



HAL
open science

**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

► **To cite this version:**

Shradha Mukherjee. 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). 2019. hal-04084202

HAL Id: hal-04084202

<https://hal.science/hal-04084202>

Preprint submitted on 27 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

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

3
4 Shradha Mukherjee, PhD Email: smukher2@yahoo.com
5 Non-ASU mentor: *William Hayes, PhD*
6 ASU Faculty Project Advisor: *Robert Greenes, MD, PhD*
7

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

10
11 **Summary Statement:** A computational pipeline for mobilization of scientific knowledge to
12 inform drug selection based on patient genetic variation profile for precision medicine in Non-
13 Small Cell Lung Carcinoma (NSCLC).

14 15 **Abstract:**

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

37 38 **Introduction:**

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

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

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

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

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

93 base. Testing this pipeline on publicly available genetic variation profile of NSCLC patients
94 revealed 3 classes of patients based on if matches were found or not in one or both NSCLC
95 knowledge base subsets, genetic variation and gene ontology. In future, this NSCLC pipeline can
96 be used to support development drug-gene pair rules-based treatment plans for a given gene
97 mutation profile of NSCLC patients.
98
99

100 **Methods:**

101 Code Availability

102 Computational code with html or pdf rendering showing input and output of code chunks is
103 available as a git local repository at <https://icedrive.net/s/Ww9Y35fjuZu6akSkYa8xaChFQTG6>
104 with all files and as a git remote repository at
105 https://gitlab.com/smukher2/nscl_drugtargetsmutations_nov2019 with large files ignored or
106 removed.
107

108 Retrieval of commonly mutated NSCLC genes

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

116 Python as the language of choice

117 All computational codes in this work were written using python because of its human readable
118 and intuitive syntax, fast performance and rich resource of libraries that enable accomplishment
119 of complex tasks with only a few lines of code (11).
120

121 NSCLC GO knowledgebase: GO annotation of commonly mutated NSCLC genes

122 Enrichr (<https://amp.pharm.mssm.edu/Enrichr/>) is an online web-tool, also available through
123 API, which can perform statistical enrichment analysis on any list of genes by comparing the
124 gene list to several biological functional annotation resources such as GO (12). The commonly
125 mutated NSCLC genes, were searched using python search API to retrieve significant gene
126 ontology terms characterized with GO terms significantly associated with NSCLC. Each GO
127 term is associated with multiple genes so the results were parsed by gene name column to create
128 a table of gene-GO term to build a population level NSCLC GO term knowledge base primed for
129 comparison with patient genes. The results were exported and saved in .excel format.
130

131 NSCLC pathogenic variants knowledgebase: Pathogenic variant annotation of commonly 132 mutated NSCLC genes

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

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

140
141 NSCLC gene-drug pair knowledgebase: Drugs to target commonly mutated NSCLC genes
142 The DGIdb (http://www.dgidb.org/search_interactions) is a drug gene interaction database built
143 from over thirty sources, with each drug-gene interaction referenced to Pubmed (14). DGIdb is
144 available as a web-tool and is also searchable through python search API. Python API search of
145 DGIdb with commonly mutated NSCLC genes provided a list of drug-gene pairs and associated
146 scores for number of citations associated with a given interaction. The resultant population level
147 NSCLC gene-drug pair knowledge base was exported and saved in .excel format.

148
149 Visualization of commonly mutated NSCLC network using BioDati Studio
150 BioDati Studio (<https://studio.demo.biodati.com/>) is powered by a well-defined and extensive
151 collection of essential gene interaction evidence lines extracted from biomedical literature in
152 Pubmed and coded into BEL statements by BEL coding experts. This extracted BEL statements
153 can be utilized by users using BioDati Studio's user-friendly web interface for building gene
154 networks. To build a network on BioDati Studio, the interactions in NSCLC pathogenic variants,
155 NSCLC GO terms and NSCLC drug-gene pairs knowledgebases were converted to BEL
156 statements. Next, the BEL statements annotated with citation urls and database urls, and other
157 metadata were saved as nanopubs (small units of knowledge represented as BEL). The nanopubs
158 were imported into BioDati Studio using the "import nanopubs" function and the network was
159 visualized using "draft network" function. This network was then visualized by clicking
160 'Visualizer' function on BioDati Studio and by using the "zoom" tool, it was possible to see the
161 genes (nodes) and connections (edges) in the network.

