RNA-Seq quantification of the human small airway epithelium transcriptome

Neil R Hackett1*, Marcus W Butler1*, Renat Shaykhiev1*, Jacqueline Salit1*, Larsson Omberg2, Juan L Rodriguez-Flores1*, Jason G Mezey1,2*, Yael Strulovici-Barel1*, Guoqing Wang1*, Lukas Didon1* and Ronald G Crystal1*

Abstract

Background: The small airway epithelium (SAE), the cell population that covers the human airway surface from the 6th generation of airway branching to the alveoli, is the major site of lung disease caused by smoking. The focus of this study is to provide quantitative assessment of the SAE transcriptome in the resting state and in response to chronic cigarette smoking using massive parallel mRNA sequencing (RNA-Seq).

Results: The data demonstrate that 48% of SAE expressed genes are ubiquitous, shared with many tissues, with 52% enriched in this cell population. The most highly expressed gene, SCGB1A1, is characteristic of Clara cells, the cell type unique to the human SAE. Among other genes expressed by the SAE are those related to Clara cell differentiation, secretory mucosal defense, and mucociliary differentiation. The high sensitivity of RNA-Seq permitted quantification of gene expression related to infrequent cell populations such as neuroendocrine cells and epithelial stem/progenitor cells. Quantification of the absolute smoking-induced changes in SAE gene expression revealed that, compared to ubiquitous genes, more SAE-enriched genes responded to smoking with up-regulation, and those with the highest basal expression levels showed most dramatic changes. Smoking had no effect on SAE gene splicing, but was associated with a shift in molecular pattern from Clara cell-associated towards the mucus-secreting cell differentiation pathway with multiple features of cancer-associated molecular phenotype.

Conclusions: These observations provide insights into the unique biology of human SAE by providing quantitative assessment of the global transcriptome under physiological conditions and in response to the stress of chronic cigarette smoking.

Background

The tracheobronchial tree, a dichotomous branching structure that begins at the larynx and ends after 23 branches at the alveoli, is lined by an epithelium comprised of 4 major cell types, including ciliated, secretory, undifferentiated columnar and basal cells [1,2]. The airway epithelium is exposed directly to environmental xenobiotics, particulates, pathogens and other toxic substances suspended in inhaled air [2-4]. Of these, chronic cigarette smoking, with its 4000 xenobiotics and > 10^{14} oxidants per puff, is a major cause of airway disease, including chronic obstructive pulmonary disease (COPD) and bronchogenic carcinoma [4-6]. It is the airway epithelium that exhibits the first abnormalities relevant to COPD and lung cancer, and it is the small airway epithelium (SAE; ≥ 6th generation) that is the primary site of the early manifestations of the majority of smoking-induced lung disease [7]. As compared to proximal airways, the small airway epithelium has unique morphologic features with a decrease in the frequency of basal cells and mucus-secreting cells accompanied by increased numbers of Clara cells, a secretory cell subtype critical for the maintenance of the structural and functional integrity at the airway-alveoli interface [1,8-10].

* Correspondence: nhackett@med.cornell.edu; mab9177@med.cornell.edu; res2003@med.cornell.edu; jas2038@med.cornell.edu; jur2014@med.cornell.edu; jgm45@cornell.edu; yas2003@med.cornell.edu; guw2003@med.cornell.edu; lud2002@med.cornell.edu, geneticmedicine@med.cornell.edu
† Contributed equally

1Department of Genetic Medicine, Weill Cornell Medical College, New York, New York, USA
Full list of author information is available at the end of the article

© 2012 Hackett et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Our group [11-13] and others [14-18] have carried out several studies using gene expression microarrays to assess the transcriptome of the human airway epithelium, demonstrating that smoking modulates the expression of hundreds of genes. The advent of RNA-Seq technology, in which the entire polyadenylated transcriptome is sequenced [19-24], is capable of building on this microarray data to provide additional insights into the transcriptome of the airway epithelium and its response to cigarette smoke. Because RNA-Seq provides direct sequencing information of all polyadenylated mRNAs and is not limited by probe design, RNA-Seq data has inherently less noise and higher specificity, and, importantly, provides quantitative information on mRNA transcript number [19]. With high sensitivity and low background, RNA-Seq has a dynamic range of $>8,000$-fold, is highly reproducible, yields digital information not requiring normalization, and can distinguish individual members of highly homologous gene families [25]. In the context of this background, the focus of this study is to utilize massive parallel sequencing to quantify the complete transcriptome of the human SAE in healthy nonsmokers and healthy smokers.

Results

Study Population and SAE sampling

SAE samples from 5 healthy nonsmokers and 6 healthy smokers were analyzed using mRNA-Seq (Additional file 1, Table S1). All individuals had no significant prior medical history and a normal physical examination. To minimize the influence of potential confounding variables, only males of African-American ancestry were assessed. The nonsmokers were younger ($p < 0.02$). There were no differences between the two groups with respect to pulmonary function criteria ($p > 0.1$, all variables). The smoking status was confirmed by urinary tobacco metabolites (see Additional Data Methods and Additional file 1, Table S1). The number of cells recovered ranged from 5.0 to $9.7 \times 10^6$, with $>97\%$ epithelial cells in all cases (Additional file 1, Table S1). There was no difference between the two groups with respect to the proportions of each of the four major cell types, with the exception of ciliated cells, which were significantly lower in the healthy smokers compared to the healthy nonsmokers ($p < 0.04$).

Data Processing and Quality Control

The cDNA generated from SAE samples was run in a single lane per subject on Illumina flow cells. A total of 182 million, 43 base pair single end reads were generated, yielding 7.8 gigabases of sequence. These sequences were aligned using Bowtie version 0.12 (see Additional file 1, Table S2 for a summary of mapping details). To correct for transcript length and depth of coverage, raw reads were converted into reads per kilobase of exon per million mapped reads (RPKM) [23]. RPKM was then assessed across the entirety of reads with reference to exons, introns and intergenic regions. A comparison was made between the expression levels of exons and intergenic regions to define a threshold value above which there was the highest confidence in the validity of the expression value (Figure 1A, B). This was performed by generating a false discovery rate for a range of expression values across all subjects, resulting in the adoption of a cut-off value of 0.125 RPKM, representing an optimal compromise between false positive and false negative values (see Methods). All subsequent analyses were based on the application of this expression threshold.

Of the 21,475 annotated genes in the Human Genome version 19 reference [26], 15,877 (73.9\%) were expressed in SAE at greater than the RPKM cut-off value of 0.125. The average expression level was 13.8 RPKM (Figure 1C) and the average among subject coefficient of variation in RKPM was 0.25. This cut off may be conservative due to overestimation of the number of intronic and intergenic reads. Based on a survey of a random intergenic domain on the genome, we estimate that ~50% of the reads mapped to intergenic and intronic regions correspond to repetitive elements. These probably represent mismapping of reads that properly belong polyadenylated mRNAs that contain the same repetitive elements. If the impact was to drop the threshold from 0.125, as used here, to 0.05, the number of expressed genes would increase from 15,877 to 16,844 (a 6.1\% increase). It is estimated that 1 RPKM corresponds to one mRNA per cell. As a result, we believe there is little significant biology lost by using a conservative cut off (consistent with literature precedent) of 0.125 (i.e., one mRNA per 8 cells) vs a more stringent cut off of 0.05 (one mRNA per 20 cells).

A subset of 12 genes with RPKM differing by $>4$ logs was confirmed by TaqMan realtime quantitative PCR with relative quantitation, using rRNA for normalization. The overall correlation between expression levels determined by these 2 methods was very strong ($r^2 = 0.81$) (Additional file 1, Figure S1).

Comparison of SAE Gene Expression to Other Tissue Transcriptomes

RNA-Seq allows absolute quantitation of mRNA levels and for the fractional contribution of individual transcripts to the total mRNA population to be assessed [24]. In some tissues, transcripts from a relatively small number of genes account for much of the total cellular poly(A)$^+$ RNA pool (Figure 2A). In the case of liver, the single most highly expressed gene contributes 10\% of the total mRNA molecules and the top ten collectively
Figure 1 Establishment of detection limit for gene expression for RNA-Seq assessment of gene expression of the small airway epithelium of healthy nonsmokers. A. Distribution of RPKM for exons (blue), introns (red), and intergenic regions (green). RPKM depends on the size and read numbers mapped in the region considered. The RPKM for introns and intergenic regions was calculated by selecting intronic and intergenic regions throughout the genome that match the size of the exons analyzed, i.e., the size is comparable for the introns, intergenic regions, and exons. B. Estimate of minimum detectable level of expression (RPKM = 0.125) determined from an estimate of false discovery rate (red) and false negative rate (purple) (24). C. Average RPKM expression levels (log10). Dashed line represents the 0.125 threshold.
Figure 2 Composition of the healthy nonsmoker small airway epithelium (SAE) transcriptome. A. Comparison of the SAE transcriptome to that of other tissues. Abscissa - the number of genes, with the genes in descending order of mRNA level. Ordinate - fraction of all mRNAs derived from these genes. The genes expressed by the small airway epithelium (blue) are compared to genes expressed by other organs as indicated [24]. Note that the SAE is similar to liver in that a few genes are expressed at very high levels. B. RNA-Seq sequence alignments for SCGB1A1 (uteroglobin; CC10), the most highly expressed gene in the SAE. The region of the genome corresponding to SCGB1A1 is shown with the read coverage depth for 5 healthy nonsmokers plotted using Partek Genomics Suite version 6.5. RPKM for whole mRNA for each subject is shown on the left. C-E. Frequency distribution of expression level for ubiquitous vs SAE-enriched genes in the small airway epithelium of healthy nonsmokers. "Ubiquitous" genes are those expressed by most tissues; "SAE-enriched" genes are those more abundant in SAE compared to other tissues (see text). For all panels, the number of genes in 1/2 log10 bins was determined starting at the detection limit (RPKM = 0.125). For each panel, the expressed genes are grouped (in 1/2 log10 bins); low (-0.9 to 1), median (> 1 to 10) and high (> 10), with the number of genes and % in each category listed and median RPKM for n = 5 healthy nonsmokers. C. All genes. D. Ubiquitous genes, representing 48% of all expressed genes. E. SAE-enriched genes representing 52% of all expressed genes. Note that the SAE-enriched genes have a much larger proportion of low level expressed genes compared to the ubiquitous genes. F, G. Comparison of coverage of RNA-Seq and microarray assessment of SAE gene expression of healthy nonsmokers. Genes assessed by RNA-Seq were divided into low (0.125-1), medium (> 1-10) and high (> 10) RPKM on the basis of median expression level in n = 5 healthy nonsmokers. Cufflinks U133 data for small airway epithelium for n = 27 African-American healthy nonsmokers [1129; Additional file 1, Table S1] were assessed based on the Affymetrix P calls in low expression (Affymetrix “present (P)” in < 50% red) or high expression in microarray ("P" in > 50% blue). Genes with no unique probe on the microarray are identified in green. F. Ubiquitous genes. G. SAE- enriched genes. For medium and high expressing genes the microarray and RNA-Seq are very similar in detecting expressed genes, but for the SAE-enriched, low level expressed genes detected by RNA-Seq, the microarrays miss a large proportion of the genes.
contribute 37% [24]. In colon, by contrast, the single top mRNA contributes only 2% of the total mRNA and the top ten contribute 9%[24]. Of the total SAE transcripts identified in healthy nonsmokers, 13% mapped to the SCGB1A1 gene (secretoglobin, family 1A, member 1 protein also known as ute-roglobin or Clara cell-specific 10 kD protein [CC10], Table 1). The top 10 genes contributed 24% of the total mRNA (p < 0.05 comparing distribution to both liver, and colon).

**Ubiquitous and SAE-enriched Genes**

To categorize the SAE-expressed genes in healthy nonsmokers as ubiquitous (i.e., expressed by most other tissues) or genes expressed in an SAE-enriched fashion, a comparison was made between the 7,897 genes identified by Ramsköld et al [24] to be ubiquitously expressed in various human tissues and the 15,877 genes expressed in human SAE. The data showed that 7,607 (96.5%) of the genes identified by Ramsköld et al [24] as ubiquitously expressed genes were also expressed by human SAE, indicating that 48% of the SAE transcriptome is comprised of ubiquitously expressed genes. The remaining 52% were designated as “SAE-enriched” genes.

The most highly expressed SAE-enriched gene (Table 1) was SCGB1A1, which is expressed primarily by Clara cells located in small airways [27-29]. RNA-Seq fragments mapped to all three exons of the SCGB1A1 gene at very high density (Figure 2B). Other highly expressed SAE-enriched genes included secretoglobin, family 3A, member 1(SCGB3A1), secretory leukocyte peptidase inhibitor (SLPI), chromosome 20 open reading frame 114 (C20orf114; also known as a long variant of the palate, lung, and nasal epithelium carcinoma associated protein PLUNC), tubulin polymerization-promoting protein family member 3 (TPPP3) and CD74.

