COPY OF MY MASTERS (MS) THESIS AS AVAILABLE IN LIBRARY OF ARIZONA STATE UNIVERSITY, U.S.: Targeted BEL network representation and characterization of commonly mutated genes in Non-Small Cell Lung Carcinoma (NSCLC)

Shradha Mukherjee

Title: Targeted BEL network representation and characterization of commonly mutated genes in Non-Small Cell Lung Carcinoma (NSCLC)

Shradha Mukherjee, PhD Email: smukher2@yahoo.com
Non-ASU mentor: *William Hayes, PhD*
ASU Faculty Project Advisor: *Robert Greenes, MD, PhD*

Source: Arizona State University (ASU) MAS Health Informatics BMI_593 Applied Project of Student Shradha Mukherjee, PhD Email: smukher2@yahoo.com.

Summary Statement: A computational pipeline for mobilization of scientific knowledge to inform drug selection based on patient genetic variation profile for precision medicine in Non-Small Cell Lung Carcinoma (NSCLC).

Abstract:
Non-Small Cell Lung Carcinoma (NSCLC) is the most common form of lung cancer and the third most common cancer overall with respect to numbers, and deaths (American Cancer Society, Cancer Facts and Figures, Feb 2019). Previous studies have shown that targeting a commonly mutated NSCLC gene EGFR with Tyrosine kinase inhibitor (TKI) drugs is successful in patients who have EGFR mutations. It is now being recognized that overall NSCLC patient outcomes can be further improved by targeting other commonly mutated NSCLC genes in addition to the widely studies and targeted EGFR. However, DNA sequencing outputs show multiple mutations or variations in genes, which are too many to interpret, study and target with drugs. To overcome this limitation, the goal of the present study was to build a NSCLC specific knowledge base that can be used to prioritize mutated genes from a given patient genetic profile and drugs to target these genes. Commonly mutated NSCLC genes reported in NSCLC cases were used as a 'bait' to 'capture' and create a computable NSCLC knowledge base using BEL (Biological Expression Language) statements. This NSCLC specific knowledge base consisted of 33 pathogenic variants, 129 significant gene functional annotations (GO, gene ontology terms) and 568 drugs BEL knowledge statements. Overlaying 17 different NSCLC patient genetic profiles against our NSCLC specific knowledge base, successfully computed and prioritized distinct genes harboring mutations, and drugs to target these genes in patients. Here we developed this computational pipeline for NSCLC and this methodology can be adapted for other diseases with known common gene mutations. Taken together, we provide a paradigm to prioritize and select drugs based on patient's genetic profile for precision medicine based clinical decision support.

Introduction:

Treatment options for NSCLC include surgery, radiation, chemotherapy, immunotherapy and targeted treatment. The advantage of targeted therapy is that unlike the other forms of treatment targeted therapy is designed to specifically attack cancer by blocking molecular targets in the cancer. EGFR is one of the most common mutations found in 10-35% of NSCLC patients and tyrosine kinase inhibitors (TKIs) drugs that target EGFR pathway specifically are successfully used in these cases (1). In addition to EGFR, there are also other gene mutations found in NSCLC that could be targeted with drugs avoiding a "one-size-fits-all" approach, and offers potential to use a combination of drugs to target multiple mutations in NSCLC for

successful treatment (2). It still remains a challenge to find an optimal combination of drugs, which targets most of the disease related cellular changes caused by gene mutation without drug overuse. One drug development strategy is to pick a combination of drugs that each directly target different regulatory hub genes in the networks, which can then initiate a 'domino effect' by modulating different cohorts of interacting genes connected to the hub gene. This strategy is the basis of the growing field of 'network medicine' (3).

Given the growing appreciation of the benefit of taking a global systems approach to target an ensemble of commonly mutated NSCLC gene with drugs, the development of suitable biomedical knowledge and big data driven informatics tools and workflows to identify key regulatory NSCLC mutations and inform clinical decision making for development of combinatorial targeted drug delivery has become increasingly important. 'Omics' or genomics data and biomedical literature text data are the two major sources for information about genes and gene interaction in the context of human health and disease. Both omics data and text data resources have been the basis of extensive research for development of gene networks and pathways to identify key regulators underlying disease phenotype and can serve as drug targets. Extensive research endeavors have led to the collection of RNA-sequencing (RNA-seq), chromatin immunoprecipitation sequencing (ChIP-seq), ES (Exome Sequencing), Whole Exome Sequencing (WES), high-throughput proteomics and other 'omics' data from patients (4). The Cancer Genome Atlas Research Network (TCGA http://cancergenome.nih.gov/), International Cancer Genome Consortium (ICGC http://icgc.org/), Alzheimer's Disease Genetics Consortium (ADGC http://www.adgenetics.org/) are some of the examples of a few projects where the omics data is collected and stored from a wide sample of patients.

Though omics datasets are fast growing, most of the biomedical research knowledge on gene interaction, effects and context is present in the form open text in biomedical literature. Pubmed https://www.ncbi.nlm.nih.gov/pubmed/ is the online site where a large volume of biomedical literature is houses and the vastness of this resource is demonstrated by the fact that a quick search with the term 'non-small cell lung cancer' shows 76958 results on Pubmed. The availability of omics data and text data has open unprecedented opportunities to study genes-gene interaction networks, called 'interactome' and the effects of gene-gene interactions on phenotype in the context of human disease. For high throughput omics data, construction of disease specific networks relies on advanced computational and statistical methods that can handle the high complexity and high dimensionality of biological networks. Initial approaches for development of omics data based regulatory networks regulated by central hubs included Boolean networks, Bayesian networks and differential equation models, which due to limitation in scalability were later replaced by Poisson graphical models and negative binomial distributions (5, 6). For open text data on gene interactions, construction of disease specific networks relies on text-mining techniques and methods, where gene-gene interaction and context information extracted from a body of text called 'coropus' is used as a foundation, converted to a computable form and manually curated to build a regulatory network (7, 8). For text-mining biomedical literature and representing the biological relations, such as gene-gene interactions in a computable form, the well-established standards available are Systems Biology Markup Language (SBML), Biological Pathway Exchange Language (BPEL) and Biological Expression Language (BEL) (9).

