Chromatin Accessibility Reveals Potential Prognostic Value of the Peak Set Associated with Smoking History in Patients with Lung Adenocarcinoma

Han Liang

BGI-Shenzhen: BGI Group

Jianlian Deng

BGI-Shenzhen: BGI Group <https://orcid.org/0000-0002-0990-9996>

Tian Luo

BGI-Shenzhen: BGI Group

Huijuan Luo

BGI-Shenzhen: BGI Group

Xuan Wu

Peking University Shenzhen Hospital

Yiwang Ye

Peking University Shenzhen Hospital

Shubin Wang

Peking University Shenzhen Hospital

Fuqiang Li

BGI-Shenzhen: BGI Group

Kui Wu

BGI-Shenzhen: BGI Group

Cong Lin (✉ lincong@genomics.cn)

BGI-Shenzhen: BGI Group

Research

Keywords: ATAC-seq, network, LUAD, smoking, prognostic

Posted Date: June 8th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-533843/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Considerable differences in molecular characteristics have been defined between non-smoker and smokers in patients with lung adenocarcinoma (LUAD), yet studies of open chromatin patterns associated with LUAD progression caused by smoking are still lacking. Here, we constructed a novel network based on correlations between each ATAC-seq peak from TCGA data using our previously developed algorithm. Subsequently, principal component analysis was performed on LUAD samples with retained peaks filtered by the correlation network and pathway analysis was conducted for potential pathways identification. We identified a set of peaks that discriminated smokers in LUAD patients according to levels of exposure to tobacco quantified in pack-years, and also significantly associated with progression-free survival and overall survival of these patients. We then investigated the gene set related to those peaks and found that the comprising genes are strongly associated with LUAD development, such as *B3GNT3*, *ACTN4* and *CLDN3*. They are consistent with the important roles for the associated pathways in LUAD oncogenesis induced by smoking, including glycosphingolipid biosynthesis and tight junction pathways. In summary, our study may provide valuable insights on exploration of ATAC-seq peaks and on smoking-related LUAD carcinogenesis from a perspective of open chromatin changes.

Introduction

Lung cancer remains the leading cause of cancer death with over 1.6 million deaths annually, and the incidence of lung cancer is still increasing worldwide (Herbst et al. 2008; Herbst et al. 2018). More than 85% of lung cancer cases are diagnosed as non-small-cell lung cancer (NSCLC), with lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) being the two main histological subtypes. LUAD alone accounts for approximately 40% of NSCLC cases, resulting in over 500,000 deaths per year globally (Herbst et al. 2008; Herbst et al. 2018). The most important risk factor for lung cancer is still cigarette smoking, which is responsible for about 85–90% of all cases (Freedman et al. 2008; Herbst et al. 2018). A cumulative exposure to tobacco of 10–20 pack-years is reported to be associated with a clinically relevant increase in morbidity (Neumann et al. 2013). In the context of lung cancer screening, smokers with a smoking history ≥ 20 pack-years is one of the major criteria recommended by the National Comprehensive Cancer Network and the American Association for Thoracic Surgery (Boiselle 2013). The other emerging risk factors include second-hand smoking and air pollution (Oberge et al. 2011), such as PM_{2.5}, which is claimed to cause lung cancer in many developing countries (Guo et al. 2019; Khilnani, Tiwari 2018). Although NSCLC is strongly associated with smoking, LUAD is more common in never-smokers (Herbst et al. 2018; Sun et al. 2007). Compelling evidence indicates that never or light - smoker patients with LUAD have a significantly better survival rate than smokers, suggesting different levels of smoke exposure may cause distinct molecular mechanisms underlying their clinical difference (Bryant, Cerfolio 2007; Casal-Mourino et al. 2019; Lofling et al. 2019).

Recent efforts have been put to characterize different molecular alterations in LUAD using high-throughput genome sequencing, which led to a comprehensive profiling of different oncogenic driver