162
163 Retrieval of patient genetic profiles
164 Full patient genetic profiles are not readily available through open access due to HIPA
165 restrictions. A truncated list of total ~3.5K validated NSCLC patient genetic variation or
166 mutation profiles were obtained for a total of 12 patients from previously published work (15).

167
168 Overlay of patient genetic profile on NSCLC knowledgebase
169 Validated patient genetic profiles for the 12 NSCLC patients were overlapped against the
170 population level NSCLC knowledge base. The computational analysis resulted in
171 recommendations of different potential drugs for each of the patients. Drug suggestions were
172 computed based on overlap of patient's genetic profile with NSCLC pathogenic variant
173 knowledgebase and NSCLC GO term knowledgebase. For visualization, patient genes that
174 overlapped with the population level NSCLC knowledgebase were zoomed on and viewed on the
175 population level NSCLC knowledgebase or nanopubs that had been built on BioDati studio.

176
177
178 **Results:**
179 Pathogenic variants and GO terms knowledgebases from commonly mutated genes in NSCLC
180 and knowledgebases
181 The distribution of frequency of cases with mutations or variants that commonly occur in
182 NSCLC was highest for TP53 (4463 cases) followed by KRAS (2637cases) and EGFR (1969
183 cases) (Fig. 2A). The frequency of cases with mutations or variants was comparable for the

184 remainder of the 28 genes commonly mutated in NSCLC (Fig. 2A). From ClinVar, 13 KRAS
185 variants, 8 EGFR variants, 4 PIK3CA variants and 8 BRAF variants reported to be pathogenic in
186 NSCLC were computationally retrieved (Fig. 2B). The report of only 4 out of 28 commonly
187 mutated genes known to be pathogenic suggests that more research is required to gather more
188 information on the pathogenicity of NSCLC genetic variation. From enrichr, a total of 130
189 significant (p-value <0.05) GO terms were obtained for the 28 commonly mutated NSCLC
190 genes, with EGFR, TP53, ERBB4, NTRK3 and KDR genes associated with >40 GO terms (Fig.
191 2C). This suggests that either these 5 highly GO term enriched genes have a multitude of
192 biological functions or have been studied more by researchers relative to the other 23 commonly
193 mutated NSCLC genes. Search of the DGIdb database for drugs that can target the 28 commonly
194 mutated NSCLC genes, retrieved 329 drugs for the NSCLC pathogenic variants knowledgebase
195 genes and 569 drugs for the NSCLC GO terms knowledgebase genes.

197 Visualization of NSCLC pathogenic variants and drug-gene interaction knowledge bases on 198 BioDati Studio

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

206 Visualization of NSCLC GO terms and drug-gene interaction knowledge bases on BioDati 207 Studio

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

222 ‘Best scenario’ patients with genetic profile overlap with NSCLC pathogenic variants and GO 223 terms knowledge bases

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

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

237

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

240 The second category of patients called 'good scenario' comprised of patients (Patient 4, 8, 9, 11
241 and 14) whose genetic profile found matches in only the NSCLC GO terms knowledge base (Fig.
242 6A). Patients 4 and 11 had only one overlapped gene, while patient 8, 9 and 11 had more than
243 one overlapped gene. Though several drugs were suggested to target the overlapped genes from
244 the patients, a score based on number of citations was used to prioritize drug selection (Fig. 6A).
245 As an illustration of the utility of the BioDati Studio for visualization of patient overlapped
246 genes, a zoomed view of Patient 1's overlapped gene KRAS on the NSCLC GO terms
247 knowledge base nanopub and NSCLC drug-gene interaction knowledgebase nanopub is shown
248 (Fig. 6B).

