Genome-wide expression changes in the normal-appearing airway during the evolution of smoking-associated lung adenocarcinoma

Jacob Kantrowitz, Ansam Sinjab, Li Xu, Tina L. McDowell, Smruthy Sivakumar, Wenhua Lang, Sayuri Nunomura-Nakamura, Junya Fukuoka, Georges Nemer, Nadine Darwiche, Hassan Chami, Arafat Tfayli, Ignacio I. Wistuba, Paul Scheet, Junya Fujimoto, Avrum E. Spira, Humam Kadara

SUPPLEMENTARY METHODS

RNA-sequencing. RNA samples with RINs ≥ 7 were selected for RNA-Seq (five mice/samples at four months following NNK, t=0; six samples at remaining time points including baseline).

Messenger RNAs were isolated from an average of 300 ng total RNA from each sample using the Dynabeads mRNA DIRECT Micro Purification Kit (Thermo Fisher Scientific) following the manufacturer's protocol. Libraries were then prepared from poly-A RNAs using the Ion Total RNA-Seq Kit v2 (Thermo Fisher Scientific) according to the manufacturer's protocol for barcoded whole-transcriptome libraries. Libraries were assayed on Agilent High Sensitivity DNA or DNA 1000 chips to assess quality (percentage of DNA between 160-1000 bp) and quantity (concentration in pM). Libraries were diluted to 50-75 pM and template reactions were performed out on the Ion Chef Instrument and using the Ion PI Hi-Q Chef kit (Thermo Fisher Scientific) according to the manufacturer's instructions. Reactions were assessed using the Ion Sphere Quality Control Kit on a Qubit 2.0 Fluorimeter before loading onto an Ion PI Chip v3 (Thermo Fisher Scientific). All samples were thereafter sequenced on an Ion Proton sequencer. On average, we obtained approximately 39.5 million reads per sample.

Gene expression analysis. Alignment of reads was performed using a two-step alignment procedure. Unaligned reads were first aligned with the STAR algorithm v2.4.1d (1). Remaining unaligned reads were then realigned with Bowtie2 v. 2.1.0 (2) after which reads from both aligners were combined. Transcripts were then quantified (mm10) using a modified version of the expectation-maximization (E/M) algorithm as described previously (3,4). The resultant transcript reads per kilobase per million (RPKM) values were first processed by adding a pseudocount to all values with RPKM < 1.0 followed by log (base 2) transformation and quantile normalization.

A linear model was used to identify gene expression profiles that were differentially expressed between airways following NNK exposure (t=0) relative to those harvested prior to exposure (baseline). The Benjamini-Hochberg method was used to control the false discovery rate (FDR) (5) and identify significant features based on a *q*-value cutoff. A linear model was also applied to identify profiles that were differentially expressed at each time point following NNK exposure (t=2, t=4 and t=6) relative to t=0. All statistical analyses and hierarchical clustering were conducted using the R statistical language environment v3.1.1 and the *limma* package v3.22.7 (6,7).

Comparative genomics analysis. Comparative genomics analysis was used to analyze mouse airway expression changes in human airway datasets. To analyze changes influenced by NNK (t=0 vs baseline), the previously reported human expression dataset by Spira and colleagues (8) consisting of airway samples from phenotypically healthy current smokers (n=34) and non-smokers (n=23) was interrogated. For concordant analysis of changes directly related to smoking, former smokers from the human airway expression dataset were excluded. The human smoking airway expression dataset was normalized using the brain array v17 CDF and the robust multi-array average (RMA) algorithm (9). Orthologues from the murine airway expression profiles that were identified to be differentially expressed at the end of NNK exposure relative to baseline (FDR < 0.001) were categorized as being up-regulated or downregulated by NNK. The genes were then ranked in the human airway smoking dataset by the *t*-statistic of the coefficient for smoking status from a linear model and then analyzed by gene set enrichment analysis (GSEA) (10) in the human smoking dataset using the GSEA v2.2.3 package in R. GSEA was used to determine profiles that were enriched, conserved and concordantly (in the same direction compared to NNK) modulated with smoking status in the human airway (current versus never).

Temporal mouse airway expression profiles that were modulated at various time points (t=2, t=4, t=6) following NNK exposure relative to t=0 were also analyzed by GSEA in the recently reported array dataset by Silvestri and colleagues (11) of bronchial brushings from smokers with suspicion of lung malignancy (253 with lung cancer and 90 cancer-free) and who had undergone endoscopic bronchoscopy for diagnosis. Following normalization of the human dataset, a linear model was applied that included terms for final lung cancer diagnosis, chronic obstructive pulmonary disease (COPD) status (with and without COPD GOLD stage 2 or worse), age, gender, and smoking status. Genes were ranked according to their *t*-statistic of the coefficient for the effect of final lung cancer diagnosis. GSEA was then used to determine temporal mouse airway expression changes that are concordantly enriched within genes that are differentially expressed in airways of human smokers with lung cancer relative to cancer-free smokers. In a separate analysis, the expression of a bronchial 232 gene-classifier, recently reported by Whitney and colleagues (12) (118 downregulated and 114 up-regulated in airways of smokers with human lung cancer), was analyzed in the post-NNK mouse airway samples. The classifier genes were ranked by the *t*-statistic of the coefficient for time-point from the linear model used to identify differentially expressed genes changing at each time point relative to t=0. Meta-gene scores from the classifier and for each time point following NNK exposure were computed using gene set variation analysis (GSVA) for microarrays and the GSVA package v1.14.1 in R. Differences in the meta-gene scores between the different time points were statistically determined by *t*-tests.

