CDK1 and HSP90AA1 appears as novel regulatory gene in Non-Small Cell Lung Cancer: A Bioinformatics Approach

Nirjhar Bhattacharyya¹, Samriddhi Gupta², Shubham Sharma³, Aman Soni³, Malini Bhattacharyya⁴, Atreyee Mukherjee⁵, Ashwini kumar Ray⁶*, Md. Zubbair Malik³*

¹School of Biotechnology, Jawaharlal Nehru University, New Delhi, India
²Department of Biochemistry, University of Hyderabad, India.
³School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India
⁴Department of Environmental Plant Biology, Hemvati Nandan Bahuguna Garhwal Central University, India.
⁵Department of Life Sciences, Presidency University, Kolkata, India
⁶Department of Environmental Studies, University Delhi, New Delhi, India

*Address for Correspondence

Md. Zubbair Malik, Ph. D.
Young Scientist
School of Computational & Integrative Sciences,
Jawaharlal Nehru University,
New Delhi-110067, India
Email: zubbairmalik@jnu.ac.in
ORCID: 0000-0001-6693-0401

Ashwini Kumar Ray, Ph.D.
Assistant professor
Department of Environmental Studies,
University Delhi, New Delhi, India
Email: aray@es.du.ac.in
Abstract

Lung cancer is one of the most invasive cancer affecting over a million of population. Non-small cell lung cancer constitutes up to 85% of all lung cancer cases. Therefore, it is important to identify prognostic biomarkers of NSCLC for therapeutic purpose. The complex behaviour of the NSCLC gene-regulatory network interaction is investigated using a network theoretical approach. We used eight NSCLC microarray datasets GSE19188, GSE118370, GSE10072, GSE101929, GSE7670, GSE33532, GSE31547, GSE31210 and meta analyse them to find differentially expressed genes (DEGs), construct protein-protein interaction (PPI) network, analysed its topological properties, significant modules using network analyser with MCODE, construct a PPI-MCODE network using the genes of the significant modules. We used topological properties such as Maximal Clique Centrality (MCC) and bottleneck from the PPI-MCODE network. We compare them with hub genes (those with highest degrees) to find key regulator (KR) gene. This result is also validated by finding of common genes among top twenty hub genes, genes with highest betweenness, closeness centrality and eigenvector values. It was found that two genes, CDK1 and HSP90AA1 were common in PPI-MCODE combined analysis, and it was also found that CDK1, HSP90AA1 and HSPA8 were common among hub and bottle neck properties and suggesting significant regulatory role of CDK1 in non-small cell lung cancer. After validation, the common genes among top twenty hubs and centrality values like Betweenness Centrality, Closeness Centrality and eigenvector properties, CDK1 again appeared as the common gene. Our study as a summary suggested CDK1 as key regulator gene in complex NSCLC network interaction using network theoretical approach and described the complex topological properties of the network.

Keyword: Non-Small Cell Lung Cancer, key regulator, differentially expressed genes, protein-protein interaction.
**Introduction**

Lung Cancer is the most prevalent cause of death from cancer with more than 1.38 million deaths worldwide [1]. Non-Small Cell Lung Cancer (NSCLC) is a kind of epithelial lung cancer other than Small Cell Lung Carcinoma (SCLC). NSCLC accounts for about 85% of all lung cancers [2]. NSCLs can be further sub categorised into adenocarcinoma (32-40%), squamous cell carcinoma (25-30%) and large cell carcinoma (8-16%) based on which cells of the lung are undergoing uncontrolled proliferation [3]. In adenocarcinomas, cells that secrete mucus proliferates uncontrollably. This type of cancer occurs mainly in current or former smokers, but also the prevalent kind of cancer in non-smokers as well [4]. It is more common in women than in men and has a higher probability to occur in younger people among other types of lung cancers. It is usually found in the outer parts of the lung and is thus likely to be found before it metastasises [5]. In squamous cell carcinoma, flat squamous cells that line the inside of the airways in the lungs undergo proliferation. Most lung cancer patients have a history of smoking [6]. The tumour is usually found in the central part of the Lungs, near bronchi. In the large cell carcinoma, the tumour can grow in any part of the lung [7]. It tends to grow and spread quickly which makes it harder to treat. The other subtypes of NSLCs such as adenosquamous carcinoma and sarcomatoid carcinoma, are much less common. One of the factor smoking is by far the most pressing risk factor for lung cancer. The cigarette smoke constitutes more than 6,000 components, most of which results in DNA damage [8]. Moreover, other sources are subjection to passive smoke, radon, exposure to materials such as soot, beryllium, nickel, chromium, asbestos, or tar, genetic predisposition of lung cancer, and air pollution [9,10]. DNA damage appears to be the primary reason for the cancer. Even though most of the damages can be repaired by the DNA repair system. Further, frequent exposure to smoke and other causative factors makes vulnerable towards it [11]. Additionally, during repair of DNA double-strand breaks, or repair of other DNA damages, incompletely finished sites of repair can cause epigenetic gene silencing [12,13]. The epigenetic gene silencing of DNA repair genes is the most common cause of NSCLC. At the minimum, nine DNA repair genes that usually function in DNA repair pathways are often suppressed by promoter hypermethylation in NSCLC. They are namely NEIL1, WRN, MGMT, ATM, MLH1, MSH2, BRCA1, BRCA2, XRCC5 [14-20]. However, FEN1, which is also a component of DNA repair pathway is expressed at an increased level due to hypomethylation at its promoter region in NSCLC [21]. There aren’t a lot of full-proof
treatment options available, even though how pressing the disease is obscure. The current treatment options include surgical intervention at the early stages namely radical mastectomy. With advancement in the stage of cancer and onset of metastasis, cisplatin-based Chemotherapy coupled with radiotherapy has been the go-to treatment. However, in certain cases multimodal approach is taken, wherein 3rd generation cytotoxic and cytostatic agents such as anti-VGFR and anti EGFR drugs are prescribed [22]. Thus, further research into recognition of new players, drugs and combinatorial therapies is necessary to expand the clinical interest to a broader patient population and to better outcomes in NSCLC. In this study, we utilise network theoretical approach to unveil complexity of non-small cell lung cancer.

