Am J Clin Exp Immunol 2024;13(2):58-67
www.ajcei.us /ISSN:2164-7712/AJCEI0156730

Original Article

Insight into NSCLC through novel analysis of gene interactions and characteristics

Eric Pan[1], Yongsheng Bai[2]

[1]*Debakey High School, Houston, TX 77030, USA;* [2]*Next-Gen Intelligent Science Training, Ann Arbor, MI 48105, USA*

Received March 21, 2024; Accepted April 23, 2024; Epub April 25, 2024; Published April 30, 2024

Abstract: Around 80 to 85% of all lung cancers are non-small cell lung cancer (NSCLC). Previous research has aimed at exploring the genetic basis of NSCLC through individual approaches, but studies have yet to investigate the results of combining them. Here we show that analyzing NSCLC genetics through three approaches simultaneously creates unique insights into our understanding of the disease. Through a combination of previous research and bioinformatics tools, we determined 35 NSCLC candidate genes. We analyzed these genes in 3 different approaches. First, we found the gene fusions between these candidate genes. Second, we found the common superfamilies between genes. Finally, we identified mutational signatures that are possibly associated with NSCLC. Each approach has its individual, unique results. Fusion relationships identify specific gene fusion targets, common superfamilies identify possible avenues to determine novel target genes, and identifying NSCLC associated mutational signatures has diagnostic and prognostic benefits. Combining the approaches, we found that gene CD74 has significant fusion relationships, but it has no association with the other two approaches, suggesting that CD74 is associated with NSCLC mainly because of its fusion relationships. Targeting the gene fusions of CD74 may be an alternative NSCLC treatment. This genetic analysis has indeed created unique insight into NSCLC genes. Both the results from each of the approaches separately and combined allow pursuit of more effective treatment strategies for this cancer. The methodology presented can also apply to other cancers, creating insights that current analytical methods could not find.

Keywords: Non-small cell lung cancer (NSCLC), gene fusions, common superfamilies, mutational signatures, *CD74*

Introduction

In recent years, there have been many genetic innovations and advancements for non-small cell lung cancer (NSCLC) treatment. For example, Guardant 360 is able to detect cell-free circulating tumor DNA (ctDNA) in blood specimens of solid tumors and assess a targeted panel of eighty three genes as shown at https://www.ncbi.nlm.nih.gov/gtr/tests/527948/methodology/ (Guardant360 - Clinical test - NIH Genetic Testing Registry (GTR) - NCBI).

However, lung cancer remains the leading cause of cancer related deaths worldwide, and NSCLC consists of around 80 to 85% of all lung cancers as mentioned by American Cancer Society (Lung Cancer Statistics, How Common is Lung Cancer? https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html). Therefore, further analysis of the genetics of NSCLC is essential for improved and alternative NSCLC treatment.

Many researchers throughout the years have analyzed the genetics of NSCLC using various approaches. This research will focus on three unique approaches: gene fusions, common superfamilies, and mutational signatures. First, gene fusions are chromosomal rearrangements that form a hybrid gene from two initially independent genes, which could create novel promoter or enhancer regions, which would then cause dysregulation of gene expression [1]. Significant gene fusions, which are found to be oncogenic drivers in cancer, in NSCLC have been identified, such as ALK-ROS1 gene fusions, and targeted treatment has significantly improved the mortality rate of NSCLC [2].

