Bulk and single-cell transcriptomics identify tobacco-use disparity in lung gene expression of ACE2, the receptor of 2019-nCov

Guoshuai Cai

Department of Environmental Health Sciences, Arnold School of Public Health, University of South Carolina, Columbia, SC 29208

Address for Correspondence:

Guoshuai Cai
Department of Environmental Health Sciences
Arnold School of Public Health
University of South Carolina
915 Greene Street
Discovery 517
Columbia, SC 29204
GCAI@mailbox.sc.edu
Phone: 803-777-4120

Abstract

In current severe global emergency situation of 2019-nCov outbreak, it is imperative to identify vulnerable and susceptible groups for effective protection and care. Recently, studies found that 2019-nCov and SARS-nCov share the same receptor, ACE2. In this study, we analyzed four large-scale bulk transcriptomic datasets of normal lung tissue and two single-cell transcriptomic datasets to investigate the disparities related to race, age, gender and smoking status in *ACE2* gene expression and its distribution among cell types. We didn't find significant disparities in *ACE2* gene expression between racial groups (Asian vs Caucasian), age groups (>60 vs <60) or gender groups (male vs female). However, we observed significantly higher *ACE2* gene expression in former smoker's lung compared to non-smoker's lung. Also, we found higher *ACE2* gene expression in Asian current smokers compared to non-smokers but not in Caucasian current smokers, which may indicate an existence of gene-smoking interaction. In addition, we found that *ACE2* gene is expressed in specific cell types related to smoking history and location. In bronchial epithelium, *ACE2* is actively expressed in goblet cells of current smokers and club cells of non-smokers. In alveoli, *ACE2* is actively expressed in remodelled AT2 cells of former smokers. Together, this study indicates that smokers especially former smokers may be more susceptible to 2019-nCov and have infection paths different with non-smokers. Thus, smoking history may provide valuable information in identifying susceptible population and standardizing treatment regimen.

Key words

Wuhan 2019-nCov, ACE2, expression, susceptibility, race, age, gender, smoking, single cell

Introduction

In the past two decades, pathogenic coronaviruses (CoVs) have caused epidemic infections, including the server acute respiratory syndrome (SARS)-CoV outbreak in 2003, the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) outbreak in 2012 and the current Wuhan 2019 Novel Coronavirus (2019-nCov) outbreak. They typically affect the respiratory tract and cause severe respiratory illnesses. We have learned from SARS-Cov and MERS-Cov that human populations showed disparities in susceptibility to these viruses. For example, epidemiology studies found that males had higher incidence and mortality rates than females.[1,2] We believe that the susceptibility to the novel 2019-nCov is also different among population groups. In current severe global emergency situation of 2019-nCov outbreak, it is imperative to identify vulnerable and susceptible groups for effective protection and care.

Recently, Xu et.al. computationally modelled protein interactions and identified a putative cell entry receptor of 2019-nCov, angiotensin-converting enzyme 2 (ACE2), which is also a receptor for SARS-nCov.[3] Zhou et.al. further confirmed this virus receptor in the HELA cell line.[4] Interestingly, Zhao et al. found *ACE2* is specifically expressed in a subset of type II alveolar cells (AT2), in which genes regulating viral reproduction and transmission are highly expressed.[5] They also found that an Asian male had much higher ratio of *ACE2*-expressing cell than other seven white and African American donors, which may indicate the higher susceptibility of Asian. However, the sample size was too small to draw conclusion on this racial disparity. Here, we analyzed four large-scale bulk transcriptomic datasets of normal lung tissue to investigate the disparities related to race, age, gender and smoking status in *ACE2* gene expression. Also, we analyzed two lung tissue single-cell transcriptomic datasets to investigate the distribution of *ACE2* gene expression among cell types, which will provide new knowledge for understanding the mechanism and population disparities of 2019-nCov infection.

