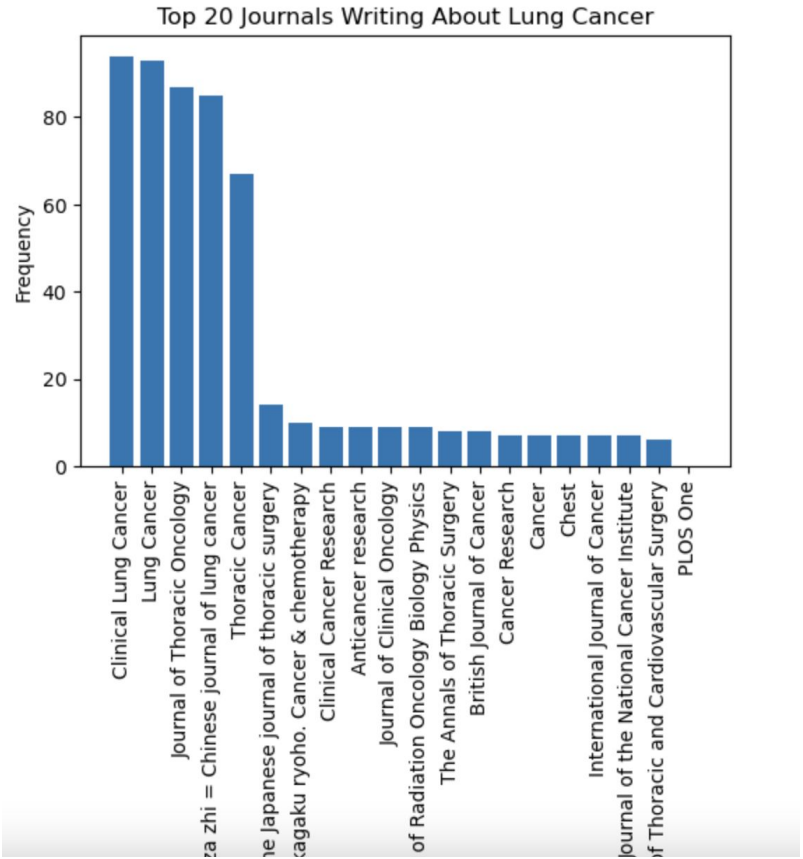
Finding Novel Gene-Disease Associations

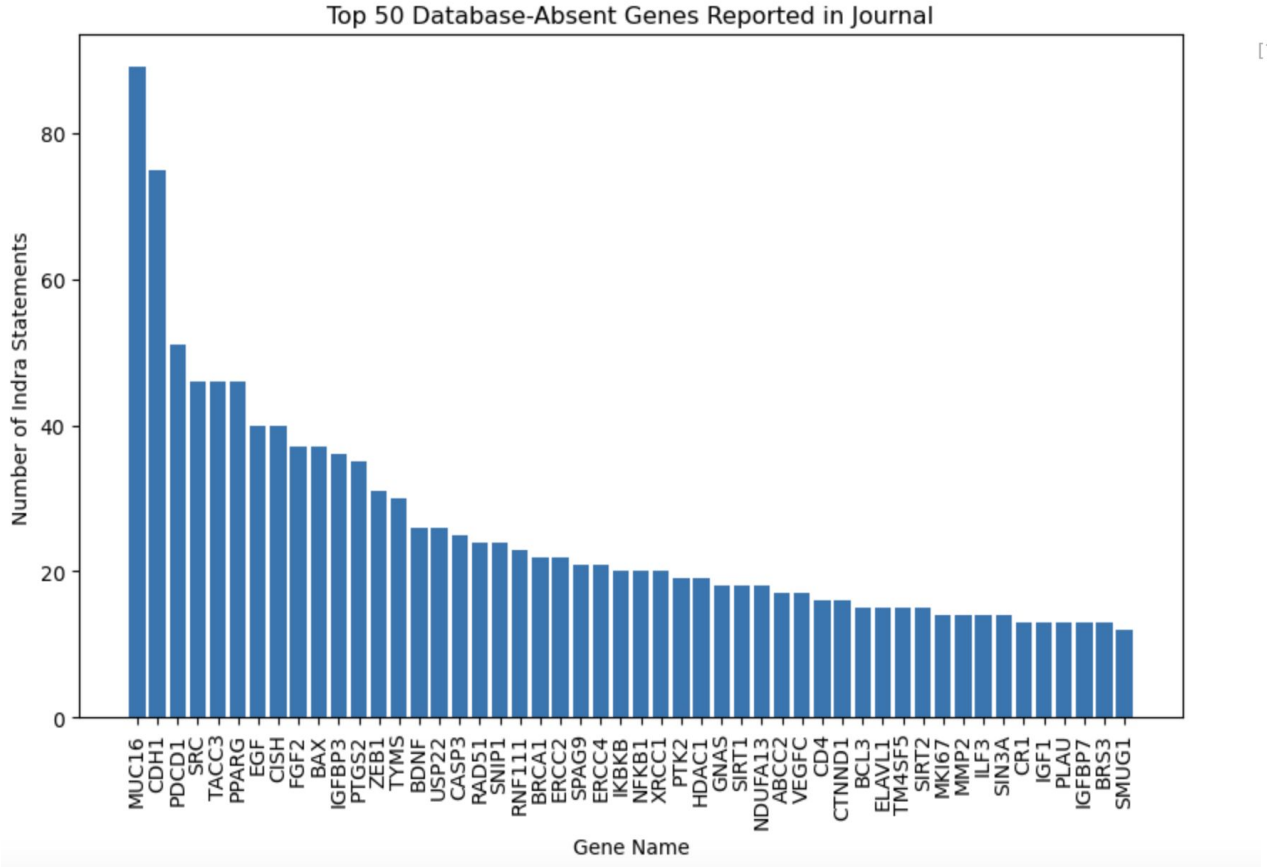
By: Aria Agarwal

Find Top Journals Researching Lung Cancer



	level_0	index	name	id	Total Pub	Lung Pub	Ratio
0	0	0	Clinical Lung Cancer	nlm:100893225	2169	2043	94
1	1	1	Lung Cancer	nlm:8800805	6772	6302	93
2	2	2	Journal of Thoracic Oncology	nlm:101274235	5522	4843	87
3	3	3	Zhongguo fei ai za zhi = Chinese journal of lu...	nlm:101126433	1995	1707	85
4	4	4	Thoracic Cancer	nlm:101531441	2486	1686	67
5	5	5	Kyobu geka. The Japanese journal of thoracic s...	nlm:0413533	12377	1788	14
6	6	6	Gan to kagaku ryoho. Cancer & chemotherapy	nlm:7810034	21263	2211	10
10	7	9	Anticancer research	nlm:8102988	24442	2218	9
9	8	10	Clinical Cancer Research	nlm:9502500	20263	1887	9
7	9	8	International Journal of Radiation Oncology Bi...	nlm:7603616	24320	2276	9
8	10	7	Journal of Clinical Oncology	nlm:8309333	25992	2454	9
11	11	11	The Annals of Thoracic Surgery	nlm:15030100R	39278	3314	8
12	12	12	British Journal of Cancer	nlm:0370635	24852	2031	8
13	13	13	Cancer Research	nlm:2984705R	53303	3832	7
14	14	14	Cancer	nlm:0374236	43649	3414	7
15	15	15	Chest	nlm:0231335	36945	2917	7
16	16	16	International Journal of Cancer	nlm:0042124	26286	1925	7
17	17	17	Journal of the National Cancer Institute	nlm:7503089	22744	1673	7
18	18	18	The Journal of Thoracic and Cardiovascular Sur...	nlm:0376343	30956	2131	6
19	19	19	PLOS One	nlm:101285081	278162	2346	0