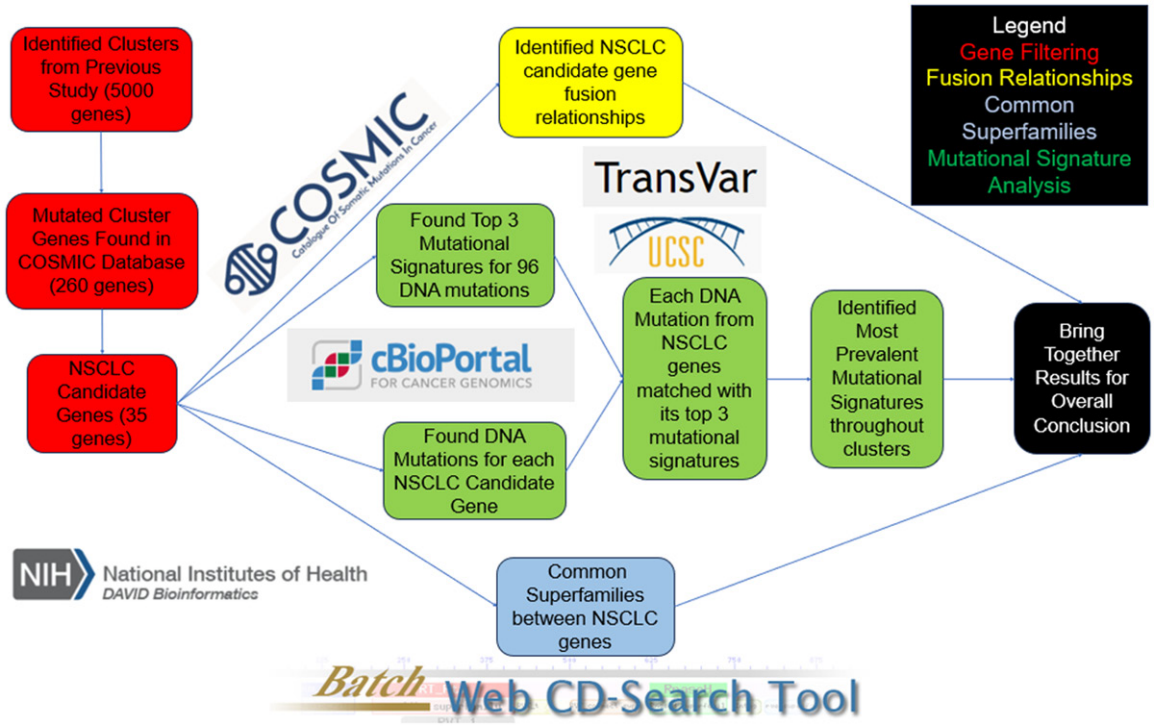

Figure 1. Methods Workflow. Bioinformatics tools used for corresponding steps in the method are illustrated with the tool's logo. Each color represents a specific section of the materials and methods.

The second approach will be common superfamilies, which are based on protein domain identification. Protein domains are the basic units of proteins which have the ability to evolve and fold independently, and these domains have proven to be essential for protein classification, structure, and biological function [3]. Furthermore, gene superfamilies are a set of conserved domain models that overlap in biological function, structure, or sequence [4]. Conserved domain models in the same superfamily are implied to have an evolutionary relationship with each other, suggesting similar annotation. Moreover, one superfamily can be associated with hundreds of genes, and one gene can be associated with multiple superfamilies.

Third, mutational signatures, a unique combination of mutation types caused by different mutational processes, have been found to be associated with certain cancers [1]. In particular, SBS4, caused by tobacco smoking, is well-known for being associated with NSCLC [1]. Identifying mutational signatures associated with cancer has the ability to serve as a biomarker for cancer prognosis [5]. Moreover, mutational signature identification can also be predictors of therapy response in cancer [6].

Previous studies have analyzed each of these three approaches separately, but there has never been research that has attempted to analyze the genetics of NSCLC using these three avenues simultaneously. In this study, we implement a novel method in analyzing the genetics of NSCLC using three different ways: gene fusions, common superfamilies, and mutational signatures. By doing so, unique insight, like novel gene targets, into NSCLC may be gained, which could be used to pave the way for improved or novel NSCLC treatment.

Materials and methods

An overview of the materials and methods is illustrated (**Figure 1**).

Selecting input genes for NSCLC analysis

Previous research used the Louvain Algorithm to detect "communities" of miRNA-mRNA pairs, which they termed as clusters [7]. They identified numerous clusters for 15 different cancers, one of which included lung adenocarcino-

ma (LUAD). Although we derived our data from Dai et al. it's important to note that Dai et al. derived their data from The Cancer Genome Atlas (TCGA), so our research inadvertently employs TCGA.

In addition, The Cancer Gene Census from COSMIC is a list of genes identified by the COSMIC database (COSMIC - Catalogue of Somatic Mutations in Cancer at https://cancer.sanger.ac.uk/) that were found to be somatically mutated and have a causal relationship with cancer [8]. Periodically updated with new genes, the Cancer Gene Census (n=743) is an important reference for this research. Furthermore, COSMIC dubs a number of genes from the Cancer Gene Census as COSMIC Classic Genes (n=272), genes that are expertly curated from COSMIC, with an emphasis on genes that are not found in any other database. Therefore, considering COSMIC Classic Genes is an exciting avenue of gene identification that may lead to novel results.

There were 9 statistically significant (FDR<0.1) LUAD clusters identified from Dai et al. [7], and additionally, we also included another cluster that wasn't identified to be statistically significant by Dai et al. [9] but had a *p*-value of 9.2E-3. The total number of the genes from the 10 clusters was approximately 5000. We then compared these 5000 genes with the 743 Cancer Gene Census genes identified from COSMIC. Only genes that were in both gene lists would be further considered for NSCLC analysis. As a result, we ended up with approximately 260 mutated LUAD genes for gene analysis.