mutations (Cancer Genome Atlas Research 2014; Weir et al. 2007). Besides *EGFR* mutations and *ALK* fusions, for which targeted therapies have become standard treatment for LUAD, several other activated oncogenes such as, *KARS*, *TP53*, *ERBB2* and *BRAF* are also found in LUAD (Imielinski et al. 2012; Wu et al. 2015). As more and more in-depth multi-omics studies are progressing, striking differences in molecular characteristics have been discovered between LUAD arising in never-smokers and smokers. For example, LUAD patients with different levels of tobacco consumption show different mutation frequencies of the *EGFR*, *TP53* and *KRAS* genes, with *EGFR* mutations occurring more frequently in never smokers (Le Calvez et al. 2005; Sun et al. 2007). In addition, gene expression analysis identified distinct patterns of dysregulated genes in smokers of LUAD, of which associated altered pathways are particularly involved in the cellular immune response and cell cycle regulation (Landi et al. 2008; Liu et al. 2018). Moreover, epigenetic studies also demonstrated clear differences between methylation profiles of LUAD in never smokers and smokers (Alexandrov et al. 2016; Divine et al. 2005; Toyooka et al. 2006). To date, however, other epigenetic studies, such as open chromatin patterns associated with LUAD progression caused by smoking are still lacking. Different from whole genome (exome) sequencing, which identify genetic risks, study of open chromatin regions can offer insights on epigenetic and regulatory modifications and thus may provide novel genes or pathways that are involved.

Recently, assay for transposase accessible chromatin sequencing (ATAC-seq) has emerged as a powerful tool for profiling chromatin accessibility in different human diseases and has exerted a profound impact on understanding the coordination in gene expression processes (Buenrostro et al. 2013; Liu et al. 2019). Until now, a few studies have explored open chromatin states in NSCLC with ATAC-seq. An elegant work by Corces et al. studied chromatin accessibility of 410 tumor samples from The Cancer Genome Atlas (TCGA), including 38 cases of NSCLC (Corces et al. 2018). More recently, an integrative analysis which linked the open chromatin variations to genomic alterations among NSCLC patients provided a comprehensive open chromatin landscape of NSCLC (Wang et al. 2019). However, emphasis has not yet been placed on linking the clinical variables, such as cigarette smoking history to open chromatin patterns in LUAD. In this study, we first generated a network based on correlations between peaks from ATAC-seq data of TCGA. Using retained peaks filtered by the correlation network, we then studied differences between never or light smokers (< 20 pack-years) and heavy smokers (≥ 20 pack-years) in LUAD patients and further identified a set of genes and their related pathways that associated with patients' progression-free survival (PFS) and overall survival (OS).

Materials And Methods

Patients and Clinical Information

Patients initially diagnosed with lung cancer at Peking University Shenzhen Hospital from July 2018 to May 2019 were included in the study. Patients' information, including pathology, tumor type and smoking history was collected for all 6 patients (**Table S1**). This study was carried out in accordance with Declaration of Helsinki and approved by the Ethics Committee of Peking University Shenzhen Hospital (Approval number: 2018033) and all patients provided written informed consent.

ATAC-seq Library Preparation and Sequencing

Isolation of nuclei from frozen tissues and transposition reaction for ATAC-seq were performed as previously described (Corces et al. 2017). Briefly, frozen tissue sections of each sample were prepared on dry ice to avoid thawing. Then the tissue was transferred into a pre-chilled 2-ml Dounce homogenizer containing 2 ml of 1× cold homogenization buffer. Tissues were then homogenized for several minutes with different sizes of the pestles and filtered by nylon mesh filter. The nuclei obtained by sucrose gradient centrifugation were transferred to a tube containing resuspension buffer and then pelleted by centrifugation at 500 rcf for 10 min. The supernatant was discarded and nuclei were resuspended in the Omni-ATAC-seq mix (Corces et al. 2017) containing Tn5 transposase. The two technical replicates were prepared with the two separated transposase reactions from the same set of nuclei of one donor. The reaction was incubated at 37 °C for 30 min in a thermomixer with shaking at 500 rpm and the tagmented DNA was isolated with MinElute PCR Purification kit (Qiagen). PCR was performed using NEBNext High-Fidelity 2X PCR master mix (72°C 5min, 98°C 30sec, 13 cycles of: [98°C 10sec, 63°C 30sec, 72°C 1min], 72°C 5min). Sizes of PCR products were selected using Ampure XP by 0.5X volume followed by a 1.0X final volume. The libraries were then prepared based on a BGISEQ-500 sequencing platform (Huang et al. 2017) and sequenced with paired-end 50-bp read lengths.

ATAC-seq Data Processing and Peak Calling

ATAC-seq data was processed using the ENCODE ATAC-seq pipeline[1]. Briefly, the fastq files were processed for adaptor trimming by cutadapt software (version 1.9.1) (Martin 2011) with “cutadapt -a TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG -A GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG -m 5 -O 5 -e 0.1” options, coming with the clean fastq files. Clean fastq files were aligned to the human hg38 (GRCh38) genome using bowtie2 (version 2.2.6) (Langmead, Salzberg 2012) with “bowtie2 -k 5 -X 2000 -mm -local -threads 4” options. Then, Samtools (Li et al. 2009) was used to generate the bam files as well as sort and retain properly paired reads with “samtools view -F 1804 -f 2” and “samtools sort” options. After this, Picard was used to remove PCR duplicates with default arguments. Mitochondrial reads were also filtered. Finally, final aligned, de-duplicated bam files were generated for the subsequent downstream analysis.