Methods

Bulk transcriptomics

Two RNA-seq datasets and two DNA microarray datasets from lung cancer patients were analyzed in this study, including a Caucasian RNA-seq dataset from TCGA (https://www.cancer.gov/tcga), an Asian RNA-seq dataset from Gene Expression Omnibus (GEO) with the accession number GSE40419[6], an Asian microarray dataset from GEO with the accession number GSE19804[7] and a Caucasian microarray dataset from GEO with the accession number GSE10072[8]. Both RNA-seq datasets were generated with the Illumina HiSeq platform and both microarray datasets were generated with the Affymetrix GeneChip Human Genome U133 Array. The details and processing of data were described in our previous study[9]. All datasets contain samples from tumor and normal pairs and we only use the normal samples in this study. In total, 54 samples in the TCGA dataset, 77 samples in the GSE40419 dataset, 60 samples in the GSE19804 dataset and 33 samples in the GSE10072 dataset were analyzed. We studied the Reads Per Kilobase per Million mapped reads (RPKM) values for RNA-seq data and Robust Multi-Array Average (RMA)[10] values for microarray data. All data were log2 transformed to improve normality. The means of data values across samples in datasets from the same platform were highly correlated (Pearson correlation coefficient r=0.9 for microarray datasets and r=0.97 for RNA-seq datasets, Fig. S1), indicating no significant system variation in datasets from the same platform.

Simple linear regressions were used to test the association of *ACE2* gene expression with each single variable of age, gender, race and smoking status. And, multiple linear regression was used to test the association of *ACE2* expression with multiple factors (age, gender, race, smoking status and data platform). Also, ordinal regression was performed to investigate the association between *ACE2* expression and ordinal categorical smoking history. All data management, statistical analyses and visualizations were accomplished using R 3.6.1.

Single-cell transcriptomics

Two single-cell RNA sequencing (scRNA-seq) datasets available in GEO with accession numbers GSE122960[11] and GSE131391[12] were downloaded and analyzed. The

GSE122960 dataset was generated from lung tissue of 8 lung transplant donors (including 1 Asian former smoker, 2 Caucasian current smoker and 5 African American non-smokers), using 3' V2 chemistry kit on 10x Genomics Chromium single cell controller. The GSE131391 study profiled bronchial epithelial cells from 6 never and 6 current smokers using CEL-Seq[13]. Counts of single cells were downloaded, and subsequent data analyses were performed using the Seurat 3.0 package[14], including data normalization, high variable feature selection, data scaling, dimension reduction and cluster identification. We also used SCANNER[15] to assist the data visualization and cell type identification.

Results

No observed disparities between race, age or gender groups

Inconsistent with the study of Zhao et al.[5], we observed no significant difference in *ACE2* expression in Caucasian lung tissue samples compared to Asian lung tissue samples in the RNAseq datasets (*p*-value=0.45, Fig 1A). In the microarray datasets, a higher *ACE2* expression was observed in Caucasian samples compared to Asian samples (*p*-value=0.001, Fig 1A). Given that the GSE19804 RNA-seq study focused on female non-smokers while the TCGA dataset includes samples from both males and females and both smokers and non-smokers, we believe that the observed disparity may be due to other factors other than race, such as smoking, gender and unknown factors. Therefore, we performed multiple linear regression on multiple independent variables (age, gender, race, smoking status and platform) and found no significant difference between racial groups (*p*-value=0.36, Fig. 1B).

Furthermore, we didn't observe a disparity between age groups (>60 vs <60) or gender groups (male vs female) in *ACE2* gene expression in each available dataset (Fig. 1C, D). Consistently, multivariate analysis didn't detect a significant difference between groups of age or gender after other variables (age/gender, race, smoking status and platforms) were adjusted (*p*-value=0.90 for age, *p*-value=0.35 for gender, Fig. 1B). We also

consistently found no difference between male and female healthy lung tissue samples from GTEx[16] (Fig. S2).