To further improve accuracy, we filtered these genes again using two different databases: COSMIC and DAVID [9, 10]. With COSMIC's concern with somatically mutated genes and DAVID's focus on gene annotation, these two databases' functions complement each other, ensuring that there would be extremely high accuracy. Genes that were found to be implicated in NSCLC from at least one of the two databases would undergo NSCLC analysis, which would ensure that all the genes would be significant in NSCLC. Of the approximately 260 genes, COSMIC identified 15 genes associated with NSCLC and DAVID identified 26 genes associated with NSCLC. DAVID identified some genes associated with NSCLC which were not in the COSMIC Cancer Gene Census but were part of the clusters identified [7]. Therefore, these genes would also undergo NSCLC analysis.

Because of the rigorous filtering, we have identified 35 mutated NSCLC genes (one of the genes *ALK* was identified by both databases), some of which have not been found in any other database. Previously overlooked genes among these 35 may serve as novel gene targets for NSCLC.

Gene fusion analysis

We employed the COSMIC database in order to determine the gene fusions of the 35 genes. COSMIC manually curates its gene fusions from peer-reviewed publications. For each gene fusion published in the database, a comprehensive literature curation is available. Each of the 35 genes was examined for all of its gene fusions in the COSMIC database, and if a gene of the 35 had a gene fusion with another gene of the 35, then the fusion relationship was recorded.

However, this gene fusion analysis has something that is unique. Many researchers have considered gene fusions as simply a relationship between two genes, but we believe that gene fusions may represent a network of gene interaction. For example, if *ROS1* had a gene fusion with *CD74*, and *CD74* was also found to have a gene fusion with *NRG1*, then it is possible that *NRG1 and ROS1* may be associated with each other through gene fusions. Therefore, in our gene fusion analysis we considered the gene fusions as connections among a network of nodes, displaying relationships that wouldn't be possible by only analyzing gene fusions as pairs.

Common gene superfamily analysis

For our common gene superfamily analysis, we used a bioinformatics tool known as CD-Batch Search [4, 11]. In contrast to Conserved Domain Search, CD-Batch Search has the remarkable ability to analyze multiple protein sequences at once. It uses the BLAST sequence of proteins and compares the BLAST sequence with conserved domain models that have been collected from a number of source databases, such as NCBI-curated domains, SMART, and

$$\text{Relative Frequency of Mutational Signature} = \frac{\text{Specific Mutational Signature Frequency}}{\text{Total \# of Mutational Signatures}}$$

Figure 2. Equation for calculating the relative frequency of a mutational signature within one cluster.

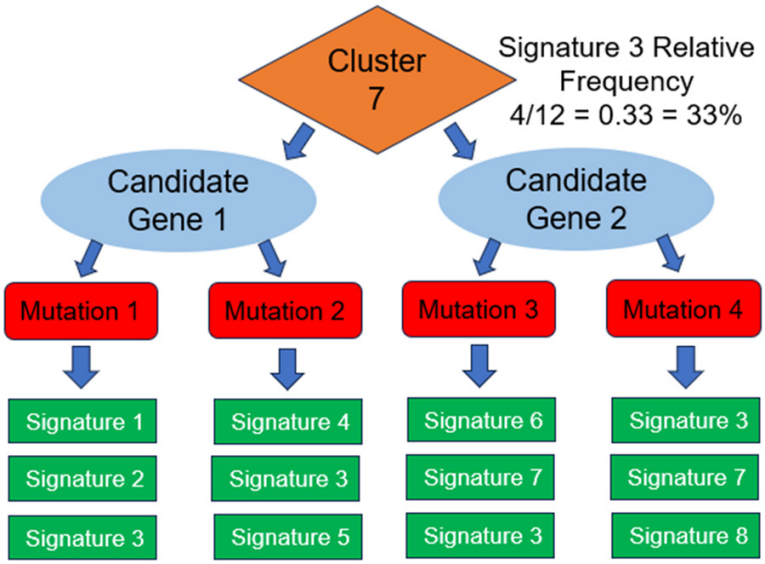

Figure 3. Methodology of mutational signature analysis.

Pfam. BLAST inputs are then processed to identify conserved domains within proteins, and it can also identify the gene superfamilies of a protein. CD-Batch Search offers both a concise graphical display and a table that displays all the results.

We obtained our BLAST sequence from the NCBI database, specifically the NCBI Reference Sequences [12]. Once CD-Batch Search identified the gene superfamilies of each protein, we analyzed and compared all the genes to determine which ones share a gene superfamily. A superfamily that contains two or more of the 35 genes is termed as a common superfamily.

Usually, the conserved domain, a domain that remains constant between proteins, is the main comparison between genes. However, analysis of the common gene superfamilies may yield unique insight into NSCLC because common gene superfamilies encompass a set of protein domains that have overlapping annotation. Common superfamilies are a broader analysis than conserved domains, which may present a unique perspective into molecular evolution that has not been previously considered.