### Table 1 Overall Most Highly Expressed Genes in the SAE of Healthy Nonsmokers

| Gene symbol | Gene title                                                                 | Expression level in SAE (RPKM) |
|-------------|---------------------------------------------------------------------------|---------------------------------|
| SCGB1A1     | secretoglobin, family 1A, member 1 (uteroglobin)                           | 38675                           |
| SCGB3A1     | secretoglobin, family 3A, member 1                                        | 7838                           |
| SLPI        | secretory leukocyte peptidase inhibitor                                    | 1602                           |
| C20orf114   | chromosome 20 open reading frame 114                                       | 1484                           |
| TPPP3       | tubulin polymerization-promoting protein family member 3                  | 1302                           |
| CD74        | CD74 molecule, major histocompatibility complex, class II invariant chain  | 947                            |
| TMEM190     | transmembrane protein 190                                                 | 945                            |
| GSTP1       | glutathione S-transferase pi 1                                             | 859                            |
| WFDCC       | WAP four-disulfide core domain 2                                           | 840                            |
| C20orf85    | chromosome 20 open reading frame 85                                        | 738                            |
| TSFAN1      | tetraspanin 1                                                              | 664                            |
| C9orf24     | chromosome 9 open reading frame 24                                         | 629                            |
| NEAT1       | nuclear paraspeckle assembly transcript 1 (non-protein coding)             | 565                            |
| S100A11     | S100 calcium binding protein A11                                           | 540                            |
| KRT19       | keratin 19                                                                 | 493                            |
| MALAT1      | metastasis associated lung adenocarcinoma transcript 1 (non-protein coding)| 461                            |
| ODF3B       | outer dense fiber of sperm tails 3B                                        | 392                            |
| CYP4B1      | cytochrome P450, family 4, subfamily B, polypeptide 1                      | 374                            |
| FOXJ1       | forkhead box J1                                                           | 363                            |
| LCN2        | lipocalin 2                                                                | 359                            |
| PIGR        | polymeric immunoglobulin receptor                                          | 351                            |
| MS4A8B      | membrane-spanning 4-domains, subfamily A, member 8B                        | 348                            |
| ALDH3B1     | aldehyde dehydrogenase 3 family, member B1                                 | 342                            |
| MSM8        | microseminoprotein, beta-                                                 | 333                            |
| RSPH1       | radial spoke head 1 homolog (Chlamydomonas)                               | 318                            |
| CLDN4       | claudin 4                                                                  | 308                            |
| AQP3        | aquaporin 3 (Gill blood group)                                             | 308                            |
| C20orf117   | chromosome 9 open reading frame 117                                        | 302                            |
| IGFBP2      | insulin-like growth factor binding protein 2, 36 kDa                      | 297                            |
| ANXA2P2     | annexin A2 pseudogene 2                                                    | 292                            |

1 Listed are the top 30 most highly expressed SAE enriched genes.
2 Median for n = 5 healthy nonsmokers. PARTEK implementation of Bowtie algorithm with parameters as described in Material and Methods.
To further characterize the SAE transcriptome, gene expression levels derived from RNA-Seq data were divided into three groups. “Low” expression was assigned as significantly expressed (i.e., > 0.125 RPKM cut off, but less than 1 RPKM, corresponding to < 1 mRNA/cell [23]). “Medium” expression was defined at between 1 and 10 RPKM and “high” expression was defined as > 10 RPKM (Figure 2C-E). Analyses of the frequency distribution of ubiquitous and SAE-enriched RefSeq-annotated genes revealed that considerably more of the SAE-enriched gene set were expressed at lower levels (median expression level 1.8 RPKM, Figure 2E) compared to the ubiquitous genes (median expression level 8.6 RPKM, Figure 2D).

Prior to the advent of RNA-Seq method, information about the transcriptome of airway epithelium was derived from gene expression microarrays [11-18]. To assess the concordance of expression pattern by RNA-Seq and microarrays, all genes expressed by RNA-Seq were evaluated as to whether they were identified as expressed in all microarrays, a subset of microarrays or not represented on the microarray (Figures 2F, G). For ubiquitous genes, the percentage of genes identified by RNA-Seq also identified as expressed in > 50% of subjects by microarray was greater for highly expressed genes (97.7%) than medium expressed genes (92.1%). Only 59.2% of ubiquitous genes with low expression identified by RNA-Seq were scored as expressed by microarray (Figure 2F). Similarly, for the SAE-enriched genes, the percent of genes identified by RNA-Seq and in > 50% of subjects by microarray was greater for highly expressed genes (89.8%), compared to medium expressed genes (70.3%), and even more so for genes with low expression (31.3%; Figure 2G). Thus, overall, the two methods are broadly in agreement, but RNA-Seq provides more information, not only quantitative data, but also identification of expression of genes with low expression levels.

**Functional Assessment of Ubiquitous and SAE-enriched Genes**

To better understand biologic functions enriched in the SAE transcriptome, the gene lists were assigned to functional categories using Gene Ontology molecular functions and the expression levels for ubiquitous and SAE-enriched genes were compared (Figure 3, Additional file 1, Table S3). In some functional categories such as signal transduction, the allotment of genes to the SAE-enriched and ubiquitous categories, as well as the distribution of expression levels within those categories, was similar to that for all genes (compare Figure 3D to Figures 2D, E). But several deviations from the expected distribution were observed. For example, for the functional category “translation”, many more genes were classified as ubiquitous compared to SAE-enriched and those that were SAE-enriched were expressed at higher levels than expected based on all genes. By contrast, in the category “immunity”, the number of genes in the SAE-enriched category was higher than expected but distribution of expression level was in line with the average expression of all genes. The median expression levels among the various categories allowed quantification of these differences. The median expression levels for the SAE-enriched genes were lower than that for the ubiquitous genes in all categories. For example, for the ubiquitous genes, the median levels ranged from 7.9 RPKM for membrane receptors to 21.7 RPKM for translation, whereas, for the SAE-enriched genes, the median levels ranged from 1.1 for membrane receptors to 5.6 RPKM for translation. On the average, the most highly expressed category was "translation, ubiquitous" genes (median 21.7 RPKM), whereas the lowest was “membrane receptors, SAE-enriched” (median 1.1 RPKM).

The human SAE is made up of 4 major cell types including ciliated cells (73% abundance in this study), undifferentiated columnar cells (9%), basal cells (10%) and secretory cells (7%, Additional file 2, Table S1). The SAE also has rare neuroendocrine cells (< 0.01%) [13,30,31]. The availability of cell type-specific gene lists, together with the ability of RNA-Seq to quantify mRNA abundance, allows the contributions of these cell types to the SAE transcriptome to be assessed quantitatively. When compared to all genes of the SAE-enriched transcriptome, genes encoding cilia components were expressed at much higher levels and genes encoding neuroendocrine cell genes were expressed at much lower levels (Figure 4). On the other hand, genes identified as representative of the basal cells and secretory cells were expressed at levels comparable with the average level for SAE-enriched genes. The highest expressed cilia-related genes included tubulin β2C and α1A subunits, tektin and a number of dynein subunits with RPKM of > 100 (Table 2). By contrast, neuroendocrine genes such as secreogranin II (SCG2) and chromogranin A (CHGA) were expressed at < 1 RPKM with the exception of enolase 2, which may not be neuroendocrine cell-specific [32]. Among the mucus-secreting cell genes, trefoil factor 3 and two mucin genes, MUC1 and MUC5B, were the most highly expressed (RPKM > 100). For the basal cell genes, when assessed in the context of all SAE genes, MALAT1 (a noncoding transcript), CST3 cystatin C (a protease inhibitor) and PFN1 profilin 1 (a ubiquitous actin monomer-binding protein) were the most highly expressed.

To obtain insight into transcriptional regulation of the SAE of nonsmokers, the SAE-enriched transcriptome was surveyed for the most highly expressed transcription factors in various structural categories (Table 3). Among
the top 30 most highly expressed, the helix-turn-helix dominated, with the basic helix-loop-helix and β-scaffold categories next. Interestingly, the top 5 most highly expressed SAE-enriched transcription factors all have previously been established as having a role in airway biology and/or lung cancer, including FOXJ1, ELF3, TRIM29, SOX2, and FOXA1 [33-43]. RNA-Seq analysis also revealed high expression levels for a variety of pathway-specific transcription factors, including two (HES6, HEY1) related to notch signaling.

To quantify the expression of the receptors and ligands that may be involved in epithelial maintenance, the most highly expressed transmembrane receptors in different structural categories were identified (Table 4). Discoidin domain receptor tyrosine kinase 1 (DDR1), a collagen receptor associated with poor prognosis in non-small cell lung cancer, was the most highly expressed transmembrane receptor [44]. There were a significant number of G protein coupled, 7 transmembrane receptors in the highly expressed category, including a homophilic cadherin-coupled receptor (CELSR1) and the complement 5a receptor (C5aR). In addition to DDR1, there were also a number of highly expressed tyrosine kinase receptors among the top 30, including fibroblast growth factor receptors (FGFR3, FGFR2) and the insulin like growth factor 1 receptor (IGF1R). Interestingly, the receptors for oxytocin and natriuretic peptide were also expressed at a high level.

With respect to ligands and growth factors, the most highly expressed included multiple chemokines MDK,
CXCL1, CX3CL1 and CXCL6 (Table 5). No known ligand of the top 10 most highly expressed receptors was expressed at RPKM > 5 in the SAE-enriched transcriptome, nor was any known receptor for the top 10 highly expressed ligands expressed in the SAE-enriched transcriptome at RPKM > 5. This observation is of interest, as it suggests that much of the biology of the SAE relates to interactions of the SAE as target (receptors) or source (ligands) of external stimuli modulating SAE function or vice versa.

Gene Families
One advantage of RNA-Seq compared to microarray is that transcripts can be unequivocally mapped to a single member of a gene family when sequence is similar but not identical. Thus, RNA-Seq can be used to identify and quantify highly homologous genes, something not possible with hybridization-based microarrays [25]. To quantify SAE expression of homologous gene families in healthy nonsmokers, we identified all gene pairs with ≥ 90% sequence identity and assessed expression level by RNA-Seq in healthy nonsmokers. For example, in the cluster on chromosome 19 containing CYP2A6, CYP2A7 and CYP2A13 (Figure 5A), it was possible to map the reads to the different genes and show that expression of CYP2A13 (median RPKM = 17) > CYP2A6 (4) > CYP2A7 (1). In the GSTA cluster on chromosome 6 (Figure 5B), a clear assignment of reads permitted the expression order of GSTA1 (249) > GSTA2 (144) > GSTA3 (16) > GSTA5 (9). As another example, in the metallothionein gene cluster on chromosome 16 (Figure 5C), it was evident that expression level for MT1E (33) > MT1L (2) > MT1M (1). Among the other highly expressed homologous gene families in the SAE were the α and β tubulins, annexins, glutathione S-transferase mu family, cytosolic phenol-preferring sulfotransferase family, α amylase, and NODAL modulator (Table 6). In all cases, the RNA-Seq allows the individual transcripts to be definitively distributed among family members.

Effect of Smoking on the SAE Transcriptome
The preceding analyses of SAE-specific and ubiquitous transcripts are based exclusively on the RNA-Seq data from n = 5 nonsmokers. RNA-Seq data was also collected for n = 6 healthy smokers, who had a mean smoking history of 35 pack-yr (range of 26 to 45 pack-yr). Extensive transcription data based on the microarray methods has shown that smoking makes a substantial impact on gene expression in airway epithelium..
| Differentiated cell type | Gene symbol | Gene title | Expression level in SAE (RPKM)² |
|--------------------------|-------------|------------|-------------------------------|
| Ciliated                 | TUBB2C      | tubulin, beta 2C | 1161                          |
|                          | ACTG1       | actin, gamma 1   | 513                           |
|                          | TUBA1A      | tubulin, alpha 1a | 342                           |
|                          | HSPA1A      | heat shock 70 kDa protein 1A | 266                           |
|                          | HSPA1B      | heat shock 70 kDa protein 1A///heat shock 70 kDa protein 1B | 260                           |
|                          | TEKT2       | tektin 2 (testicular) | 146                           |
|                          | DYNLT1      | dynein, light chain, Tctex-type 1 | 131                           |
|                          | DNAI1       | dynein, axonemal, intermediate chain 1 | 108                           |
|                          | DNAI1I      | dynein, axonemal, light intermediate chain 1 | 108                           |
|                          | DNAI2       | dynein, axonemal, intermediate chain 2 | 105                           |
|                          | SPAG6       | sperm associated antigen 6 | 101                           |
|                          | DYNLRB2     | dynein, light chain, roadblock-type 2 | 100                           |
|                          | CROCC       | ciliary rootlet coiled-coil, rootletin | 84                            |
|                          | PPP2R1A     | protein phosphatase 2, regulatory subunit A, alpha | 73                            |
|                          | DNAH9       | dynein, axonemal, heavy chain 9 | 63                            |
|                          | CCDC146     | coiled-coil domain containing 146 | 60                            |
|                          | RSPH4A      | radial spoke head 4 homolog A (Chlamydomonas) | 59                            |
|                          | CALM3       | calmodulin 3 (phosphorylase kinase, delta) | 56                            |
|                          | TCTEX1D2    | Tctex1 domain containing 2 | 50                            |
|                          | IFT140      | intraflagellar transport 140 homolog (Chlamydomonas) | 49                            |
| Mucins and mucus components | AGR2       | anterior gradient homolog 2 | 166                           |
|                          | TFF3        | trefoil factor 3 (intestinal) | 149                           |
|                          | MUC1        | mucin 1, cell surface associated | 123                           |
|                          | MUC5B       | mucin 5B, oligomeric mucus/gel-forming | 118                           |
|                          | MUC4        | mucin 4, cell surface associated | 93                            |
|                          | MUC15       | mucin 15, cell surface associated | 28                            |
|                          | MUC20       | mucin 20, cell surface associated | 27                            |
|                          | MUC16       | mucin 16, cell surface associated | 20                            |
|                          | MUC13       | mucin 13, cell surface associated | 15                            |
|                          | MUC14L1     | mucin-like 1 | 3.90                          |
|                          | TFF1        | trefoil factor 1 | 209                           |
|                          | PARM1       | prostate androgen-regulated mucin-like protein 1 | 1.85                          |
|                          | EMR2        | egf-like module containing, mucin-like, hormone receptor-like 2 | 1.38                          |
|                          | GCNT3       | glucosaminyl (N-acetyl) transferase 3, mucin type | 0.81                          |
|                          | MUC2        | mucin 2, oligomeric mucus/gel-forming | 0.60                          |
|                          | MUC6        | mucin 6, oligomeric mucus/gel-forming | 0.58                          |
|                          | MUC12       | mucin 12, cell surface associated | 0.46                          |
| Basal                    | MALAT1      | metastasis associated lung adenocarcinoma transcript 1 (non-protein coding) | 461                           |
|                          | CST3        | cystatin C | 295                           |
|                          | PFN1        | profilin 1 | 224                           |
|                          | ALDOA       | aldolase A, fructose-bisphosphate | 183                           |
|                          | SQSTM1      | sequestosome 1 | 106                           |
|                          | MT2A        | metallothionein 2A | 91                            |
|                          | ENO1        | enolase 1, (alpha) | 89                            |
|                          | KRT7        | keratin 7 | 83                            |
|                          | MYL12A      | myosin, light chain 12A, regulatory, non-sarcomeric///myosin, light chain 12B, regulatory | 70                            |
|                          | FLNB        | filamin B, beta | 69                            |
|                          | BRI3        | brain protein I3///hypothetical protein LOC644975 | 62                            |
Quantitative comparison by RNA-Seq of SAE gene expression of healthy smokers vs healthy non-smokers showed there was no gross effect of smoking on the overall distribution of the SAE transcriptome in nonsmokers and smokers (Figure 6A). However, there were changes in expression of individual genes with smoking, constituting 8 to 13% of the ubiquitous genes and 9 to 14% of the SAE-enriched genes (Figure 6B). In both categories, smoking responsiveness was slightly more noticeable in genes with medium and high expression than in genes with low expression.