Smokers especially former smokers have upregulated *ACE2* in lung

We found a significant higher *ACE2* gene expression in smoker (including current smoker and former smoker) samples compared to non-smoker samples in the TCGA (*p*-value=0.05) and GSE40419 RNA-seq datasets (*p*-value=0.01, Fig. 2A). Smokers in GSE10072 showed a higher mean of *ACE2* gene expression than non-smokers. The difference is not significant (*p*-value=0.18), which may be due to the small sample size of this study (n=33) with insufficient power to detect the difference. The GSE19804 data which has only non-smoker samples available was not included into the analysis. Adjusted by other factors (age, gender, race and platforms) in multivariate analysis, smoking still shows a significant disparity in *ACE2* gene expression (*p*-value=0.01, Fig. 1B). Furthermore, we investigated ordinal categorized smoking history (non-smoker, former smoker quit more than 15yrs, former smoker quit less than15yrs, and current smoker) and showed results in Figure 2B. In the GSE19804 dataset, we found a significant trend of *ACE2* gene expression regulation associated with the smoking history that current smokers had the highest expression, non-smokers had the lowest expression and former smokers had that in-between (ordinal regression *p*-value=0.01, Fig. 2B). We found a similar but non-significant trend in the GSE10072 due to its small size (ordinal regression *p*-value=0.11). We didn't observe such a tread in the TCGA dataset. Instead, we found a higher average expression in recent quitters (<=15 years) compared to non-smokers, current smokers and former smokers who have quit for longer durations (>15 years). This may indicate a difference in *ACE2* gene expression between Caucasian and Asian current smokers, but it is not statistically significantly detected due to the limited sample size in this study (p-value=0.43). Compared to non-smokers, multivariate analysis on all data showed a significant higher *ACE2* expression in former smokers (p-value=0.05) and a higher mean of *ACE2* expression in current smokers, which did not reach statistical significance from current analysis though (p-value=0.13).

Smokers have upregulated *ACE2* in remodelled cell types

Duclos G et.al. studied human bronchial epithelial cells using single-cell RNA sequencing and found smokers showed a remodeled cell composition in bronchial epithelium with a loss of club cells and extensive hyperplasia of goblet cells.[13] We confirmed their analysis in this study based on the same set of cell makers, including *KRT5* for basal cells, *FOXJ1* for ciliated cells, *SCGB1A1* for club cells, *MUC5AC* for goblet cells and *CD45* for WBCs (white blood cells) (Fig. 3A,C, Fig. S3). We also identified a smoking related basal cell subpopulation which might be pro-goblet precursor cells and has been discussed in the study of Duclos G et.al.[13] Interestingly, we found *ACE2* is mainly expressed in goblet cells in current smokers and club cells in non-smokers (Fig. 3B), indicating 2019-nCov may infect different cell types in bronchial epithelium of current smokers and non-smokers.

Further, we analysed a scRNA-seq dataset of whole-lung tissue. Based on the expression of a set of markers shown in Figure 4C and Figure S4, we identified 13 distinct cell populations including alveolar type I (AT1) cells, alveolar type II (AT2) cells, endothelial cells, ciliated cells, club cells, fibroblast, monocytes, macrophages, B, T cell or natural killer T cell (T/NKT), dendritic cells, a former smoker-specific subpopulation of AT2 cells (AT2-reformed) and a current smoker-specific subpopulation of AT2 cells (AT2-smoking) (Fig. 4A). Significantly, current smoker and former smoker showed differently remodelled AT2 cells (AT2-smoking and AT2-reformed, respectively). Also, we found *ACE2* is most actively expressed in AT2-reformed cells in former Asian smokers but not in Caucasian current smokers and African American non-smokers (Fig. 4B, D), which is consistent with our finding from above large-scale bulk transcriptome analysis. In addition, we observed an expression of *MUSC5AC* (the marker of goblet cells) in a subpopulation of club cells of current smokers, indicating the smoking related tissue remodelling with club cell loss and goblet cell hyperplasia (Fig. S5 top). This is consistent with our finding in bronchial epithelium. However, we failed to observe an active expression of *ACE2* in the goblet cells of current smokers in a way similar to what we found in bronchial epithelial cells (Fig. S5 bottom). The limited cell number of this cluster in this dataset might be the reason.

Discussion

In this study, we investigated the disparities related to race, age, gender and smoking status in *ACE2* gene expression by analyzing bulk and single-cell transcriptome data. We found significantly higher *ACE2* gene expression in lung tissue of former smokers compared to that of non-smokers. This may explain the reason why more males (56% of 425 cases) were found in a recent epidemiology report of 2019-nCov early transmission by China CDC[17]. We didn't observe significant disparities in *ACE2* gene expression between racial groups (Asian vs Caucasian), age groups (>60 vs <60) or gender groups (male vs female). However, Asian current smokers may have higher *ACE2* gene expression than Caucasian current smokers. The different is not statistically significant in this study but may indicate an existence of gene-smoking interaction.

