Systems biology analysis of publicly available transcriptomic data reveals a critical link between AKR1B10 gene expression, smoking and occurrence of lung cancer.

Juan M. Cubillos-Angulo¹,²,³, Eduardo R. Fukutani, Luís A. B. Cruz¹,³,⁴, María B. Arriaga¹,²,³, João Victor Lima⁵, Bruno B. Andrade¹,²,³,⁴,⁵,⁶a, Artur T. L. Queiroz¹a, Kiyoshi F. Fukutani¹,³,⁴a

¹. Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Bahia, Brazil
². Faculdade de Medicina, Universidade Federal da Bahia, Salvador, Bahia, Brazil
³. Multinational Organization Network Sponsoring Translational and Epidemiological Research (MONSTER) Initiative, Fundação José Silveira, Salvador, Bahia, Brazil
⁴. Curso de Medicina, Faculdade de Tecnologia e Ciências, Salvador, Bahia, Brazil
⁵. Universidade Salvador (UNIFACS), Laureate Universities, Salvador, Bahia, Brazil
⁶. Escola Bahiana de Medicina e Saúde Pública (EBMSP), Salvador, Bahia, Brazil

#Corresponding Authors: bruno.andrade@fiocruz.br (BBA), arturlopo@gmail.com (ATLQ) and ferreirafk@gmail.com (KFF)

aBBA, ATLQ and KFF equally contributed to the work
Abstract

**Background:** Cigarette smoking is associated with increased risk of developing respiratory diseases and various types of cancer. Early identification of such unfavorable outcomes in patients who smoke is critical for optimizing personalized medical care.

**Methods:** Here, we perform a comprehensive analysis using Systems Biology tools of publicly available data from a total of six transcriptomic studies, which examined different specimens of lung tissue and/or cells of smokers and nonsmokers to identify potential markers associated with lung cancer.

**Results:** Expression level of twenty-two genes was capable of classifying smokers from non-smokers. A machine learning algorithm revealed that *AKR1B10* was the most informative gene among the 22 DEGs accounting for the classification of the clinical groups. *AKR1B10* expression was higher in smokers compared to non-smokers in datasets examining small and large airway epithelia, but not in the data from a study of sorted alveolar macrophages. We next tested whether *AKR1B10* expression could be useful in identification of cancer tissue in patients who were not exposed to smoking. *AKR1B10* expression was substantially higher in lung cancer specimens compared to matched healthy tissue obtained from nonsmoking individuals (accuracy: 80%, p<0.0001). Finally, we searched the expression of 11 single nucleotide polymorphisms (SNPs) of *AKR1B10* worldwide. We found that the SNP rs782881 was the most frequent mutant allele in the majority of continents. Africa was the continent which exhibited higher frequency of SNPs associated with lower *AKR1B10* expression and displayed lower lung cancer incidence and deaths attributable to tobacco.

**Conclusion:** The systematic analysis of transcriptomic studies performed here revealed a potential critical link between *AKR1B10* expression, smoking and occurrence of lung cancer.
Introduction

Worldwide, cigarette smoking is a life-style habit of approximately 1.1 billion individuals and is associated with more than six million deaths annually [1]. The immunological responses in persons chronically exposed to smoke from cigarette are characterized by protracted secretion of inflammatory factors and by accumulation of several leukocytes in lung tissue and production of pro-fibrotic mediators such as transforming growth factor (TGF)-β [2, 3]. These inflammatory perturbances likely result in increased risk development of tobacco associated morbidity including several types of cancer [4], autoimmune disorders [5], chronic obstructive pulmonary diseases[6] and respiratory infections [7].

The role of tobacco smoking in the induction of disturbances in cell/tissue homeostasis and gene mutations, broadly or specifically associated with several types of tumors, have been investigated. Smoking-related malignancies have been reported to be associated with DNA methylation [8] and mutations in several proto-oncogenes, such as *p53*, *KRAS*, *BRCA-1*, *BRCA-2*, *GPX2*, *GABP*, *TCF3*, *CRX*, *CYP2A13*, *CYP2A6*, *CYP2B6*, among others [9-12]. In addition, it has also been reported that components of cigarette smoking modulate immune cell functions, which could lead to loss of T-cell proliferation and antibody responses [13]. Furthermore, chromosomal instability, epigenomic alterations and several mutations have been reportedly associated with lung cancer in particular [14].

Thus, in general, all of these events ultimately culminate with altered gene expression, even though the conversion of carcinogens to DNA adducts is more efficient in some individuals than in others [15]. Therefore, understanding the expression of these gene is
important to fully understand the link between smoking exposure and risk of cancer
development.

Identification of genetic markers predictive of cancer development is of utmost importance
for promoting personalized medicine [16]. Such markers could be implemented as
screening strategy for patients who exhibit strong risk factors for cancer, such as cigarette
smoking. To identify such potential markers, we performed a systematic analysis of
publicly available data from transcriptomic studies performed in lung tissue and/or cells
and found that, among most of the studies investigated, increased expression of the gene
AKR1B10 was strongly associated with cigarette smoking as well as lung cancer.

Development of a point-of-care assay to assess AKR1B10 expression in individuals
exposed to cigarette smoking may serve as a relevant tool to identify those with high risk
of cancer.
Methods

Ethics Statement

There were no patients directly involved in the research. The present study used publicly available gene expression data from previous studies to perform a meta-transcriptome analysis. All information given to the research team were de-identified.