249

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

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

256

257

258 **Discussion:**

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

268

269 The biochemical process by which cells communicate and coordinate activities in
270 response to stimuli from their environment is called cell signaling. Here we found several cells
271 signaling pathways, especially kinase activity and kinase activity regulation through
272 phosphorylation as the major theme underlying the significant GO terms enriched in NSCLC.
273 This is consistent with the presence of several protein kinases, 11 out of 28 genes, among the
274 commonly mutated NSCLC genes, EGFR, STK11, ATM, BRAF, MET, ALK, EPHA3, ERBB4,
275 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

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

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

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

311
312

313 **Bibliography:**

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

322 4. Scelfo C, Galeone C, Bertolini F, Caminati M, Ruggiero P, Facciolongo N, et al.
323 Towards precision medicine: The application of omics technologies in asthma management.
324 F1000Research. 2018;7:423.

325 5. Jia B, Xu S, Xiao G, Lamba V, Liang F. Learning gene regulatory networks from next
326 generation sequencing data. *Biometrics*. 2017;73(4):1221-30.

327 6. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis
328 of complex disease: integrative methods for multi-omics data. *Briefings in bioinformatics*.
329 2018;19(6):1370-81.

330 7. Jurca G, Addam O, Aksac A, Gao S, Ozyer T, Demetrick D, et al. Integrating text
331 mining, data mining, and network analysis for identifying genetic breast cancer trends. *BMC*
332 *research notes*. 2016;9:236.

333 8. Luo Y, Riedlinger G, Szolovits P. Text mining in cancer gene and pathway prioritization.
334 *Cancer informatics*. 2014;13(Suppl 1):69-79.

335 9. Domingo-Fernandez D, Mubeen S, Marin-Llao J, Hoyt CT, Hofmann-Apitius M.
336 PathMe: merging and exploring mechanistic pathway knowledge. *BMC bioinformatics*.
337 2019;20(1):243.

338 10. Kusnoor SV, Koonce TY, Levy MA, Lovly CM, Naylor HM, Anderson IA, et al. My
339 Cancer Genome: Evaluating an Educational Model to Introduce Patients and Caregivers to
340 Precision Medicine Information. *AMIA Joint Summits on Translational Science proceedings*
341 *AMIA Joint Summits on Translational Science*. 2016;2016:112-21.

342 11. Ekmekci B, McAnany CE, Mura C. An Introduction to Programming for Bioscientists: A
343 Python-Based Primer. *PLoS computational biology*. 2016;12(6):e1004867.

344 12. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive
345 and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*. 2013;14:128.

346 13. Perez-Palma E, Gramm M, Nurnberg P, May P, Lal D. Simple ClinVar: an interactive
347 web server to explore and retrieve gene and disease variants aggregated in ClinVar database.
348 *Nucleic acids research*. 2019;47(W1):W99-W105.

349 14. Cotto KC, Wagner AH, Feng YY, Kiwala S, Coffman AC, Spies G, et al. DGIdb 3.0: a
350 redesign and expansion of the drug-gene interaction database. *Nucleic acids research*.
351 2018;46(D1):D1068-D73.

352 15. Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, et al. Genomic
353 landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*. 2012;150(6):1121-
354 34.

355 16. Fujimoto N, Wislez M, Zhang J, Iwanaga K, Dackor J, Hanna AE, et al. High expression
356 of ErbB family members and their ligands in lung adenocarcinomas that are sensitive to
357 inhibition of epidermal growth factor receptor. *Cancer research*. 2005;65(24):11478-85.

358 17. Sattler M, Reddy MM, Hasina R, Gangadhar T, Salgia R. The role of the c-Met pathway
359 in lung cancer and the potential for targeted therapy. *Therapeutic advances in medical oncology*.
360 2011;3(4):171-84.

361 18. Auliac JB, Bayle S, Vergnenegre A, Le Caer H, Falchero L, Gervais R, et al. Patients
362 with non-small-cell lung cancer harbouring a BRAF mutation: a multicentre study exploring
363 clinical characteristics, management, and outcomes in a real-life setting: EXPLORE GFPC 02-
364 14. *Current oncology (Toronto, Ont)*. 2018;25(5):e398-e402.