Presently, a NSCLC specific knowledge base was built to prioritize genes to target with drugs from patient genetic variation profile. Drugs for a given patient genetic profile was suggested based on overlap with gene variant and/or gene ontology NSCLC specific knowledge

base. Testing this pipeline on publicly available genetic variation profile of NSCLC patients revealed 3 classes of patients based on if matches were found or not in one or both NSCLC knowledge base subsets, genetic variation and gene ontology. In future, this NSCLC pipeline can be used to support development drug-gene pair rules-based treatment plans for a given gene mutation profile of NSCLC patients.

Methods:

Code Availability

Computational code with html or pdf rendering showing input and output of code chunks is available as a git local repository at https://icedrive.net/s/Ww9Y35fjuZu6akSkYa8xaChFQTG6 with all files and as a git remote repository at https://gitlab.com/smukher2/nsclc_drugtargetsmutations_nov2019 with large files ignored or removed.

Retrieval of commonly mutated NSCLC genes

Mycancergenome (https://www.mycancergenome.org/) is an online open source curated knowledge resource that has current information on cancer related mutations, clinical trials, drugs, pathways and biomarkers designed to facilitate development of precision medicine (10). A list of commonly mutated genes in NSCLC reported in this site (https://www.mycancergenome.org/content/disease/non-small-cell-lung-carcinoma/) were copied and stored in excel format. A total of 28 genes commonly mutated in NSCLC and the number of cases for each were retrieved in this manner.

Python as the language of choice

All computational codes in this work were written using python because of its human readable and intuitive syntax, fast performance and rich resource of libraries that enable accomplishment of complex tasks with only a few lines of code (11).

NSCLC GO knowledgebase: GO annotation of commonly mutated NSCLC genes

Enrichr (https://amp.pharm.mssm.edu/Enrichr/) is an online web-tool, also available through API, which can perform statistical enrichment analysis on any list of genes by comparing the gene list to several biological functional annotation resources such as GO (12). The commonly mutated NSCLC genes, were searched using python search API to retrieve significant gene ontology terms characterized with GO terms significantly associated with NSCLC. Each GO term is associated with multiple genes so the results were parsed by gene name column to create a table of gene-GO term to build a population level NSCLC GO term knowledge base primed for comparison with patient genes. The results were exported and saved in .excel format.

NSCLC pathogenic variants knowledgebase: Pathogenic variant annotation of commonly mutated NSCLC genes

Simple ClinVar (http://simple-clinvar.broadinstitute.org/) is an online web-tool that curates reported variants associated with diseases as pathogenic, probable pathogenic, benign and unknown (13). Simple ClinVar was searched with search term “non-small cell lung cancer” to retrieve a table of NSCLC associated variants. The results were exported as excel file and used as input in the Python code to keep only the pathogenic NSCLC variants. The results from this

analysis comprised population level NSCLC pathogenic variants knowledge base primed for comparison with patient genes. The results were exported and saved in .excel format.

NSCLC gene-drug pair knowledgebase: Drugs to target commonly mutated NSCLC genes

The DGIdb (http://www.dgidb.org/search_interactions) is a drug gene interaction database built from over thirty sources, with each drug-gene interaction referenced to Pubmed (14). DGIdb is available as a web-tool and is also searchable through python search API. Python API search of DGIdb with commonly mutated NSCLC genes provided a list of drug-gene pairs and associated scores for number of citations associated with a given interaction. The resultant population level NSCLC gene-drug pair knowledge base was exported and saved in .excel format.

Visualization of commonly mutated NSCLC network using BioDati Studio

BioDati Studio (https://studio.demo.biodati.com/) is powered by a well-defined and extensive collection of essential gene interaction evidence lines extracted from biomedical literature in Pubmed and coded into BEL statements by BEL coding experts. This extracted BEL statements can be utilized by users using BioDati Studio's user-friendly web interface for building gene networks. To build a network on BioDati Studio, the interactions in NSCLC pathogenic variants, NSCLC GO terms and NSCLC drug-gene pairs knowledgebases were converted to BEL statements. Next, the BEL statements annotated with citation urls and database urls, and other metadata were saved as nanopubs (small units of knowledge represented as BEL). The nanopubs were imported into BioDati Studio using the "import nanopubs" function and the network was visualized using "draft network" function. This network was then visualized by clicking 'Visualizer' function on BioDati Studio and by using the "zoom" tool, it was possible to see the genes (nodes) and connections (edges) in the network.

Retrieval of patient genetic profiles

Full patient genetic profiles are not readily available through open access due to HIPA restrictions. A truncated list of total ~3.5K validated NSCLC patient genetic variation or mutation profiles were obtained for a total of 12 patients from previously published work (15).

Overlay of patient genetic profile on NSCLC knowledgebase

Validated patient genetic profiles for the 12 NSCLC patients were overlapped against the population level NSCLC knowledge base. The computational analysis resulted in recommendations of different potential drugs for each of the patients. Drug suggestions were computed based on overlap of patient's genetic profile with NSCLC pathogenic variant knowledgebase and NSCLC GO term knowledgebase. For visualization, patient genes that overlapped with the population level NSCLC knowledgebase were zoomed on and viewed on the population level NSCLC knowledgebase or nanopubs that had been built on BioDati studio.

Results:

Pathogenic variants and GO terms knowledgebases from commonly mutated genes in NSCLC and knowledgebases

The distribution of frequency of cases with mutations or variants that commonly occur in NSCLC was highest for TP53 (4463 cases) followed by KRAS (2637cases) and EGFR (1969 cases) (Fig. 2A). The frequency of cases with mutations or variants was comparable for the

remainder of the 28 genes commonly mutated in NSCLC (Fig. 2A). From ClinVar, 13 KRAS variants, 8 EGFR variants, 4 PIK3CA variants and 8 BRAF variants reported to be pathogenic in NSCLC were computationally retrieved (Fig. 2B). The report of only 4 out of 28 commonly mutated genes known to be pathogenic suggests that more research is required to gather more information on the pathogenicity of NSCLC genetic variation. From enrichr, a total of 130 significant (p-value <0.05) GO terms were obtained for the 28 commonly mutated NSCLC genes, with EGFR, TP53, ERBB4, NTRK3 and KDR genes associated with >40 GO terms (Fig. 2C). This suggests that either these 5 highly GO term enriched genes have a multitude of biological functions or have been studied more by researchers relative to the other 23 commonly mutated NSCLC genes. Search of the DGIdb database for drugs that can target the 28 commonly mutated NSCLC genes, retrieved 329 drugs for the NSCLC pathogenic variants knowledgebase genes and 569 drugs for the NSCLC GO terms knowledgebase genes.