Mutational signature analysis

The goal of this mutational signature analysis is to identify mutational signatures that are possibly associated with NSCLC. First, we identified the top three mutational signatures that were most likely associated with each of the 96 possible DNA mutations, considering only the pyrimidines of the Watson-Crick pairs [10, COSMIC - Catalogue of Somatic Mutations in Cancer at https://cancer.sanger.ac.uk/]. We based our top three mutational signatures for each DNA mutation based on the relative frequency values of the DNA mutation in each mutational signature. COSMIC provides the relative frequency for each of the 96 DNA mutations for all the single base substitution (SBS) mutational signatures. Whichever three mutational signatures have the highest relative frequency of a DNA mutation was termed as the "top three" mutational signatures. This was done for all 96 possible SBS DNA mutations, and the results will be our "annotation table", which will be referenced later.

On another note, using both COSMIC and CBioPortal, we identified the SBS DNA mutations for each of the 35 candidate genes. COSMIC, a somatic mutation cancer database, can identify amino acid substitutions for a certain gene. However, CbioPortal [13-15], a cancer genomics database, has the additional feature of focusing only on NSCLC amino acid mutations, so we also used CBioPortal to identify the amino acid substitutions for each of the 35 candidate genes. Only amino acid substitutions that were present in both databases were employed for the mutational signature analysis. Then, using Transvar [16], a multi-way annotator on genetic elements, and UCSC Genome Browser [17, https://genome.ucsc.edu/index.html], a visualization of the human genome, we

Table 1. The 35 NSCLC candidate genes with their associated cluster

Cluster	Gene
1	EGFR
1	ERBB4
1	KIF5B
2	PIK3R3
2	PLCG2
2	PRKCB
2	STK4
5	KEAP1
5	PTPN13
5	SLC34A2
7	ALK
7	CDK6
7	BRAF
7	FOXO3
7	STAT3
7	STAT5B
7	NRG1
7	CD74
7	ROS1
10	CCND1
10	RB1
10	PIK3R2
11	EML4
12	JAK3
12	MET
12	RET
12	TP63
14	TPM3
18	CDKN2A
18	LRIG3
18	BAK1
18	E2F1
18	E2F2
18	E2F3

converted these amino acid substitutions into DNA mutations.

Using our "annotation table", which identified the top three mutational signatures for all possible 96 DNA mutations, we were able to connect the 35 NSCLC candidate genes with the mutational signatures via the DNA mutations. It is important to note that the 35 NSCLC candidate genes were dispersed throughout their clusters, previously identified from Dai et al. [7]. We implemented a formula to determine the mutational signatures that had the highest relative frequency for each cluster (**Figure 2**). For example, hypothetically, in cluster 7, there were 4 SBS DNA mutations identified, so there would be 12 total mutational signatures in cluster 7. Since 4 of the 12 mutational signatures in cluster 7 were "signature 3", then the relative frequency of "signature 3" in cluster 7 is 33% (**Figure 3**). The 5 most prevalent mutational signatures based off of the relative frequency throughout the clusters were identified, suggesting that they may be associated with NSCLC.

Combining analyses

Each of the three approaches contributes important insight into the genetics of NSCLC, and in order to analyze the genetics of NSCLC in a more holistic approach, we identified certain patterns that were present throughout the three approaches. Specifically, certain genes throughout the three approaches exhibited unusual characteristics that made them possible gene targets for NSCLC.

Moreover, this holistic approach has also identified possible ways to target these NSCLC genes, something that is not present in previous research. The combination of diverse approaches to form overall conclusions provides a more comprehensive, holistic analysis of the genetics of NSCLC.

Results

Identification of NSCLC candidate genes

We filtered down the approximately 5000 original genes, dispersed throughout 10 clusters, identified by Dai et al. [7] using the Cancer Gene Census, COSMIC, and DAVID. Of these approximately 5000 original genes, 35 genes were identified to be associated with NSCLC (**Table 1**). Moreover, 14 of the 35 genes are found to be in the COSMIC Classic Genes, suggesting that these genes may be potentially overlooked yet significant NSCLC genes.

Gene fusion analysis results

Although some genes have no fusion relationships, there are numerous gene fusions throughout the 35 candidate genes. This is one way that the NSCLC genes interact with each