365 19. Petersen BS, Fredrich B, Hoepfner MP, Ellinghaus D, Franke A. Opportunities and
366 challenges of whole-genome and -exome sequencing. *BMC genetics*. 2017;18(1):14.

367 20. Mani A. Pathogenicity of De Novo Rare Variants: Challenges and Opportunities.
368 Circulation Cardiovascular genetics. 2017;10(6).

369 21. Papadimitriou S, Gazzo A, Versbraegen N, Nachtegaele C, Aerts J, Moreau Y, et al.
370 Predicting disease-causing variant combinations. Proceedings of the National Academy of
371 Sciences of the United States of America. 2019;116(24):11878-87.

372

373 **Access to Codes:**

374 <https://icedrive.net/s/Ww9Y35fjuZu6akSkYa8xaChFQTG6> and

375 https://gitlab.com/smukher2/nsclc_drugtargetsmutations_nov2019. Please cite this paper if you
376 use these codes. Thank you.

377

Figure 1: Workflow showing all steps of project

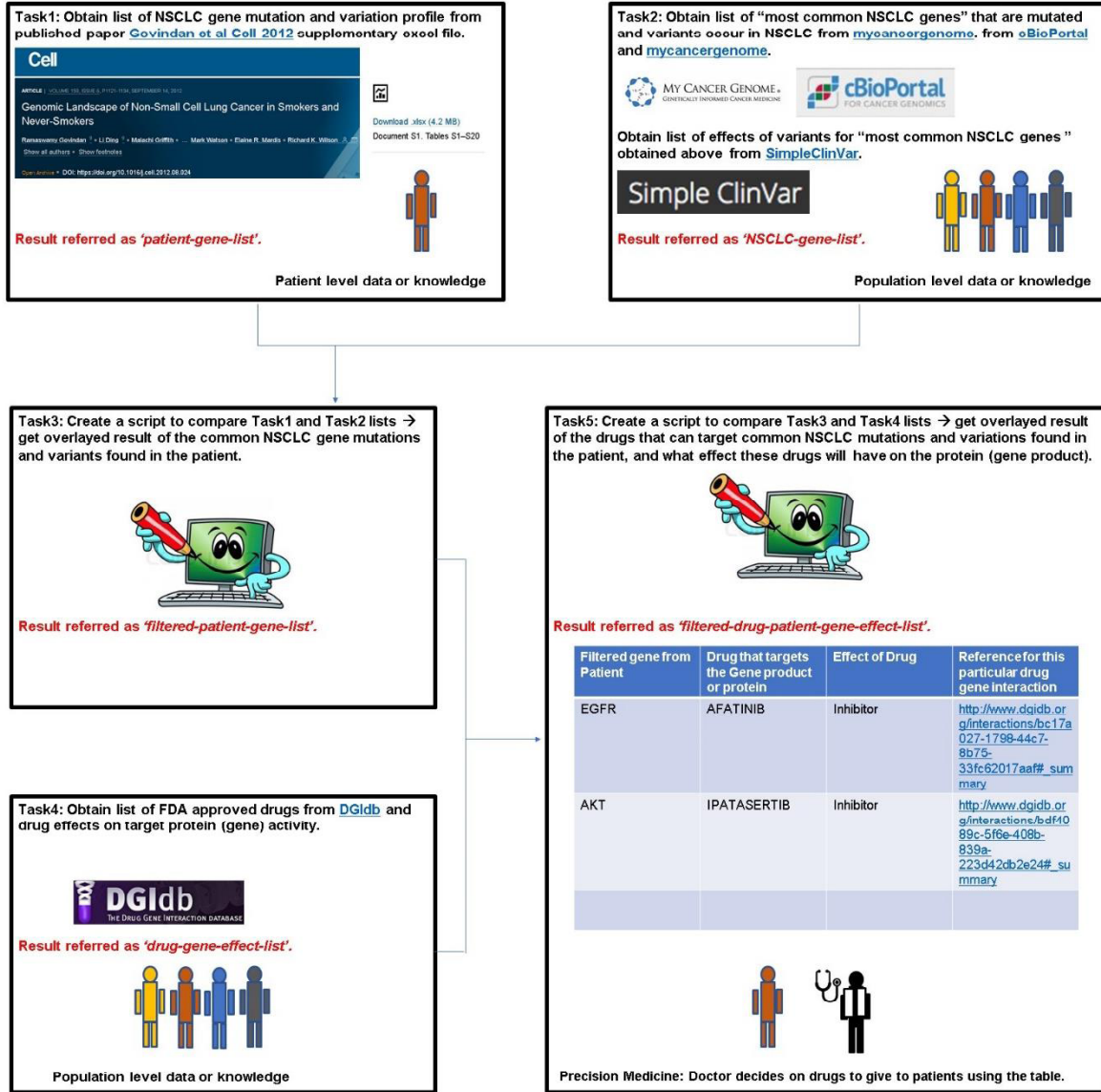
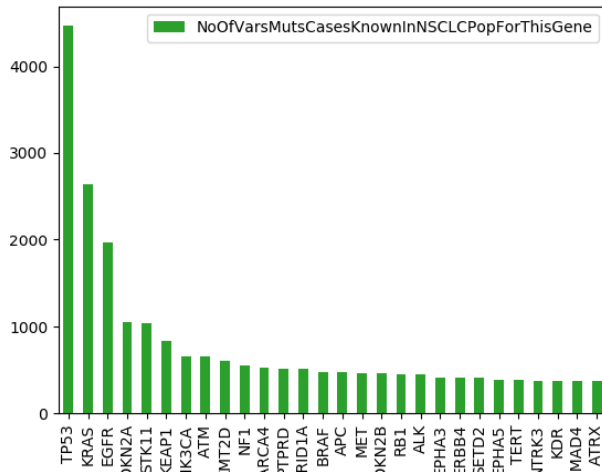