We also found that *ACE2* gene is expressed in specific cell types related to smoking history and location. In bronchial epithelium, ACE2 is actively expressed in goblet cells of current smokers and club cells of non-smokers. In alveoli, ACE2 is actively expressed in remodelled AT2 cells of former smokers. This may indicate that 2019-nCov infect respiratory tract through different paths in smokers, former smokers and non-smokers, and this may partially lead to different susceptibility, disease severity and treatment outcome.

This study has several limitations. First, the data analysed in this study were from the normal lung tissue of patients with lung adenocarcinoma, which may be different with the lung tissue of healthy people. Although we observed no difference between male and female healthy samples from GTEx, further validation studies are required for other factors. Second, the small sample size of current single-cell transcriptome datasets has limited power in studying multiple factors involved in this question.

Whether *ACE2* is the only or major receptor of 2019-nCov is unknown. The reason(s) for the tobacco-related disparity in *ACE2* expression is unknown. Studies found smoke significantly increased *ACE2* expression in the lung of rats[18] and cigarette smoke exposure increased pulmonary ACE2 activities in mice[19]. Controversially, other studies showed chronic cigarette smoke and nicotine decreased *ACE2* expression in rats[20,21]. Thus,

substance other than nicotine might contribute to the smoking-related upregulation of *ACE2* found in this study. Further studies are required to find the answer. Despites current limited knowledge, this study indicates that smokers especially former smokers may be more susceptible to 2019-nCov and have infection paths different with non-smokers. Thus, smoking history may provide valuable information in identifying susceptible population and standardizing treatment regimen. Wuhan, stay strong.

Ethical oversight

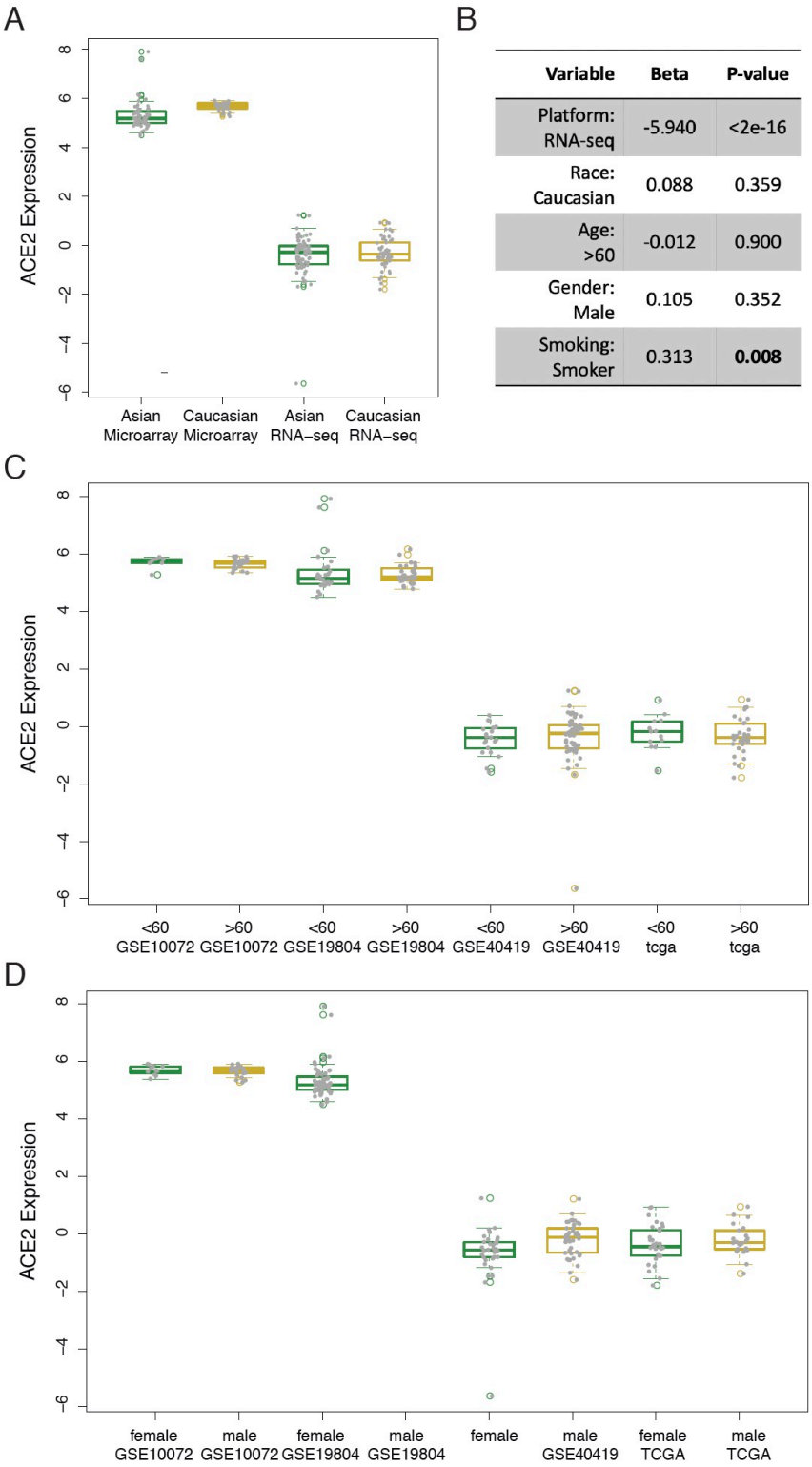
There is no direct involvement of human subjects in this study. All the data use existing de-identified biological samples and data from prior studies. Therefore, ethical oversight and patient consent were not handled in this project.