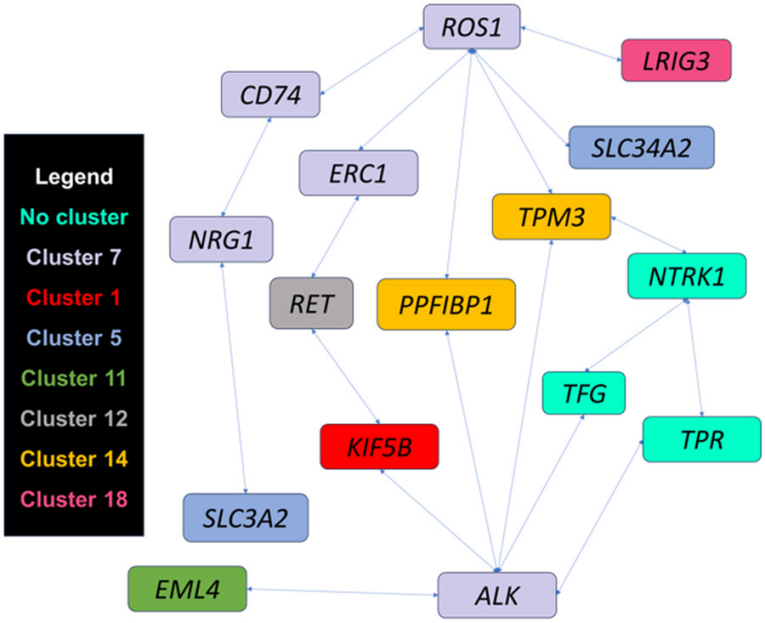

Figure 4. Fusion relationships throughout the candidate genes. A blue arrow represents a gene fusion between the two genes (CD74-NTRK1 gene fusion not shown). Each color represents a certain cluster identified from [7]. Genes are shown to be within a certain cluster by distinguishing them with a certain color.

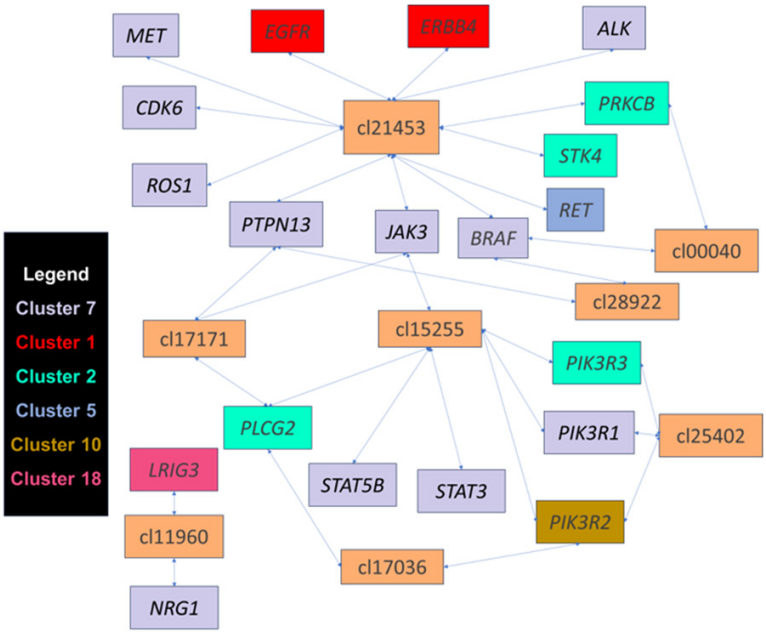

Figure 5. Peach colored boxes are superfamilies. The other boxes are genes. Lines show that a gene is part of that superfamily. Each color represents a certain cluster identified from [7]. Genes are shown to be within a certain cluster by distinguishing them with a certain color.

other. *ALK* and *ROS1* have the most fusion relationships, suggesting that *ALK* and *ROS1* play a particularly important role in NSCLC through fusion relationships. *ALK,* a 1620 amino acid tyrosine kinase receptor, and *ROS1,* another tyrosine kinase receptor, are genes that are widely recognized for playing a critical role in NSCLC [18]. By targeting these two genes, one cuts off NSCLC gene interaction, which is a possible NSCLC treatment for some patients, something which is verified from this research.

It is important to not disregard the other NSCLC candidate genes and their fusion relationships. For example, gene *NTRK1*, which is another receptor tyrosine kinase that plays an essential role in the development of the nervous system, has four fusion relationships, which demonstrates a possibility that other NSCLC gene fusion targets could also be important in the context of fusion relationships [18, 19]. *CD74* is also a gene of interest because it is implicated in NSCLC from many pieces of literature [20, 21]. Therefore, paying attention to which genes *CD74* fuses with is important as this could be the basis of an alternative method to inhibit or treat NSCLC for patients with *CD74* positive lung cancer. Nevertheless, we believe that every single identified gene fusion is possibly significant in NSCLC (**Figure 4**). While there are some well-known ones, such as ALK-KIF5B, there are some gene fusions present in the figure that were previously overlooked.