To assess the quantitative effects of smoking, a modified volcano plot was devised in which the absolute change was plotted as a function of p value (Figure 6C, D). The data show that, for both the ubiquitous and for the SAE-enriched transcriptome, the number of genes down-regulated by smoking was substantially higher than that number of genes up-regulated by smoking. This was particularly noticeable among the ubiquitous genes.

To assign function to smoking dependent genes, Gene Ontology searches of the Biological Process term were used to classify the expression of the smoking-suppressed and smoking-repressed genes (Figure 8). This analysis also showed that, in almost all categories, the expression of more genes was suppressed than induced. Of interest, a comparison of function in the ubiquitous and SAE-enriched categories pointed to some contrasts of potential significance. For example, among genes involved in transport, there were more smoking-induced SAE-enriched genes than ubiquitous genes and also a higher proportion of smoking-inducible genes in the

| Neuro- endocrine |
|-----------------|
| ENO2 enolase 2 (gamma, neuronal) | 23 |
| GRP gastrin-releasing peptide | 0.82 |
| UCHL1 ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase) | 0.65 |
| SCG2 secretogranin II | 0.48 |
| ASCL1 achaete-scute complex homolog 1 (Drosophila) | 0.34 |
| CHGA chromogranin A (parathyroid secretory protein 1) | 0.32 |

1 List includes genes known to be enriched in expression in ciliated cells (Dvorak et al. 2010) [52], secretory cells (consisting of all mucins and mucin components), basal cells (Hackett et al. 2011) [30], and neuroendocrine cells (Carolan et al. 2008) [31]. The small airway epithelium expression level was determined and the top 20 highly expressed (or all genes expressed from mucin list and neuroendocrine list, i.e. above threshold of 0.125) were tabulated in descending order of expression level.

2 Median for n = 5 nonsmokers.
SAE-enriched group compared to ubiquitous category. Similarly, among proteases and anti-proteases, there was some smoking-inducibility in the SAE-enriched genes, but none for ubiquitous genes.

In contrast to microarray data that suggests that cytochrome P450 genes and oxidant-related genes are those most highly induced by smoking [11-18,45,47], quantitative RNA-Seq analysis demonstrated the largest up-regulation of a gene by smoking was \( b \)-microseminoprotein (MSMB) and chromosome 20 open reading frame 114 (C20orf114, Table 7; see Additional file 1, Figure S1 for examples of this and other RNA-Seq-identified smoking-related genes). The most smoking-repressed genes were SCGB1A1 and SCGB3A1, which were also the two most highly expressed SAE genes in nonsmokers. The smoking-induced down-regulation of SCGB1A1 gene expression was dramatic, with an absolute decrease in median RPKM from 38,675 (13.1% of total mRNAs) to 17,244 (6.5% of mRNAs).

To detect novel smoking-dependent genes, we exploited the ability of RNA-Seq to quantify expression of genes with low expression. Among the novel smoking-inducible and smoking-suppressed genes with low level expression were the smoking-inducible genes that had been previously identified by microarray (e.g., AKR1B10, CYP1B1) [11,47], but also newly identified smoking-induced genes, such as the oxido-reductase AKR1B15 and transcription factor TPRXL (Table 7). Similarly, new smoking-repressed genes were identified, including transcription factor PAX1 and AZU1, an inflammatory mediator [48].

Based on the number of SAE-enriched, smoking-dependent transport genes (Figure 7), we examined the expression of ion transport genes with low overall

| Table 3 Highly Expressed SAE-enriched Transcription Factors$^1$ |
|---------------|-----------------|-----------------|
| Category      | Gene symbol     | Gene title      | Median expression level (RPKM) |
| Basic helix-loop -helix | ATF6B           | activating transcription factor 6 beta | 264 |
| BHLHE40       | basic helix-loop-helix family, member e40 | 16.2 |
| RFX3          | regulatory factor X, 3 (influences HLA class II expression) | 15.6 |
| HES6          | hairy and enhancer of split 6 (Drosophila) | 13.6 |
| FOXC1         | forkhead box C1 | 9.3 |
| CEBPA         | CCAAT/enhancer binding protein (C/EBP), alpha | 9.1 |
| HEY1          | hairy/enhancer-of-split related with YRPW motif 1 | 9.1 |
| Zinc finger   | TRIM29          | tripartite motif-containing 29 | 77.3 |
| KLF5          | kruppel-like factor 5 (intestinal) | 42.0 |
| RREB1         | ras responsive element binding protein 1 | 7.2 |
| KLF4          | kruppel-like factor 4 (gut) | 7.1 |
| Helix-turn-helix | FOXJ1           | forkhead box J1 | 363.1 |
| ELF3          | E74-like factor 3 (ets domain transcription factor, epithelial-specific) | 170.1 |
| FOXA1         | forkhead box A1 | 62.6 |
| EHF           | ets homologous factor | 39.1 |
| TBX1          | T-box 1         | 19.7 |
| SATB1         | SATB homeobox 1 | 19.3 |
| MYB           | v-myb myeloblastosis viral oncogene homolog (avian) | 15.9 |
| SIX2          | SIX homeobox 2  | 13.9 |
| NKX2-1        | NK2 homeobox 1  | 11.6 |
| PHTF1         | putative homeodomain transcription factor 1 | 10.9 |
| TEAD3         | TEA domain family member 3 | 10.9 |
| ETV6          | ets variant 6   | 10.6 |
| FOXA2         | forkhead box A2 | 6.9 |
| $\beta$-scaffold | SOX2            | SRY (sex determining region Y)-box 2 | 71.3 |
| RUNX1         | runt-related transcription factor 1 | 16.7 |
| SOX4          | SRY (sex determining region Y)-box 4 | 11.4 |
| TFCP2         | transcription factor CP2 | 10.7 |
| SOX21         | SRY (sex determining region Y)-box 21 | 9.0 |
| NFATC1        | nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 1 | 7.8 |
| SOX9          | SRY (sex determining region Y)-box 9 | 7.7 |

$^1$ The top 30 DNA-binding transcription factors were identified in the highly expressed SAE-enriched list for \( n = 5 \) healthy nonsmokers. They were categorized by transcription factor family and sorted in descending order of expression within each category.
The significance of this gene group is evident in the fact that polymorphisms in the cystic fibrosis transmembrane conductance regulator (CFTR) gene, a chloride transporter, result in cystic fibrosis, a lethal hereditary disorder with a dramatic pulmonary phenotype [49]. CFTR expression levels were in the range of 2 to 4 RPKM corresponding to an average of ~2 to 4 mRNA molecules per cell, a similar value to that estimated by polymerase chain reaction methodology [50]. There was no difference in CFTR expression level between nonsmokers and smokers (Figure 9A). In contrast, there were smoking-inducible transporters including the CFTR related ATP-binding cassette, subfamily C, member 3 (ABCC3), L-type calcium channel, voltage-dependent, gamma subunit 4 (CACNG4), and cyclic nucleotide gated channel, beta 1 (CNGB1, Figure 9B-D). There were also significantly smoking-suppressed ion transporters, including solute carrier family 13, member 2 (sodium dependent dicarboxylate transporter, SLC13A2) and potassium voltage-gated channel, Shaw-related subfamily, member 4 (KCNC4) detected (Figure 9E, F).

### Effect of Smoking on Alternative Splicing

Among the advantages of RNA-Seq is the ability to easily quantify different isoforms of one gene generated by alternate splicing. The frequencies of reads that span one or more splice junction were used to assess the relative levels of different isoforms separately in smokers and nonsmokers. Interestingly, comparison of the

### Table 4 Highly Expressed SAE-enriched Transmembrane Receptors

| Category                        | Gene symbol | Gene title                                           | Median expression level (RPKM) |
|---------------------------------|-------------|-----------------------------------------------------|-------------------------------|
| G protein coupled/7 transmembrane | CELSR1      | cadherin, EGF LAG seven-pass G-type receptor 1 (flamingo homolog, Drosophila) | 63.0                          |
|                                 | CSAR1       | complement component 5a receptor 1                  | 27.6                          |
|                                 | GPR110      | G protein-coupled receptor 110                       | 21.7                          |
|                                 | GPRC5C      | G protein-coupled receptor, family C, group 5, member C | 20.3                          |
|                                 | OXTR        | oxytocin receptor                                    | 19.3                          |
|                                 | LPAR3       | lysophosphatidic acid receptor 3                     | 17.6                          |
|                                 | FZD6        | frizzled homolog 6 (Drosophila)                      | 12.5                          |
|                                 | PTGER4      | prostaglandin E receptor 4 (subtype EP4)             | 9.2                           |
|                                 | VIPR1       | vasoactive intestinal peptide receptor 1             | 8.8                           |
|                                 | ADRB1       | adrenergic, beta-1, receptor                         | 7.8                           |
|                                 | GPR116      | G protein-coupled receptor 116                       | 7.5                           |
|                                 | FZD8        | frizzled homolog 8 (Drosophila)                      | 7.1                           |
|                                 | ADRA2A      | adrenergic, alpha-2A, receptor                       | 6.8                           |
|                                 | PTGFR       | prostaglandin F receptor (FP)                        | 6.6                           |
| Cyclase related                 | NRP2        | neuropilin 2                                         | 12.0                          |
|                                 | NPR2        | natriuretic peptide receptor B/guanylate cyclase B (atrinatriuretic peptide receptor B) | 10.7                          |
|                                 | CRCP        | CGRP receptor component                              | 10.4                          |
| IgG like                        | SCARA3      | scavenger receptor class A, member 3                 | 21.8                          |
|                                 | PTPTRT      | protein tyrosine phosphatase, receptor type, T       | 13.0                          |
|                                 | IL1R1       | interleukin 1 receptor, type 1                       | 6.5                           |
| Ion channel                     | GABRP       | gamma-aminobutyric acid (GABA) A receptor, pi       | 15.7                          |
| Serine kinase                   | TGFB2       | transforming growth factor, beta receptor II (70/80 kDa) | 8.8                           |
| Tyrosine kinase                 | DDR1        | discoid domain receptor tyrosine kinase 1            | 140.3                         |
|                                 | FGF3        | fibroblast growth factor receptor 3                  | 19.9                          |
|                                 | IGF1R       | insulin-like growth factor 1 receptor                | 15.2                          |
|                                 | PTK7        | PTK7 protein tyrosine kinase 7                       | 14.5                          |
|                                 | FGF2        | fibroblast growth factor receptor 2                  | 11.6                          |
|                                 | MET         | met proto-oncogene (hepatocyte growth factor receptor) | 8.5                           |
|                                 | EGFR        | epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian) | 7.2                           |
| Other                           | SORL1       | sortilin-related receptor, L(DLR class) A repeats-containing | 7.7                           |

1 The top 30 transmembrane receptors were identified in the highly expressed SAE-enriched list for n = 5 healthy nonsmokers. They were categorized by structural family and sorted in descending order of expression within each category.
splicing events between nonsmokers and smokers to the expected distribution revealed no divergence (Figure 10). This was true for both the ubiquitous and SAE-enriched genes, suggesting there are no major smoking-dependent differences in splicing patterns.