References

1. Karlberg J, Chong DS, Lai WY. Do men have a higher case fatality rate of severe acute respiratory syndrome than women do? Am J Epidemiol 2004;159:229-31.
2. Alghamdi IG, Hussain, II, Almalki SS, Alghamdi MS, Alghamdi MM, El-Sheemy MA. The pattern of Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive epidemiological analysis of data from the Saudi Ministry of Health. Int J Gen Med 2014;7:417-23.
3. Xu X, Chen P, Wang J, et al. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. SCIENCE CHINA Life Sciences 2020.
4. Zhou P, Yang X-L, Wang X-G, et al. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. bioRxiv 2020:2020.01.22.914952.
5. Zhao Y, Zhao Z, Wang Y, Zhou Y, Ma Y, Zuo W. Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov. bioRxiv 2020:2020.01.26.919985.
6. Seo JS, Ju YS, Lee WC, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome research 2012;22:2109-19.
7. Lu TP, Tsai MH, Lee JM, et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 2010;19:2590-7.
8. Landi MT, Dracheva T, Rotunno M, et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PloS one 2008;3:e1651.
9. Cai G, Xiao F, Cheng C, Li Y, Amos CI, Whitfield ML. Population effect model identifies gene expression predictors of survival outcomes in lung adenocarcinoma for both Caucasian and Asian patients. PLoS One 2017;12:e0175850.

10. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249-64.
11. Reyfman PA, Walter JM, Joshi N, et al. Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. Am J Respir Crit Care Med 2019;199:1517-36.
12. Duclos GE, Teixeira VH, Autissier P, et al. Characterizing smoking-induced transcriptional heterogeneity in the human bronchial epithelium at single-cell resolution. Sci Adv 2019;5:eaaw3413.
13. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep 2012;2:666-73.
14. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. Cell 2019;177:1888-902 e21.
15. Cai G, Xiao F. SCANNER: A Web Server for Annotation, Visualization and Sharing of Single Cell RNA-seq Data. bioRxiv 2020:2020.01.25.919712.
16. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580-5.
17. Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med 2020.
18. Yilin Z, Yandong N, Faguang J. Role of angiotensin-converting enzyme (ACE) and ACE2 in a rat model of smoke inhalation induced acute respiratory distress syndrome. Burns 2015;41:1468-77.
19. Hung YH, Hsieh WY, Hsieh JS, et al. Alternative Roles of STAT3 and MAPK Signaling Pathways in the MMPs Activation and Progression of Lung Injury Induced by Cigarette Smoke Exposure in ACE2 Knockout Mice. Int J Biol Sci 2016;12:454-65.
20. Oakes JM, Fuchs RM, Gardner JD, Lazartigues E, Yue X. Nicotine and the renin-angiotensin system. Am J Physiol Regul Integr Comp Physiol 2018;315:R895-R906.
21. Han SX, He GM, Wang T, et al. Losartan attenuates chronic cigarette smoke exposure-induced pulmonary arterial hypertension in rats: possible involvement of angiotensin-converting enzyme-2. Toxicol Appl Pharmacol 2010;245:100-7.