Common gene superfamily analysis results

One similar characteristic that many NSCLC genes have between each other within and between clusters is superfamilies, which are

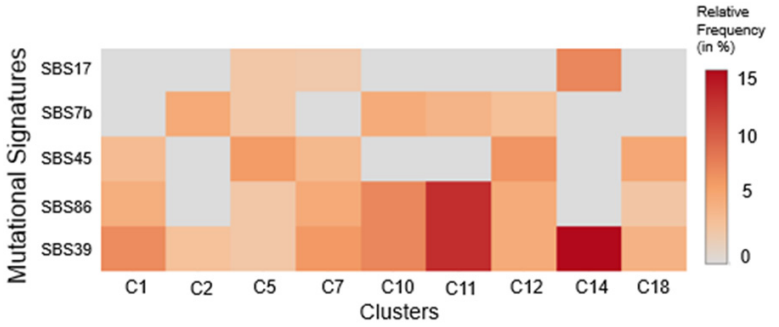

Figure 6. Heatmap of the 5 mutational signatures that are found to be associated with NSCLC. SBS39 and SBS86 are found to be the most prevalent.

usually represented with the prefix cl-. Genes can have multiple superfamilies, and if genes share a superfamily, they could be evolutionarily related to each other [4]. If a particular superfamily has many genes which are associated with NSCLC, focusing on the genes in those superfamilies may lead to identification of novel gene targets.

One can see that superfamily cl21453 is a particularly prevalent superfamily throughout many of the NSCLC candidate genes (**Figure 5**). Another superfamily cl15255 is also quite prevalent. Since many of the NSCLC candidate genes are within these two superfamilies, investigating the hundreds of genes within those superfamilies could lead to the identification of multiple, novel gene targets of NSCLC treatment.

In addition, if there is a specific gene that is significant in NSCLC, then along with targeting the fusion relationships, it could also be possible to consider the superfamily that it contains as well. Because a common superfamily consists of multiple conserved domains of overlapping annotation, genes in the same superfamily also may have similar structure and/or function, which may suggest vulnerability to similar treatments [11].

Overall, the common superfamilies between NSCLC genes are an important characteristic that should be considered. In conjunction with the results from gene fusions and mutational signatures, some NSCLC target genes and alternative treatments may be identified.

Mutational signature analysis results

The prevalence of the mutational signatures throughout the NSCLC candidate genes was determined based on the relative frequency. Essentially, how frequently a mutational signature was associated with NSCLC genes. 5 mutational signatures were identified as possibly being associated with NSCLC: SBS39, SBS86, SBS45, SBS7b, and SBS17. It is important to note that SBS45 is a possible sequencing artifact, suggesting that it plays no notable role in the pathology of NSCLC [8]. Nevertheless, SBS39 and SBS86 are the most prevalent mutational signatures throughout the NSCLC candidate genes, implying that SBS39 and SBS86 are associated with NSCLC (**Figure 6**).

Results of the combined analyses

By analyzing patterns and characteristics of the 35 candidate genes throughout the three approaches, we found two main conclusions. First of all, the characteristics of *ALK* and *ROS1* throughout the three approaches implicate them in NSCLC. *ALK* and *ROS1* have the most fusion relationships out of all NSCLC candidate genes, with both well-known (*ALK-EML4)* and obscure (*ROS1-LRIG3*) ones. They both share the protein kinase superfamily cl21453. *ALK* is associated with two prevalent NSCLC mutational signatures (SBS39 and SBS86), and *ROS1* is also associated with two (SBS7b and SBS17). Through gene fusions, superfamilies, and mutational signatures, *ALK* and *ROS1* are particularly important genes in NSCLC, suggesting that these genes can act as gene targets for NSCLC. The importance of *ALK* and *ROS1* in the pathogenesis of NSCLC is quite well known [2]. However, this is important because it demonstrates that the application of this novel method for analyzing the genetics of NSCLC is valid. The conclusion gained by this novel method matches those of previous research.

In addition to *ALK* and *ROS1*, we also found that *CD74* exhibited some unusual characteristics throughout the three analyses. Gene *CD74* is presented on the surface of antigen-presenting cells as a membrane protein. In addition, inside the endoplasmic reticulum, *CD74,* plays a role in the formation of major histocompatibility class II molecules [20]. Furthermore, *CD74*

has already been identified as an important gene in the pathology of NSCLC [22]. However, our research presents unique insight into the processes of *CD74* in NSCLC. *CD74* has fusion relationships with *NRG1, ROS1,* and *NTRK1,* which are all significant genes in NSCLC. However, *CD74* has no common superfamilies with any other NSCLC candidate genes. Despite extensive investigation into CBioPortal and COSMIC, no significant *CD74* mutations or NSCLC associated mutational signatures were found in NSCLC. Our research shows the only characteristic of *CD74* that causes it to be associated with NSCLC is the fusion relationships. Based on our multi-pronged approach to NSCLC, *CD74* is dependent on its fusion relationships to play a role in NSCLC. Targeting *CD74* gene fusions in *CD74* positive individuals may impair *CD74's* effects. This may be an alternative approach to NSCLC treatment.