According to data quality control standards of ENCODE[2], only samples with transcription start site (TSS) enrichment value >5 were classified as qualified and used for downstream analysis. Peak calling was carried out to generate high quality fixed-width peaks from the final bam files of QC-passed samples, each coupled with technical replicates to improve sensitivity and the accuracy. Peak calling was performed using MACS2 with “macs2 callpeak -t bamfile -cap-num-peak 300000 -shift -75 -extsize 150 -nomodel -call-summits -nolambda -keep-dup all -p 0.01” options in each replicate. Only reproducible peaks between two technical replicates identified by Irreproducible Discovery Rate (IDR) analysis (threshold 0.05) were retained for downstream analysis. Fixed-width peak set was then generated by extending the peak summits by 250 bp on either side to a final width of 501 bp. We filtered the ENCODE hg38 (GRCh38) blacklist[3] manually to avoid the interference of the high signal regions such as

centromeres, telomeres, and satellite repeats. Peaks that beyond the ends of chromosomes after extending were also filtered (Corces et al. 2018). Coverage of each defined peak was evaluated using BEDTools (v2.26.0) (Quinlan, Hall 2010) with “multicov” argument, while raw count matrix was normalized using “DESeq2” package in R by Reads Per Million aligned reads (RPM). The ATAC-seq data from each replicate of our primary tumor samples generated 48~141M (median = 101M) reads, in which two samples did not reach TSS value of 5, leaving 4 samples from LUAD patients for further analysis.

ATAC-seq Data Analysis

We used previously described analysis method on peaks selecting and exploring (Liang et al. 2020)⁴. Briefly, we selected the available data from TCGA according to peak's quality. A peak would be considered low-quality if it has a same value in more than 5% patients from single type of cancer, as the repeat values were likely produced by nonsense 0s before normalization. Eventually, we obtained 64316 peaks across 386 samples from TCGA dataset. To further reduce the scale of data, we then applied the previously developed algorithm on the correlation network construction with retained peaks from TCGA, in which two peaks would be connected if their direct or indirect correlation is significant (Liang et al. 2020) (**Fig. 1**). We considered the direct correlation between two peaks is significant, if the absolute value of correlation coefficient is not less than 0.4, considering the noisy level is around 0.2. Furthermore, we considered the indirect correlation between the two peaks is significant, if their direct correlations with the third peak are both significant. Here we allowed the indirect correlation to amplify peaks' aggregation effect in the network. To assess the correlation between peaks more reliably, outlier values were removed before calculating correlation (Hearst 1999). In practice, we used the function “cor” from R package “stats” V3.6.2 with default arguments. We selected the 10% most frequently-connected peaks for the further analysis, as those peaks were more likely to be the hub peaks in the network with high connectivity.

The unsupervised classification method principal component analysis (PCA) was used to analyze the selected ATAC-seq peaks from 22 LUAD patients in TCGA dataset. We used the function “PCA” from R package “FactoMineR” V1.34 with default arguments which produced 5 most important components (Lê et al. 2008). The association between components and smoking status were checked by the distribution of samples classified by components. To statistically assess the difference between distribution distance of samples from LUAD patients with different smoking histories (based on pack-years of smoking, status of smoking exposure was classified as light smoker (<20 pack-years) and heavy smoker (≥ 20 pack-years)), two-tailed unpaired t-tests was performed. $P < 0.05$ were considered significant. We used the function “survdif” from R package “survival” (Harrington, Fleming 1982), in which Chi-Squared Test was used to assess the PFS and OS analysis. Then the function “pchisq” from R package “stats” was used to calculate P value with arguments $df=1$, lower tail=F.

Pathway Analysis

For the pathway analysis, all linc-RNA genes (such as RP11- genes) have been removed as they are excluded in KEGG database. We searched the potential pathways using DAVID 6.8. with the non-linc-RNA genes related to top 100 highest contributing peaks of targeted component. Pathways were found with p value less than 0.1, which is the recommended value of DAVID (Huang da et al. 2009).