Description of discovery datasets

We accessed the Gene Expression Omnibus (GEO-NCBI - https://www.ncbi.nlm.nih.gov/geo/) and have looked for datasets with tobacco smoking information in human biopsies or tissues, without diagnosis of other comorbidities. Five datasets have been found in GEO. Thus, two previously published microarray datasets were selected to be used as a discovery set (available from the GEO under accession no. GSE4498 [17] and GSE3320 [18]) and 3 have been used as validation set (GSE20257 [19], GSE17905 [20] and GSE13931 [21]). The Dataset GSE4498 [17] was designed with samples of human small airway bronchial epithelium of smokers (n=10) compared to matched samples from non-smokers (n=12). The dataset GSE3320 [18] was extracted from samples of human small airway bronchial epithelium to assess gene expression in phenotypically smokers (n=6) compared to matched non-smokers (n=5). These included datasets using the same method to collect the samples, by fiberoptic bronchoscopy and brushing. In addition, these studies used a similar transcriptional protocol using the platform Affymetrix Array, making possible to combine both datasets in a discovery set.
In silico validation

We next performed validation of the differentially expressed genes detected in the first phase of the investigation using three distinct datasets selected by examination of gene expression by smoking status: (i) GSE20257 was published by Shaykhievet al [19]. In this study, they used samples of small airway epithelium collected from individuals who were smokers (n=51) and also from those who did not smoke (n=42) and performed an analysis of microarray assays in these samples. (ii) GSE17905 was published by Wang et al [20]. The authors used large airways samples collected by bronchoscopy of 31 smokers and 21 non-smoker individuals and also performed a microarray analysis. (iii) GSE13931 was published by Carolan et al [21]. The investigators used alveolar macrophages collected by broncho-alveolar lavage of 30 smokers and 19 non-smokers and performed a microarray analysis. (iv) Finally, GSE19804 was available in a publication from Lu et al [22]. This dataset had information of gene expression (assessed by microarray) of sixty (n=60) pairs of lung cancer tissue and adjacent normal lung tissue from female patients who were not exposed to cigarette smoking.

The datasets were obtained using the GEOquerry [23] package and raw expression data of 22 samples present on GSE4498 and 11 samples on GSE3320 were normalized and log2 transformed by preprocesscoreR package [24]. Duplicated probes were collapsed by collapserows function in WGCNA package [25] and all common genes to both datasets were kept and used to merge the datasets. The expression data was submitted to a correction procedure of batch effect using an empirical Bayes framework implemented in the COMBAT function available in SVApackage [26].
Statistical analysis

The DEGs were identified by applying the absolute ≥1.0 log2-fold-change threshold and p-value corrected with FDR adjustment for multiple testing (FDR = 5%), from limmapackage [27]. A volcano plot we used to identify changes in gene expression, the significance versus fold-change on the y and x axes, respectively. We use Venn diagrams to visualize all possible logical relations between all the DEGs between smokers and non-smokers in all datasets evaluated. The modular analysis was performed using the Cemitool package [28]. It is based on Weighted correlation network analysis (WGCNA) and default parameters was employed (Beta Parameter = 7). The module annotation was performed with the Kegg database v6.2 [29] and Gene Set Enrichment Analysis (GSEA) algorithm is available internally in the Cemitool package and the Single sample Gene Set Enrichment analysis (ssGSEA) was performed with GSVA package [30]. The significant and annotated pathways were clustered using Euclidean distance as dissimilarity measure and average linkage for between-cluster separation (hclust function in the stats package in R 3.2.2). All The heatmap was generated in R via the heatmap.2 function from the gplots package, using the “scale = “row” switch to Z-score standardize the rows [31]. PCA was performed in order to compare and visualize the expression values of all genes to estimate the variance of the global gene expression with the function prcomp a native package in R. The decision trees were employed to validate and identify the minimal gene set that correctly classifies the smokers from nonsmokers from the 22-gene signature [32]. To estimate the decision tree models accuracy, we performed a 10-fold cross validation. The partition procedure was applied to avoid bias in the training/test sets
sampling. Thus, the training set was used to tune the parameters, learning and building
a model. The validation set was used to test the classifier performance. The sensibility
and specificity were measured from the confusion matrix and visualized in the receiver–
operating characteristic curve (ROC) [32]. Accuracy was evaluated by area under the
curve of ROC plot. We used the genotype-tissue expression project (GTEx) test
association of the SNPs in the AKR1B10 gene with tissues of the respiratory tract [33].
Results

Meta-transcriptome signature of smoking

After preprocessing and merging the datasets, we applied a Principal component analysis (PCA) algorithm using the expression values of all genes to estimate the variance of the global gene expression. This analysis revealed that the subgroups of smokers and non-smokers could not be separated, and two main groups containing both smoker and non-smoker individuals were observed (Figure 1A). To visualize the overall profile of individual gene expression, we used a volcano plot (Figure 1B). This approach indicated presence of a total of 800 statistically significant genes (p<0.05, corrected by Benjamini–Hochberg false discovery rate [FDR]), of which 375 genes were upregulated and 425 genes were downregulated (Figure 1B). Additional analyses identified 22 the differentially expressed genes (DEGs), defined here and genes which exhibited more than ± 1 fold-difference variation (smokers vs. non-smokers) and a significant p-value after FDR adjustment (p<0.05). Such DEGs were inputted in an unsupervised two-way hierarchical clustering analysis. The results demonstrated that when considered together, the 22-gene signature was capable of classifying smokers from non-smokers into completely separate clusters (Figure 1C).