Visualization of NSCLC pathogenic variants and drug-gene interaction knowledge bases on BioDati Studio

Visualization of the NSCLC pathogenic variants knowledge base nanopub and NSCLC drug-gene interaction knowledgebase nanopub revealed a complex network of interactions between the genes and drugs, with 1119 edges and 629 nodes (Fig. 3). The 4 NSCLC variant knowledgebase genes KRAS, EGFR, PIK3CA and BRAF genes, and the drugs they interact with accounted for ~50% of all nodes (sub-node count of 333), with the drugs targeting one or more of these 4 genes (Fig. 3).

Visualization of NSCLC GO terms and drug-gene interaction knowledge bases on BioDati Studio

Visualization of the NSCLC GO terms knowledge base nanopub and NSCLC drug-gene interaction knowledgebase nanopub revealed a complex network of interactions between the genes and drugs, with 1697 edges and 725 nodes (Fig. 4). The 28 NSCLC GO terms enriched genes and the drugs they interact with accounted for ~80% of all nodes (sub-node count of 597), with the drugs targeting one or more of these 4 genes (Fig. 4). The other 20% of the nodes were comprised of the GO terms themselves and with the most significant GO terms of the knowledge base also accounted for the highest number of connections. These highly connected and most significant GO terms all belonged to regulation of cell signaling, namely “negative regulation of cell communication” (GO:0010648, p-value = 9.98E-12),
“negative regulation of response to stimulus” (GO:0048585, p-value = 5.54E-12),
“negative regulation of signaling” (GO:0023057, p-value = 5.92E-12),
“regulation of kinase activity” (GO:0043549, p-value = 2.72E-11), and
“positive regulation of protein phosphorylation” (GO:0001934, p-value = 3.24E-10) (Fig. 4).

‘Best scenario’ patients with genetic profile overlap with NSCLC pathogenic variants and GO terms knowledge bases

Comparison of patient genetic profile with the NSCLC knowledge bases computationally revealed three types of overlaps and formed the basis of classification of the patients. The first category of patients called ‘best scenario’ comprised of patients (Patient 1, 2, 6, 10 and 13) whose genetic profile found matches in both the NSCLC pathogenic variants and GO terms knowledge bases (Fig. 5A). Patients 1, 2 and 6 had only one overlapped gene, while patient 10 and 13 had more than one overlapped gene. Though several drugs were suggested to target the

overlapped genes from the patients, a score based on number of citations was used to prioritize drug selection (Fig. 5A). As an illustration of the utility of the BioDati Studio for visualization of patient overlapped genes, a zoomed view of Patient 1's overlapped gene KRAS on the NSCLC pathogenic variants knowledge base nanopub and NSCLC drug-gene interaction knowledgebase nanopub is shown (Fig. 5B). Also shown is a zoomed view of Patient 1's overlapped gene KRAS on the NSCLC GO terms knowledge base nanopub and NSCLC drug-gene interaction knowledgebase nanopub is shown (Fig. 5C).

'Good scenario' patients with genetic profile overlap with NSCLC pathogenic variants and GO terms knowledge bases

The second category of patients called 'good scenario' comprised of patients (Patient 4, 8, 9, 11 and 14) whose genetic profile found matches in only the NSCLC GO terms knowledge base (Fig. 6A). Patients 4 and 11 had only one overlapped gene, while patient 8, 9 and 11 had more than one overlapped gene. Though several drugs were suggested to target the overlapped genes from the patients, a score based on number of citations was used to prioritize drug selection (Fig. 6A). As an illustration of the utility of the BioDati Studio for visualization of patient overlapped genes, a zoomed view of Patient 1's overlapped gene KRAS on the NSCLC GO terms knowledge base nanopub and NSCLC drug-gene interaction knowledgebase nanopub is shown (Fig. 6B).

'No Good scenario' patients with no genetic profile overlap with NSCLC pathogenic variants and GO terms knowledge bases

The third category of patients called 'no good scenario' comprised of patients (Patient 4 and 16) whose genetic profile found no matches in either the NSCLC pathogenic variants and GO terms knowledge bases (no data to show). These patients potentially house novel variants on genes not common in the population and require further investigative research.

Discussion:

There is increasing demand for global systems based unbiased approaches for identification of druggable targets for development of safer and effective combinatorial patient treatment. NSCLC is the most common form of lung cancer, accounting for 85% of lung cancers and the third most prevalent cause of cancer deaths (American Cancer Society, Cancer Facts and Figures, Feb 2019). Therefore, there is a great need to understand NSCLC mechanism and drug development. In the context of NSCLC, geftinib that targets EGFR has become the most widely used drug, but development of drug resistance is NSCLC has made it critical to develop alternative approaches to treatment of NSCLC (2). In the present project, an informatics workflow for 'network medicine' using different existing web-tools is demonstrated.