[1] <https://github.com/ENCODE-DCC/atac-seq-pipeline>

[2] <https://www.encodeproject.org/atac-seq/>. Retrieved 24/12, 2019

[3] <https://www.encodeproject.org/files/ENCFF356LFX/>

⁴ <https://www.biorxiv.org/content/10.1101/2020.03.30.015180v1.full>

Results

A correlation network is built based on ATAC-seq data from TCGA

High-quality ATAC-seq data of 410 tumor samples across 23 cancer types was downloaded and collected from TCGA. The extended peak summits with fixed width of 501 bp were extracted from the dataset and used to perform the further analysis (Corces et al. 2018). We first constructed a correlation network with retained peak summits from TCGA, in which two peaks would be connected if their peak values were correlated with each other significantly (**Fig. 1**). To this end, correlations between each peak value in all samples were calculated and the 10% most frequently-connected peaks were chosen after removing the low-quality peaks and outliers. Using PCA on those chosen peaks across all cancer types, we identified distinct clusters labeled based on the different cancer-type enrichment (**Fig. 2**). We found strong concordance between our clustering results and the t-distributed stochastic neighbor embedding (t-SNE) results from TCGA (Corces et al. 2018).

Identification of the smoking associated peak set in LUAD patients and related pathways

We next focused on the analysis of 22 LUAD patients in the TCGA dataset, using the same correlation network and analyzing methods described above. PCA result indicated that the first principal component (PC1) explained 16.9% of the variability, while PC2 explained 12.8% of the variability in the peaks with all LUAD samples (**Fig. 3A**). The LUAD samples did not appear to form distinct patterns within the two dimensions generated by PCA analysis, interestingly however, it seemed that samples with less smoking exposure are closer to PC2 axis rather than PC1 (**Fig. 3A**). The x or y values of each sample in Fig.3A were generated automatically in PCA and only represented the level of variety towards PC1 or 2-axis, which are not true peaks values. Therefore, a smaller absolute x or y value of a sample meant its PC1 or 2-related peaks are less variable. Using each sample's absolute distance from the PC1 or 2-axis, we compared the difference between heavy smokers (patients with at least 20 pack-years smoking history) and the rest (**Fig. 3B and C, Table S2**). The results showed that the ≥ 20 pack-years group indeed had a significant

longer absolute distance from PC2 axis compared to the <20 pack-years group, suggesting the ATAC-seq peaks associated with PC2 rather than PC1 were influenced by smoking history. Additionally, shorter absolute distance from PC2 axis, representing more stable PC2-related peak values, was significantly associated with both better PFS and OS of patients, indicating a potential prognostic value for the corresponding peaks of PC2 (**Fig. 4**). We thus further studied the gene set related to those peaks according to the defined peak-gene relationships (Corces et al. 2018)(**Table S3**) and explored the associated pathways. Consequently, we identified three potential pathways, including glycosphingolipid biosynthesis, tight junction and metabolic pathways (**Table 1**).

To verify the smoking associated peak set identified, we performed ATAC-seq on the primary tumors from 6 patients with LUAD and were eventually left with 4 samples after check for TSS values (**Table S4 and S5**). We calculated their PC2 values according to weights of PC2-related peaks showed in LUAD samples from TCGA data, so that the PC2 values of these samples resembled that seen as absolute distance from PC2 axis in **Fig. 3A**. The results seemed to be in agreement with our previous observation from TCGA ATAC-seq data (**Fig. S1**), however, the intention of verifying our findings with an independent primary dataset was not fulfilled due to the very small sample size.

Discussion

During the past decade, many important alterations in genomes and signaling pathways caused by smoking in LUAD were unveiled by in-depth analyses, yet the precise chromatin accessibility alterations induced by smoking remained obscure(Landi et al. 2008; Sun et al. 2007; Toyooka et al. 2006). Previous work on open chromatin landscape of NSCLC divided LUAD samples into three sub-clusters based on open chromatin peaks and also found correlations between the clusters and smoking. However, associated genes or pathways were not further studied(Wang et al. 2019). In the present study, we first constructed a correlation network with ATAC-seq data from TCGA using the algorithm we previously developed, and analyzed 22 LUAD samples with peaks selected by the network. We determined smoking history related peaks with potential prognostic value and found associated pathways based on the defined peak-gene relationships.

Different from the classic analysis on correlation of gene expression and chromatin accessibility (Corces et al. 2018; Wang et al. 2019), we solely focused on correlations between each peak and constructed the network based on the plausible theory that peaks which are highly connected with many other peaks would be more likely to have a crucial function in gene regulatory processes. The correlation network assisted us with critical peaks selection and enabled a more precise analysis on further peak classification. With this novel conception, smoking-related peak set in this study was effectively identified from the massive ATAC-seq data. The effect of this method was also reflected by the identification of specific mitosis-related expression pattern in the previous work on data mining of transcriptomes from 5,001 cancer patients cross 22 cancer types(Liang et al. 2020). Therefore, we believe that this method is practical in revealing crucial factors from complicated datasets and will additionally increase analytical effects on studies of future multi-omics data.