Materials and Methods

Data collection

A non-small cell lung cancer data sets were retrieved from Gene Expression Omnibus (GEO), NCBI (GEO, https://www.ncbi.nlm.nih.gov/geo/). An eight dataset such as GSE19188 of 65 control and 91 cancer patients’ samples, GSE118370 of 6 control and 6 cancer patients samples, GSE10072 of 49 control and 58 cancer patients’ samples, GSE101929 of 34 control and 32 cancer patients samples, GSE7670 of 27 control and 27 cancer patients samples, GSE33532 of 20 control and 80 cancer patients samples, GSE31547 of 20 control and 30 cancer patients samples, GSE31210 of 20 control and 226 cancer patients samples were taken into this study. The description of datasets and controls and patients numbers are described in Table 1.

Table 1: List of datasets used in the meta-analysis.

| Microarray datasets | Platforms | Control | Cases | References                  |
|---------------------|-----------|---------|-------|-----------------------------|
| GSE19188            | GPL570    | 65      | 91    | Hou J et al., 2010          |
| GSE118370           | GPL570    | 6       | 6     | Xu L et al., 2018           |
| GSE10072            | GPL96     | 49      | 58    | Landi MT et al., 2008       |
| GSE101929           | GPL570    | 34      | 32    | Mitchell KA et al., 2017    |
| GSE7670             | GPL96     | 27      | 27    | Su LJ et al., 2007; Chen CH et al., 2009 |
| GSE33532            | GPL570    | 20      | 80    | Michael M et al., 2014      |
| GSE31547            | GPL96     | 20      | 30    | -                           |
| GSE31210            | GPL570    | 20      | 226   | Okayama H et al., 2012; Yamauchi M et al., 2012 |
Meta-analysis of the datasets

The summarises the methodology of the integrated analysis of this study (Fig. 1). Meta-analysis of the datasets was conducted using the codes obtained from ImaGEO server R.4.0.2 [23]. The codes obtained from ImaGEO were further modified using quartile normalisation between arrays. The meta-analysis was conducted using maxP value method with Benjamini Hochberg FDR value less than 0.05. The DEGs between cases and control were identified using the R limma program, with a $|\log_{2}\text{FC}| > 1$ and a P-value $< 0.05$ deemed statistically significant.

Gene ontology analysis

The GO annotation and KEGG pathway enrichment details of the DEGs reported previously were obtained using the Database for Annotation Visualization and Integrated Discovery (DAVID; https://david.ncifcrf.gov/). Statistical significance was attributed to findings with a P value less than 0.05. Gene ontology analysis of Up regulated and downregulated genes was carried out using g: Profiler web server with Benjamini Hochberg (FDR $< 0.05$)[24].
**PPI network construction**

To find out good association among different protein on whole genome parameter, PPI network analysis is critical point and providing deep information about the protein function [25]. The analysis of PPI association of DEGs was done with STRING data base (http://string-db.org/) [26]. The differentially expressed genes (DEGs) were submitted to STRING database for PPI network construction. Furthermore, all DEGs were uploaded to the Cytoscape programme for gene network visualisation and key regulator gene recognition, which was accomplished using the CytoHubba plugin.

**Network analysis**

The analysis of topological properties of the network was carried out using an Cytohubba plugin network analyser in Cytoscape 3.7. Another plug-in CytoNCA was used to calculate the EigenVector values. Modules of the primary network was calculated using “leading eigen vector method” algorithm using “igraph” package in R.4.0.2 [27]. Additionally, using default...
parameters (degree cutoff 2, node score cutoff 2, K-core 2, and max depth = 100), the MCODE plugin of Cytoscape was used to define possible functional modules in the PPI network [28].

Key regulator gene identification

A Cytohubba, a Cytoscape plugin was applied for identifying key regulator genes (KR), by using several topological properties. The top ten genes with highest degree, closeness centrality, Bottleneck, MCC (Maximal Clique Centrality), MNC Maximum (Neighborhood Component), radiality, EPC (Edge Percolated Component) and eigenvector topological properties were used, in Cytohubba to extract key regulator genes. Furthermore, top modules were identified by applying MCODE, a Cytoscape plugin (MCODE score ≥ 5) to generate PPI-MCODE network. The key regulator genes were further traced in the PPI-MCODE network to check the progression of signal in the network.

Survival plot analysis of the driver gene

The GEPIA data tool (http://gepia.cancer-pku.cn/) enable researchers to examine functional genomic datasets for associations between genomic and/or phenotypic variables. This tool was used to see whether the expression of hub genes was linked to the survival of PCa patients from the TCGA data. The lung cancer patients were divided into two classes based on median gene expression values, with their overall survivals (OSs) evaluated using the Kaplan– Meier approach with a log-rank test for obtaining the survival plot of the key regulator or driver gene. We validated the DEGs with box plot, analysed the pathological stage and transcript per million of DEGs. The P <0.05 was considered statistically significant.

Results

Identification of differential expressed genes and meta-analysis

Meta-analysis was performed using the ImaGeo server of eight microarray datasets of non-small cell lung cancer yielded a total of 4535 official genes and the construction of gene expression matrix. In this meta-analysis, wilkinsons method or maximum P value method was considered which is most restrictive and used to identify the most robust genes [29]. It considers maximum P value across the datasets, max (p1, p2, …., pi, pk), which is distributed as beta (k,1) under null hypothesis [23, 30]. The fold change obtained after meta-analysis is the average of fold changes across all studies. Therefore, the fold change of each gene gets reduced after meta-analysis as compared to the individual datasets. We considered a P value
less than 0.05 and an arbitrary threshold of \( logFC > 0.1 \) and \( logFC < -0.1 \) to denote up and downregulated genes. The limma package was used for the filtration of differentially expressed gene. After analysis between NSCLC (non-small cell lung cancer) and NC (non-cancer), a total of 2700 differentially expressed gene were identified including 1842 genes were found to be upregulated and 858 genes were found to be downregulated.