Discussion

The small airway epithelium, the cell population lining the bronchial tree ≥ 6 generations, plays a central role in normal lung function and in the pathogenesis of many lung disorders [7]. Among the most common SAE-associated diseases are those caused by cigarette smoking, including chronic obstructive pulmonary disease (COPD) and lung adenocarcinoma. The development of massive parallel RNA sequencing (RNA-Seq) technology permits quantitative assessment of poly(A)* mRNA levels to a high degree of sensitivity [19-24]. Compared to hybridization-based methodologies of transcriptome analysis, RNA-Seq has low background, broad dynamic range and high specificity [22]. Using this approach, we have built upon the body of microarray-generated data to provide quantitative characterization of the transcriptome of the normal healthy human SAE and characterize how it changes with smoking [11-18]. By comparing the SAE RNA-Seq data to that of other tissues and organs, the present study grouped the SAE transcriptome into 2 categories: (1) ubiquitous genes, i.e., SAE genes shared with other organs and tissues, and (2) SAE-enriched genes, i.e., those expressed by the SAE, but not in the majority of other organs and tissues. Using this classification, and based on the capacity of RNA-Seq to provide quantification of mRNA, we further characterized the effect of smoking on the SAE transcriptome.

**Table 5** Highly Expressed SAE-enriched Signaling Ligands and Growth Factors

| Gene symbol | Gene title                                      | Median expression level (RPKM) |
|-------------|-------------------------------------------------|-------------------------------|
| MDK         | midkine (neurite growth-promoting factor 2)     | 598                           |
| CXCL1       | chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha) | 49.3                          |
| CX3CL1      | chemokine (C-X-C motif) ligand 1                | 32.9                          |
| TNFSF10     | tumor necrosis factor (ligand) superfamily, member 10 | 31.8                          |
| FSTL1       | follistatin-like 1                              | 27.9                          |
| CXCL6       | chemokine (C-X-C motif) ligand 6 (granulocyte chemotactic protein 2) | 18.6                          |
| FGF14       | fibroblast growth factor 14                     | 16.5                          |
| DLL1        | delta-like 1 (Drosophila)                       | 13.6                          |
| JAG2        | jagged 2                                        | 11.9                          |
| IL8         | interleukin 8                                   | 8.6                           |
| CCL15       | C-C motif chemokine 15                          | 7.8                           |
| PDGFA       | platelet-derived growth factor alpha polypeptide | 6.1                           |
| TNFSF12     | tumor necrosis factor (ligand) superfamily, member 12 | 5.9                           |
| CCL23       | chemokine (C-C motif) ligand 23                 | 5.3                           |
| NMB         | neuromedin B                                    | 4.9                           |
| CCL5        | chemokine (C-C motif) ligand 5                  | 4.9                           |
| NPFF        | neuropeptide FF-amide peptide precursor         | 4.6                           |
| FAS         | fas (TNF receptor superfamily, member 6)        | 3.6                           |
| WIF1        | WNT inhibitory factor 1                         | 3.6                           |
| CCL18       | chemokine (C-C motif) ligand 18 (pulmonary and activation-regulated) | 3.5                           |
| LT8         | lymphotoxin beta (TNF superfamily, member 3)   | 3.0                           |
| LIF         | leukemia inhibitory factor (cholinergic differentiation factor) | 2.9                           |
| ERBB4       | v-erb-a erythoblastic leukemia viral oncogene homolog 4 (avian) | 2.7                           |
| PTCH1       | patched homolog 1 (Drosophila)                  | 2.6                           |
| WNTSA       | wingless-type MMTV integration site family, member 5A | 2.1                           |
| CCL17       | chemokine (C-C motif) ligand 17                 | 1.7                           |
| NRTN        | neurturin                                       | 1.7                           |
| AREG        | amphiregulin                                    | 1.4                           |
| CCL4        | chemokine (C-C motif) ligand 4                  | 1.3                           |
| NTS         | neurotensin                                     | 1.2                           |

¹ The top 30 signaling ligands and growth factors were identified in the highly expressed SAE-enriched list for n = 5 healthy nonsmokers.
A. CYP2A6, 2A7 and 2A13
Chr 19

B. GSTA1, A2, A3 and A5
Chr 6

C. MT1E, 1L and 1M
Chr 16

Figure 5 Examples of RNA-Seq quantification of small airway epithelium (SAE) expression levels of genes within gene families of ≥ 90% homology. To identify gene families expressed by the SAE, the % identity between gene pairs expressed by the healthy nonsmoker SAE was determined using BLAST, where each gene was blasted against a database of all human RefSeq mRNA [26]. Gene families were defined as genes for which the alignments yielded ≥ 90% identity and the alignment length was at least 50% of both sequences. A. CYP2A6, CYP2A7 and CYP2A13, B. GSTA1, GSTA2, GSTA3 and GSTA5, and C. MT1E, MT1L and MT1M.
| Category                        | Gene name                        | Gene symbol | Median expression level (RPKM) |
|---------------------------------|----------------------------------|-------------|--------------------------------|
| Cilia                           | tubulin, alpha                   | TUBA1A      | 341.7                          |
|                                 |                                  | TUBA1B      | 162.6                          |
|                                 |                                  | TUBA1C      | 120.5                          |
|                                 | tubulin, beta                    | TUBB2A      | 38.7                           |
|                                 |                                  | TUBB2B      | 23.3                           |
|                                 |                                  | TUBB4       | 14.8                           |
| Annexin (signal transduction)   | annexin A2                       | ANXA2       | 362.1                          |
|                                 |                                  | ANXA2P2     | 292.3                          |
|                                 |                                  | ANXA2P1     | 42.6                           |
|                                 |                                  | ANXA2P3     | 26.1                           |
| Glutathione S-transferase alpha | glutathione S-transferase alpha  | GSTA1       | 248.9                          |
|                                 |                                  | GSTA2       | 143.9                          |
|                                 |                                  | GSTA3       | 15.9                           |
|                                 |                                  | GSTA5       | 9.3                            |
| Glutathione S-transferase mu    | glutathione S-transferase mu     | GSTM2       | 35.9                           |
|                                 |                                  | GSTM1       | 20.8                           |
|                                 |                                  | GSTM4       | 8.8                            |
| Sulfotransferase - phenol preferring | sulfotransferase family, cytosolic, 1A, phenol-preferring | SULT1A1     | 25.7                           |
|                                 |                                  | SULT1A4     | 14.6                           |
|                                 |                                  | SULT1A3     | 10.3                           |
|                                 |                                  | SULT1A2     | 5.4                            |
| Amylase                         | amylase, alpha                   | AMY1A       | 43.4                           |
|                                 |                                  | AMY1B       | 18.6                           |
|                                 |                                  | AMY1C       | 15.1                           |
|                                 |                                  | AMY2B       | 13.8                           |
|                                 |                                  | AMY2A       | 7.3                            |
| Polarity/left right signaling   | NODAL modulator                  | NOMO2       | 22.4                           |
|                                 |                                  | NOMO1       | 16.0                           |
|                                 |                                  | NOMO3       | 14.3                           |
| Metallothionein                 | metallothionein                  | MT1E        | 33.2                           |
|                                 |                                  | MT1L        | 2.4                            |
|                                 |                                  | MT1M        | 1.1                            |
|                                 |                                  | MT1JP       |                                |
| P450                            | cytochrome P450, family 2, subfamily A, polypeptides | CYP2A13     | 17.2                           |
|                                 |                                  | CYP2A6      | 4.2                            |
|                                 |                                  | CYP2A7      | 1.4                            |
| Aldo-keto reductase             | aldo-keto reductase family 7     | AKR7A2      | 34.1                           |
|                                 |                                  | AKR7A3      | 3.0                            |
|                                 |                                  | AKR7L       | 1.5                            |
|                                 |                                  | AKR7A2P1    |                                |
| Aldehyde dehydrogenase          | alcohol dehydrogenase            | ADH1C       | 52.2                           |
|                                 |                                  | ADH1B       | 4.3                            |
|                                 |                                  | ADH1A       | 2.6                            |
| Short chain dehydrogenase       | dehydrogenase/reductase          | DHR59       | 68.4                           |
Table 6 Different Expression Levels Among Members of Common Gene Families Expressed in the SAE\(^1\) (Continued)

| Gene     | Expression Level |
|----------|------------------|
| MUC20    | 27.5             |
| VPS53    | 4.0              |
| SMU1     | 3.9              |
| FAM153B  | 0.4              |
| LEP      | 0.0              |

\(^1\) List of gene families identified using BLAST with ≥ 90% identity and alignment lengths of at least 50% in both sequences. They were categorized by gene family and sorted in descending order within each category by median expression level for \(n = 5\) nonsmokers.

Figure 6 Overall impact of smoking on small airway epithelium gene expression. Shown are comparisons of RNA-Seq assessment of genes expressed in the small airway epithelium (SAE) in nonsmokers (\(n = 5\)) vs smokers (\(n = 6\)). A. Cumulative frequency of expression levels as a function of increasing RPKM. The data is shown as cumulative frequency in 1/2 log\(_{10}\) bins starting at the lower limit (RPKM 0.125, log\(_{10}\) = -0.9) for healthy nonsmokers (blue) and healthy smokers (red). On an overall basis assessing all genes, there is no difference in the nonsmokers vs smokers. B. Comparison of expression of the subset of smoking-responsive vs non-responsive genes for the ubiquitous and SAE-enriched genes. Each category is divided into low, medium and high expression groups using the same criteria as in Figures 2, 3, with smoking-responsive genes defined as \(p < 0.05\). Ordinate - number of genes; abscissa - smoking responsive (red) and smoking non-responsive (blue) for ubiquitous and small airway epithelium (SAE)-enriched genes. Note that for both ubiquitous and SAE-enriched genes, only a small fraction, and approximately the same proportion (8-14%; low, medium, high), are smoking-responsive. C, D. Modified volcano plot showing absolute change in expression level (RPKM smoker - RPKM nonsmoker) vs -log \(p\) value for ubiquitous and SAE-enriched genes. C. Ubiquitous genes. D. SAE-enriched genes. Note that for both ubiquitous and SAE-enriched genes, more genes are down-regulated by smoking than up-regulated.
Figure 7 Comparison of smoking dependent genes observed by microarray and RNA-Seq. 

A. Microarray-determined smoking-responsive genes. The data includes all significant genes (Benjamini Hochberg corrected p value < 0.05; Additional file 1, Table S6) with > 1.5-fold different in mean expression level between n = 12 smokers and n = 12 nonsmokers, as determined by microarray. For each probeset the corresponding genes was assessed by RNA-Seq for n = 5 nonsmokers an n = 6 smokers and fold-change by microarray is plotted against the fold-change by RNA-Seq.

B. RNA-Seq-determined smoking-responsive genes. The data includes the fold-change of all genes significantly impacted by smoking (uncorrected p < 0.005, 1.5-fold-change cut off), as assessed by RNA-Seq for n = 5 nonsmokers an n = 6 smokers. n = 12 nonsmokers and n = 12 healthy smokers were assessed by microarray and the fold-change for RNA-Seq is plotted against the fold-change for the probeset with largest change.
The smoking-responsive genes (p < 0.05) of the ubiquitous and SAE-enriched groups were assigned function based on Gene Ontology classification and searches of NCBI databases. The 9 functional categories chosen accounted for the largest fraction of the genes that could be assigned functional categories. For each category, the genes were divided into smoking-induced and smoking-repressed and the log10 of absolute change (RPKM smoker - RPKM nonsmoker) was plotted. Genes are divided by category and ordered within each category by decreasing change in expression level. Red - down-regulated; blue - up-regulated. Note that for both ubiquitous and SAE-enriched genes, a higher fraction of genes in most categories are down-regulated, and that most up-regulated genes are in the SAE-enriched subgroup.
### Table 7: Small Airway Epithelium Expressed Genes Most Affected by Smoking