The biochemical process by which cells communicate and coordinate activities in response to stimuli from their environment is called cell signaling. Here we found several cells signaling pathways, especially kinase activity and kinase activity regulation through phosphorylation as the major theme underlying the significant GO terms enriched in NSCLC. This is consistent with the presence of several protein kinases, 11 out of 28 genes, among the commonly mutated NSCLC genes, EGFR, STK11, ATM, BRAF, MET, ALK, EPHA3, ERBB4, EPHA5, NTRK3 and KDR. Phosphorylation of target proteins by kinases is one of the major mechanisms of on-off switch for regulation of signaling pathways. Several studies have

demonstrated the importance of maintenance of protein kinase activity in NSCLC or other cancers. NSCLC cells that depend on EGFR for survival, constitutively maintain activation of EGFR through its overexpression and overexpression its binding partners from the ERBB family such as ERBB4 (16). Overexpression, gene amplification and increases activation of MET has been reported to be associated with poor prognosis in NSCLC (17). BRAF mutations that occur in NSCLC either enhance its kinase activity, which increases activation of its target MAPK pathway and PI3K pathway, while other mutation in BRAF render it completely inactive (18). Thus, modulation of protein phosphorylation by kinases that makes them hyperactive or inactive can both be detrimental to the cell and promote NSCLC. Presently, we have developed a computational pipeline that can prioritize patient genetic variants in kinases that are commonly mutated in NSCLC and suggest drugs to target these kinases.

In this study, we built a NSCLC specific knowledge base to filter and prioritize patient genetic variation data aimed towards personalized medicine. The pipeline successfully suggested drugs that could be used for 10 out of 12 patients whose genetic profile overlapped with the NSCLC knowledge bases built for this work. However, two patients' genetic profiles showed no overlap with any of the genes present in the NSCLC knowledge bases built for this work. One possible solution is to expand the avenues of NSCLC knowledge bases to increase the probability of overlap with patients' genetic profiles. During the data mining for reported pathogenic variants of genes in NSCLC we retrieved reported pathogenic variants for only 4 out of the 28 genes commonly mutated in NSCLC, while other genes had unknown or benign reported variants. Thus, more research is required into pathogenicity of variants, which will help expand the NSCLC pathogenic variants knowledgebase. Modern high-throughput sequencing tools such as Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) have provided a plethora of data to analyze for identification of disease related variants (19). However, gene variant disease pathogenicity are challenging to study because any given gene-disease association may involve multiple genetic interactions and depend on context related variables (20). Recently big data analytics methodologies are being applied and developed, such as VarCoPP specifically to predict the causal role of combination of genetic variants and their combined pathogenicity in diseases (21).

This project provides a paradigm for utilization of BEL statements derived from biomedical literature to build networks and identify hubs/modules in diseases such as NSCLC. Furthermore, the workflow presented here shows the utility of online web-tools for characterization of the modules and identification of drugs that can target the hub genes. Taken together the power of a systems based global approach for network building, network characterization and identifying drug targets at network hubs is demonstrated.

Bibliography:

1. Zhang H. Three generations of epidermal growth factor receptor tyrosine kinase inhibitors developed to revolutionize the therapy of lung cancer. Drug design, development and therapy. 2016;10:3867-72.
2. Terlizzi M, Colarusso C, Pinto A, Sorrentino R. Drug resistance in non-small cell lung Cancer (NSCLC): Impact of genetic and non-genetic alterations on therapeutic regimen and responsiveness. Pharmacology & therapeutics. 2019.
3. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature reviews Genetics. 2011;12(1):56-68.

4. Scelfo C, Galeone C, Bertolini F, Caminati M, Ruggiero P, Facciolongo N, et al. Towards precision medicine: The application of omics technologies in asthma management. F1000Research. 2018;7:423.
5. Jia B, Xu S, Xiao G, Lamba V, Liang F. Learning gene regulatory networks from next generation sequencing data. Biometrics. 2017;73(4):1221-30.
6. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Briefings in bioinformatics. 2018;19(6):1370-81.
7. Jurca G, Addam O, Aksac A, Gao S, Ozyer T, Demetrick D, et al. Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends. BMC research notes. 2016;9:236.
8. Luo Y, Riedlinger G, Szolovits P. Text mining in cancer gene and pathway prioritization. Cancer informatics. 2014;13(Suppl 1):69-79.
9. Domingo-Fernandez D, Mubeen S, Marin-Llao J, Hoyt CT, Hofmann-Apitius M. PathMe: merging and exploring mechanistic pathway knowledge. BMC bioinformatics. 2019;20(1):243.
10. Kusnoor SV, Koonce TY, Levy MA, Lovly CM, Naylor HM, Anderson IA, et al. My Cancer Genome: Evaluating an Educational Model to Introduce Patients and Caregivers to Precision Medicine Information. AMIA Joint Summits on Translational Science proceedings AMIA Joint Summits on Translational Science. 2016;2016:112-21.
11. Ekmekci B, McAnany CE, Mura C. An Introduction to Programming for Bioscientists: A Python-Based Primer. PLoS computational biology. 2016;12(6):e1004867.
12. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics. 2013;14:128.
13. Perez-Palma E, Gramm M, Nurnberg P, May P, Lal D. Simple ClinVar: an interactive web server to explore and retrieve gene and disease variants aggregated in ClinVar database. Nucleic acids research. 2019;47(W1):W99-W105.
14. Cotto KC, Wagner AH, Feng YY, Kiwala S, Coffman AC, Spies G, et al. DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. Nucleic acids research. 2018;46(D1):D1068-D73.
15. Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell. 2012;150(6):1121-34.
16. Fujimoto N, Wislez M, Zhang J, Iwanaga K, Dackor J, Hanna AE, et al. High expression of ErbB family members and their ligands in lung adenocarcinomas that are sensitive to inhibition of epidermal growth factor receptor. Cancer research. 2005;65(24):11478-85.
17. Sattler M, Reddy MM, Hasina R, Gangadhar T, Salgia R. The role of the c-Met pathway in lung cancer and the potential for targeted therapy. Therapeutic advances in medical oncology. 2011;3(4):171-84.
18. Auliac JB, Bayle S, Vergnenegre A, Le Caer H, Falchero L, Gervais R, et al. Patients with non-small-cell lung cancer harbouring a BRAF mutation: a multicentre study exploring clinical characteristics, management, and outcomes in a real-life setting: EXPLORE GFPC 02-14. Current oncology (Toronto, Ont). 2018;25(5):e398-e402.
19. Petersen BS, Fredrich B, Hoeppner MP, Ellinghaus D, Franke A. Opportunities and challenges of whole-genome and -exome sequencing. BMC genetics. 2017;18(1):14.