Figure 1. Differentially expressed genes associated with cigarette smoking. We analyzed publicly available data of two datasets of small airways transcriptome (RNAseq). (A) A principal component analysis (PCA) model of 13,516 genes was used to distinguish smokers from nonsmokers. (B) Volcano plot of all genes (smoker vs. nonsmokers). (C) 22 differentially expressed genes (DEGs), defined as p<0.05 after 1%FDR and 1.0-fold change expression, were found and together were able to discriminate the clinical conditions.
To delineate the gene pathways from which the overall transcription profile in smokers vs. non-smokers were involved, we used the CemiTool package[28]. We detected three distinct co-expressed gene modules, annotated in Kyoto Encyclopedia of Genes and Genomes (Kegg) database (Module [M] 2, M7, M14). Two modules were enriched in the non-smoking samples compared to smokers, based on the normalized enrichment scores (NES) (Figure 2A). The first module (M2), found to be enriched in non-smokers was Glycosaminoglycan biosynthesis chondroitin (log10 p = 1.57). A second module (M7) was overrepresented in smokers compared to non-smokers and showed to be enriched in the Peroxisome proliferator-activated receptor (PPAR) signaling pathway (log10 p-value = 1.5). A third module, also more representative in smokers encompassed colorectal cancer (p< 5.34e-38) and basal cell carcinoma (log10 p-value = 3.2). We next calculated the NES for each top ranked pathway identified per individual study subject and found that, when considered together, such pathways were not able to cluster smokers and non-smokers separately (Figure 2C).

Figure 2. Gene pathway analysis in smokers and nonsmokers. (A) Co-expressed modules of all genes. Circle sizes are proportional to the normalized enrichment scores (NES). (B) The modules were annotated using Keg package for R. Dashed lines represent significance threshold. (C) Hierarchical cluster analysis (Ward’s method) using the NES scores for each annotated module and calculated for each person was employed test discrimination between smokers and nonsmokers.

AKR1B10 gene expression level in lung tissue, but not in alveolar macrophages, exhibits high accuracy in discriminating smoking from nonsmoking individuals.

To validate our discoveries, we tested the 22 DEGs identified in our analyses in 3 distinct datasets that compared smokers and non-smokers: (i) GSE20257, that was composed by data from small airways samples, (ii) GSE17905, which compared gene expression from large airways samples and (iii) GSE13931, which used data from alveolar
macrophages. Discriminant analyses using Receiver Operating Characteristic (ROC) curves were able to reveal high accuracy of such gene signature to distinguish smokers from nonsmokers in the two datasets that large and small airway samples (GSE20257 Area under the curve [AUC]: 0.862, p<0.0001; GSE17905 AUC: 0.864, p<0.0001). The same approach indicated that when a dataset from alveolar macrophages was considered, the 22-gene signature was not able to distinguish the study groups (GSE13931AUC: 0.607, p=0.150) (Figure 3A). We next employed a machine-learning approach using decision trees to identify which markers from the 22-gene signature would exhibit more robust discrimination power in each dataset evaluated. Of note, the gene AKR1B10 was the most informative gene in the discovery set and also in the two distinct datasets that used large or small airway tissue (Figure 3B). In the dataset that used gene expression values from alveolar macrophages, AKR1B10 was not shown to be relevant in discrimination, and a combination of two other genes (VGLL3 and TAS2R10) accounted for the differences between smokers and non-smokers (Figure 3B). AKR1B10 expression was higher in smokers compared to non-smokers in all datasets evaluated, except again in the GSE13931, which used data on alveolar macrophages (Figure 3C). Furthermore, we plotted Venn diagrams of all the DEGs between smokers and non-smokers in each dataset to verify overlaps. We confirmed that AKR1B10 was a DEG commonly shown in the discovery set as well as in the databanks which used airway tissue samples, but not in the alveolar macrophage dataset (Figure 3D). The two other DEGs found in smokers were CYP1A1 and HS3ST3A1 (Figure 3D). CYP1A1 encodes a protein that localizes at the endoplasmic reticulum and its expression is induced by
polycyclic aromatic hydrocarbons, some of which are found in cigarette smoke [34].

*HS3ST3A1* is a member of the heparan sulfate biosynthetic enzyme family [35].