**Figure 2:** DEGs identification with volcano plot in a microarray represented the top DEGs in lung cancer cases and controls. The top DEGs are those that follow the logFC value and P 0.05 criterion.

Using log FC value, a volcano plot shows an overview of differentially expressed genes. (Fig. 2). The top 30 upregulated and downregulated DEGs based on fold changes are listed in Table 1.

**Gene ontology analysis of differentially expressed genes**

The DAVID online software was used to for the functional and pathway enrichment analysis of selected DEGs, categorized as biological process (BP), cellular component (CC), and molecular function (MF). The results were deemed as statistically significant if P < 0.05, and the top 30 GO terms of the downregulated and upregulated DEGs are compiled in Supplementary File 1 and Supplementary File 2 respectively.
Figure 3: Gene enrichment analyses of upregulated DEGs (A) The upregulated DEGs' top 30 enriched GO terms with logFC > |0.1| for each of the biological process (BP) (B) The upregulated DEGs' top 30 enriched GO cellular component (CC) (C) The upregulated DEGs' top 30 enriched GO molecular function (MF) (D) The upregulated DEGs' top 30 enriched GO KEGG pathway (KP). Gene Ratio is defined as the ratio between intersection size and query size. Number refers to interaction size, i.e., number of genes corresponding ontology term.

The bubble plot demonstrated that topmost upregulated genes mainly involved in metabolic process, organic substance metabolic process, cellular metabolic process (Fig. 3A). Furthermore, the topmost upregulated gene were present in intracellular, organelle, membrane bound organelle etc. (Fig. 3B). Topmost upregulated genes were involved in binding, protein binding, organic cycle compound binding etc. (Fig. 3C). Apart from above, topmost upregulated gens were involved in metabolic pathway, Alzheimer disease pathway, Huntington disease pathway etc. (Fig. 3D). The topmost downregulated genes were involved in multicellular organismal process, developmental process, localization process etc. under gene ontology BP (Fig. 4A). The downregulated genes were present in cytoplasm, cell periphery,
plasma membrane etc. under gene ontology CC (Fig. 4B) and the topmost downregulated gene involved in protein binding, anion binding, enzyme binding under gene ontology MF (Fig. 4C). and topmost downregulated genes involved in cancer pathways, MAPK kinase signaling pathway, Ras signaling pathway etc. under KEGG pathway analysis (Fig. 4D).

Figure 4: The gene enrichment analyses (A) The downregulated DEGs' top 30 enriched GO terms with logFC > |0.1| for each of the biological process (BP) (B) The downregulated DEGs' top 30 enriched GO cellular component (CC) (C) The downregulated DEGs' top 30 enriched GO molecular function (MF) (D) The downregulated DEGs' top 30 enriched GO KEGG pathway (KP). Gene Ratio is defined as the ratio between intersection size and query size. Number refers to interaction size, i.e., number of genes corresponding ontology term.

Protein-protein interaction network construction analysis

A 2700 DEGs were submitted to STRING for the construction and analysis of network. The constructed network analysis demonstrated that the protein-protein interaction network had 2637 interacting nodes and 31349 edges (Fig. 5 A). Further analysis of network was done by the results which retrieved from Cytoscape. After analysis of network, we were found top 10 hub genes including GAPDH, IL6, CDK1, HSP90AA1, EGF, PTEN, CASP3, HSPA8, CDC20, PLK1 etc as per the decreasing order of highest degree value and also found top 10
bottle neck genes including EIF4A1, HSPA8, GNL3, SPN, CDK, ZEB, HSP90AA1, EPRS, TARS2, DICER1 etc. as per the deceasing order of highest bottleneck score (Fig. 5B, 5C & Supplementary File 3).

Figure 5: (A) network of DEGs protein-protein interactions retrieved from Cytoscape tool. Blue colour node showing upregulated genes and yellow colour node showing downregulated genes of Non-small cell lung cancer vs normal cancer (B) Top 10 hub & (C) bottleneck genes of PPI network retrieved from significant modules (dark colour showing higher value in analysis, all are upregulated except blue which represent downregulated genes) (D-E) Top two modules after integrated analysis of PPI network (D) Modules 1 (E) Module 2. Node showing genes, lines denoted interaction between genes.

A total of 65 modules were identified, out of which 5 cluster significant modules were screened from PPI network with MCODE (Molecular Complex Detection) plugin of the Cytoscape tool (Modules with MCODE score ≥ 5 were considered as significant modules). The modules A was having 68 nodes and 2078 edge with 62.03 MCODE score, modules B was having 161 nodes and 2417 edge with 30.213 MCODE score, modules C was having 31 nodes and 291 edge with 19.4 MCODE score, modules D was having 36 nodes and 308 edge with 17.6 MCODE score, modules E was having 19 nodes and 150 edge with 16.66 MCODE
Supplementary file contains the list of MCODE scores and network parameters. **Supplementary File 4 contains details of modules identified by MCODE** (Fig. 5D & 5E)

The topological properties of this network, namely, probability of degree distribution $P(k)$, neighbourhood connectivity $C_N(k)$ and clustering coefficient $C(k)$ follows power law characters as a function of $k$.

Node degree distribution, clustering coefficient, neighbourhood connectivity follows hierarchical structure of the network. Centrality values, betweenness centrality, closeness centrality and eigen vector supported the hierarchical nature of the network. All six distribution of topological properties were fitted with power laws $y = A_0x^{-A_1}$ (Fig. 6).