20. Mani A. Pathogenicity of De Novo Rare Variants: Challenges and Opportunities. Circulation Cardiovascular genetics. 2017;10(6).
21. Papadimitriou S, Gazzo A, Versbraegen N, Nachtegael C, Aerts J, Moreau Y, et al. Predicting disease-causing variant combinations. Proceedings of the National Academy of Sciences of the United States of America. 2019;116(24):11878-87.

Access to Codes:

https://icedrive.net/s/Ww9Y35fjuZu6akSkYa8xaChFQTG6 and https://gitlab.com/smukher2/nsclc_drugtargetsmutations_nov2019. Please cite this paper if you use these codes. Thank you.

Figure 1: Workflow showing all steps of project

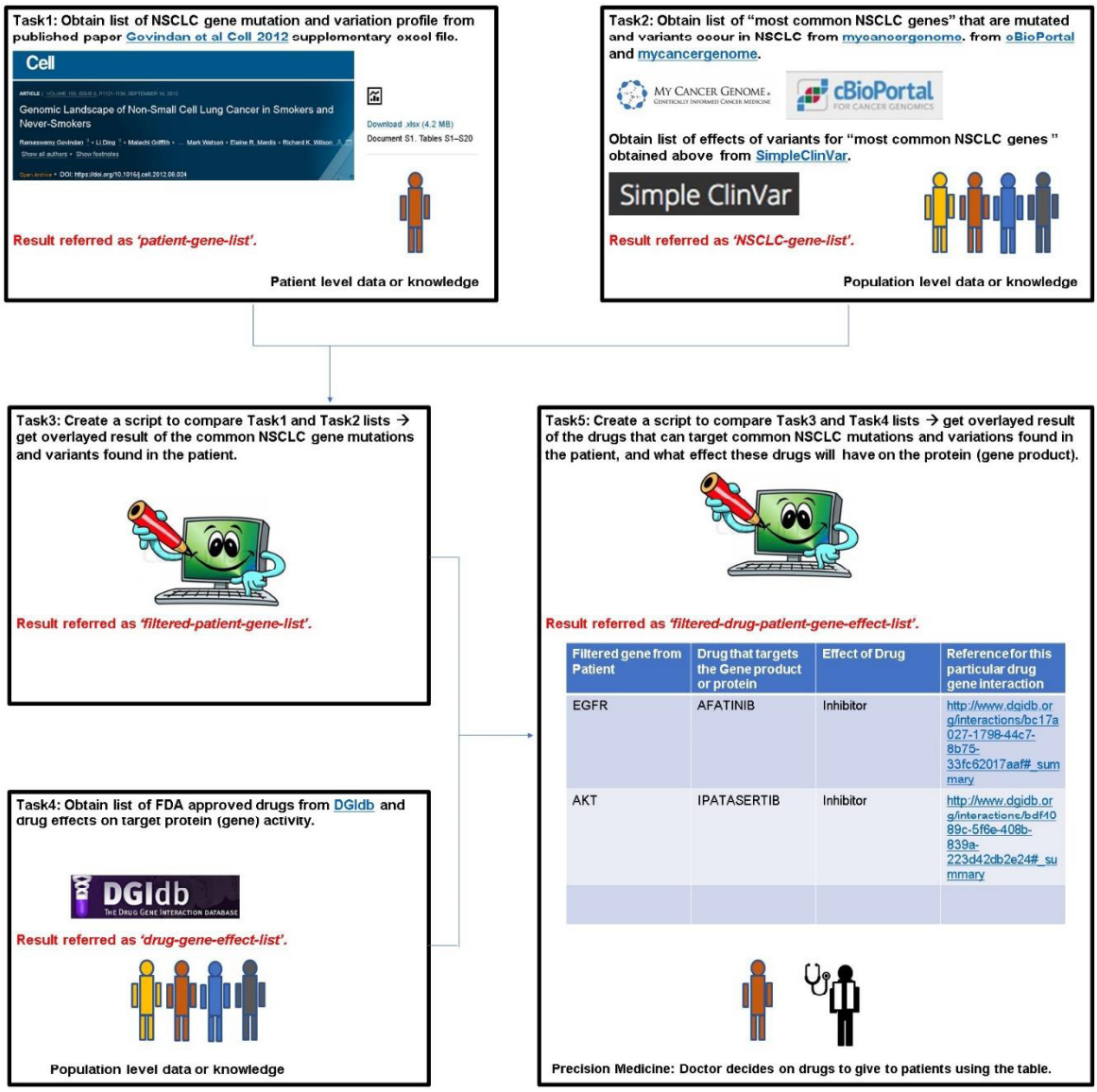
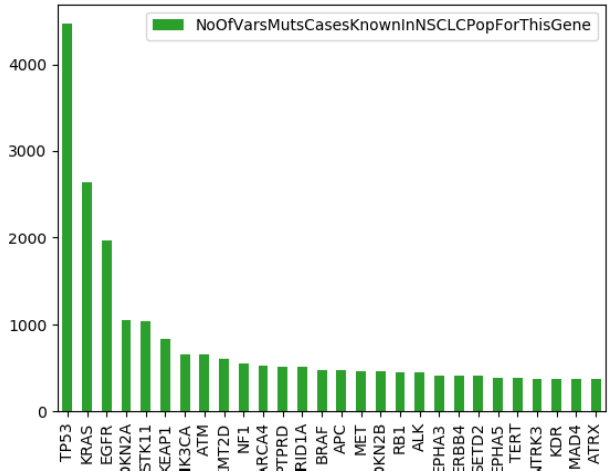
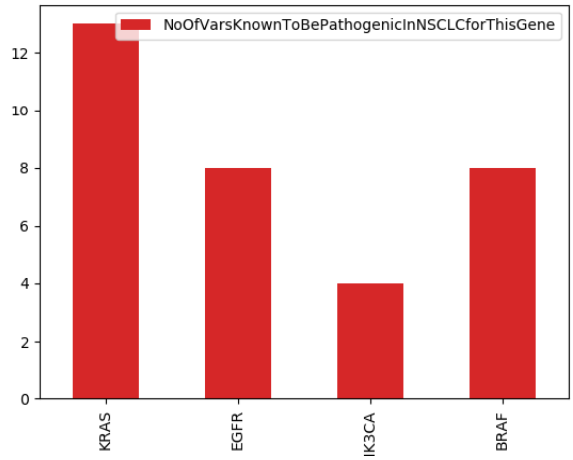