The peaks selected by the correlation network were further analyzed with PCA, and the PC2-related peaks we identified from TCGA ATAC-seq data effectively differentiated heavy smokers (≥ 20 pack-years) from the rest of patients with LUAD tumors. Intriguingly, PC2 was also associated with both PFS and OS of these LUAD patients, pointing to genes that corresponding to PC2-related peaks as potential effectors in carcinogenesis of LUAD caused by smoking. Indeed, the pathways significantly associated with PC2-related gene set, including glycosphingolipid biosynthesis and tight junction pathways, are proven to be important in tumor progression(Soini 2012; Zhuo et al. 2018). Glycosphingolipid is a crucial part of the cell membrane, which is formed by a ceramide backbone linked to a glycan moiety. The aberrations in glycosphingolipid metabolism can lead to abnormal cell rearrangements and thus contribute to proliferation, metastatic cellular invasion and even resistance to chemotherapy in tumor cells (Russo et al. 2018; Ogretmen 2018; Zhuo et al. 2018). In the context of LUAD progression, several previous findings showed that the exposure to cigarette smoking results in increased levels of ceramide, which in turn promotes aberrant EGFR activation and leads to elevated tumorigenic signaling(Goldkorn et al. 2013). Additionally, dysregulated expressions of genes that are related to glycosphingolipid pathways, for instance *B4GAT1* and *B3GNT3* identified by PC2-related peaks, have also been shown to correlated with the observed malignancy(Gao et al. 2018; Willer et al. 2014). Among those genes, overexpression of *B3GNT3* is specifically associated with unfavorable OS in NSCLC patients(Gao et al. 2018). As for the tight junction pathway, alterations in expressions of its proteins induced by smoking are involved in increased lung epithelial permeability and epithelial-mesenchymal transition (EMT), thus influencing cancer progression(Shaykhiev et al. 2011; Soini 2012). The genes we identified through PC2-related peaks, such as *ACTN4* and *CLDN3*, are consistent with the important role for this pathway in LUAD oncogenesis(Tentler et al. 2019; Zhang et al. 2017). Specifically, *ACTN4* amplification can be induced by smoking and is also associated with increased mortality risk among patients with early stage of LUAD(Noro et al. 2013). Therefore, the gene set we identified through ATAC-seq peaks comprises genes that are strongly associated with smoking status and may have potential prognostic value in patients with LUAD. However, before drawing conclusions on potential clinical implications of these genes, it is important to elucidate their function with respect to LUAD development.

In conclusion, our study provided a novel method to explore ATAC-seq peaks and identified a set of peaks with potential prognostic value based on chromatin accessibility alterations induced by smoking in LUAD patients. These findings provide information on smoking-related LUAD carcinogenesis from a perspective of open chromatin changes and may have impact on future clinical application, although further studies are warranted with datasets of LUAD patients to confirm the effects of PC2-related peaks exerted in this research.

Declarations

Funding: This research was funded by the Science, Technology, and Innovation Commission of Shenzhen Municipality (grant number JCYJ20170817145454378, JCYJ20160531193931852) and the Guangdong Enterprise Key Laboratory of Human Disease Genomics (grant number 2020B1212070028).

Conflict of interests: The authors declare that they have no competing interests.

Ethics approval: This study was carried out in accordance with Declaration of Helsinki and approved by the Ethics Committee of Peking University Shenzhen Hospital (Approval number: 2018033).

Consent to participate: All patients provided written informed consent for participation.

Consent for publication: All patients provided written informed consent for publication.

Availability of data and material: The data that support the findings of this study have been deposited into CNGB Sequence Archive (CNSA: <https://db.cngb.org/cnsa/>) of CNGBdb with accession number CNP0001103.

Code availability: The code that used in current study would be available upon reasonable request.

Authors' contributions: CL, HL, and JD conceived the study. HL and JD performed the ATAC-seq data analysis; TL and HJL carried out the experiments; XW, YY and SW conducted sample collection; FL contributed to the ENCODE pipeline localization; KW instructed the study; CL supervised the study and wrote the manuscript.