Discussion

Because of the rigorous standards we used to identify the 35 mutated NSCLC candidate genes, we believe that these 35 genes may be possible NSCLC gene targets, and since we used the COSMIC Classic Genes list, some of the 35 candidate genes may have been overlooked by previous research, suggesting possible alternative NSCLC treatment.

Identification of significant gene fusions are essential because if these significant fusion relationships are disabled, then that could possibly impair the progression of NSCLC. Additionally, if a specific gene is implicated in NSCLC, then targeting this gene through fusion relationships may be a possible approach. Overall, this gene fusion approach identifies specific gene fusions that may be important in the pathogenesis of NSCLC, suggesting alternative treatment of NSCLC.

The common gene superfamilies are also an important aspect of NSCLC. The identification of common gene superfamilies among the 35 NSCLC candidate genes highlights the similarities in biological function, structure, or sequence between these genes. These findings set the backbone for possible identification of future novel gene targets for NSCLC. Additionally, since many of the NSCLC candidate genes are within two specific superfamilies, our results suggest that analyzing the genes in superfamily cl21453 and cl15255 could be a possible reservoir of new novel NSCLC gene targets.

Identifying mutational signatures associated with NSCLC can provide both therapeutic and prognostic benefits [23]. Mutational signatures are particularly valuable because of their possible role as biomarkers for predictors of drug or therapy response [23]. For example, presence of mutational signatures in patients has been associated with poly (ADP)-ribose polymerase (PARP) inhibitor sensitivity [23]. The identification of mutational signatures associated with NSCLC can thus advance the field of personalized medicine for NSCLC treatment, among other things [6].

While common superfamily analysis is a relatively under-investigated field of study, there have been numerous studies of gene fusions and mutational signatures. Previous mutational signature analyses have typically sequenced tumor DNA samples and then performed mutational signature profiling using COSMIC's SigProfiler Tools. We believe our research is the first attempt at utilizing an "annotation table" and employing several different databases and bioinformatics tools to identify NSCLC associated mutational signatures. Our usage of the COSMIC Classic Gene list for our gene fusion analysis possibly allows us to identify novel, previously overlooked NSCLC gene targets. Finally, combining the results from the three approaches both validated our novel method and identified a possible alternative NSCLC treatment. Compared to other studies, our unique methods of analyzing each of the three approaches and combining them bring unique insight into the genetics of NSCLC.

Conclusion

We believe that our research introduces a novel way to analyze the genetics of NSCLC. Next, we plan to apply this novel method to other cancers, such as glioblastoma, melanoma, or pancreatic cancer. Additionally, we believe that analyzing more approaches may also provide more unique insight into NSCLC. For example, using Timer, a bioinformatics tool, we can find the association between immune infiltration and gene expression. With this approach, some of the 35 NSCLC candidate genes may be found to play a significant role in immune infiltration.

By utilizing various bioinformatics tools and databases and considering all the three approaches simultaneously, we provide a more comprehensive and holistic way to analyze the genetics of NSCLC. This will likely contribute to improved or alternative NSCLC treatments, and its applicability to a various range of cancers and the possible addition of more approaches highlights the potential of this novel method.

Acknowledgements

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Disclosure of conflict of interest

None.

Address correspondence to: Eric Pan, 18 Pin Oak Estates CT, Bellaire, TX 77401, USA. E-mail: ericyspan@gmail.com

References

[1] Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, Islam SMA, Lopez-Bigas N, Klimczak LJ, McPherson JR, Morganella S, Sabarinathan R, Wheeler DA, Mustonen V; PCAWG Mutational Signatures Working Group; Getz G, Rozen SG and Stratton MR; PCAWG Consortium. The repertoire of mutational signatures in human cancer [published correction appears in Nature 2023; 614: E41]. Nature 2020; 578: 94-101.

[2] Takeuchi K, Soda M, Togashi Y, Suzuki R, Sakata S, Hatano S, Asaka R, Hamanaka W, Ninomiya H, Uehara H, Lim Choi Y, Satoh Y, Okumura S, Nakagawa K, Mano H and Ishikawa Y. RET, ROS1 and ALK fusions in lung cancer. Nat Med 2012; 18: 378-381.

[3] Wang Y, Zhang H, Zhong H and Xue Z. Protein domain identification methods and online resources. Comput Struct Biotechnol J 2021; 19: 1145-1153.

[4] Marchler-Bauer A and Bryant SH. CD-Search: protein domain annotations on the fly. Nucleic Acids Res 2004; 32: W327-W331.

[5] Drummond RD, Defelicibus A, Meyenberg M, Valieris R, Dias-Neto E, Rosales RA and da Silva IT. Relating mutational signature exposures to clinical data in cancers via signeR 2.0. BMC Bioinformatics 2023; 24: 439.