Figure 2: A: Commonly mutated genes in NSCLC obtained from mycancergenome B: Pathogenic variants reported in ClinVar for commonly mutated NSCLC genes C: Significant Gene Ontology (GO) terms from enrichr for commonly mutated NSCLC genes

A



B



C

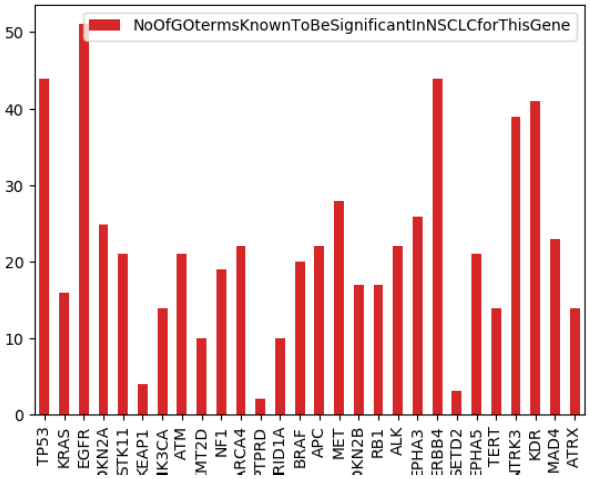
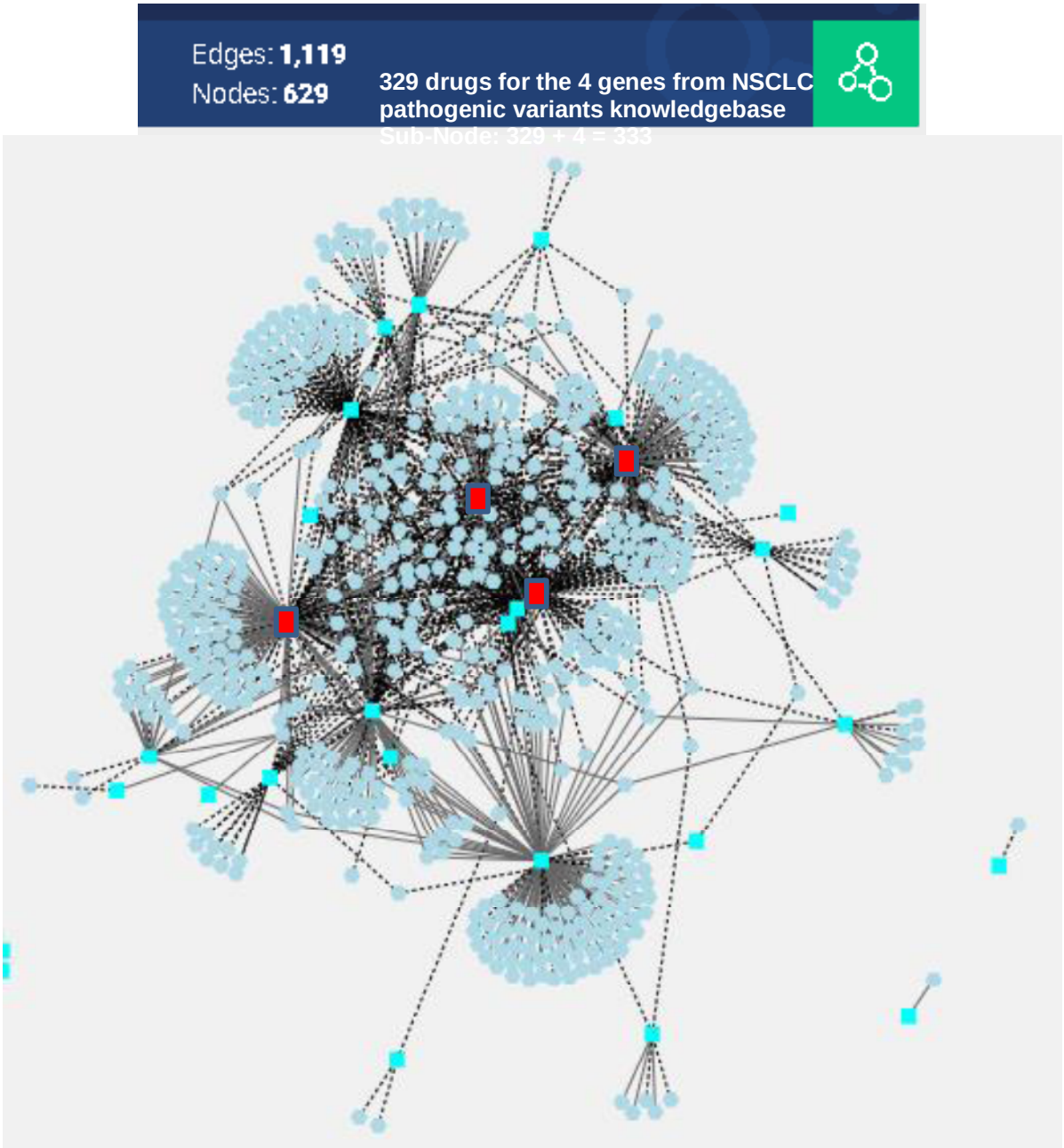
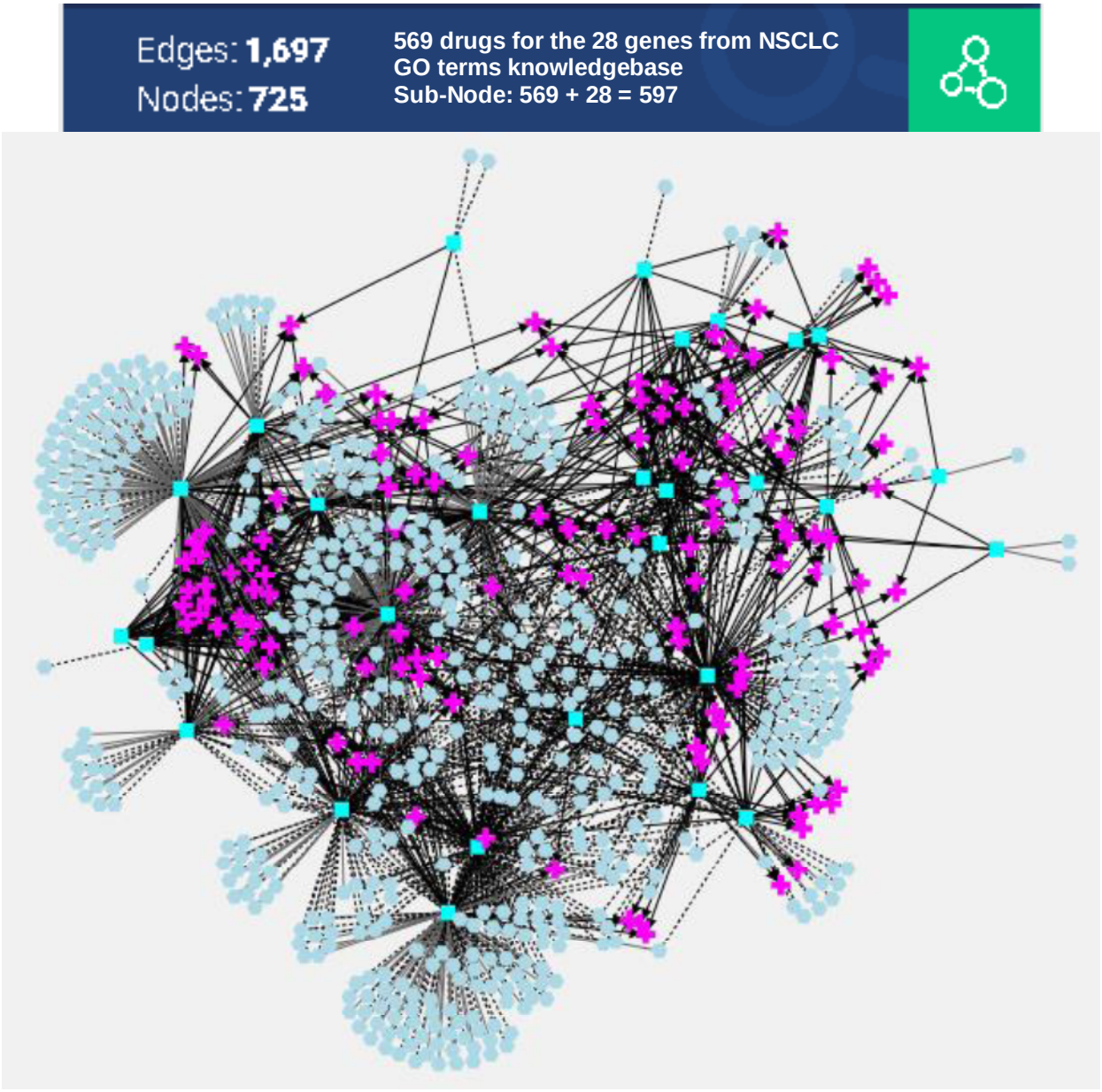