Finding Novel Genes Using for DisGeNET using INDRA Database



[7]:

	name	gene_id	indra statements
18	MUC16	hgnc:15582	89
22	CDH1	hgnc:1748	75
34	PDCD1	hgnc:8760	51
40	SRC	hgnc:11283	46
41	TACC3	hgnc:11524	46
43	PPARG	hgnc:9236	46
49	EGF	hgnc:3229	40
50	CISH	hgnc:1984	40
58	FGF2	hgnc:3676	37
59	BAX	hgnc:959	37

-Got a list of the top 50 genes that were mentioned in the journal but did not appear in the DisGenet curated dataset

-Investigated genes by reading relevant INDRA statements and manually searching through scientific literature (Pubmed) to determine lung cancer association

Results From Database

Gene Name	ID	# INDRA statement	Association Based on Literature	Associated based on INDRA statements	Is it listed in DisgeNET for lung cancer?	Source the association comes from
MUC16	hgnc: 15582	89	-MUC16 mutations lead to overexpression -> leads to lung cancer	Doesn't mention lung cancer, but does mention pancreatic cancer which is also not in the database	Yes	BEFREE, LHGDN
CDH1	hgnc: 1748	75	-reduced expression of this gene leads to lung cancer	Statements mentions lung cancer	yes	Animal Models, BEFREE, LHGDN, RGD
PDCD1	hgnc: 8760	51	-decreased expression, higher chance of lung cancer -also therapeutic target	Did not mention lung cancer, but does mention tumors	Yes	BEFREE
SRC	hgnc: 11283	46	-promotes metastasis of lung cancer through overexpression	Does mention lung cancer	yes	BEFREE
TACC3	hgnc: 11524	46	Gene can be a prognostic marker for lung cancer	Does not mention lung cancer	Yes	BEFREE
PPARG	hgnc: 9236	46	Increased expression decreases risk of developing lung cancer, being looked into as a target to treat	Does mention lung cancer	yes	BEFREE, LGHDN

Analysis

-Seemingly found many genes that were missing from the INDRA database that were related to lung cancer

- However, upon looking at the processor code, found that a curated dataset for gene-disease associations was being used, which explains why many genes were missing

- The curated dataset only includes gene-disease associations that Disgenet classifies as "strong" based on their GDA scores and other statistical factors

Finding Novel Genes Using the DisGeNET Database

Gene Name	Association	In Disgenet?	Articles	In New Disgenet?
CD4	Yes	No	https://pubmed.ncbi.nlm.nih.gov/23384671/ -> published 2013	yes
SRBD1	Yes	No	https://pubmed.ncbi.nlm.nih.gov/32010555/ -> published 2019	yes
TKT	Yes	No	https://pubmed.ncbi.nlm.nih.gov/35711845/ -> published 2022	yes
CTNNA2	Yes	No	https://pubmed.ncbi.nlm.nih.gov/34163353/ -> published 2021	yes
SETD3	supports association for SETD3-ALK fusion gene	No	https://pubmed.ncbi.nlm.nih.gov/36495785/ -> published 2022	-> fusion gene not listed in database
FSIP2	supports association for FSIP2-ALK fusion gene	No	https://pubmed.ncbi.nlm.nih.gov/33419583/ -> published 2021	-> fusion gene not listed in database
NBEA	Yes - NBEA-ALK fusion gene	No	https://pubmed.ncbi.nlm.nih.gov/34763158/ -> published in 2021	No
KLHDC2	Minimal	No	protein atlas	No
LIG3	Yes	No	https://pubmed.ncbi.nlm.nih.gov/17108146/ -> 2006	No
ETV6	Minimal	comment	research gate link	Yes - under ETV-NTRK3 fusion gene
FH	Yes	No	https://pubmed.ncbi.nlm.nih.gov/34737838/ -> 2021	Yes

- Using the older DisGeNET database, was able to find many genes missing from the database that had scientific literature proving associations, mostly published after 2021

- However upon finding new database, many of the genes had been added after they updated the dataset

- Still 4 genes were missing, 3 fusion genes (SETD3-ALK, FSIP2-ALK, AND NBEA-ALK), and LIG3