**Figure 6:** Topological properties of network. (A) Probability of node degree distribution. (B) Clustering coefficient vs degree distribution. (C) Neighbourhood connectivity vs degree distribution.
Key regulator gene identification

It was clear from Fig. 5 that the upregulated genes were outnumbered than downregulated gene. Most of the hub genes in the network like GAPDH, IL6, CDK1, PTEN etc. are either associated with non-small cell lung cancer itself or other forms of cancer. For identification of key regulators in the PPI network, seven topological properties of top 10 genes were taken to compared with each other. After deep analysis of seven topological algorithm, CDK1 (Cyclin-dependent kinase 1) and HSP90AA1 (heat shock protein 90 alpha family class A member 1) were common and suggesting significant regulatory role of CDK1 and HSP90AA1 in non-small cell lung cancer. CDK1 was top most gene in the all the topological properties after closer inspection of table. Supplementary file showing that three genes including CDK1, HSP90AA1, HSPA8 were common in both hub and bottle neck. This result was further verified by checking the common genes among top twenty hubs and centrality values like Betweenness Centrality, Closeness Centrality and eigenvector values (supplementary File 4 and Fig. 7).
Survival plot analysis of the driver gene

The mRNA expression of CDK1 and HSP90AA1 were found significantly connected with overall survival through the GEPIA online software (Fig. 8A & 9A). The survival plot of the both CDK1 (CDC2) and HSP90AA1 genes were obtained from GEPIA tool and analysed that high expression of both the gene CDK1 and HSP90AA1 were found to be connected with poorer overall survival. The Kaplan Meier (KM) plot also showed that CDK1 and HSP90AA1 expression significantly associated with OS (Fig. 8B & 9B). Moreover, GEPIA tool was used to validate expression of CDK1 and HSP90AA1 gene expression between lung cancer and control tissue in LUAD and LUSC cohort from TGCA data. The mRNA expression of both gene CDK1 and HSP90AA1 were significantly upregulated in both the datasets LUAD and LUSC between lung cancer and non-cancer patients (Fig. 8C & 9C). Further, the relationship was evaluated between expression of CDK1, HSP90AA1 and tumor stage (pathological stage plot) in lung cancer patients via online tool GEPIA. The data revealed that CDK1 expression levels showed strong association with the stage of tumor in
lung cancer patients as well as HSP90AA1 also showed strong association with the stage of tumor in lung cancer patients (Fig. 8D & 9D). The TPM's (Transcripts per million) analysis of RNA-seq data from GEPIA with default parameters revealed that CDK1 was overexpressed in both dataset LUAD and LUSC along with HSP90AA1 was also found overexpressed in both cohorts. (Fig. 8 E & 9E).

Figure 8: The survival, expressional and pathological analysis of CDK1. (A) Survival plot for CDK1 obtained from GEPIA (gepia.cancer-pku.cn). (B) Survival plot of CDK1 (CDC2) from KM-Plotter. (C) Expression profile of CDK1 in normal (grey) and tumour (red) samples of LUAD and LUSC datasets (D) Pathological stage condition for CDK1 genes using LUAD and LUSC datasets. (E) Expression of CDK1 gene based on RNA-seq data from TCGA database assessed by TPM (Transcripts per million) from GEPIA.
**Figure 9**: The survival, expressional and pathological analysis of HSP90AA1. (A) Survival plot for HSP90AA1 obtained from GEPIA (gepia.cancer-pku.cn). (B) Survival plot of HSP90AA1 from KM-Plotter. (C) Expression profile of HSP90AA1 in normal (grey) and tumour (red) samples of LUAD and LUSC datasets. (D) Pathological stage condition for HSP90AA1 genes using LUAD and LUSC datasets. (E) Expression of HSP90AA1 gene based on RNA-seq data from TCGA database assessed by TPM (Transcripts per million) from GEPIA.

**Discussion**

Non-Small Cell Lung Cancer (NSCLC) is a cancer that is highly malignant and has a low 5-year survival rate [31, 32]] as well as NSCLC is responsible for a significant percentage of cancer-related deaths [33]. In this NSCLC study for the identification of potent biomarker, DEGs were profiled through the comparison between lung cancer and normal lung cancer tissue samples of 550 Lung Cancer patients in eight datasets from Gene Expression Omnibus (GEO). According to this functional enrichment analysis, the cell cycle, metabolic pathways, pathways in cancer and the MAPK signalling pathway were all associated to the carcinogenesis of LUAD. In this present study, the most of the DEGs were found to be upregulated. This study revealed two very important gene CDK1 and HSP90AA1 and chosen
for the further analysis. It was found that the expression of both gene CDK1 and HSP90AA1 was upregulated in NSCLC and both genes expression was validated in LUAD as well as LUSC cohort independently. CDK1 and HSP90AA1 were found to be linked to overall survival in a Kaplan-Meier study (OS).