**Figure 3. Defining the molecular signatures of smoking.** (A) Data on the 22 DEGs found in our discovery analyses were used to validate discrimination between smokers and nonsmokers in 3 different previously published datasets. (B) Machine-learning decision trees were built for each dataset to describe the most relevant genes driving discrimination. Of note, the gene AKR1B10 was found to be the main discriminator in 3 out of the 4 datasets examined. (C) Scatter plots of the AKR1B10 gene expression in the 4 datasets. (D) Venn diagram of the DEGs in each dataset shows AKR1B10 in the intersection of 3 datasets extracted from lung tissue specimens but not included among DEGs from alveolar macrophages. *p<0.05 (Student’s t-test).*

**AKR1B10 is able to discriminate lung cancer patients who do not smoke**

The results described above demonstrate that higher AKR1B10 expression hallmarks tissue airways from smokers. Given that smoking is a well-established risk factor of lung cancer[9], we hypothesized that AKR1B10 gene expression could also be useful in identification of cancer patients who were not exposed to smoking. To test this hypothesis, we downloaded the dataset GSE19804, which included tissue samples from non-small cell lung cancer as well as ipsilateral healthy lung tissue obtained from patients who did not present history of cigarette smoking. The AKR1B10 gene expression was substantially higher in the specimen collected from the tumor compared to the healthy lung tissue in the same patients (Figure 4A). ROC curve analysis indicated that AKR1B10 gene expression value was able to discriminate cancer and non-cancer tissue with relative high accuracy (AUC 0.80, p<0.0001) (Figure 4B).

**Figure 4. In nonsmokers, higher AKR1B10 expression is a biomarker of lung cancer.** (A) We analyzed AKR1B10 gene expression values in a published dataset of neoplastic lung tissue microarray in nonsmoking individuals who were diagnosed with lung cancer and compared to ipsilateral healthy lung tissue specimens (controls.) Scatter plots of AKR1B10 gene expression in the groups. *p<0.05 (Student’s t-test).* (B) Receiver Operator Characteristics (ROC) indicated a high accuracy to discriminate cancer tissue from controls.
Distribution of Single Nucleotide Polymorphisms of AKR1B10, of lung cancer incidence and of deaths attributed to tobacco worldwide

We next searched for all the single nucleotide polymorphisms (SNPs) of AKR1B10 that have been published so far and found 11, which have been specific in the gene region (Figure 5A). The allele I was denominated the normal nucleotide whereas allele II was the SNP. In Europe, Asia and in the American continents, the most frequent mutant allele was rs782881 (48%, 54% and 42% of the population from these continents, respectively) (Figure 5A). In Africa, the most common mutant allele was rs782510 (78% of the population Figure 5A). Of note, analysis of two SNPs rs782510 and rs782881 with RegulomeDB [36] revealed that both SNPs are likely to affect binding and linked to expression of a gene target e-QTLs (expression quantitative trait loci).

Figure 5. Allele frequencies of Single Nucleotide Polymorphisms of AKR1B10 and distribution of lung cancer incidence and of deaths attributed to tobacco worldwide. (A) Single nucleotide polymorphisms (SNPs) of the AKR1B10 which were published in the Genotype-Tissue Expression (GTEx) project. The allele I was denominated the normal nucleotide whereas allele II was the SNP. (B) Lung cancer incidence in both sexes obtained from the Global Cancer Observatory in 2018. (C) Graph obtained from "Our World in Data" which depicts cancer deaths attributable tobacco in 2016.

Importantly, when lung cancer incidence in 2018 in both sexes was ranked by continent, we observed that Asia accounted for the majority of cases, followed by Europe and North America, whereas incidence of lung cancer in the African continent was relatively low (Figure 5B). Finally, while assessing the Global Burden of Disease dataset [37] we found that these three continents had increased share of cancer deaths attributable tobacco whereas this variable was less representative in the African continent (Figure 5C).
Discussion

In the present study, we examined a number of publicly available transcriptome data to identify a gene signature that could distinguish lung tissue specimens from smokers vs. non-smoking individuals. The most relevant finding of the initial part of analyses was the \textit{AKR1B10} expression level was the most informative in such discrimination among three different datasets obtained from lung tissue, but not in the transcriptome data originated from alveolar macrophages. Moreover, \textit{AKR1B10} gene expression level was able to reliably identify neoplastic from healthy lung tissue in persons not exposed to cigarette smoking. These findings are important because they have identified \textit{AKR1B10} as a reliable biomarker which expression is triggered by cigarette smoking and is simultaneously observed in lung cancer specimens. It is possible that such gene may be involved in carcinogenesis associated with cigarette smoking. In fact, among the multiple carcinogens from cigarette smoke, the nitrosamine 4-(methylnitrosamino)-1-(3-pyridyl)-1-butane (NNK) is described to play a critical role in lung carcinogenesis [38]. Carbonyl reduction takes place in both microsomal and cytosolic fractions from different human tissues such as lung and liver [39]. Within these subcellular fractions, several enzymes have been described to mediate NNK reduction, including the protein encoded by \textit{AKR1B10}, which is from the aldo-keto reductase superfamily (AKR) [40]. Our findings not only confirm an association between \textit{AKR1B10} and occurrence of lung cancer but also argues that expression of such gene may occur before onset of neoplastic lesions in smokers and may be suitable to be used for assessment of risk of cancer in highly exposed individuals. Additional studies are warranted to directly test this hypothesis.
The gene *AKR1B10* found differentially expressed in lung tissue from smokers vs. non-smokers has been previously described in experimental studies to play an important role in the pathophysiology of lung cancer [41]. *AKR1B10* is a regulator of the synthesis of fatty acid and participates in the metabolic pathway of lipids and isoprenoids [42]. In addition, the protein encoded by *AKR1B10* exhibits a high retinaldehyde reductase activity [43]. Importantly, *AKR1B10* can metabolize specific substrates, such as aldo-ketoreductases; farnesal, geranylgeraniol, retinal and carbonyls[44]. Such activity is associated with promotion of carcinogenesis [45]. *AKR1B10* has also been shown to promote cancer cell survival by two distinct studies [46, 47]. These previous investigations revealed that knocking down *AKR1B10* expression induces cancer cell apoptosis and inhibited cancer cell proliferation, suggesting *AKR1B10* could serve as a potential therapeutic target.

