://doi.org/10.1038/s41419-020-02858-3>

Lim, S. Bin, Tan, S. J., Lim, W. T., & Lim, C. T. (2018). A merged lung cancer transcriptome dataset for clinical predictive modeling. *Scientific Data*.
<https://doi.org/10.1038/sdata.2018.136>

Liu, L., Li, Y., & Tollefsbol, T. O. (2008). Gene-environment interactions and epigenetic basis of human diseases. In *Current Issues in Molecular Biology* (Vol. 10, Issue 1, pp. 25–36). NIH Public Access. <https://doi.org/10.21775/cimb.010.025>

Manuck, S. B., & McCaffery, J. M. (2014). Gene-environment interaction. In *Annual Review of Psychology* (Vol. 65, pp. 41–70). Annual Reviews Inc. <https://doi.org/10.1146/annurev->

psych-010213-115100

Myers, D. J., & Wallen, J. M. (2020). Cancer, Lung, Adenocarcinoma. In *StatPearls*. StatPearls Publishing. <http://www.ncbi.nlm.nih.gov/pubmed/30137862>

Myler, L. R., Gallardo, I. F., Zhou, Y., Gong, F., Yang, S. H., Wold, M. S., Miller, K. M., Paull, T. T., & Finkelstein, I. J. (2016). Single-molecule imaging reveals the mechanism of Exo1 regulation by single-stranded DNA binding proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 113(9), E1170–E1179. <https://doi.org/10.1073/pnas.1516674113>

Rao, V. S., Srinivas, K., Sujini, G. N., & Kumar, G. N. S. (2014). Protein-Protein Interaction Detection: Methods and Analysis. *International Journal of Proteomics*, 2014, 1–12. <https://doi.org/10.1155/2014/147648>

Wu, M., Pang, J.-S., Sun, Q., Huang, Y., Hou, J.-Y., Chen, G., Zeng, J.-J., & Feng, Z.-B. (2019). The clinical significance of CHEK1 in breast cancer: a high-throughput data analysis and immunohistochemical study. *International Journal of Clinical and Experimental Pathology*, 12(1), 1–20. <http://www.ncbi.nlm.nih.gov/pubmed/31933717>

Zhao, Z. M., Yost, S. E., Hutchinson, K. E., Li, S. M., Yuan, Y. C., Noorbakhsh, J., Liu, Z., Warden, C., Johnson, R. M., Wu, X., Chuang, J. H., & Yuan, Y. (2019). CCNE1 amplification is associated with poor prognosis in patients with triple negative breast cancer. *BMC Cancer*, 19(1), 96. <https://doi.org/10.1186/s12885-019-5290-4>

Zheng, M. (2016). Classification and Pathology of Lung Cancer. *Surgical Oncology Clinics of NA*, 25, 447–468. <https://doi.org/10.1016/j.soc.2016.02.003>