CDKs (cyclin-dependent kinases) are crucial proteins that regulate the cell cycle [34]. Although the cell cycle is a crucial and strictly controlled mechanism that controls cell proliferation and growth but the deviation from normal cell cycle regulation results in uncontrolled proliferation and cancer [35]. Malignancy is indicated by the absence of typical cell-cycle regulation [36]. To control the cell cycle in cancers, several therapeutic approaches have been developed [37]. CDK1 is the catalytic subunit of a protein kinase complex that triggers cell division. The transformation from G1 to S phase and from G2 to Mitotic phase of the cell cycle is regulated by CDK1 gene [34]. CDK1 activity has been shown to be dysregulated frequently in various cancers. Latest data revealed that CDK1 is promising diagnostic and prognostic biomarker as well as target for lung cancer since it is participating in cell cycle progression and dysregulation [38]. Previous study suggesting that targeting the CDK1 has shown reduced apoptotic resistance in colorectal cancer [39]. Another study showed that protein phosphatase 1 alpha mobilisation of CDK1 mediated by positive feedback loop has shown to drive androgen receptor in prostate cancer [40]. It has been also speculated by the study that accumulated cytoplasmic Cdk1 has been correlated with growth and survival rate of epithelial ovarian cancer [41]. Kubo et al observed that the prognosis of stage I and stage II non-small cell lung cancer are found to be dependent on cyclin dependent kinases [42]. Another study showed that the lung carcinogenesis is promoted activation of the GP130/STAT3 signalling pathway iron-dependent activity of CDK1 [43]. It was reported by shi, et al that predictive and prognostic values of CDK1 and MAD2L1 has been described in lung adenocarcinoma as higher expression of CDK1 is correlated to poor survival and higher chances of cancer recurrences [44]. A Cdc2/Cdk1 inhibitor, purvalanol A, has shown to enhance the cytotoxic effects of taxol by Op18/stathmin in in-vitro NSCLC [45]. Furthermore, Dinaciclib has been shown to downregulate proliferation of lung cancer by inhibition of cyclin-dependent kinases 1 and 2, while Cucurbitacin-D-induced upregulation of CDK1 mRNA level has shown to arrest the proliferation of non-small cell lung cancer [46, 47]. Therefore, Li et al and his co-worker concluded that CDK1 appears as a prognostic biomarker and therapeutic target in lung cancer [48]. Nevertheless, further characterisation of CDK1 as a biomarker of NSCLC is essential for the better treatment of lung cancer. Our
study uses network theoretical approach to decipher the complex architecture of NSCLC network and identifies CDK1 as a potent biomarker and therapeutic target in progression of non-small cell lung cancer.

Heat associated protein 90 such as heat shock protein (HSP90) is well known and universal protein. HSP90 consist flexible homodimer and its molecular structure consist of three region such as N-terminal domain, middle domain, and C-terminal domain. The gene of HSP90AA1 (commonly known as HSP90) is situated on the chromosomes 14q32.2 [49]. In the current cancer therapy, HSP90 is centre of attraction for its capability to inhibit to multiple signalling pathway simultaneously [50]. A study reported by prominent scientist that IPI 504, a powerful HSP90 inhibitor, has shown therapeutic response in patients with NSCLC, especially those with anaplastic lymphoma kinase gene rearrangements [51]. HSP90 has been highly expressed in multiple cancer like lung, ovarian, endometrial and pancreatic cancer, additionally oropharyngeal SCC and various myeloma [52-55]. Latest study revealed that high HSP90 expression was found poor prognosis marker in various cancer such as lung cancer, melanoma, esophageal cancer, bladder cancer as well as leukaemia [56-59]. Apart from above, HSP90 has been confirmed to play a role in a variety of cancers as well as overexpression of HSP90 has been suggested to improve carcinogenesis and influence patient prognosis [60-64]. The current study found that the presence or absence of HSP90 expression has an effect on the OS rate of NSCLC patients. Another study found that HSP90AA1 protein product HSP90α is thought to play a key role in tumour invasion and migration regulation [65, 66]. From our study, it is possible to conclude that high HSP90AA1 mRNA expression is associated with a poorer prognosis of OS in all NSCLC patients. Current research suggested that HSP90 could be a critical drug target for the treatment of NSCLC. As for example, mitochondrial HSP90 network repression can be an effective treatment for refractory tumours [67] as well as introduction of HSP90 inhibitor 17AAG prevents epidermal hyperplasia and SCC [68]. HSP90AA1 affects the patient’s reaction with systemic lupus erythematosus to glucocorticoids and the lipopolysaccharide-induced inflammatory response, including tumour necrosis factor secretion by monocytes, according to previously published findings [69]. Moreover, the role of HSP90AA1 in cancer pathogenesis is a hot topic of research right now. According to previous study, high HSP90AA1 expression was an independent factor correlated with mortality in breast cancer patients (triple-negative and human epidermal growth factor receptor 2-negative/estrogen receptor-positive subtypes) [70].
Conclusions

In conclusion, a present stratified research recognised the multiple important genes including CDK1, HSP90AA1 which showed differential expression in non-small cell lung cancer patients compared to control through DEG analysis and network building. Our integrated analysis suggested that CDK1 and HSP90AA1 could be significant potential biomarker for the diagnosis and prognosis of NSCLC as well as both CDK1 and HSP90AA1 could be a potential therapeutic drug target in the treatment of lung cancer. However, there are some drawbacks to the current research that need to be addressed. Moreover, the current findings were extracted by the application of bioinformatics techniques and more sample size, efficient research is needed to fully understand the underlying molecular mechanisms by in vivo experiment in future.

Acknowledgement

M.Z.M. acknowledges financial assistance from Department of Health Research, Ministry of Health and Family Welfare, Government of India under Young Scientist scheme (Sanction File No. R.12014/01/2018-HR, FTS No. 3146887). We acknowledge Jordi Martorell-Marugán (Genyo Bioinformatics Unit) for his immense support for every query despite his busy schedule.

Author contributions

M.Z.M. conceived the model. N.B., A.K.R. and M.Z.M. and conducted numerical experiments. N.B. and M.Z.M. prepared figures of the numerical results. N.B., S.G., S.S., A.S., M.B., A.M., A.K.R. and M.Z.M. analysed and interpreted the simulation results and wrote the manuscript. A.K.R. and M.Z.M jointly supervised the study, and all authors approved the final draft.

Competing interests

The authors declare no competing interests.

References

1. Jemal, A. et al. Cancer statistics, 2008. CA Cancer J Clin. 58(2), 71-96 (2008).
2. Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E. & Adjei, A. A. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc. 83(5), 584-94 (2008).
3. Chansky, K. et al. IASLC Staging and Prognostic Factors Committee, Advisory Boards, and Participating Institutions. The IASLC Lung Cancer Staging Project: External Validation of the Revision of the TNM Stage Groupings in the Eighth Edition of the TNM Classification of Lung Cancer. J Thorac Oncol. 12(7), 1109-1121 (2017).
4. Subramanian, J. & Govindan, R. Lung cancer in never smokers: a review. *J Clin Oncol.* **25**(5), 561-70 (2007).