| Category       | Gene Symbol | Gene Title                                      | RNA-Seq | Micorarray |
|----------------|-------------|------------------------------------------------|---------|------------|
|                |             | Non-smoker median | Smoker median | Absolute difference | Fold-change | Fold-change | p-value |
| Largest absolute increase |             |                   |               |                  |            |            |         |
| Ubiquitous     | FTL         | ferritin, light polypeptide | 371.1   | 843.4 | 472.2 | 2.3 | 1.6 | 0.1784 |
|                | PRDX1       | peroxiredoxin 1   | 187     | 398.6 | 211.6 | 2.1 | 1.6 | 0.1083 |
|                | FTH1        | ferritin, heavy polypeptide 1 | 349.2   | 551.5 | 202.3 | 1.6 | 1.8 | 0.0602 |
| SAE-enriched   | TUBB2C      | tubulin, beta 2C  | 1161.3  | 1331.9 | 170.5 | 1.1 | 1.1 | 0.8995 |
|                | CLU         | clustatin       | 498     | 665.2 | 167.3 | 1.3 | 1.5 | 0.2157 |
|                | NQO1        | NAD(P)H dehydrogenase, quinone 1 | 38.3  | 198.7 | 160.4 | 5.2 | 4.8 | <0.0001 |
|                | UBB         | ubiquitin B      | 615.1   | 711.1 | 96 | 1.2 | -1.0 | 0.9760 |
|                | GSN         | gelsolin        | 93.1    | 174.9 | 81.8 | 1.9 | 1.4 | 0.4772 |
|                | TUBA1A      | tubulin, alpha 1a | 341.7   | 423.2 | 81.5 | 1.2 | 1.1 | 0.7340 |
|                | ALDOA       | aldolase A, fructose-bisphosphate | 182.9  | 263.4 | 80.5 | 1.4 | 1.5 | 0.1670 |
|                | MSMB        | microseminoprotein, beta | 333.1  | 3112.7 | 2779.6 | 9.3 | 2.1 | 0.0820 |
|                | C20orf114   | chromosome 20 open reading frame 114 | 1484.3 | 4102.7 | 2618.5 | 9.3 | 2.1 | 0.6685 |
|                | ALDH3A1     | aldehyde dehydrogenase 3 family, member A1 | 226.9  | 2077.9 | 1851 | 9.2 | 9.8 | <0.0001 |
|                | TFF3        | trefoil factor 3 (intestinal) | 1494    | 697.9 | 596.8 | 4.7 | 2.7 | 0.1771 |
|                | WFDC2       | WAP four-disulfide core domain 2 | 840.3  | 1327 | 486.7 | 1.6 | 1.3 | 0.5735 |
|                | TPP3        | tubulin polymerization-promoting protein family member 3 | 1301.7 | 1604.9 | 303.2 | 1.2 | -1.0 | 0.9493 |
|                | TSPAN1      | tetraspanin 1    | 663.9   | 960.2 | 296.6 | 1.4 | 1.1 | 0.8884 |
|                | S100P       | S100 calcium binding protein P | 79.6   | 291.2 | 211.5 | 3.7 | 1.8 | 0.4832 |
|                | GSTA2       | glutathione S-transferase alpha 2 | 143.9  | 373.5 | 236.2 | 1.6 | 1.3 | 0.3080 |
|                | PLUNC       | palate, lung and nasal epithelium associated | 5.4   | 186 | 180.6 | 34.5 | 1.6 | 0.8375 |
| Largest absolute decrease |             |                   |               |                  |            |            |         |
| Ubiquitous     | CRIP1       | cysteine-rich protein 1 (intestinal) | 1014.1  | 587.7 | -426.3 | -1.7 | -1.4 | 0.4842 |
|                | RPLP1       | ribosomal protein, large, P1 | 916.1   | 516 | -400.1 | -1.8 | -1.2 | 0.3845 |
|                | CAPS        | calyciphosine    | 2197.4  | 1994.7 | -202.8 | -1.1 | -1.5 | 0.1733 |
|                | PRDX5       | peroxiredoxin 5 | 1022.4  | 823.3 | -199.1 | -1.2 | -1.2 | 0.3653 |
|                | RPS11       | ribosomal protein S11 | 537.2  | 357.5 | -179.7 | -1.5 | -1.3 | 0.3209 |
|                | RPLP2       | ribosomal protein, large, P2 | 347.5  | 181.2 | -166.4 | -1.9 | -1.2 | 0.6557 |
|                | RPL8        | ribosomal protein L8 | 631.5  | 468.9 | -162.6 | -1.3 | -1.2 | 0.5028 |
|                | TPT1        | tumor protein, translationally-controlled 1 | 655.2  | 500.7 | -154.5 | -1.3 | -1.1 | 0.6701 |
|                | S100A6      | S100 calcium binding protein A6 | 758.3  | 617.2 | -141 | -1.2 | -1.1 | 0.9250 |
|                | CD81        | CD81 molecule | 248.3   | 120.5 | -127.8 | -2.1 | -1.3 | 0.3499 |
| SAE-enriched   | SCGB1A1     | secretoglobin, family 1A, member 1 (uteroglobin) | 38675.4 | 17244 | -21431.5 | -2.2 | -1.1 | 0.4670 |
|                | SCGB3A1     | secretoglobin, family 3A, member 1 | 7838.2  | 2947.3 | -4890.8 | -2.7 | -1.3 | 0.1509 |
|                | CD74        | CD74 molecule, major histocompatibility complex, class II invariant chain | 947.2  | 723.7 | -223.5 | -1.3 | -2.1 | 0.2881 |
|                | C9orf24     | chromosome 9 open reading frame 24 | 628.7  | 488 | -140.7 | -1.3 | -1.3 | 0.2256 |
|                | CYP4B1      | cytochrome P450, family 4, subfamily B, polypeptide 1 | 373.5  | 259.3 | -114.2 | -1.4 | -1.6 | 0.1826 |
|                | C2orf85     | chromosome 20 open reading frame 85 | 738.1  | 625.3 | -112.8 | -1.2 | -1.2 | 0.4756 |
|                | KRT19       | keratin 19 | 493.4   | 396.8 | -96.6 | -1.2 | -1.1 | 0.7722 |
|                | RPS18       | ribosomal protein S18 | 285   | 207.8 | -77.3 | -1.4 | -1.2 | 0.3731 |
|                | ALDH3B1     | aldehyde dehydrogenase 3 family, member B1 | 341.9  | 265.1 | -76.7 | -1.3 | -1.5 | 0.3595 |
|                | TMEM190     | transmembrane protein 190 | 945.4  | 868.8 | -76.6 | -1.1 | -1.2 | 0.5710 |

1. Hackett et al. BMC Genomics 2012, 13:82
2. http://www.biomedcentral.com/1471-2164/13/82
(SAGE) allows the identification of ubiquitous and tissue specific genes [24,51]. By comparing to the RNA-Seq data obtained for other organs and tissues, we found that among 15,877 genes expressed in the SAE, 52% of genes are enriched in the SAE in a relatively selective manner and 48% of genes are ubiquitous. Interestingly, the SAE transcriptome includes more tissue-characteristic RNAs than other epithelial (breast, kidney, colon) and non-epithelial (heart, brain, skeletal muscle, adipose tissue, lymph node) tissues, where ubiquitous genes contribute to 65 to 85% of the transcriptome [24]. This may reflect the high purity of the epithelial cells obtained by bronchial brushing (i.e. they are not contaminated by endothelium, connective tissue or inflammatory cells and therefore do not appear to express genes that are expressed by contaminating cell types). Notably, SAE genes with low expression levels contributed to 36% of the SAE-enriched and only 2% of ubiquitous genes, indicating that molecular uniqueness of the SAE is determined to a considerable degree by the transcripts with a low abundance. From the functional perspective, ubiquitous SAE genes dominated in the categories related to housekeeping biologic processes such as translation and transcription, whereas SAE-enriched genes were prevalent in more specific categories such as immunity, signal transduction, adhesion, and ion transport.

### Table 7 Small Airway Epithelium Expressed Genes Most Affected by Smoking

| Novel genes up-regulated by smoking | Ubiquitous | SAE-enriched |
|------------------------------------|------------|--------------|
| Ubiquitous AHRR aryl-hydrocarbon receptor repressor | 0.1 | 1.3 | 1.2 | 20.8 | 3.7 | 0.0054 |
| Ubiquitous AKR1B10 aldo-keto reductase family 1, member B10 (aldose reductase) | 0.3 | 28.5 | 28.2 | 94.8 | 56.6 | < 0.0001 |
| Ubiquitous CABYR calcium binding tyrosine-(Y)-phosphorylation regulated | 1 | 12.5 | 11.6 | 12.7 | 9.4 | < 0.0001 |
| Ubiquitous SPP1 secreted phosphoprotein 1 | 0.8 | 10.6 | 9.8 | 12.9 | 8.5 | 0.0021 |
| Ubiquitous CYP1B1 cytochrome P450, family 1, subfamily B, polypeptide 1 | 0.2 | 9.2 | 9 | 43.3 | 55.0 | < 0.0001 |
| Ubiquitous AKR1B15 aldo-keto reductase family 1, member B15 | 0.1 | 6.5 | 6.3 | 50 | Not in U133 |
| Ubiquitous B3GNT6 UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 6 (core 3 synthase) | 0.3 | 4.4 | 4.1 | 16 | 4.5 | 0.0025 |
| Ubiquitous NOS3 nitric oxide synthase 3 (endothelial cell) | 0.6 | 4 | 3.4 | 6.5 | "P" < 25% |
| Ubiquitous TPRXL tetra-peptide repeat homeobox-like | 0.3 | 3.2 | 2.9 | 10.2 | 3.8 | 0.0761 |
| Ubiquitous SFRP2 secreted frizzled-related protein 2 | 0.3 | 2.7 | 2.4 | 9.8 | 9.7 | 0.0071 |
| Ubiquitous FAM177B family with sequence similarity 177, member B | 0.2 | 2.5 | 2.3 | 11 | Not in U133 |

| Low level expressed genes suppressed by smoking |
|-----------------------------------------------|
| Ubiquitous PANK1 pantothenate kinase 1 | 1.4 | 0.7 | -0.7 | -2.1 | -1.9 | 0.0391 |
| Ubiquitous LYPD2 LY6/PLAUR domain containing 2 | 23.9 | 1.7 | -22.3 | -14.5 | Not in U133 |
| Ubiquitous LYNX1 Ly6/neurotoxin 1 | 8.6 | 2 | -6.6 | -4.3 | -1.8 | 0.3058 |
| Ubiquitous AZU1 azurocidin 1 | 6.1 | 1.9 | -4.3 | -3.3 | -2.1 | 0.2405 |
| Ubiquitous ITM2A integral membrane protein 2A | 5.2 | 1.3 | -4 | -2.6 | 0.0071 |
| Ubiquitous ITM2A integral membrane protein 2A | 5.2 | 1.3 | -4 | -2.6 | 0.0071 |
| Ubiquitous SAA4 serum amyloid A4, constitutive | 4.8 | 1.4 | -3.4 | -3.4 | -3.4 | 0.0147 |
| Ubiquitous GAL3ST2 galactose-3-O-sulfotransferase 2 | 4.7 | 1.6 | -3.1 | -2.9 | "P" < 25% |
| Ubiquitous NEU4 sialidase 4 | 3.7 | 0.7 | -3 | -5.4 | -1.2 | 0.6483 |
| Ubiquitous PANX1 paired box 1 | 3.6 | 1 | -2.6 | -3.5 | -1.9 | 0.3312 |
| Ubiquitous ERp27 endoplasmic reticulum protein 27 | 2.4 | 0.5 | -2 | -5.2 | -4.4 | 0.0016 |

1 The top 20 genes were identified in four categories of genes most affected by smoking; largest absolute increase, largest absolute decrease, novel genes up-regulated by smoking and low level expressed genes suppressed by smoking. They were sorted in descending order of absolute difference in expression within each category and identified as being from the ubiquitous or small airway epithelium-enriched groups.

2 Absolute difference = smoker median - nonsmoker median

3 Fold change = mean in smokers/mean in non smokers

4 Benjamini-Hochberg corrected p value. Where the gene is not represented on the microarray or the Affymetrix Present call "P" is less than 25% of subjects, it is so indicated.
Figure 9 RNA-Seq quantification of examples of smoking-dependent changes in expression of ion channel-related genes expressed in the small airway epithelium. Smoking responsiveness of selected ion channels. A. CFTR - unchanged; B-D. Up-regulated. B. ABCC3; C. CACNG4; and D. CNGB1. E-F. Down-regulated. E. SLC13A2; and F. KCNC4. In all panels, each data point represents one individual.
expressed in distinct cell populations typical for each tissue. The human SAE is composed of various cell types, including ciliated, secretory (mostly Clara cells but also surface epithelium mucus-producing cells), basal, undifferentiated columnar, and rare neuroendocrine cells [1,2,13,31,52]. Although most of the SAE-enriched genes are represented by low expressed transcripts, the top 30 highly expressed SAE-enriched genes accounted for about 20% of the total SAE mRNA, suggesting that a limited number of genes may dictate the specific pattern of biological processes dominating in the SAE under steady-state conditions. Detailed analysis of the most highly expressed SAE-enriched genes revealed a unique pattern of epithelial differentiation and molecular functions.

Secretory differentiation
Genes related to secretory epithelial differentiation dominated the most highly expressed SAE-enriched genes. Of the total SAE transcripts identified, 13% mapped to the secretoglobin SCGB1A1. The high level of expression of SCGB1A1 is expected in the SAE, where Clara cells are enriched and play an important role in the pulmonary host defense [27-29]. SCGB1A1, is involved in regulation of critical processes in the distal airways such as protection against oxidative stress, maintenance of the normal airway lining fluid homeostasis, regulation of inflammation and airway reactivity during respiratory infection, and control of macrophage activation in the lung [53-57]. Another secretoglobin, SCGB3A1, originally called HIN-1, was the second-highest expressed SAE-enriched gene. Previous studies have identified the lungs as major site of SCGB3A1 in humans [58]. Expression of SCGB3A1 is induced during epithelial differentiation and restricted to terminally differentiated airway epithelial cells and down-regulated in cancer [58,59]. There is evidence that SCGB3A1 is also produced by Clara cells [60] and exerts growth-inhibitory activities [61]. Consistent with its putative tumor-suppressor function, SCGB3A1, is aberrantly methylated in various types of cancer, including lung carcinomas [62]. Based on previous observations, the quantitative data in the present study suggests that SCGB3A1 could be a major steady-state tumor-suppressor gene in the human SAE.

![Quantile-quantile plot of significance of difference in splice junction usage between smokers and nonsmokers. Normalized reads supporting splicing in the smokers and nonsmoker samples were compared. The data shows that smoking caused no significant difference in the splicing for either the ubiquitously expressed genes (blue) or SAE-enriched genes (green).](image)
High expression of Clara cell-associated secretoglobin genes in the SAE was accompanied with enrichment of transcription factors forkhead box A1 (FOXA1), NK2 homeobox 1 (NKX2-1), FOXA2, and CCAAT/enhancer binding protein, alpha (C/EBPα), transcription factors that constitute a major regulatory network for the development and maintenance of SAE and Clara cell differentiation [43,63-65]. NKX2-1 interacts with FOXA1 [34], FOXA1 and FOXA2 complement each other [35], and both NKX2-1 and FOXA2 are thought to act upstream of C/EBPα in lung epithelial differentiation [65,66]. A number of secretory genes, not previously described for the human SAE, were identified by RNA-Seq as highly abundant SAE-enriched genes, including tetraspanin-1 (TSPAN1), a protein involved in secretory pathways in glandular cells [67], cytochrome CYP4B1, a CYP family member localized within the secretory granules of the respiratory mucosa [68], and microseminoprotein-beta (MSMB), an androgen-responsive secretory protein regulating cell growth and apoptosis [69].