References

1. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S et al. Mutational signatures associated with tobacco smoking in human cancer. *Science*. 2016;354(6312):618-22. doi:10.1126/science.aag0299.
2. Boisselle PM. Computed tomography screening for lung cancer. *Jama*. 2013;309(11):1163-70. doi:10.1001/jama.2012.216988.
3. Bryant A, Cerfolio RJ. Differences in epidemiology, histology, and survival between cigarette smokers and never-smokers who develop non-small cell lung cancer. *Chest*. 2007;132(1):185-92. doi:10.1378/chest.07-0442.
4. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*. 2013;10(12):1213-8. doi:10.1038/nmeth.2688.
5. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511(7511):543-50. doi:10.1038/nature13385.
6. Casal-Mourino A, Valdes L, Barros-Dios JM, Ruano-Ravina A. Lung cancer survival among never smokers. *Cancer letters*. 2019;451:142-9. doi:10.1016/j.canlet.2019.02.047.
7. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W et al. The chromatin accessibility landscape of primary human cancers. *Science*. 2018;362(6413). doi:10.1126/science.aav1898.
8. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature*

- methods. 2017;14(10):959-62. doi:10.1038/nmeth.4396.
9. Divine KK, Pulling LC, Marron-Terada PG, Liechty KC, Kang T, Schwartz AG et al. Multiplicity of abnormal promoter methylation in lung adenocarcinomas from smokers and never smokers. *International journal of cancer*. 2005;114(3):400-5. doi:10.1002/ijc.20761.
 10. Freedman ND, Leitzmann MF, Hollenbeck AR, Schatzkin A, Abnet CC. Cigarette smoking and subsequent risk of lung cancer in men and women: analysis of a prospective cohort study. *The Lancet Oncology*. 2008;9(7):649-56. doi:10.1016/S1470-2045(08)70154-2.
 11. Gao L, Zhang H, Zhang B, Zhu J, Chen C, Liu W. B3GNT3 overexpression is associated with unfavourable survival in non-small cell lung cancer. *Journal of clinical pathology*. 2018;71(7):642-7. doi:10.1136/jclinpath-2017-204860.
 12. Goldkorn T, Chung S, Filosto S. Lung cancer and lung injury: the dual role of ceramide. *Handbook of experimental pharmacology*. 2013(216):93-113. doi:10.1007/978-3-7091-1511-4_5.
 13. Guo H, Chang Z, Wu J, Li W. Air pollution and lung cancer incidence in China: Who are faced with a greater effect? *Environment international*. 2019;132:105077. doi:10.1016/j.envint.2019.105077.
 14. Harrington DP, Fleming TR. A Class of Rank Test Procedures for Censored Survival Data. *Biometrika*. 1982;69(3):553-66.
 15. Hearst MA, editor. Untangling text data mining. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*; 1999.
 16. Herbst RS, Heymach JV, Lippman SM. Lung cancer. *N Engl J Med*. 2008;359(13):1367-80. doi:10.1056/NEJMr0802714.
 17. Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. *Nature*. 2018;553(7689):446-54. doi:10.1038/nature25183.
 18. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211.
 19. Huang J, Liang X, Xuan Y, Geng C, Li Y, Lu H et al. A reference human genome dataset of the BGISEQ-500 sequencer. *GigaScience*. 2017;6(5):1-9. doi:10.1093/gigascience/gix024.
 20. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*. 2012;150(6):1107-20. doi:10.1016/j.cell.2012.08.029.
 21. Khilnani GC, Tiwari P. Air pollution in India and related adverse respiratory health effects: past, present, and future directions. *Current opinion in pulmonary medicine*. 2018;24(2):108-16. doi:10.1097/MCP.0000000000000463.
 22. Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PloS one*. 2008;3(2):e1651. doi:10.1371/journal.pone.0001651.
 23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods*. 2012;9(4):357-9. doi:10.1038/nmeth.1923.