Figure 3: BioDati Studio network for population level NSCLC drug-gene interaction knowledgebase (blue-green and red square) and NSCLC pathogenic variant knowledgebase genes (red square).



Legend:
Blue circles=drugs
Blue-green squares=genes and their variants

Figure 4: BioDati Studio network for population level NSCLC drug-gene interaction knowledgebase (blue-green squares) and NSCLC significant GO terms knowledgebase genes (blue-green squares).



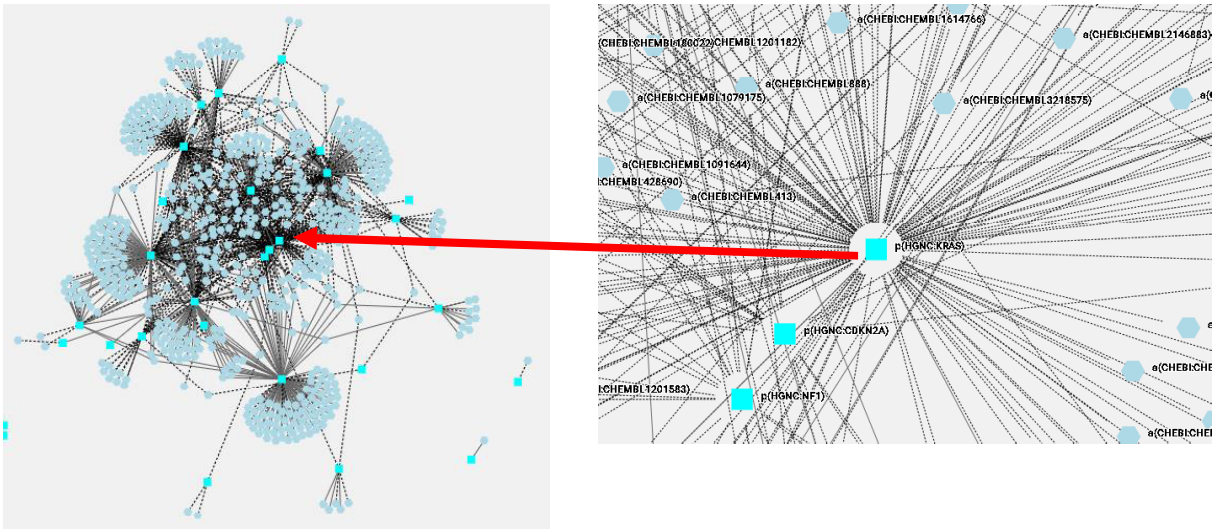
Legend:
Blue circles=drugs
Blue-green squares=genes
Pink+ = GO terms

Figure 5: A: 2of2Match, Best scenario: Five patients' genetic profile found overlaps with population level NSCLC pathogenic variants, GO terms and drug-gene knowledgebases. B: BioDati Studio visualization showing Patient 1 overlapping gene in the population level NSCLC pathogenic variants and drug-gene knowledgebases. C: BioDati Studio visualization showing Patient 1 overlapping gene in the population level NSCLC GO terms and drug-gene knowledgebases.