Mucosal host defense
Secretory leukocyte peptidase inhibitor (SLPI) and polymeric immunoglobulin receptor (PIGR), two key components of the mucosal defense system, were among the most highly expressed SAE-enriched genes. SLPI has multiple contributions to pulmonary defense, including its ability to neutralize neutrophil elastase, one of the major mediators of lung derangement associated with inflammatory diseases, direct antimicrobial and anti-inflammatory activities, and augmentation of anti-oxidant resistance by increasing glutathione levels in the respiratory surface fluid [70-73]. PIGR is essential for the transepithelial basal-to-apical transport of the polymeric immunoglobulin IgA to the epithelial surface, where it functions to sample and neutralize luminal pathogens [74]. Lipocalin 2 (LCN2), a siderophore-binding antimicrobial protein secreted by pulmonary epithelial cells [75], and the whey acid protein four-disulfide core domain 2 (WFDC2), a SLPI-related gene with potential innate immune functions [76], were also among the most abundant genes enriched in the SAE. Among the most highly expressed genes in the SAE was ELF3, a helix-loop-helix transcription factor expressed in diverse epithelial tissues implicated in the regulation of inflammatory responses [39]. In the context that the airway epithelium is at the interface of the environment (the apical surface) and potential inflammatory/immune mediators (the basolateral surface), the host defense genes identified in the present study as the most abundant SAE genes may play a central role in both mediating and controlling the responses of the airway to environmental xenobiotics and pathogens.

Anti-oxidant protection
The ability to resist deleterious effects of the oxidative stress is critical for the SAE, continuously interacting with oxidants present in the inhaled air. Apart from the secretory genes with anti-oxidant functions such as SGCB1A1 and SLPI, a number of other genes directly related to the protection from oxidative stress, including glutathione S-transferases pi 1 and alpha 1, and glutathione peroxidase 1 (GPX1), were identified as highly expressed SAE-enriched genes. One of these components, GPX1, also known as Clara cell-specific protein CC26, is selectively expressed by Clara cells [55], suggesting that high abundance of both secretory and oxidative stress-related features in the SAE might reflect a secretory cell origin of at least some of the anti-oxidant mechanisms in the human SAE.

Mucociliary differentiation
Consistent with the abundance of ciliated cells in the human SAE, transcription factor FOXJ1, the major regulator of ciliogenesis and ciliated cell differentiation in the airway epithelium [41,42], was among the top 20 SAE-enriched genes and the most highly expressed transcription factor. Other cilia-related genes enriched in the SAE were tektin-1 and -2, structural determinants of the basal bodies of cilia [77], cilia apical structure protein sentan [78], dynein chains DNAI1, DNALI1, DNAI2 and sperm associated antigen SPAG6, the classic components of motile cilia [79]. In addition to these well-known genes, RNA-Seq analysis revealed that several recently discovered cilia-related genes were highly enriched in the human SAE, including the member of the membrane-spanning 4-domain family MS4A8B, which has high sequence homology to cilia-associated gene L985P [80]. Surprisingly, a considerable number of mucus-related genes, such as trefoil factor 3 [81], mucin 1 and mucin 5B [82,83], were highly expressed in the SAE transcriptome along with AGR2, a secretory factor related to goblet cell differentiation [84,85]. Of note, as compared to the large airways, where secreted polymeric mucins are abundant [86], the SAE transcriptome was enriched in membrane-tethered mucins such as MUC1, MUC4, MUC15, MUC20, MUC16, and MUC13, which have various signaling functions [87].

Stem/progenitor cell features
Although Clara cells are considered to be stem/progenitor cells of the mouse bronchiolar epithelium [8,88], the identity of stem/progenitor cell population of the SAE in humans is unknown. Several genes related to stem/progenitor cells were identified in the present study as SAE-enriched genes, including aquaporin-3, a marker of basal cell and suprabasal cell populations with progenitor cell function described for the human
tracheobronchial epithelium [89] and aldehyde dehydrogenase ALDH1, a marker of normal and malignant stem cells in various tissues [90,91]. It is notable that among the top 5 highly-expressed SAE-enriched transcription factors were ELF3, which promotes epithelial morphogenesis [92], and embryonic stem cell-related gene SOX2, recently shown to be important for the progenitor cell function of the airway basal cells and Clara cells and induction of the airway epithelial cell phenotype in mice [36–38]. Due to its high sensitivity, RNA-Seq analysis also identified markers of the putative stem/progenitor cells previously found in the airway epithelium with relatively low frequency, such as keratin 14, a marker of a basal cell subpopulation [93], and surfactant protein C, a gene ascribed to a unique population of bronchoalveolar stem cells in mice [88]. Together, the RNA-Seq data of the present study demonstrates SAE expression of multiple pathways potentially relevant for the maintenance of human SAE via local stem/progenitor cell activity.

Transmembrane receptors, signaling ligands and growth factors

The most highly expressed transmembrane receptor in SAE of nonsmokers was DDR1 (discoidin domain receptor 1), a receptor tyrosine kinase [44]. Expression of the DDR1 protein is located on the basolateral surface of human bronchial epithelium, where it interacts with type IV collagen with consequent activation of its tyrosine kinase activity. The second most abundant SAE-enriched receptor was CELSR1 (cadherin, EGF LAG seven-pass G-type receptor 1), a G protein coupled receptor known to be critical for branching morphogenesis in mouse lung [94]. The most highly expressed SAE-enriched ligand was midkine (MDK), which has a role in lung morphogenesis and is believed to be essential for vascular maintenance in the adult lung [95]. In mouse, midkine expression is controlled by Nkx2-1 [96] which, as mentioned above, is also highly expressed in the human SAE. Among the highly expressed ligands, there was a clear prevalence of chemokines such as MDK, CXCL1 and CX3CL1. Consistent with this observation, expression of diverse cytokines by airway epithelium and cell lines derived from airway epithelium is well established and epithelial derived chemokines are recognized to play an important role in attracting immune and inflammatory cells [97,98].

The RNA-Seq data also pointed to potentially novel aspects of cell signaling in epithelial biology. For example, the oxytocin receptor (OXTR) was expressed at high levels in all subjects who were male. This was initially surprising due to roles of oxytocin in childbirth, lactation and brain biology [99] but, relevant to the airway epithelium, a role for oxytocin in autocrine signaling in small cell lung cancer has been described [100].

Gene Family Members

RNA-Seq offers the potential advantage of distinguishing expression levels among different members of closely related gene families with potentially different functions, whereas cross hybridization among probes often makes this a challenge using microarrays [25]. For example, the RNA-Seq analysis permitted quantification of the expression levels of 3 highly different homologous members of the cytochrome P450 family 2, subfamily A, CYP2A6, CYP2A7 and CYP2A13. RNA-Seq allowed the transcripts to be unambiguously attributed primarily to CYP2A13 which is responsible for metabolism of the cigarette smoke specific carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol [101]. On the other hand, the family member CYP2A6 has different substrate specificity dictated by critical amino acid differences between these otherwise closely related proteins [102]. The significance of these differences are underscored by the variant CYP2A13*2, which is associated with decreased incidences of lung adenocarcinoma in smokers [103].

SAE Transcriptome Response to Smoking

Extensive microarray studies have identified a dramatic effect of smoking on the gene expression profile of human airway epithelium [11–18]. By using RPKM quantification as a measure of smoking-dependent changes in SAE transcript levels, the present study expands the insights into the airway epithelial response to smoking. The quantitative RNA-Seq analysis revealed that smoking suppressed the expression level of a greater number of genes than it induced. Interestingly, among the up-regulated genes, smoking has a larger effect on SAE-enriched genes rather than the ubiquitous genes. From the functional perspective, the SAE-enriched smoking-up-regulated genes were related to transcription, signal transduction, protease/antiprotease homeostasis, and immunity.

The top 2 SAE-enriched genes, the Clara cell associated genes SCGB1A1 and SCGB3A1, were both down-regulated by smoking with a large magnitude of change in expression levels. Smoking and especially COPD have been associated with the loss of Clara cells and the levels of SCGB1A1 in both induced sputum and serum are lower in smokers with COPD as compared to both nonsmokers and healthy smokers [104–107]. It is possible that down-regulation of Clara cell secretoglobins, with their anti-oxidant, anti-inflammatory and tumor-suppressor activities [60,108,109], is a critical component of smoking-related development of COPD. The decreased number of SCGB1A1-expressing Clara cells in smokers is generally accompanied by an increased frequency of mucin-secreting cells [29]. Indeed, a subset of highly expressed SAE-enriched genes, such as
C20orf114 (also known as long PLUNC1), and MSMB, both associated with mucin-producing secretory cell phenotype [110,111], were among the smoking-induced genes with the highest amplitude of up-regulation. Other genes related to a secretory phenotype such as WDFC2, TSPAN1, TFF3, S100P, and short PLUNC, were also induced by smoking; each of these genes has been associated with epithelial carcinogenesis [67,112-115]. Thus, a broad induction of a mucin-producing cell secretory program, characteristic of epithelial malignancies, may represent an early molecular phenotype relevant to smoking-induced carcinogenesis in the distal airways.

Other smoking-induced changes among the highly expressed SAE-enriched genes included up-regulation of oxidative stress-related genes ALDH3A1 and GSTA2, also associated with cancer development [116,117], and down-regulation of genes associated with epithelial differentiation such as CD74, C9orf24 (also known as ciliated bronchial epithelium 1), and luminal cell-associated keratin 19[118,119]. Some of these changes have not been previously detected by microarrays, likely due to microarray saturation of signal with high levels of expression and/or higher sensitivity of the RNA-Seq methodology to gene expression changes with relatively low overall fold-difference between the groups. The ability of RNA-Seq to assess genes with low steady-state expression was utilized in the present study to characterize the effect of smoking on the expression of low abundance SAE genes. Although some of changes, such as up-regulation of oxidative stress-responsive genes AKR1B10, CAV1YR, and CYP1B1 have been previously reported [11,45-47], RNA-Seq quantification revealed a number of novel smoking-responsive genes, including smoking-induced NOS3, a gene encoding nitric oxide isoform usually expressed by endothelial cells but induced in the airway epithelium in association with squamous differentiation [120], and smoking-suppressed Ly6/neurotoxin 1 (LYNX1), an allosteric modulator of nicotinic acetylcholine receptors [121].

Functional classification of the low level, smoking-related genes also identified the class of ion transport genes as being modulated by smoking. One example was CNGB1, a smoking-induced gene that encodes a cyclic nucleotide gated channel that was first identified for its role in light activated cellular polarization in retinal photoreceptor cells [122] and linked to olfactory receptor function [123]. The discovery that airway ciliated cells have olfactory receptors that operate by the same signal transduction pathways as visual rhodopsin [124] suggests a role for CNGB1 in airway epithelial function. Also notable among smoking-dependent genes were a series of ion channels whose overall low expression level in the SAE may reflect expression predominantly in neuroendocrine cells which constitute < 0.01% of total airway epithelium. For example, CACNG4, the gamma subunit of a voltage gated calcium channel, is a smoking-induced gene. Previous reports suggest that this gamma subunit is expressed primarily in brain [125] but expression of voltage gated calcium channels in neuroendocrine cells and neuroendocrine-derived tumors has been demonstrated [126].

**Splicing**

The mRNA sequence reads across exon junctions permit quantitative assessment of the splicing pattern for all genes. By comparing splice events for smokers and nonsmokers, we were able to demonstrate there are no overall smoking-dependent changes in the patterns of splicing for either ubiquitous or SAE-enriched genes. This was surprising, since there are known to be substantial genome wide differences in splicing between normal airway epithelium and lung cancer [127,128], suggesting those splicing-related changes are late events and are not represented in non-transformed airway epithelial cells.

**Conclusions**

RNA-Seq method provides wide dynamic range and low noise. Application of RNA-Seq to SAE allowed the unequivocal identification of highly expressed ubiquitous and SAE-enriched genes. Functional assignment of highly expressed genes showed Clara cell specific genes were most abundantly expressed. But genes characteristic of minor cell types such as neuroendocrine cells were also evident. Comparison of the transcriptome of nonsmokers to that of healthy smokers allowed the response of SAE to cigarette smoke to be quantified and novel smoking-responsive genes to be identified.

**Methods**

**Study Population**

Following approval by the Weill Cornell Medical College Institutional Review Board, healthy nonsmokers and healthy smokers, who responded to local advertisements regarding a research study to assess lung health, were assessed in the Weill Cornell National Institutes of Health Clinical and Translational Sciences Center and Department of Genetic Medicine Clinical Research Facility. Prior to study enrollment, each individual provided written informed consent. The study population included healthy nonsmokers (n = 5) and healthy smokers (n = 6), phenotyped by a standardized screening assessment consisting of a history, physical examination, complete blood count, coagulation profile, liver function tests, urine studies, chest X-ray, EKG, and lung function tests (see Additional Data Methods for inclusion/exclusion criteria; Additional file 1, Table S1 for detailed
demographics). Urinary nicotine and cotinine were used to verify the self-reported smoking status of smokers. For comparison between RNA-Seq and microarray data, 27 healthy nonsmokers from a previous study were assessed [129] (see Additional file 1, Table S1 for demographic details).