Figure 2: A: Commonly mutated genes in NSCLC obtained from mycancergenome B:

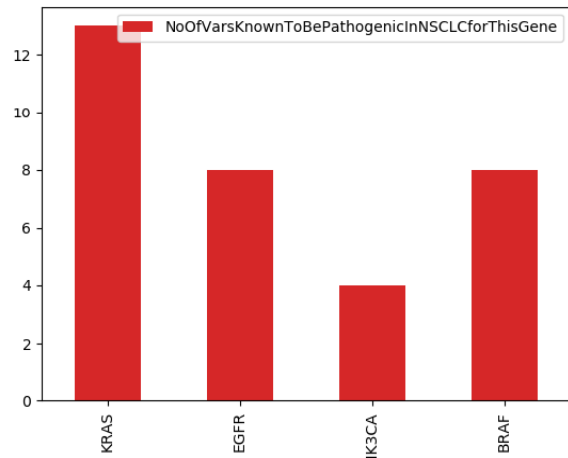
Pathogenic variants reported in ClinVar for commonly mutated NSCLC genes C:

Significant Gene Ontology (GO) terms from enrichr for commonly mutated NSCLC genes

A



B



C

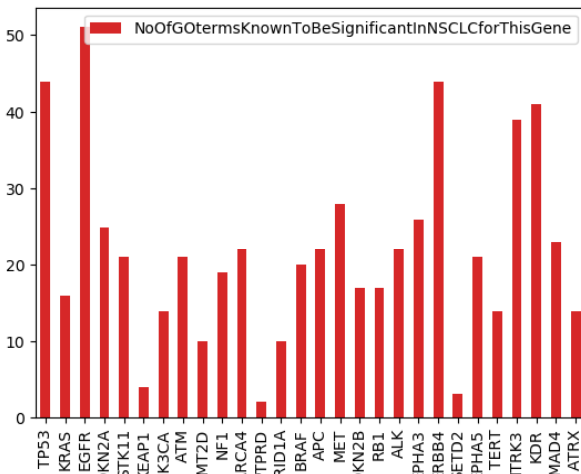