Aside from being associated with lung carcinogenesis, *AKR1B10* expression has also been linked to development of several additional types of cancers. In hepatocellular carcinoma (HCC), *AKR1B10* expression is found upregulated, and experimental deletion of such gene inhibited the proliferation of HCC cells tumor growth in a xenograft mice model [48]. In HCT-8, a human colon adenocarcinoma cell line, and NCI-H460, a human lung carcinoma cell line, *AKR1B10* gene deletion has been shown to induce cell apoptosis and mitochondrial degeneration, leading to oxidative stress [46]. Furthermore, higher *AKR1B10* expression has been observed in squamous cell lung carcinoma (SCC) associated with smoking [49]. Finally, our findings indicate that *AKR1B10* is overexpressed in lungs of healthy people who smoke but had no cancer as well as in lung
carcinoma from non-smokers. These observations argue that cigarette smoking already modifies the microenvironment of the lung epithelium probably creating a favorable scenario for carcinogenesis. This idea corroborates with previously published studies which demonstrated that smoking per se mediates upregulation of AKR1B10 expression in the airway epithelia of healthy smokers with no evidence of lung cancer [50]. Thus, there is strong evidence to suggest that cigarette smoking-induced upregulation of AKR1B10 may represent an initial critical step in the cascade of events leading to lung cancer.

In addition to AKR1B10, our analyzes revealed that two additional genes, CYP1A1 and HS3ST3A1, overlapped in the datasets as DEGs capable of discriminating smokers from nonsmokers. Of note, CYP1A1 has also been described to induce carcinogenesis, by promoting CYP-catalyzed epoxidation reactions, resulting in the formation of reactive metabolites that can cause DNA [51, 52]. Moreover, CYP1A1 polymorphisms in smokers increase susceptibility to stomach cancer [53]. Furthermore, HS3ST3A1gene encodes the enzyme 3-O-sulfotransferase, which catalyzes the biosynthesis of a specific subtype of heparan sulfate (HS), 3-O-sulfated heparan sulfate, which is found to be upregulated in human lung cancer specimens and to contribute to its elevated metastatic potential [35]. Thus, the three genes found commonly differentiate regulated in individuals exposed to cigarette smoking are all know to favor development of cancer and could be used as early biomarker of disease progression in high risk populations, but future studies specifically designed to test this hypothesis are necessary.
Our investigations ultimately demonstrated that the $AKR1B10$ SNP rs782881, which is linked to potential increase in gene expression, is the most frequent in Europe, Asia and American continents, where the incidence of lung cancer is high and cigarette smoking is also more frequent. The presence of this SNP has been correlated with increased number of cigarettes consumed per day [54]. Noteworthy, the African continent has much lower frequency of SNP rs782881 and coincidently a relatively lower incidence of lung cancer and consumption of cigarettes compared to the other continents. These findings make us hypothesize that there are SNPs in the $AKR1B10$ gene that may protect against lung cancer. It would be important to conduct studies directly testing the association between different $AKR1B10$ SNPs and occurrence of lung (or other) cancers in distinct populations worldwide.

Our study has several strengths such as the large number of samples evaluated, the use of discovery and validation datasets using different lung tissue/cellular types and different clinical conditions, as well as the detailed search of information about the distribution of different $AKR1B10$ SNPs in the world. An important limitation was the low number of studies included, which was dependent of publicly available datasets. In addition, we have not performed validation in experimental systems. Regardless, by performing a systematic analysis of publicly available data from transcriptomic studies of lung tissue and cells, our study provides strong evidence to support the critical role of $AKR1B10$ in smoking-associated lung cancer.
Acknowledgments

We thank Universidade Salvador, Fundação José Silveira and Fundação Oswaldo Cruz for the support. Mr. Olival Rocha, Mr. Jose Lima, Mr. Luiz Matos, Mr. Getúlio Pacheco, Mr. Humberto and Mr. Edvan Santana (Universidade Salvador) for the technical support. This study was supported by the Intramural Program of Fundação Oswaldo Cruz (FIOCRUZ), Fundação José Silveira and by the Brazilian National Council for Scientific and Technological Development (CNPq). K.F.F. received a fellowship from the Programa Nacional de Pós-Doutorado, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (Finance Code 001). The work of B.B.A. was supported by grants from the NIH (U01AI115940, R01AI069923-08, R01AI20790-02). BBA and A.T.L.Q are senior investigators from CNPq. J. M. C.-A. was supported by the Organization of American States - Partnerships Program for Education and Training (OAS-PAEC) and his study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. M.B.A. received PhD fellowship from Fundação de Amparo à Pesquisa da Bahia (FAPESB) and FIOCRUZ. L.A.B.C. was supported by a research fellowship from CNPq. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
References

1. WHO. World Health Organization global report on trends in prevalence of tobacco smoking. Publications of the World Health Organization; 2015.