5. Mitchell, E. D., Pickwell-Smith, B. & Macleod, U. Risk factors for emergency presentation with lung and colorectal cancers: a systematic review. *BMJ Open.* **5**(4), e006665 (2015).

6. Kenfield, S. A., Stampfer, M. J., Rosner, B. A. & Colditz, G. A. Smoking and smoking cessation in relation to mortality in women. *JAMA.* **299**(17), 2037-47 (2008).

7. Popper, H. H. Progression and metastasis of lung cancer. *Cancer Metastasis Rev.* **35**(1), 75-91 (2016).

8. Liu, X. *et al.* Cigarette smoke extract induces DNA damage but not apoptosis in human bronchial epithelial cells. *Am J Respir Cell Mol Biol.* **33**(2), 121-9 (2005).

9. Coultas, D. B. & Samet, J. M. Occupational lung cancer. *Clin Chest Med.* **13**(2), 341-54 (1992).

10. Edling, C., Kling, H. & Axelson, O. Radon in homes--a possible cause of lung cancer. *Scand J Work Environ Health.* **10**(1), 25-34 (1984).

11. Kastan, M. B. DNA damage responses: mechanisms and roles in human disease: 2007 G.H.A. Clowes Memorial Award Lecture. *Mol Cancer Res.* **6**(4):517-24 (2008).

12. O'Hagan, H. M., Mohammad, H. P. & Baylin, S. B. Double strand breaks can initiate gene silencing and SIRT1-dependent onset of DNA methylation in an exogenous promoter CpG island. *PLoS Genet.* **4**(8), e1000155 (2008).

13. Cuozzo, C. *et al.* DNA damage, homology-directed repair, and DNA methylation. *PLoS Genet.* **3**(7), e110 (2007).

14. Do, H. *et al.* A critical re-assessment of DNA repair gene promoter methylation in non-small cell lung carcinoma. *Sci Rep.* **4**, 4186 (2014).

15. Agrelo, R. *et al.* Epigenetic inactivation of the premature aging Werner syndrome gene in human cancer. *Proc Natl Acad Sci U S A.* **103**(23), 8822-7 (2006).

16. Wolf, P., Hu, Y. C., Doffek, K., Sidransky, D. & Ahrendt, S. A. O(6)-Methylguanine-DNA methyltransferase promoter hypermethylation shifts the p53 mutational spectrum in non-small cell lung cancer. *Cancer Res.* **61**(22), 8113-7 (2001).

17. Ekim, M., Caner, V., Büyükpınarbaşılı, N., Tepeli, E., Elmas, L. & Bağci, G. Determination of O(6)-methylguanine DNA methyltransferase promoter methylation in non-small cell lung cancer. *Genet Test Mol Biomarkers.* **15**(5), 357-60 (2011).

18. Safar, A. M. *et al.* Methylation profiling of archived non-small cell lung cancer: a promising prognostic system. *Clin Cancer Res.* **11**(12), 4400-5 (2005).

19. Gomes, A., Reis-Silva, M., Alarcão, A., Couceiro, P., Sousa, V. & Carvalho, L. Promoter hypermethylation of DNA repair genes MLH1 and MSH2 in adenocarcinomas and squamous cell carcinomas of the lung. *Rev Port Pneumol.* **20**(1), 20-30 (2014).

20. Lee, M. N. *et al.* Epigenetic inactivation of the chromosomal stability control genes BRCA1, BRCA2, and XRCC5 in non-small cell lung cancer. *Clin Cancer Res.* **13**(3), 832-8 (2007).

21. Nikolova, T., Christmann, M. & Kaina, B. FEN1 is overexpressed in testis, lung and brain tumors. *Anticancer Res.* **29**(7), 2453-9 (2009).

22. Zarogoulidis, K. *et al.* Treatment of non-small cell lung cancer (NSCLC). *J Thorac Dis.* **5** Suppl 4(Suppl 4), S389-96 (2013).

23. Toro-Domínguez, D., Villatoro-García, J. A., Martorell-Marugán, J., Román-Montoya, Y., Alarcón-Ríquelmé, M. E. & Carmona-Sáez, P. A survey of gene expression meta-analysis: methods and applications. *Brief Bioinform.* **22**(2):1694-1705 (2021).

24. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**(W1), W191-W198 (2019).

25. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**(6), 957–968 (2005).

26. Von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork. P. & Snel, B. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research* **31**(1), 258–261 (2003).