Figure



Variable	Beta	P-value
Platform: RNA-seq	-5.940	<2e-16
Race: Caucasian	0.088	0.359
Age: >60	-0.012	0.900
Gender: Male	0.105	0.352
Smoking: Smoker	0.313	**0.008**

Figure 1. *ACE2* gene expression profiles in groups.
A, C and D shows groups in race (Caucasian vs Asian), age (>60 vs <60) and gender (male vs female). B shows the result from multivariate analysis with all factors including age, gender, race, smoking and platforms.

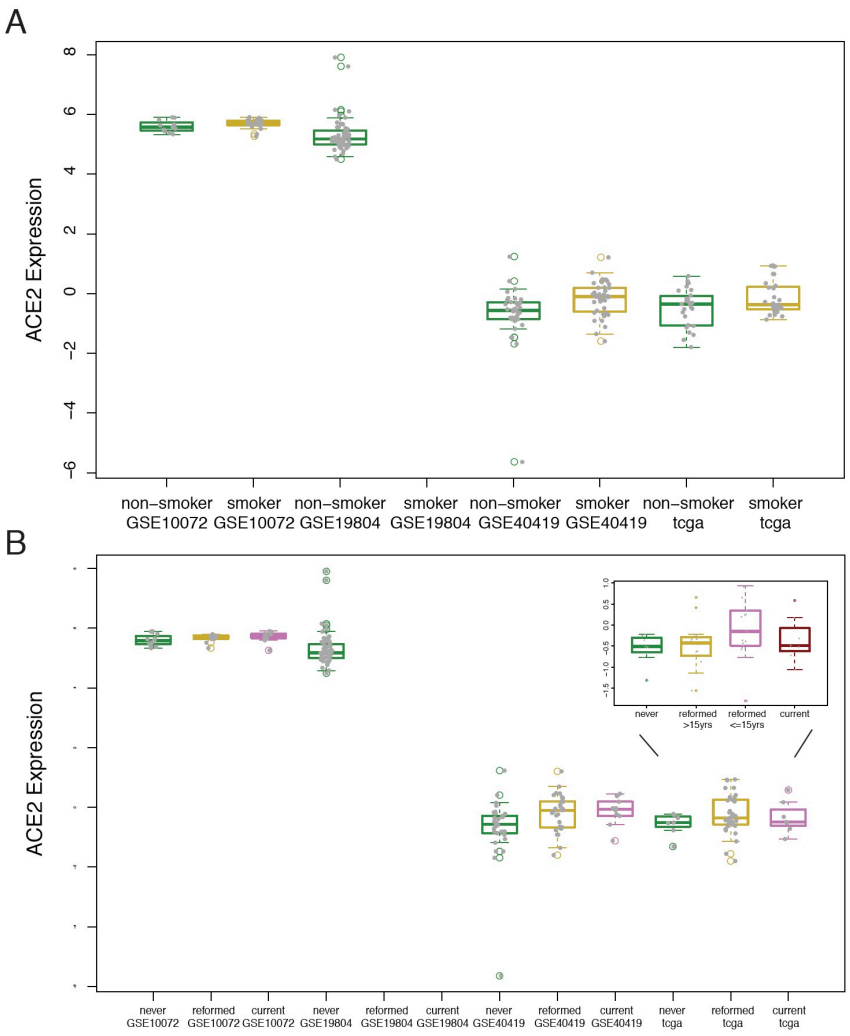


Figure 2. *ACE2* gene expression profiles in smoking groups.
A shows groups of smoker and non-smoker. B shows groups of never-smoker/non-smoker, reformed/former smoker and current smoker. TCGA dataset has more categories of smoking history, including never-smoker, smoker reformed more than 15 years, smoker reformed less than 15 years and current smoker.