24. Le Calvez F, Mukeria A, Hunt JD, Kelm O, Hung RJ, Tanriere P et al. TP53 and KRAS mutation load and types in lung cancers in relation to tobacco smoke: distinct patterns in never, former, and current smokers. *Cancer research*. 2005;65(12):5076-83. doi:10.1158/0008-5472.CAN-05-0551.
25. Lê S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*. 2008;25(1):1-18.
26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9. doi:10.1093/bioinformatics/btp352.
27. Liang H, Lin C, Hou Y, Li F, Wu K. Identification of the associated expression patterns as potential predictive markers for cancer prognosis. *bioRxiv*. 2020.
28. Liu L, Leng L, Liu C, Lu C, Yuan Y, Wu L et al. An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. *Nature communications*. 2019;10(1):364. doi:10.1038/s41467-018-08244-0.
29. Liu ZH, Lian BF, Dong QZ, Sun H, Wei JW, Sheng YY et al. Whole-exome mutational and transcriptional landscapes of combined hepatocellular cholangiocarcinoma and intrahepatic cholangiocarcinoma reveal molecular diversity. *Biochimica et biophysica acta Molecular basis of disease*. 2018;1864(6 Pt B):2360-8. doi:10.1016/j.bbadis.2018.01.027.
30. Lofling L, Karimi A, Sandin F, Bahmanyar S, Kieler H, Lambe M et al. Clinical characteristics and survival in non-small cell lung cancer patients by smoking history: a population-based cohort study. *Acta oncologica*. 2019;58(11):1618-27. doi:10.1080/0284186X.2019.1638521.
31. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal*. 2011;17(1):3. doi:10.14806/ej.17.1.200.
32. Neumann T, Rasmussen M, Heitmann BL, Tønnesen H. Gold standard program for heavy smokers in a real-life setting. *International journal of environmental research and public health*. 2013;10(9):4186-99. doi:10.3390/ijerph10094186.
33. Noro R, Honda K, Tsuta K, Ishii G, Maeshima AM, Miura N et al. Distinct outcome of stage I lung adenocarcinoma with ACTN4 cell motility gene amplification. *Annals of oncology : official journal of the European Society for Medical Oncology*. 2013;24(10):2594-600. doi:10.1093/annonc/mdt293.
34. Oberg M, Jaakkola MS, Woodward A, Peruga A, Pruss-Ustun A. Worldwide burden of disease from exposure to second-hand smoke: a retrospective analysis of data from 192 countries. *Lancet*. 2011;377(9760):139-46. doi:10.1016/S0140-6736(10)61388-8.
35. Ogretmen B. Sphingolipid metabolism in cancer signalling and therapy. *Nature reviews Cancer*. 2018;18(1):33-50. doi:10.1038/nrc.2017.96.
36. Quinlan AR, Hall IMJB. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-2. doi:10.1093/bioinformatics/btq033.
37. Russo D, Capolupo L, Loomba JS, Sticco L, D'Angelo G. Glycosphingolipid metabolism in cell fate specification. *Journal of cell science*. 2018;131(24). doi:10.1242/jcs.219204.
38. Shaykhiev R, Otaki F, Bonsu P, Dang DT, Teater M, Strulovici-Barel Y et al. Cigarette smoking reprograms apical junctional complex molecular architecture in the human airway epithelium in vivo.

- Cellular and molecular life sciences : CMLS. 2011;68(5):877-92. doi:10.1007/s00018-010-0500-x.
39. Soini Y. Tight junctions in lung cancer and lung metastasis: a review. *International journal of clinical and experimental pathology*. 2012;5(2):126-36.
 40. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers—a different disease. *Nature reviews Cancer*. 2007;7(10):778-90. doi:10.1038/nrc2190.
 41. Tentler D, Lomert E, Novitskaya K, Barlev NA. Role of ACTN4 in Tumorigenesis, Metastasis, and EMT. *Cells*. 2019;8(11). doi:10.3390/cells8111427.
 42. Toyooka S, Tokumo M, Shigematsu H, Matsuo K, Asano H, Tomii K et al. Mutational and epigenetic evidence for independent pathways for lung adenocarcinomas arising in smokers and never smokers. *Cancer research*. 2006;66(3):1371-5. doi:10.1158/0008-5472.CAN-05-2625.
 43. Wang Z, Tu K, Xia L, Luo K, Luo W, Tang J et al. The Open Chromatin Landscape of Non-Small Cell Lung Carcinoma. *Cancer research*. 2019;79(19):4840-54. doi:10.1158/0008-5472.CAN-18-3663.
 44. Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhi R et al. Characterizing the cancer genome in lung adenocarcinoma. *Nature*. 2007;450(7171):893-8. doi:10.1038/nature06358.
 45. Willer T, Inamori K, Venzke D, Harvey C, Morgensen G, Hara Y et al. The glucuronyltransferase B4GAT1 is required for initiation of LARGE-mediated alpha-dystroglycan functional glycosylation. *eLife*. 2014;3. doi:10.7554/eLife.03941.
 46. Wu K, Zhang X, Li F, Xiao D, Hou Y, Zhu S et al. Frequent alterations in cytoskeleton remodelling genes in primary and metastatic lung adenocarcinomas. *Nat Commun*. 2015;6:10131. doi:10.1038/ncomms10131.
 47. Zhang L, Wang Y, Zhang B, Zhang H, Zhou M, Wei M et al. Claudin-3 expression increases the malignant potential of lung adenocarcinoma cells: role of epidermal growth factor receptor activation. *Oncotarget*. 2017;8(14):23033-47. doi:10.18632/oncotarget.14974.
 48. Zhuo D, Li X, Guan F. Biological Roles of Aberrantly Expressed Glycosphingolipids and Related Enzymes in Human Cancer Development and Progression. *Frontiers in physiology*. 2018;9:466. doi:10.3389/fphys.2018.00466.