27. Newman, M. E. Modularity and community structure in networks. *Proc Natl Acad Sci U S A.* **103**(23), 8577-82 (2006).
28. Bandettini, W. P. et al. MultiContrast Delayed Enhancement (MCODE) improves detection of subendocardial myocardial infarction by late gadolinium enhancement cardiovascular magnetic resonance: a clinical validation study. *J Cardiovasc Magn Reson.* **14**(1), 83 (2012).
29. Heard, N. & Rubin-Delanchy, P. Choosing between methods of combining p-values. *Biometrika* **105**, 239–46 (2018).
30. Li, J. & Tseng, G. C. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann Appl Stat.* **5**, 994–1019 (2011).
31. Minna, J. D., Roth, J. A. & Gazdar, A. F. Focus on lung cancer. *Cancer Cell.* **1**(1):49-52 (2002 Feb).
32. Lukas, R. V., Gondi, V., Kamson, D. O., Kumthekar, P. & Salgia, R. State-of-the-art considerations in small cell lung cancer brain metastases. *Oncotarget.* **8**(41), 71223-71233 (2017).
33. Zhao, Y. et al. Sirtuin 7 promotes non-small cell lung cancer progression by facilitating G1/S phase and epithelial–mesenchymal transition and activating AKT and ERK1/2 signaling. *Onco Rep.* **44**(3), 959-972 (2020).
34. Malumbres, M. & Barbacid, M. Cell cycle, CDKs and cancer: a changing paradigm. *Nat Rev Cancer.* **9**(3), 153-66 (2009).
35. Matera, R. & Saif, M. W. New therapeutic directions for advanced pancreatic cancer: cell cycle inhibitors, stromal modifiers and conjugated therapies. *Expert Opin Emerg Drugs.* **22**(3), 223-233 (2017).
36. Dominguez-Brauer, C., Thu, K. L., Mason, J. M., Blaser, H., Bray, M. R. & Mak, T. W. Targeting Mitosis in Cancer: Emerging Strategies. *Mol Cell.* **60**(4), 524-36 (2015).
37. He, Y. C. et al. Curcumin Nicotinate Selectively Induces Cancer Cell Apoptosis and Cycle Arrest through a P53-Mediated Mechanism. *Molecules.* **24**(22), 4179 (2019).
38. Zhang, L., Peng, R., Sun, Y., Wang, J., Chong, X. & Zhang, Z. Identification of key genes in non-small cell lung cancer by bioinformatics analysis. *PeerJ.* **7**, e8215 (2019).
39. Liu, X. et al. Positive feedback loop mediated by protein phosphatase 1α mobilization of P-TEFb and basal CDK1 drives androgen receptor in prostate cancer. *Nucleic Acids Res.* **45**(7), 3738-3751 (2017).
40. Yang, W. et al. Accumulation of cytoplasmic Cdk1 is associated with cancer growth and survival rate in epithelial ovarian cancer. *Oncotarget.* **7**(31), 49481-49497 (2016).
41. Kubo, H. et al. Cyclin-dependent kinase-specific activity predicts the prognosis of stage I and stage II non-small cell lung cancer. *BMC Cancer.* **14**, 755 (2014).
42. Shi, Y. X. et al. Iron-dependent CDK1 activity promotes lung carcinogenesis via activation of the GP130/STAT3 signaling pathway. *Cell Death Dis.* **10**(4), 297 (2019).
43. Chen, X., Liao, Y., Long, D., Yu, T., Shen, F. & Lin X. The Cdc2/Cdk1 inhibitor, purvalanol A, enhances the cytotoxic effects of taxol through Op18/stathmin in non-small cell lung cancer cells in vitro. *Int J Mol Med.* **40**, 235-242 (2017).
44. Danilov, A. V. et al. Dinaciclib Induces Anaphase Catastrophe in Lung Cancer Cells via Inhibition of Cyclin-Dependent Kinases 1 and 2. *Mol Cancer Ther.* **15**(11), 2758-2766 (2016).
45. Jacquot, C. et al. Cucurbitacin-D-induced CDKI mRNA up-regulation causes proliferation arrest of a non-small cell lung carcinoma cell line (NSCLC-N6). *Anticancer Res.* **34**(9):4797-806. PMID: 25202060 (2014).
46. Li, M., He, F., Zhang, Z., Xiang, Z. & Hu. D. CDK1 serves as a potential prognostic biomarker and target for lung cancer. *J Int Med Res.* **48**(2), 300060519897508 (2020).
47. Gallegos Ruiz, M. I. et al. Integration of gene dosage and gene expression in non-small cell lung cancer, identification of HSP90 as potential target. *PLoS One.* **3**(3), e0001722 (2008).
50. Workman, P., Burrows, F., Neckers, L. & Rosen, N. Drugging the cancer chaperone HSP90: Combinatorial therapeutic exploitation of oncogene addiction and tumor stress. *Ann NY Acad Sci.* **1113**, 202-216 (2007).

51. Sequist, L. V. *et al.* Activity of IPI-504, a novel heat shock protein 90 inhibitor, in patients with molecularly defined non-small cell lung cancer. *J Clin Oncol* **28**, 4953-4960 (2010).

52. Burrows, F., Zhang, H. & Kamal, A. Hsp90 activation and cell cycle regulation. *Cell Cycle.* **3**, 1530-1536 (2004).

53. Kolosenko, I., Grander, D. & Tamm, K.P. IL-6 activated JAK/STAT3 pathway and sensitivity to Hsp90 inhibitors in multiple myeloma. *Curr Med Chem.* **21**, 3042-3047 (2014).

54. Patel, K. *et al.* Heat shock protein 90 (HSP90) is overexpressed in p16-negative oropharyngeal squamous cell carcinoma, and its inhibition in vitro potentiates the effects of chemoradiation. *Cancer Chemother Pharmacol.* **74**, 1015-1022 (2014).

55. Shi, Y. *et al.* Plasma levels of heat shock protein 90 alpha associated with lung cancer development and treatment responses. *Clin Cancer Res.* **20**, 6016-6022 (2014).

56. Huang, T. *et al.* Expression of Hsp90alpha and cyclin B1 were related to prognosis of esophageal squamous cell carcinoma and keratin pearl formation. *Int J Clin Exp Pathol.* **7**, 1544-1552 (2014).

57. Tian, W. L. *et al.* High expression of heat shock protein 90 alpha and its significance in human acute leukemia cells. *Gene* **542**, 122-128 (2014).

58. McCarthy, M. M. *et al.* HSP90 as a marker of progression in melanoma. *Ann Oncol.* **19**, 590-594 (2008).

59. Záčková, M., Moučková, D., Lopotová, T., Ondračková, Z., Klamová, H. & Moravcová, J. Hsp90-a potential prognostic marker in CML. *Blood Cells Mol Dis.* **50**, 184-189 (2013).

60. Calderwood, S. K. & Neckers, L. Hsp90 in cancer: Transcriptional roles in the nucleus. *Adv Cancer Res.* **129**, 89-106 (2016).

61. Zhang, B. *et al.* Aberrantly upregulated TRAP1 is required for tumorigenesis of breast cancer. *Oncotarget.* **6**, 44495-44508 (2015).