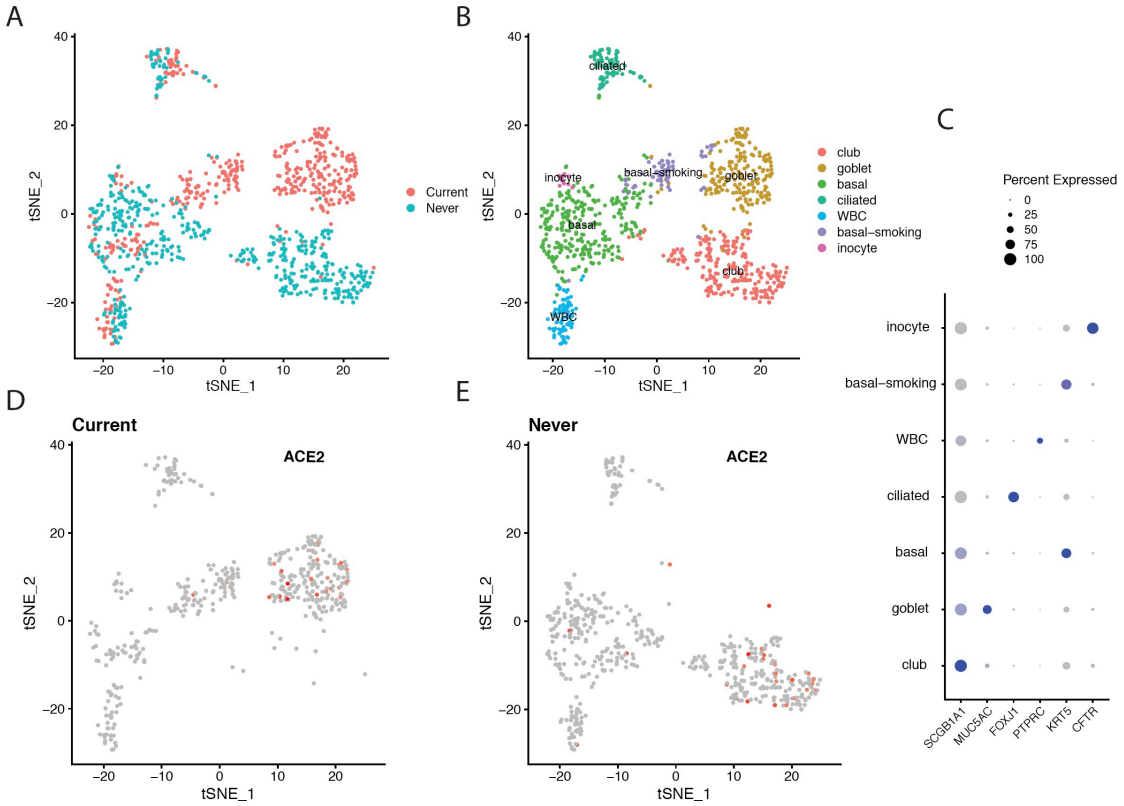


Figure 3. single-cell transcriptomics of bronchial epithelium cells.

A shows tSNE plots of single-cell transcriptome profiles from never smokers and current smokers. B shows identified cell types. C shows detection rates of markers in each cell cluster. D and E show *ACE2* expression in cells of current smokers and never smokers separately.