Tables

Table 1. Three pathways related to PC2 high contributing peaks found with DAVID 6.8

Pathways	p value	Genes
hsa00601: Glycosphingolipid biosynthesis - lacto and neolacto series	0.0080	<i>B4GAT1, FUT3, B3GNT3</i>
hsa04530: Tight junction	0.010	<i>ACTN4, CLDN3, TJP3, LLGL2</i>
hsa01100: Metabolic pathways	0.091	<i>DGKA, GALNT3, POLE2, HSD17B2, B4GAT1, AKR1B1, FUT3, B3GNT3, GALE, AGMAT, GAL3ST1</i>

Figures

Figure 1

The graphical abstract of analyses performed in this study To reduce the complexity of the large data and only emphasize on important peaks, a correlation network was constructed. Peaks were connected to each other if their direct or indirect correlations are significant (blue dots). We chose the 10% most frequently-connected peaks as the important peaks for the further PCA analysis. We then analyzed the association between components obtained from PCA and patient’s survival.

All cancer types (N=385)

Figure 2

Principal Component Analysis on the selected peaks from TCGA ATAC-data The unsupervised PCA for the 10% most frequently-connected peaks selected from TCGA ATAC-seq data across all cancer types. Each dot represents a given sample. Color represents the cancer type shown in the Figure.

Figure 3

Identification of the smoking associated peak set from LUAD patients using TCGA ATAC-data A. The unsupervised PCA on LUAD samples (N=22) from TCGA. Dots present samples and their colors present patients' different smoking histories. The left and right sides of the orange rectangle were defined by the $\pm$ mean of all LUAD samples' distances from the y axis. The upper and lower sides of the blue rectangle area were defined by the $\pm$ mean of all LUAD samples' distances from the x axis. The x and y values of each sample were generated in PCA, which are not true peaks values. They represent the level of variety towards PC1 or 2-axis. The samples within the two rectangle areas thus have relatively more stable PC1 or 2-related peaks. B. The absolute distance from PC1 axis of LUAD patients were compared between groups of <20 pack-years (N=5) and ≥ 20 pack-years (N=12). One outlier from group of ≥ 20 pack-years had been removed, according to Grubbs' test. Pack-years information of four patients is not available in TCGA data. Each sample's absolute distance from PC1 axis (the line $x=0$ in Fig.3A) was measured and two-tailed unpaired t-tests was performed to compare the difference between the two groups. PC1 related peak set identified from TCGA ATAC-data in LUAD patients was not associated with smoking, as the difference is not significant. C. The absolute distance from PC2 axis of LUAD patients were compared between groups of <20 pack-years (N=4) and ≥ 20 pack-years (N=13). One outlier from group of <20 pack-years had been removed, according to Grubbs' test. Each sample's absolute distance from PC2 axis (the line $y=0$ in Fig.3A) was measured and two-tailed unpaired t-tests was performed to compare the difference between the two groups. The difference is statistically significant as p value=0.029.

Figure 4

The PFS and OS of LUAD patients with different absolute distance towards PC1 or PC2 Samples were divided into two groups based on whether it locates in the rectangle with orange borders in Fig. 3A, which were defined by the $\pm$ mean of all LUAD samples' distances from PC1 axis. The PFS (A) and OS (B) were compared between the group of samples outside rectangle (N=9) and the group of samples inside the rectangle (N=13). Also, samples were divided into two groups based on whether it locates in the blue rectangle in Fig. 3A, which borders were defined by the $\pm$ mean of all LUAD samples' distances from PC2 axis. The PFS (C) and OS (D) were compared between the group of samples outside rectangle (N=8) and the group of samples inside the rectangle (N=14). Differences were measured by Chi-Squared Test, $P < 0.05$ were considered significant. Patients survived but stopped being tracked are indicated by crosses.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Fig.S1.tif](#)
- [SupplementaryTable1.xlsx](#)
- [SupplementaryTable2.xlsx](#)
- [SupplementaryTable3.xlsx](#)
- [SupplementaryTable4.xlsx](#)
- [SupplementaryTable5.xlsx](#)