62. Bieszczad, C. M., Mancini, M., Benazzi, C. & Della Salda, L. Heat shock protein 90 is associated with hyperplasia and neoplastic transformation of canine prostatic epithelial cells. *J Comp Pathol.* **150**, 393-398 (2014).

63. Badowska-Kozakiewicz, A. M. & Malicka, E. Immunohistochemical evaluation of expression of heat shock proteins HSP70 and HSP90 in mammary gland neoplasms in bitches. *Pol J Vet Sci.* **15**, 209-214 (2012).

64. Lee, J. H. *et al.* Differential expression of heat shock protein 90 isoforms in small cell lung cancer. *Int J Clin Exp Pathol* **8**, 9487-9493 (2015).

65. Wong, D. S. & Jay, D. G. Emerging roles of extracellular Hsp90 in cancer. *Adv Cancer Res.* **129**, 141-163 (2016).

66. Sims, J. D., McCready, J. & Jay, D. G. Extracellular heat shock protein (Hsp)70 and Hsp90α assist in matrix metalloproteinase-2 activation and breast cancer cell migration and invasion. *PLoS One.* **6**, e18848 (2011).

67. Siegelin, M. D. Inhibition of the mitochondrial Hsp90 chaperone network: A novel, efficient treatment strategy for cancer? *Cancer Lett.* **333**, 133-146 (2013).

68. Singh, A. *et al.* Topically applied Hsp90 inhibitor 17AAG inhibits UVR-induced cutaneous squamous cell carcinomas. *J Invest Dermatol* **135**, 1098-1107 (2015).

69. Zou, Y. F. *et al.* Single nucleotide polymorphisms of HSP90AA1 gene influence response of SLE patients to glucocorticoids treatment. *Springerplus.* **5**, 222 (2016).

70. Cheng, Q. *et al.* Amplification and high-level expression of heat shock protein 90 marks aggressive phenotypes of human epidermal growth factor receptor 2 negative breast cancer. *Breast Cancer Res.* **14**, R62 (2012).
List of Figures

Figure 1: Flowchart of the methodology showing network analysis of non-small cell lung cancer.

Figure 2: DEGs identification with volcano plot in a microarray represented the top DEGs in lung cancer cases and controls. The top DEGs are those that follow the logFC value and P 0.05 criterion.

Figure 3 (A-D): The upregulated DEGs enrichment analyses (A) The upregulated DEGs' top 30 enriched GO terms with logFC > |0.1| for each of the biological process (BP) (B) The upregulated DEGs' top 30 enriched GO cellular component (CC) (C) The upregulated DEGs' top 30 enriched GO molecular function (MF) (D) The upregulated DEGs' top 30 enriched GO KEGG pathway (KP). Gene Ratio is defined as the ratio between intersection size and query size. Number refers to interaction size, i.e., number of genes corresponding ontology term.

Figure 4 (A-D): The downregulated DEGs enrichment analyses (A) The downregulated DEGs' top 30 enriched GO terms with logFC > |0.1| for each of the biological process (BP) (B) The downregulated DEGs' top 30 enriched GO cellular component (CC) (C) The downregulated DEGs' top 30 enriched GO molecular function (MF) (D) The downregulated DEGs' top 30 enriched GO KEGG pathway (KP). Gene Ratio is defined as the ratio between intersection size and query size. Number refers to interaction size, i.e., number of genes corresponding ontology term.

Figure 5 (A-E): (A) network of DEGs protein-protein interactions retrieved from Cytoscape tool. Blue colour node showing upregulated genes and yellow colour node showing downregulated genes of Non-small cell lung cancer vs normal cancer (B) Top 10 hub & (C) bottleneck genes of PPI network retrieved from significant modules (dark colour showing higher value in analysis, all are upregulated except blue which represent downregulated genes) (D-E) Top two modules after integrated analysis of PPI network (D) Modules 1 (E) Module 2. Node showing genes, lines denoted interaction between genes.
Figure 6 (A-F): Topological properties of network. (A) Probability of node degree distribution. (B) Clustering coefficient vs degree distribution. (C) Neighbourhood connectivity vs degree distribution. (D) Betweenness centrality vs degree distribution. (E) Closeness centrality vs degree distribution. (F) Eigenvector value vs degree distribution

Figure 7 (A-C): Identifying common genes among top twenty hubs, betweenness, closeness centrality and eigenvector values and common genes among hubs, MCODE-PPI MCC and bottleneck genes.

Figure 8(A-E): The survival, expressional and pathological analysis of CDK1. (A) Survival plot for CDK1 obtained from GEPIA (gepia.cancer-pku.cn). (B) Survival plot of CDK1 (CDC2) from KM-Plotter. (C) Expression profile of CDK1 in normal (grey) and tumour (red) samples of LUAD and LUSC datasets (D) Pathological stage condition for CDK1 genes using LUAD and LUSC datasets. (E) Expression of CDK1 gene based on RNA-seq data from TCGA database assessed by TPM (Transcripts per million) from GEPIA.

Figure 9(A-E): The survival, expressional and pathological analysis of HSP90AA1. (A) Survival plot for HSP90AA1 obtained from GEPIA (gepia.cancer-pku.cn). (B) Survival plot of HSP90AA1 from KM-Plotter. (C) Expression profile of HSP90AA1 in normal (grey) and tumour (red) samples of LUAD and LUSC datasets (D) Pathological stage condition for HSP90AA1 genes using LUAD and LUSC datasets. (E) Expression of HSP90AA1 gene based on RNA-seq data from TCGA database assessed by TPM (Transcripts per million) from GEPIA.

Table Caption

Table 2: List of datasets used in the meta-analysis.

Supplementary Files

Supplementary File1: Gene enrichment analysis of upregulated DEGs in lung datasets.

Supplementary File2: Gene enrichment analysis of down regulated DEGs in lung datasets.

Supplementary File 3: Top ten topological properties of genes.
**Supplementary File 4:** List of genes according to rank of topological properties.