2. Barnes PJ. Inflammatory mechanisms in patients with chronic obstructive pulmonary disease. J Allergy Clin Immunol. 2016;138(1):16-27. Epub 2016/07/05. doi: 10.1016/j.jaci.2016.05.011. PubMed PMID: 27373322.

3. Sohal SS, Ward C, Danial W, Wood-Baker R, Walters EH. Recent advances in understanding inflammation and remodeling in the airways in chronic obstructive pulmonary disease. Expert Rev Respir Med. 2013;7(3):275-88. Epub 2013/06/06. doi: 10.1586/ers.13.26. PubMed PMID: 23734649.

4. Ranjit S, Kumar S. Recent advances in cancer outcomes in HIV-positive smokers. F1000Res. 2018;7. Epub 2018/06/28. doi: 10.12688/f1000research.12068.1. PubMed PMID: 29946425; PubMed Central PMCID: PMCPMC5998002.

5. Perricone C, Versini M, Ben-Ami D, Gertel S, Watad A, Segel MJ, et al. Smoke and autoimmunity: The fire behind the disease. Autoimmun Rev. 2016;15(4):354-74. Epub 2016/01/17. doi: 10.1016/j.autrev.2016.01.001. PubMed PMID: 26772647.

6. Cockcroft DW. Environmental Causes of Asthma. Semin Respir Crit Care Med. 2018;39(1):12-8. Epub 2018/02/11. doi: 10.1055/s-0037-1606219. PubMed PMID: 29427981.

7. Feldman C, Anderson R. Cigarette smoking and mechanisms of susceptibility to infections of the respiratory tract and other organ systems. J Infect. 2013;67(3):169-84. Epub 2013/05/28. doi: 10.1016/j.jinf.2013.05.004. PubMed PMID: 23707875.

8. Lee KW, Pausova Z. Cigarette smoking and DNA methylation. Front Genet. 2013;4:132. Epub 2013/07/25. doi: 10.3389/fgene.2013.00132. PubMed PMID: 23882278; PubMed Central PMCID: PMCPMC3713237.

9. Gibbons DL, Byers LA, Kurie JM. Smoking, p53 mutation, and lung cancer. Mol Cancer Res. 2014;12(1):3-13. Epub 2014/01/21. doi: 10.1158/1541-7786.MCR-13-0539. PubMed PMID: 24442106; PubMed Central PMCID: PMCPMC3925633.

10. Blackford A, Parmigiani G, Kensler TW, Wolfgang C, Jones S, Zhang X, et al. Genetic mutations associated with cigarette smoking in pancreatic cancer. Cancer Res. 2009;69(8):3681-8. Epub 2009/04/09. doi: 10.1158/0008-5472.CAN-09-0015. PubMed PMID: 19351817; PubMed Central PMCID: PMCPMC2669837.
11. Hecht SS. Progress and challenges in selected areas of tobacco carcinogenesis. Chem Res Toxicol. 2008;21(1):160-71. Epub 2007/12/07. doi: 10.1021/tr7002068. PubMed PMID: 18052103; PubMed Central PMCID: PMCPMC2556958.

12. Jin Y, Xu P, Liu X, Zhang C, Tan C, Chen C, et al. Cigarette Smoking, BPDE-DNA Adducts, and Aberrant Promoter Methyllations of Tumor Suppressor Genes (TSGs) in NSCLC from Chinese Population. Cancer Invest. 2016;34(4):173-80. Epub 2016/04/05. doi: 10.3109/07357907.2016.1156689. PubMed PMID: 27042875.

13. Sopori M. Effects of cigarette smoke on the immune system. Nat Rev Immunol. 2002;2(5):372-7. Epub 2002/05/30. doi: 10.1038/nri803. PubMed PMID: 12033743.

14. Tan Q, Wang G, Huang J, Ding Z, Luo Q, Mok T, et al. Epigenomic analysis of lung adenocarcinoma reveals novel DNA methylation patterns associated with smoking. Onco Targets Ther. 2013;6:1471-9. Epub 2013/11/10. doi: 10.2147/OTT.S51041. PubMed PMID: 24204162; PubMed Central PMCID: PMCPMC3818101.

15. Peluso M, Munnia A, Piro S, Armillis A, Ceppi M, Matullo G, et al. Smoking, DNA adducts and number of risk DNA repair alleles in lung cancer cases, in subjects with benign lung diseases and in controls. J Nucleic Acids. 2010;2010:386798. Epub 2010/10/27. doi: 10.4061/2010/386798. PubMed PMID: 20976253; PubMed Central PMCID: PMCPMC2952824.

16. Hamburg MA, Collins FS. The path to personalized medicine. N Engl J Med. 2010;363(4):301-4. Epub 2010/06/17. doi: 10.1056/NEJMtp1006304. PubMed PMID: 20551152.

17. Tilley AE, Harvey BG, Heguy A, Hackett NR, Wang R, O'Connor TP, et al. Down-regulation of the notch pathway in human airway epithelium in association with smoking and chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2009;179(6):457-66. Epub 2008/12/25. doi: 10.1164/rccm.200705-795OC. PubMed PMID: 19106307; PubMed Central PMCID: PMCPMC2654975.

18. Harvey BG, Heguy A, Leopold PL, Carolan BJ, Ferris B, Crystal RG. Modification of gene expression of the small airway epithelium in response to cigarette smoking. J Mol Med (Berl). 2007;85(1):39-53. Epub 2006/11/23. doi: 10.1007/s00109-006-0103-z. PubMed PMID: 17115125.