A

		NoOfVars In This Patient For This Gene	NoOfVarsMuts Cases Known In NSCLC Pop For This Gene	NoOfVars Known To Be Pathogenic In NSCLC for This Gene	NoOfGOterms Known To Be Significant In NSCLC for This Gene	NoOfDrugs for This Gene	Best inhibitor action drug selected by highest score
Patient1	KRAS	1	2637	13	16	148	CHEMBL1614701, SELUMETINIB, score 44
Patient2	BRAF	1	474	8	20	100	CHEMBL1229517, VEMURAFENIB, score 169
Patient6	EGFR	1	1969	8	51	146	CHEMBL1173655, AFATINIB, score 141
Patient10	KRAS	1	2637	13	16	148	CHEMBL1614701, SELUMETINIB, score 44
	ERBB4	1	407	0	44	24	CHEMBL2110732, DACOMITINIB, score 6
	ATM	1	656	0	21	25	CHEMBL1221601, (no common name), score 1
Patient13	KRAS	1	2637	13	16	148	CHEMBL1614701, SELUMETINIB, score 44
	EPHA3	1	412	0	26	2	CHEMBL24828, VANDETANIB, score 1

B



C

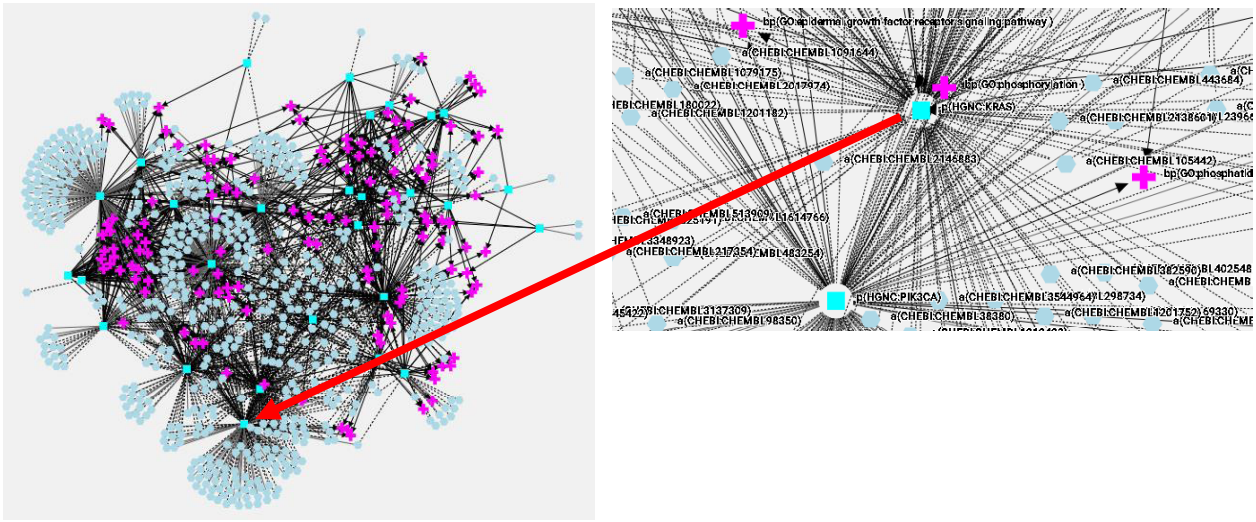

bp(GO:epidermal growth factor receptor signaling pathway)
a(CHEBI:CHEMBL1091644)
a(CHEBI:CHEMBL1079175)
a(CHEBI:CHEMBL2019974)
bp(GO:phosphorylation)
a(CHEBI:CHEMBL443684)
a(CHEBI:CHEMBL1201182)
p(HGNC:KRAS)
a(CHEBI:CHEMBL2145883)
a(CHEBI:CHEMBL105442)
a(CHEBI:CHEMBL3348923)
a(CHEBI:CHEMBL217354)
a(CHEBI:CHEMBL483254)
p(HGNC:PIK3CA)
a(CHEBI:CHEMBL3544964)
a(CHEBI:CHEMBL98350)
a(CHEBI:CHEMBL38380)

Figure 6: A: 1of2Match, Good scenario: Five patients' genetic profile found overlaps with population level NSCLC GO terms and drug-gene knowledgebases, but not with NSCLC pathogenic variants knowledgebase. B: BioDati Studio visualization showing Patient 8 overlapping gene in the population level NSCLC GO terms and drug-gene knowledgebases.

A

		NoOfVars In This Patient For This Gene	NoOfVarsMuts Cases Known In NSCLC Pop For This Gene	NoOfVars Known To Be Pathogenic In NSCLC for This Gene	NoOfGOterms Known To Be Significant In NSCLC for This Gene	NoOfDrugs for This Gene	Best inhibitor action drug selected by highest score
Patient4	APC	1	472	0	22	23	No known inhibitor
Patient8	TP53		4463	0	44	108	CHEMBL325041, BORTEZOMIB, score 1
	EPHA3		412	0	26	2	CHEMBL24828, VANDETANIB, score 1
Patient9	TP53	1	4463	0	44	108	CHEMBL325041, BORTEZOMIB, score 1
	MET	1	460	0	28	81	CHEMBL601719, CRIZOTINIB, score 52
	NF1	1	547	0	19	22	No known inhibitor
	TERT	1	379	0	14	7	CHEMBL129, ZIDOVUDINE, score 2
	EPHA3	2	412	0	26	2	CHEMBL24828, VANDETANIB, score 1
	SMARCA4		525	0	22	2	No known inhibitor
Patient11	SETD2	1	402	0	3	1	CHEMBL299763,(no common name), score 1
Patient14	ATM	1	656	0	21	25	CHEMBL1221601,(no common name), score
	ARID1A	1	505	0	10	3	No known inhibitor

B