Connecting mouse airway field of injury gene expression to mouse and human tumor expression profiles. Murine RPKM transcript-level data of six lung tumors (from five *Gprc5a*$^{-/-}$ mice at four to seven months following NNK) along with five adjacent normal lung tissues from ongoing RNA-Seq efforts were collapsed into gene level by summation followed by log(base 2) transformation. The downregulated mouse airway field of injury genes were also examined in publicly available datasets of human LUAD or NSCLC and adjacent normal lung tissues. Raw data (CEL files) from the gene expression omnibus (GEO) repository for GSE27262 (13), GSE44077 (14) and GSE7670 (15) were obtained and analyzed as described above using RMA along with Entrez gene-specific probe set mapping (17.0.0) from the Molecular and Behavioral Neuroscience Institute (Brainarray) to produce gene-level expression values. A linear model for paired samples was employed to compute a t-statistic to generate a ranked list for use with GSEA. Genes were ranked according to the beta coefficient of their t-statistic for tissue type (mouse or human lung tumor versus paired adjacent normal). GSEA was used to determine whether murine airway field of injury genes that were downregulated at t=2 versus t=0 (n=68 genes) were enriched with genes also decreased in the murine and human lung tumors compared to their respective adjacent normal lung tissues.

REFERENCES

1. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al.* STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) **2013**;29:15-21
2. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods **2012**;9:357-9
3. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics (Oxford, England) **2010**;26:493-500
4. Xing Y, Yu T, Wu YN, Roy M, Kim J, Lee C. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. Nucleic acids research **2006**;34:3150-60

5. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Statistics in medicine **1990**;9:811-8

6. Diboun I, Wernisch L, Orengo CA, Koltzenburg M. Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. BMC genomics **2006**;7:252

7. 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 research **2015**;43:e47

8. Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X*, et al.* Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proceedings of the National Academy of Sciences of the United States of America **2004**;101:10143-8

9. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U*, et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **2003**;4:249-64

10. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America **2005**;102:15545-50

11. Silvestri GA, Vachani A, Whitney D, Elashoff M, Porta Smith K, Ferguson JS*, et al.* A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. N Engl J Med **2015**;373:243-51

12. Whitney DH, Elashoff MR, Porta-Smith K, Gower AC, Vachani A, Ferguson JS*, et al.* Derivation of a bronchial genomic classifier for lung cancer in a prospective study of patients undergoing diagnostic bronchoscopy. BMC Med Genomics **2015**;8:18

13. Wei TY, Juan CC, Hisa JY, Su LJ, Lee YC, Chou HY*, et al.* Protein arginine methyltransferase 5 is a potential oncoprotein that upregulates G1 cyclins/cyclin-dependent kinases and the phosphoinositide 3-kinase/AKT signaling cascade. Cancer science **2012**;103:1640-50

14. Kadara H, Fujimoto J, Yoo SY, Maki Y, Gower AC, Kabbout M*, et al.* Transcriptomic architecture of the adjacent airway field cancerization in non-small cell lung cancer. Journal of the National Cancer Institute **2014**;106:dju004

15. Su LJ, Chang CW, Wu YC, Chen KC, Lin CJ, Liang SC*, et al.* Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. BMC genomics **2007**;8:140

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Lung tumor development in NNK-exposed *Gprc5a*$^{-/-}$ mice. *Gprc5a*$^{-/-}$ mice were exposed to NNK and followed up for lung tumor development as described in the Materials and Methods section. Following mice euthanasia. Histopathological analysis of mouse lungs was performed to enumerate lung tumor burdens for each mouse at each time point prior to (baseline), at the end of NNK exposure (t=0) and at two (t=2), four (t=4) and six (t=6) months post- NNK exposure. Lung tumor burdens across the different time points were statistically assessed using ANOVA.

Figure S2. Evolutionarily conserved patterns of tobacco-carcinogen mediated gene expression in murine and human airway epithelia. A. GSEA enrichment plots of 362 matched orthologous genes revealing up- and down-regulated subsets that were significantly and concordantly modulated in airways of human cancer-free current smokers versus never-smokers. **B**. The leading edge set of genes identified by GSEA was then analyzed by hierarchical semi-supervised clustering analysis in human airway brushings. Columns represent samples and rows denote gene features (red, higher; blue, lower expression).

Figure S3. Early and persistent changes in mouse airway gene expression *in vivo* following tobacco carcinogen exposure and during lung oncogenesis. Genes differentially expressed in airways of mice at two, four and six months following completion of NNK relative to airways at the end of treatment (t=0) were statistically determined as described in the Materials and Methods section and in Supplementary Methods. Airway profiles were analyzed by hierarchical clustering. Columns represent samples and rows denote gene features; red and blue: higher and lower expression, respectively.

Figure S4. Gene profiles downregulated in the mouse normal-appearing airway field of injury are overall suppressed in lung tumors relative to adjacent normal lung. Genes that were downregulated in the mouse airway field of injury at t=2 relative to t=0 (Table S3) were analyzed by GSEA in the indicated human publicly available cohorts comprised of LUADs or NSCLCs and adjacent normal lung tissues (13-15) and in a set of *Gprc5a*$^{-/-}$ mouse lung tumors and paired normal lungs as described in the Supplementary Methods (*, $P < 0.05$).