19. 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. Cell Mol Life Sci. 2011;68(5):877-92. Epub 2010/09/08. doi: 10.1007/s00018-010-0500-x. PubMed PMID: 20820852; PubMed Central PMCID: PMCPMC3838912.

20. Wang G, Wang R, Ferris B, Salit J, Strulovici-Barel Y, Hackett NR, et al. Smoking-mediated up-regulation of GAD67 expression in the human airway epithelium. Respir Res.
21. Carolan BJ, Harvey BG, Hackett NR, O'Connor TP, Cassano PA, Crystal RG. Disparate oxidant gene expression of airway epithelium compared to alveolar macrophages in smokers. Respir Res. 2009;10:111. Epub 2009/11/19. doi: 10.1186/1465-9921-10-111. PubMed PMID: 19919714; PubMed Central PMCID: PMCPMC2787510.

22. Lu TP, Tsai MH, Lee JM, Hsu CP, Chen PC, Lin CW, et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. Cancer Epidemiol Biomarkers Prev. 2010;19(10):2590-7. Epub 2010/08/31. doi: 10.1158/1055-9965.EPI-10-0332. PubMed PMID: 20802022.

23. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007;23(14):1846-7. Epub 2007/05/15. doi: 10.1093/bioinformatics/btm254. PubMed PMID: 17496320.

24. Bolstad B. preprocessCore: A Collection of Pre-Processing Functions. R package. 2018.

25. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. Epub 2008/12/31. doi: 10.1186/1471-2105-9-559. PubMed PMID: 19114008; PubMed Central PMCID: PMCPMC2631488.

26. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27. Epub 2006/04/25. doi: 10.1093/biostatistics/kxj037. PubMed PMID: 16632515.

27. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. Epub 2015/01/22. doi: 10.1093/nar/gkv007. PubMed PMID: 25605792; PubMed Central PMCID: PMCPMC4402510.

28. Russo PST, Ferreira GR, Cardozo LE, Burger MC, Arias-Carrasco R, Maruyama SR, et al. CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. BMC Bioinformatics. 2018;19(1):56. Epub 2018/02/21. doi: 10.1186/s12859-018-2053-1. PubMed PMID: 29458351; PubMed Central PMCID: PMCPMC5819234.

29. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30. Epub 1999/12/11. doi: 10.1093/nar/28.1.27. PubMed PMID: 10592173; PubMed Central PMCID: PMCPMC102409.

30. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7. Epub 2013/01/18. doi: 10.1186/1471-2105-14-7. PubMed PMID: 23323831; PubMed Central PMCID: PMCPMC3618321.
31. Wickham H. ggplot2 Elegant Graphics for Data Analysis: Springer-Verlag; 2016. 260 p.
32. Fukutani KF, Kasprzykowski JI, Paschoal AR, Gomes MS, Barral A, de Oliveira CI, et al. Meta-Analysis of Aedes aegypti Expression Datasets: Comparing Virus Infection and Blood-Fed Transcriptomes to Identify Markers of Virus Presence. Front Bioeng Biotechnol. 2017;5:84. Epub 2018/01/30. doi: 10.3389/fbioe.2017.00084. PubMed PMID: 29376049; PubMed Central PMCID: PMCPMC5768613.
33. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5. Epub 2013/05/30. doi: 10.1038/ng.2653. PubMed PMID: 23715323; PubMed Central PMCID: PMCPMC4010069.
34. Pavanello S, Clonfero E. Biological indicators of genotoxic risk and metabolic polymorphisms. Mutat Res. 2000;463(3):285-308. Epub 2000/10/06. PubMed PMID: 11018745.
35. Nakano T, Shimizu K, Kawashima O, Kamiyoshihara M, Kakegawa S, Sugano M, et al. Establishment of a human lung cancer cell line with high metastatic potential to multiple organs: gene expression associated with metastatic potential in human lung cancer. Oncol Rep. 2012;28(5):1727-35. Epub 2012/08/28. doi: 10.3892/or.2012.1972. PubMed PMID: 22922681.
36. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22(9):1790-7. Epub 2012/09/08. doi: 10.1101/gr.137323.112. PubMed PMID: 22955989; PubMed Central PMCID: PMCPMC3431494.
37. Ritchie MRaH. Share of cancer deaths attributed to smoking 2019 Available from: https://ourworldindata.org/smoking.
38. Akopyan G, Bonavida B. Understanding tobacco smoke carcinogen NNK and lung tumorigenesis. Int J Oncol. 2006;29(4):745-52. Epub 2006/09/12. PubMed PMID: 16964372.
39. Maser E, Stinner B, Atalla A. Carbonyl reduction of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) by cytosolic enzymes in human liver and lung. Cancer Lett. 2000;148(2):135-44. Epub 2000/03/01. PubMed PMID: 10695989.
40. Stapelfeld C, Neumann KT, Maser E. Different inhibitory potential of sex hormones on NNK detoxification in vitro: A possible explanation for gender-specific lung cancer risk. Cancer Lett. 2017;405:120-6. Epub 2017/07/27. doi: 10.1016/j.canlet.2017.07.016. PubMed PMID: 28743530.
41. Kang MW, Lee ES, Yoon SY, Jo J, Lee J, Kim HK, et al. AKR1B10 is associated with smoking and smoking-related non-small-cell lung cancer. J Int Med Res. 2011;39(1):78-85. Epub 2011/06/16. doi: 10.1177/147323001103900110. PubMed PMID: 21672310.
42. Ma J, Yan R, Zu X, Cheng JM, Rao K, Liao DF, et al. Aldo-keto reductase family 1 B10 affects fatty acid synthesis by regulating the stability of acetyl-CoA carboxylase-alpha in breast cancer cells. J Biol Chem. 2008;283(6):3418-23. Epub 2007/12/07. doi: 10.1074/jbc.M707650200. PubMed PMID: 18056116.