**Collection of SAE**

Fiberoptic bronchoscopy was used to sample SAE cells as previously described [13]. After routine anesthesia, a 2 mm disposable brush (Wiltek Medical, Winston-Salem, NC) was inserted into the working channel of the bronchoscope and advanced to the airways distal to the orifice of the desired lobar bronchus. Small airway epithelial samples were obtained by lightly wedging the brush 7 to 10 cm distal to the 3rd generation bronchial airway (i.e., the 10th to 12th order bronchi), and sliding the brush back and forth on the epithelium 10 to 20 times in 8 to 10 sites. For each brush, after withdrawing from the bronchoscope, the cells were dislodged from the brush by flicking the brush tip in 5 ml of ice-cold Bronchial Epithelium Basal Medium (BEBM, Lonza, Basel, Switzerland). A 1 ml aliquot of all airway epithelial samples was used to quantify the percentage of epithelial and inflammatory cells and the proportions of basal, ciliated, secretory and undifferentiated columnar cells by centrifuging 2 × 10⁴ cells per slide (Cytospin 11, Shandon Instruments, Pittsburgh, PA) and using Diff-quick staining reagents (Dade Behring, Newark, NJ); a portion of this aliquot was also used to quantify the number of cells recovered from airway brushings using a hemocytometer. The remaining 4 ml of sample was immediately processed for RNA extraction.

**RNA Extraction and Sample Preparation**

The freshly acquired small airway epithelial samples were treated with TRIzol (Invitrogen Carlsbad, CA) to extract total RNA, and residual DNA was removed by RNeasy MinElute RNA purification kit (Qiagen, Valencia, CA), resulting in a yield of between 2 and 4 µg RNA per 10⁶ cells. To assess the integrity of the RNA, an aliquot of each sample of RNA was analyzed with the Agilent Bioanalyzer (Santa Clara, CA), and the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE) was used to determine the RNA concentration. Samples were then stored in RNAsecure (Ambion, Austin, TX) until further analysis. Using the reagents of the mRNA Sample Prep Kit and in accordance with the Illumina poly(A)⁺ mRNA selection protocol provided by Illumina, poly(A)⁺ mRNA was selected out from the total RNA samples using Sera-mag magnetic oligo(dT) beads. An RNA fragmentation kit (Ambion, Austin, TX) was used to fragment the mRNA, followed by first- and second-strand cDNA synthesis using random hexamer primers. An “end repair” reaction to blunt the ends of all fragments was then performed with Klenow polymerase and T4 DNA polymerase, and 3'-to-5' exo-nuclease was used to create a 3' adenine overhanging tail, facilitating the ligation of the amplification adapters. Ligation products were then separated on a 2% tris-acetate-EDTA-agarose gel for size selection, followed by purification with a gel extraction kit. The purified ligation products were then PCR amplified with complementary primers and the resultant cDNA was purified with QIAquick PCR kit (Qiagen), and the concentration was measured by the NanoDrop spectrophotometer. Samples were then loaded onto Illumina flow cells for single end, 43 nucleotide, sequencing reactions.

**Data Filtering, Read Mapping and Quantification of Gene Expression**

Images acquired by the Illumina Genome Analyzer 2 were analyzed by Firecrest and bases called by Bustard (both part of Illumina RTA pipeline version 1.6). All lanes of data were required to show low overall error rate (< 1.5%), low inter-base phasing (< 1.0), and all reads passed the Illumina GA quality filters (PF = Y). Resultant reads were aligned to the reference genome build UCSC hg19 using Bowtie v 0.12 [130]. The Bowtie default parameters used were those in which the first seed alignment of size > 28 nt, allowing up to 2 mismatches, is chosen at random, and is used if it yields an alignment quality sum of > 70 with a maximum of 125 backtraces. Thus, multiple alignments are not specifically assessed nor scored. The data was then processed with Python scripts to assign aligned reads to the coordinates of exons and genes. Mean read density values for exons, introns, and intergenic regions were computed in units of reads per kilobase of exon/intron/intergenic per million mapped reads [23]. Reads were mapped to the annotated transcribed strand of the genome, because the protocol for sequencing used in the current study was not strand specific. Reads per kilobase of exon per million mapped reads (RPKM) are indicative of actual mRNA concentration when samples have relatively uniform sequencing coverage across the entire gene model [23].

To determine the minimum detectable level of expression, a false discovery rate (FDR) and false negative rate (FNR) was estimated by comparing the expression levels of known exons to intergenic regions (Figure 1). This was done in accordance with the method described by Ramsköld et al. [24]. The distribution of exon expression levels was compared to the expression levels of intergenic regions based on the criteria: (1) no annotated genes according to the NCBI Reference Sequence (RefSeq: http://www.ncbi.nlm.nih.gov/RefSeq/) and Ensembl http://www.ensembl.org databases; and (2) no
known expressed sequence tags in the GenBank sequence database http://www.ncbi.nlm.nih.gov/genbank. In order to avoid a bias due to changes in the size distribution of intergenic regions and exons, the intergenic regions were chosen at random to have the same size distribution of the expressed exons. The FDR was calculated for different expression levels as the normalized ratio of number of intergenic regions to number of exons at each expression level. The FNR for different expression levels was estimated from the cumulative ratio of the true positive rate (as estimated from the product of number of expressed exons and the FDR) and the total fraction of expressed exons. Based on this analysis, the optimal expression value as defined by the intersection of the FDR and FNR in all non-smoker samples was 0.125 RPKM.

All sequence read data have been submitted to the Short Read Archive (SRA) section of the NCBI SRA database (SRA accession #SRP005411); and U133 data submitted to GEO (GSE27681).

Data Analysis

A cut off value of RPKM 0.125 was established, below which expression was considered as noise (Figure 1). Genes for which the median expression in nonsmokers was > 0.125 RPKM were scored as expressed. Genes expressed by the small airway epithelium (SAE) were categorized as “ubiquitous” and “SAE-enriched” as follows. Ubiquitous genes were defined as described by Ramsköld et al. [24] based on expression in 11 of 12 tissues surveyed. The SAE expressed genes were grouped as “ubiquitous” if also expressed in at least 11 of 12 other tissues or “SAE-enriched” (if not in the “ubiquitous” list) [24]. Based on the median expression level in nonsmokers, expressed genes were further divided into “low” (RPKM 0.125 to 1), “medium” (RPKM between 1 and 10) and “high” expression genes (RPKM > 10).

The abundance of the transcripts from individual genes in the total mRNA pool in different tissues was assessed by building a frequency distribution. For our SAE data and for published RNA-Seq data from various tissue [24], all genes were ranked by transcript abundance and then the fraction of total mRNA contributed by gene #1, genes #2-10, genes #11-100, genes #101-1000 and # > 1000 was determined. This gives a frequency distribution that could be compared among tissues with assessment by Fisher’s exact test.

To compare the RNA-Seq data to that of microarrays, Human Genome U133 Plus 2.0 microarray data (Affymetrix, Santa Clara, CA) from 27 healthy, African-American nonsmokers was used ([129]; Additional file 1, Table S1). The microarray CEL files were analyzed by Affymetrix Suite software and the “P” calls for each probeset were totaled for all subjects. The gene list from RNA-Seq was systematically reviewed in comparison to the microarray data. Where there was a corresponding probeset on the microarray data, the percentage of subjects with “P” call was determined. When there was > 1 probeset corresponding to a single named gene, the probeset with the highest percentage P call was used.

To further characterize the healthy SAE transcriptome, the data from the healthy nonsmokers were assessed for: (1) the overall most highly expressed genes; (2) the most highly expressed genes of differentiated cell types (ciliated, secretory, basal and neuroendocrine cells), using lists of genes characteristic of these differentiated cell types (Additional file 1, Table S4) [30,31,52]; (3) genes coding for transcription factors; (4) genes coding for transmembrane receptors; and (5) genes coding for signaling ligands and growth factors. In all cases, the most highly expressed was based on the median for all nonsmokers. Gene families expressed by the SAE of healthy nonsmokers were identified using Basic Local Alignment Search Tool (BLAST; http://www.ncbi.nlm.nih.gov/BLAST/). All RefSeq genes expressed by nonsmokers were aligned against a database of all human RefSeq mRNA [26]. Gene families were defined as groups of genes for which the alignments yielded ≥ 90% identity and the alignment length was at least 50% of both the query and matched sequences. Changes in gene expression of the family members were assessed as described above.

Smoking responsive genes were assigned on the basis of comparing RPKM level in 5 nonsmokers to that in 6 smokers by t-test with no correction for multiple comparisons. All genes with a p value of < 0.05 were deemed to be smoking-dependent regardless of any cut off in absolute change or fold-change (smoker/nonsmoker expression ratio).

The effect of smoking on alternative splicing was estimated by comparing normalized splice junction usage. To accomplish this, all reads that failed to align to the reference genome were aligned (using Bowtie) to a database of all RefSeq annotated exon-exon boundaries generated such that each junction required reads to overlap each exon by at least 3 nucleotides. By normalizing the number of reads at each junction by the length of each junction in kilobases, number of reads in the sample in millions of reads and the expression level of neighboring exons, it was possible to compare junction usage even in genes with different expression levels. To re-due the false positive rate, filtering included exclusion of all junctions with expression levels (RPKM) below 0.125, all junctions with less than 2 spliced reads in both the smoker and non-smoker samples, as well as any genes where the standard error in RPKM across all samples was greater than 0.5. A t-test was used to estimate the significance of the difference in splice junction usage of the
filtered junctions between smokers and nonsmokers. The data were analyzed using multiple test corrections with evaluation by Q-Q plot.

Additional material

Additional file 1: Additional Data Methods. Additional Table S1. Demographics of the study population and biologic samples. Additional Table S2. Mapping summary. Additional Table S3. Comparison of the median expression levels of different categories of genes in the small airway epithelium of healthy nonsmokers and healthy smokers. Additional Table S4. Cell type-specific gene lists. Additional Table SS. Reproducibility of smoking-responsive genes discovered by microarray using RNA-Seq method. Additional Table S6. Reproducibility of smoking-responsive genes discovered by RNA-Seq using microarray method. Additional Figure Legends. Additional figures S1 and S2.

List of abbreviations

RNA-Seq: high throughput sequencing of mRNA fragments; SAE: small airway epithelium; COPD: chronic obstructive pulmonary disease; FDR: false discovery rate; FNR: false negative rate; CDNA: complementary DNA; RPMV: reads per kilobase of exon per million mapped reads.

Acknowledgements

We thank P. Schweitzer and W. Wang, Cornell Biotechnology Lab, Cornell University for performing RNA-Seq and N. Mohamed for help in preparing this manuscript. These studies were supported, in part, by P50 HL084936 and ULI-RR024996, and the Starr Foundation/Starr Cancer Consortium. JLRF was supported, in part, by NIH 1T32HL094284.

Author details

1 Department of Genetic Medicine, Weill Cornell Medical College, New York, New York, USA. 2 Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, USA.

Authors' contributions

NRH, MWB and RS analyzed data, wrote article, mined data for biological meaning; JS, JRFL, JGM, YS-B performed bioinformatic and statistical analyses; LG analyzed splicing; GW analyzed and interpreted data related to secretory protein (CC16): features as a peripheral lung biomarker. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 13 October 2011 Accepted: 29 February 2012 Published: 29 February 2012