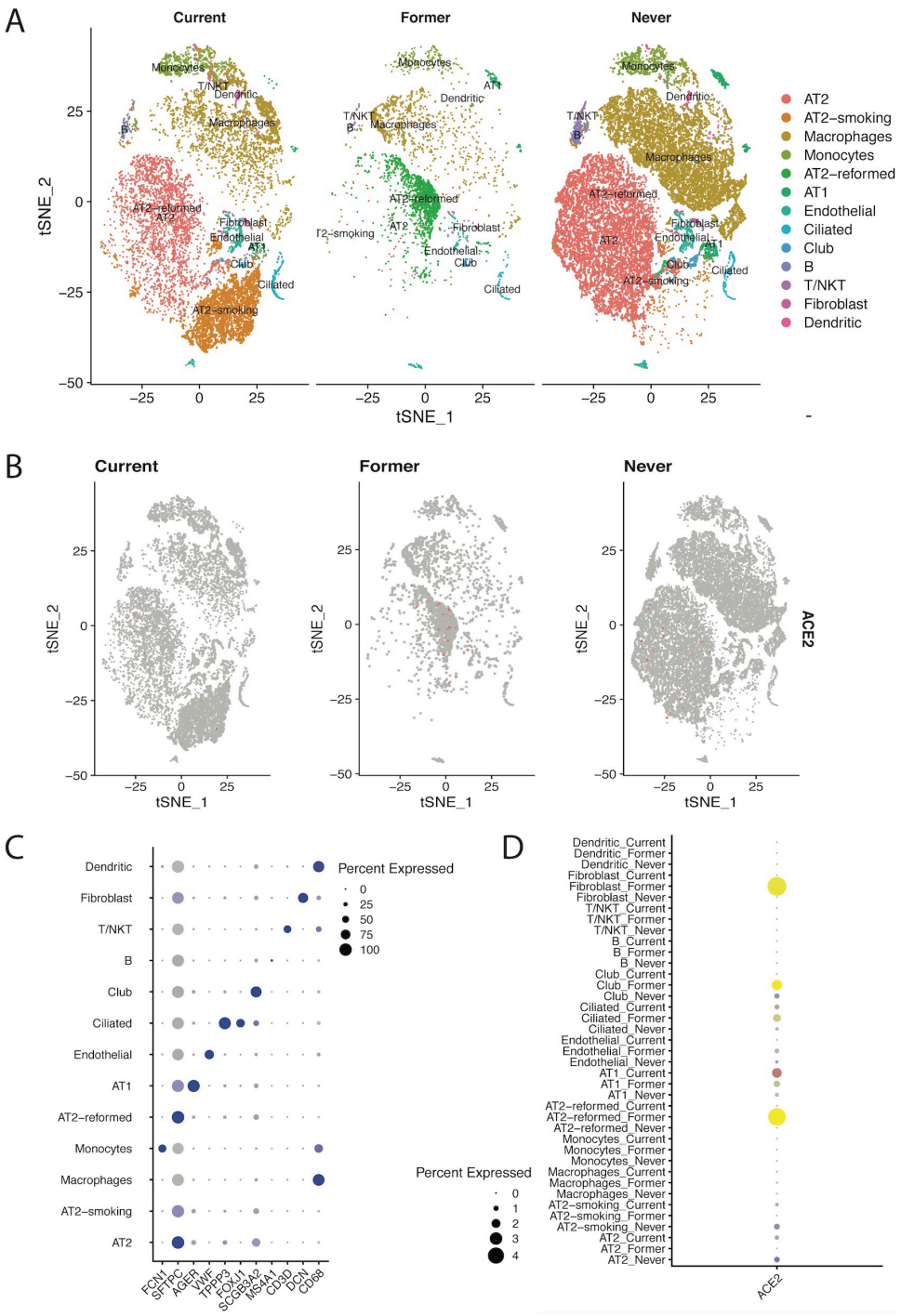


Figure 4. single-cell transcriptomics of lung cells.

A shows tSNE plots of single-cell transcriptome profiles and identified cell types from never smokers, former smokers and current smokers. B shows *ACE2* expression in cells from never smokers, former smokers and current smokers separately. C shows detection rates of markers in each cell cluster. D shows detection rates of *ACE2* in each type of cells from never smokers, former smokers and current smokers.

Supplementary File

Figure S1. Correlation of four datasets.
Lower panel shows pairwise scatter plots of data mean across samples in each dataset.

Upper panel shows their corresponding Pearson correlation coefficients.

Figure S2. *ACE2* gene expression in GTEx female and male lung tissues.
y-axix shows the log10 scaled RNA-seq Transcript Per Million (TPM) values.

Figure S3. Expression profiles of bronchial epithelium cell markers.

Figure S4. Expression profiles of lung tissue cell markers.

Figure S5. Expression profiles of *MUSC5AC* (top) and *ACE2* (bottom).