43. Gallego O, Ruiz FX, Ardevol A, Dominguez M, Alvarez R, de Lera AR, et al. Structural basis for the high all-trans-retinaldehyde reductase activity of the tumor marker AKR1B10. Proc Natl Acad Sci U S A. 2007;104(52):20764-9. Epub 2007/12/19. doi: 10.1073/pnas.0705659105. PubMed PMID: 18087047; PubMed Central PMCID: PMCPMC2410076.

44. Martin HJ, Maser E. Role of human aldo-keto-reductase AKR1B10 in the protection against toxic aldehydes. Chem Biol Interact. 2009;178(1-3):145-50. Epub 2008/11/18. doi: 10.1016/j.cbi.2008.10.021. PubMed PMID: 19013440.

45. Berndt N, Hamilton AD, Sebti SM. Targeting protein prenylation for cancer therapy. Nat Rev Cancer. 2011;11(11):775-91. Epub 2011/10/25. doi: 10.1038/nrc3151. PubMed PMID: 22020205; PubMed Central PMCID: PMCPMC4037130.

46. Wang C, Yan R, Luo D, Watabe K, Liao DF, Cao D. Aldo-keto reductase family 1 member B10 promotes cell survival by regulating lipid synthesis and eliminating carbonyls. J Biol Chem. 2009;284(39):26742-8. Epub 2009/08/01. doi: 10.1074/jbc.M109.022897. PubMed PMID: 19643728; PubMed Central PMCID: PMCPMC2785362.

47. Chung YT, Matkowskyj KA, Li H, Bai H, Zhang W, Tsao MS, et al. Overexpression and oncogenic function of aldo-keto reductase family 1B10 (AKR1B10) in pancreatic carcinoma. Mod Pathol. 2012;25(5):758-66. Epub 2012/01/10. doi: 10.1038/modpathol.2011.191. PubMed PMID: 22222635; PubMed Central PMCID: PMCPMC3323665.

48. Satow R, Shitashige M, Kanai Y, Takeshita F, Ojima H, Jigami T, et al. Combined functional genome survey of therapeutic targets for hepatocellular carcinoma. Clin Cancer Res. 2010;16(9):2518-28. Epub 2010/04/15. doi: 10.1158/1078-0432.CCR-09-2214. PubMed PMID: 20388846.

49. Fukumoto S, Yamauchi N, Moriguchi H, Hippo Y, Watanabe A, Shibahara J, et al. Overexpression of the aldo-keto reductase family protein AKR1B10 is highly correlated with smokers' non-small cell lung carcinomas. Clin Cancer Res. 2005;11(5):1776-85. Epub 2005/03/10. doi: 10.1158/1078-0432.CCR-04-1238. PubMed PMID: 15755999.

50. Wang R, Wang G, Ricard MJ, Ferris B, Strulovici-Barel Y, Salit J, et al. Smoking-induced upregulation of AKR1B10 expression in the airway epithelium of healthy individuals. Chest. 2010;138(6):1402-10. Epub 2010/08/14. doi: 10.1378/chest.09-2634. PubMed PMID: 20705797; PubMed Central PMCID: PMCPMC2998206.
51. Badal S, Delgoda R. Role of the modulation of CYP1A1 expression and activity in chemoprevention. J Appl Toxicol. 2014;34(7):743-53. Epub 2014/02/18. doi: 10.1002/jat.2968. PubMed PMID: 24532440.

52. Buterin T, Hess MT, Luneva N, Geacintov NE, Amin S, Kroth H, et al. Unrepaired fjord region polycyclic aromatic hydrocarbon-DNA adducts in ras codon 61 mutational hot spots. Cancer Res. 2000;60(7):1849-56. Epub 2000/04/15. PubMed PMID: 10766171.

53. Li H, Chen XL, Li HQ. Polymorphism of CYPIA1 and GSTM1 genes associated with susceptibility of gastric cancer in Shandong Province of China. World J Gastroenterol. 2005;11(37):5757-62. Epub 2005/11/05. doi: 10.3748/wjg.v11.i37.5757. PubMed PMID: 16270381; PubMed Central PMCID: PMCPMC4479672.

54. Rohde K, Federbusch M, Horstmann A, Keller M, Villringer A, Stumvoll M, et al. Genetic variants in AKR1B10 associate with human eating behavior. BMC Genet. 2015;16:31. Epub 2015/04/19. doi: 10.1186/s12863-015-0189-9. PubMed PMID: 25887478; PubMed Central PMCID: PMCPMC4379593.
Figure 1
Figure 3
Figure 4
Figure 5