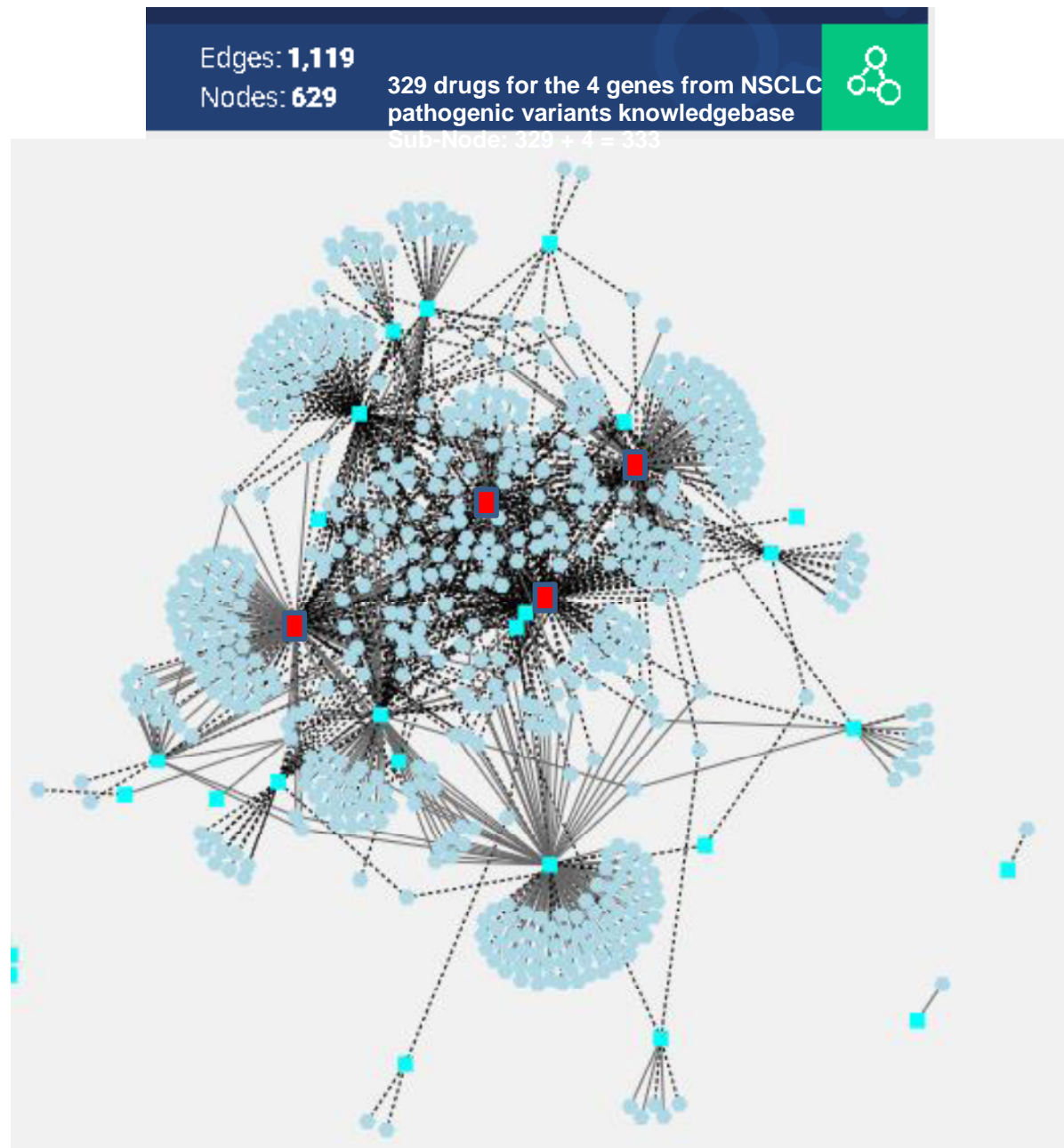
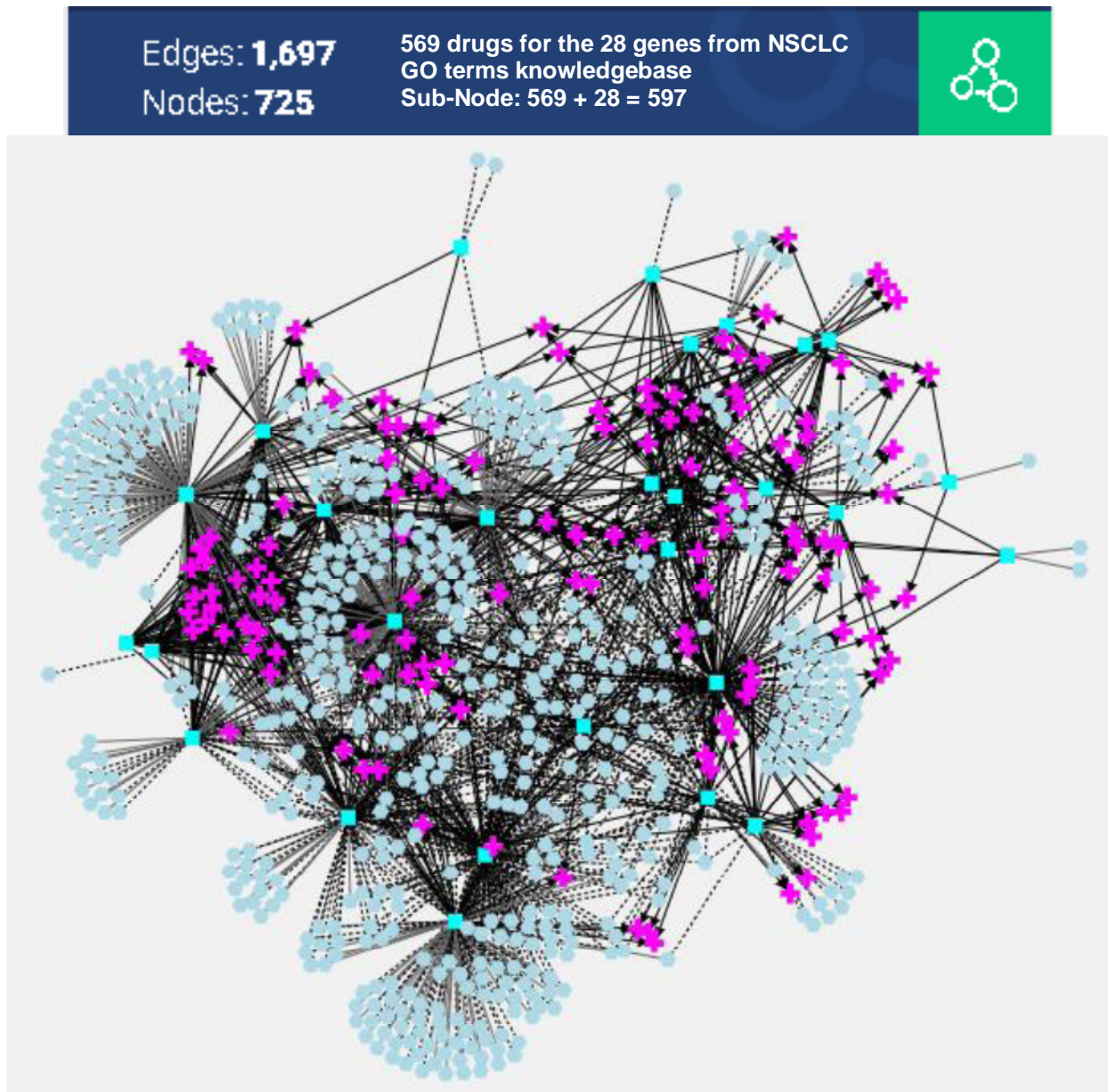


Figure 3: BioDati Studio network for population level NSCLC drug-gene interaction knowledgebase (blue-green and red square) and NSCLC pathogenic variant knowledgebase genes (red square).



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

		<u>NoOfVars In This Patient For This Gene</u>	<u>NoOfVarsMuts Cases Known In NSCLC Pop For This Gene</u>	<u>NoOfVars Known To Be Pathogenic In NSCLC for This Gene</u>	<u>NoOfGOterms Known To Be Significant In NSCLC for This Gene</u>	<u>NoOfDrugs for This Gene</u>	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