[6] van den Heuvel GRM, Kroeze LI, Ligtenberg MJL, Grünberg K, Jansen EAM, von Rhein D, de Voer RM and van den Heuvel MM. Mutational signature analysis in non-small cell lung cancer patients with a high tumor mutational burden. Respir Res 2021; 22: 302.

[7] Dai X, Ding L, Liu H, Xu Z, Jiang H, Handelman SK and Bai Y. Identifying Interaction Clusters for MiRNA and MRNA Pairs in TCGA Network. Genes (Basel) 2019; 10: 702.

[8] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ and Forbes SA. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res 2019; 47: D941-D947.

[9] Huang da W, Sherman BT and Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009; 4: 44-57.

[10] Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T and Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res 2022; 50: W216-W221.

[11] Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C and Bryant SH. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 2011; 39: D225-D229.

[12] Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD and Sherry ST. Database resources of the national center for biotechnology information. Nucleic Acids Res 2022; 50: D20-D26.

[13] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C and Schultz N. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data [published correction appears in Cancer Discov 2012; 2: 960]. Cancer Discov 2012; 2: 401-404.

[14] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C and Schultz N. Integrative analysis of complex cancer ge-

nomics and clinical profiles using the cBioPortal. Sci Signal 2013; 6: pl1.
[15] de Bruijn I, Kundra R, Mastrogiacomo B, Tran TN, Sikina L, Mazor T, Li X, Ochoa A, Zhao G, Lai B, Abeshouse A, Baiceanu D, Ciftci E, Dogrusoz U, Dufilie A, Erkoc Z, Garcia Lara E, Fu Z, Gross B, Haynes C, Heath A, Higgins D, Jagannathan P, Kalletla K, Kumari P, Lindsay J, Lisman A, Leenknegt B, Lukasse P, Madela D, Madupuri R, van Nierop P, Plantalech O, Quach J, Resnick AC, Rodenburg SYA, Satravada BA, Schaeffer F, Sheridan R, Singh J, Sirohi R, Sumer SO, van Hagen S, Wang A, Wilson M, Zhang H, Zhu K, Rusk N, Brown S, Lavery JA, Panageas KS, Rudolph JE, LeNoue-Newton ML, Warner JL, Guo X, Hunter-Zinck H, Yu TV, Pilai S, Nichols C, Gardos SM, Philip J; AACR Project GENIE BPC Core Team, AACR Project GENIE Consortium; Kehl KL, Riely GJ, Schrag D, Lee J, Fiandalo MV, Sweeney SM, Pugh TJ, Sander C, Cerami E, Gao J and Schultz N. Analysis and visualization of longitudinal genomic and clinical data from the AACR Project GENIE biopharma collaborative in cBioPortal. Cancer Res 2023; 83: 3861-3867.
[16] Zhou W, Chen T, Chong Z, Rohrdanz MA, Melott JM, Wakefield C, Zeng J, Weinstein JN, Meric-Bernstam F, Mills GB and Chen K. TransVar: a multilevel variant annotator for precision genomics. Nat Methods 2015; 12: 1002-1003.
[17] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D. The human genome browser at UCSC. Genome Res 2002; 12: 996-1006.
[18] Araghi M, Mannani R, Heidarnejad maleki A, Hamidi A, Rostami S, Safa SH, Faramarzi F, Khorasani S, Alimohammadi M, Tahmasebi S and Akhavan-Sigari R. Recent advances in non-small cell lung cancer targeted therapy; an update review. Cancer Cell Int 2023; 23: 162.
[19] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA and Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018; 68: 394-424. Erratum in: CA Cancer J Clin 2020; 70: 313.
[20] Vargas J and Pantouris G. Analysis of CD74 occurrence in oncogenic fusion proteins. Int J Mol Sci 2023; 24: 15981.
[21] Izumi M, Lee M, Shibahara D, Kobayashi IS, Plotnick D and Kobayashi SS. Targeting the MIF-CD74 axis to overcome resistance to tyrosine kinase inhibitors in non-small cell lung cancer [abstract]. In: Proceedings of the American Association for Cancer Research Annual Meeting 2023; Part 1 (Regular and Invited Abstracts); 2023 Apr 14-19; Orlando, FL. Philadelphia (PA): AACR; Cancer Res 2023; 83 Suppl: Abstract nr 3460.
[22] McClelland M, Zhao L, Carskadon S and Arenberg D. Expression of CD74, the receptor for macrophage migration inhibitory factor, in non-small cell lung cancer. Am J Pathol 2009; 174: 638-646.
[23] Brady SW, Gout AM and Zhang J. Therapeutic and prognostic insights from the analysis of cancer mutational signatures. Trends Genet 2022; 38: 194-208.