References

1. Crystal RG, Randell SH, Engelhardt JF, Voyvod J, Sunday ME. Airway epithelial cells: current concepts and challenges. Proc Am Thorac Soc 2008, 5:772-777.
2. Knight DA, Holgate ST. The airway epithelium: structural and functional properties in health and disease. Respir Physiol 2003, 134:452-466.
3. Hecht SS. Tobacco carcinogens, their biomarkers and tobacco-induced cancer. Nat Rev Cancer 2003, 3:733-744.
4. Center for Disease Control and Prevention. 2004 Surgeon General’s report - The health consequences of smoking. Centers for Disease Control and Prevention 2004 [http://www.cdc.gov/tobacco/data_statistics/sgr/2004/].
5. Pryor WA, Hales BJ, Premovic PI, Church DF. The radicals in cigarette tar: their nature and suggested physiological implications. Science 1983, 220:425-427.
6. Rodgman A, Perfetti TA. The chemical components of tobacco and tobacco smoke. CRC Press, Boca Raton; 2009, DOI:10.1201/9781420078848.
7. Hogg JC, Chu F, Utokaparch S, Woods R, Elliot WM, Buzatu L, Cherniack RM, Rogers RM, Scruib FA, Coxson HO, Pare PD. The nature of small-airway obstruction in chronic obstructive pulmonary disease. N Engl J Med 2004, 350:3645-3653.
8. Rawlins EL, Okubo T, Xue Y, Bratt DM, Auten RL, Hasegawa H, Wang F, Hogan BL. The role of Sgpb1α+ Clara cells in the long-term maintenance and repair of lung airway, but not alveolar, epithelium. Cell Stem Cell 2009, 4:525-534.
9. Rawlins EL, Hogan BL. Epithelial stem cells of the lung: privileged few or opportunities for many? Development 2006, 133:2465-2466.
10. Reynolds SD, Malkinova AA. Clara cell: progenitor for the bronchiolar epithelium. Int J Biochem Cell Biol 2010, 42:1-4.
11. Hackett NR, Heguy A, Harvey BG, O’Connor TP, Luetchtich K, Flieder DB, Kaplan R, Crystal RG. Variability of antioxidant-related gene expression in the airway epithelium of cigarette smokers. Am J Respir Cell Mol Biol 2003, 29:331-345.
12. Ammoura Z, Hackett NR, Butler MW, Raman T, Dolgalev I, O’Connor TP, Harvey BG, Crystal RG. Variability in small airway epithelial gene expression among normal smokers. Chest 2008, 133:1344-1353.
13. 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 2007, 85:59-53.
14. Spira A, Beane J, Shah N, Liu G, Schembri F, Yang X, Palma J, Brody JS. Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc Natl Acad Sci USA 2004, 101:10143-10148.
15. Beane J, Sebastiano P, Liu G, Brody JS, Libnmg ME, Spira A. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. Genome Biol 2007, 8:R201.
16. Chan R, Loneragan KM, Ng RT, Macalau C, Lam WL, Lam S. Effect of active smoking on the human bronchial epithelial transcriptome. BMC Genomics 2007, 8:297.
17. Zhang L, Lee JJ, Tang H, Fan YH, Xiao L, Ren H, Kurie J, Monroe RC, Hong WK, Mao L. Impact of smoking cessation on global gene expression in the bronchial epithelium of chronic smokers. Cancer Prev Res (Phila) 2008, 1:112-118.
18. Pierrou S, Broberg P, Olofsson RA, Pawlovski K, Virtalı R, Lindqvist E, Richter A, Wilson SJ, Angco G, Moller S, Bergstrand H, Koopmann W, Wiselander E, Stromstedt PE, Holgate ST, Davies DE, Lund J, Djukanovic R. Expression of genes involved in oxidative stress responses in airway epithelial cells of smokers with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2007, 175:577-586.
19. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009, 10:57-63.
20. Costa V, Angelini C, De F, I, Ciccodicola A. Uncovering the complexity of transcriptomes with RNA-Seq. J Biomed Biotechnol 2010, 2010:353916.
21. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y. Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 2010, 464:768-772.
22. Manoni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 2008, 18:1509-1517.
23. Montazav A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008, 5:628-628.
24. Ramskold D, Wang ET, Burge CB, Sandberg R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol 2009, 5:e1000598.
25. Bloom JS, Khan Z, Kugiyaki L, Singh M, Caudy AA. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics 2009, 10:221.
26. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2003, 31:39-53.
27. Spira A, Beane J, Sebastiani P, Liu G, Brody JS, Libnmg ME, Spira A. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. Genome Biol 2007, 8:R201.
28. Montazav A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008, 5:628-628.
29. Lumsdon AB, McLean A, Lamb D. Goblet and Clara cells of human distal airways: evi-dence for smoking induced changes in their numbers. Thorax 1984, 39:844-849.

30. Hackett NR, Shaykhiev R, Walters MS, Wang R, Zwicky RK, Ferris B, Witbrot O, Salti J. Crystal RG. The human airway epithelial basal cell transcriptome. PLoS One 2011, 6:e18378.

31. Carolan BJ, Harvey BG, De BP, Vanni H, Crystal RG. Decreased expression of inhibin B in the human airway epithelium of smokers compared to nonsmokers. J Immunol 2008, 181:5760-5767.

32. Haimoto H, Takanashi Y, Koshikawa T, Nagura H, Kato K. Immunohistochemical localization of gamma-enolase in normal human tissues other than nervous and neuroendocrine tissues. Lab Invest 1985, 52:257-263.

33. Ring BZ, Seitz RS, Beck RA, Shasteen WJ, Sotterman A, Arborgat S, Robert F, Schroeder MT, Ross DT. A novel five-antibody immunohistochemical test for subclassification of lung carcinoma. Mod Pathol 2009, 22:1032-1043.

34. Minoo P, Hu L, Xing Y, Zhu NL, Chen H, Li M, Borok Z, Li C. Physical and functional interactions between homeodomain NKX2.1 and winged helix/forkhead FOXA1 in lung epithelial cells. Mol Cell Biol 2007, 27:2155-2165.

35. Han W, Dingle S, Xu Y, Besnard V, Kaestner KH, Ang SL, Wert SE, Stahlman MT. Coordinate control of expression of Nrf2-modulated genes in FOXA1 and FOXA2 during lung morphogenesis. J Biol Chem 2005, 280:13809-13816.

36. Tompkins DH, Besnard V, Lange AW, Wert SE, Kieser AR, Smith AN, Lang R, Whitsett JA. Sox2 is required for maintenance and differentiation of bronchiolar Clara, ciliated, and goblet cells. PLoS One 2009, 4:e8248.

37. Tompkins DH, Besnard V, Lange AW, Wert SE, Bruno MD, Whitsett JA. Sox2 Activates Cell Proliferation and Differentiation in the Respiratory Epithelium. Am J Respir Cell Mol Biol 2011, 45:101-110.

38. Que J, Okubo T, Goldengarten JR, Nam KT, Kurotani R, Morrisey EE, Taranova O, Pevny LH, Hogan BL. Multiple dose-dependent roles for Sox2 in the patterning and differentiation of anterior foregut endoderm. Development 2007, 134:2521-2531.

39. Wu J, Duan R, Cao H, Field D, Newnham CM, Koehler DR, Zamel N, Devidas S, Guggino WB. CFTR: domains, structure, and function. J Bioenerg Biomembr 1997, 29:443-451.

40. Trapnell BC, Chu CS, Paakko PK, Banks TC, Yoshimura K, Ferrans VJ, Chernick MS, Crystal RG. Expression of the cystic fibrosis transmembrane conductance regulator gene in the respiratory tract of normal individuals and individuals with cystic fibrosis. Proc Natl Acad Sci USA 1991, 88:6565-6569.

41. Veluculescu VE, Madden SL, Zhang L, Lash AE, Yu J, Rago C, Lal A, Wang CJ, Beaudry GA, Cimino KM, Cook BP, Dufault MR, Ferguson AT, Gao Y, He TC, Hermeking H, Hirakdo SK, Hwang PM, Lopez MA, Luderer HF, Mathews B, Petrozello JM, Polyak K, Zawel L, Kinder KW. Analysis of human transcriptomes. Nat Genet 1999, 23:387-388.

42. Ovchar T, Iliey AE, Shaykhiev R, Wang R, Crystal RG. Do Airway Epithelium Air-liquid Cultures Represent the In Vivo Airway Epithelium Transcriptome? Am J Respir Cell Mol Biol 2010, 44:465-473.

43. Mango GW, Johnston CJ, Reynolds SD, Finkelstein JN, Plopper CG, Stripp BR. Clara cell secretory protein deficiency increases oxidant stress response in conducting airways. Am J Physiol 1998, 275:C438-C536.

44. Ramsay PL, DeMayo FJ, Hegeman SE, Wearden ME, Smith CV, Welty SE. Clara cell secretory protein oxidation and expression in premature infants who develop bronchopulmonary dysplasia. Am J Respir Crit Care Med 2001, 164:155-161.

45. Stripp BR, Reynolds SD, Boe IM, Lund J, Power JH, Copps JT, Wong V, Reynolds PR, Plopper CG. Clara cell secretory protein deficiency alters clara cell secretory apparatus and the protein composition of airway lining fluid. Am J Respir Cell Mol Biol 2002, 27:170-178.

46. Wang SZ, Rosenberger CL, Bao YX, Stark JM, Harrod KD. Clara cell secretory protein modulates lung inflammatory and immune responses to respiratory syncytial virus infection. J Immunol 2003, 171:1051-1060.

47. Snyder JC, Reynolds SD, Hollingsworth JW, Li Z, Kaminiski N, Stripp BR. Clara cells attenuate the inflammatory response through regulation of macrophage behavior. Am J Respir Cell Mol Biol 2010, 42:161-171.

48. Porter D, Laht-Domenici J, Torres-Arayus M, Chin L, Polyak K. Expression of high in normal-1 (HIN-1) and uteroglobin related protein-1 (UGRP-1) in adult and developing tissues. Mech Dev 2002, 114:201-204.

49. Krop IE, Sgroi D, Porter DA, Dunetta KL, LeVange R, Seth P, Kaelin CM, Rhee E, Rosenberg M, Schmitt S, Marks JR, Pagon Z, Belina D, Razumov J, Polyak K. HIN-1, a putative cytotoxic highly expressed in normal but not cancerous mammary epithelial cells. Proc Natl Acad Sci USA 2001, 98:9796-9801.

50. Reynolds SD, Reynolds PR, Pfyffer GS, Finder JD, Stripp BR. Secretoglobin SCGB3A1 and SCGB3A2 define secretory cell subsets in mouse and human airways. Am J Respir Crit Care Med 2002, 166:1498-1509.

51. Krop I, Parker MT, Blouschantch-Qimron N, Porter D, Gelman R, Sasaki H, Maurer M, Terry MB, Parsons R, Polyak K. HIN-1, an inhibitor of cell growth, invasion, and AKT activation. Cancer Res 2005, 65:9659-9669.

52. Shimematsu H, Suzuki M, Takahashi T, Miyajima K, Toyooka S, Shivapurkar N, Czerniai B, Fujisawa T, Shimizu N, Gasdar AF. Aberrant methylation of HIN-1 (high in normal-1) is a frequent event in many human malignancies. Int J Cancer 2003, 110:600-604.

53. Savaya PL, Stripp BR, Whitsett JA, Luse DS. The lung-specific CC10 gene is regulated by transcription factors from the AP-1, octamer, and hepatocyte nuclear factor 3 families. Mol Cell Biol 1999, 19:3860-3871.

54. Ray MK, Chen CY, Schwartz RJ, DeMayo FJ. Transcriptional regulation of a mouse Clara cell-specific protein (mCC10) gene by the Nkx transcription factor family members thyroid transcription factor 1 and cardiac muscle-specific homeobox protein (CSX). Mol Cell Biol 1996, 16:2056-2064.

55. Mantis PC, Whitsett JA, Xu T, Perl AK, Han H, Ikegami M, CEBP alpha is a therapeutic target. Clin Cancer Res 2005, 11:139-141.

56. Scholz CJ, Kurzeder C, Schreeder MT, Ross DT, Reynolds PR, Schreeder MT, Reynolds PR, Plopper CG, Stripp BR. Clara cells attenuate the inflammatory response through regulation of macrophage behavior. Am J Respir Cell Mol Biol 2010, 42:161-171.

57. Porter D, Laht-Domenici J, Torres-Arayus M, Chin L, Polyak K. Expression of high in normal-1 (HIN-1) and uteroglobin related protein-1 (UGRP-1) in adult and developing tissues. Mech Dev 2002, 114:201-204.

58. Krop IE, Sgroi D, Porter DA, Dunetta KL, LeVange R, Seth P, Kaelin CM, Rhee E, Rosenberg M, Schmitt S, Marks JR, Pagon Z, Belina D, Razumov J, Polyak K. HIN-1, a putative cytotoxic highly expressed in normal but not cancerous mammary epithelial cells. Proc Natl Acad Sci USA 2001, 98:9796-9801.

59. Reynolds SD, Reynolds PR, Pfyffer GS, Finder JD, Stripp BR. Secretoglobin SCGB3A1 and SCGB3A2 define secretory cell subsets in mouse and human airways. Am J Respir Crit Care Med 2002, 166:1498-1509.

60. Krop I, Parker MT, Blouschantch-Qimron N, Porter D, Gelman R, Sasaki H, Maurer M, Terry MB, Parsons R, Polyak K. HIN-1, an inhibitor of cell growth, invasion, and AKT activation. Cancer Res 2005, 65:9659-9669.

61. Shinematsu H, Suzuki M, Takahashi T, Miyajima K, Toyooka S, Shivapurkar N, Tominoshin GE, Wangxie D, Fai J, Birmabla E, Sathyanarayana UK, Czerniai B, Fujisawa T, Shimizu N, Gasdar AF. Aberrant methylation of HIN-1 (high in normal-1) is a frequent event in many human malignancies. Int J Cancer 2003, 110:600-604.

62. Savaya PL, Stripp BR, Whitsett JA, Luse DS. The lung-specific CC10 gene is regulated by transcription factors from the AP-1, octamer, and hepatocyte nuclear factor 3 families. Mol Cell Biol 1999, 19:3860-3871.

63. Ray MK, Chen CY, Schwartz RJ, DeMayo FJ. Transcriptional regulation of a mouse Clara cell-specific protein (mCC10) gene by the Nkx transcription factor family members thyroid transcription factor 1 and cardiac muscle-specific homeobox protein (CSX). Mol Cell Biol 1996, 16:2056-2064.

64. Mantis PC, Whitsett JA, Xu T, Perl AK, Han H, Ikegami M, CEBP alpha is a therapeutic target. Clin Cancer Res 2005, 11:139-141.
80. Bangur CS, Johnson JC, Switzer A, Wang YH, Hill B, Fanger GR, Wang T, Nettesheim P: Regulation of MUC5 and MUC1 gene expression: correlation with airway mucous differentiation. J Biol Chem 1996, 270:2716-2719.

79. Zariwala MA, Knowles MR, Omran H: Genetic defects in ciliary structure and function. Cell Motil Cytoskeleton 2004, 59:338-346.

78. Kubo A, Yuba-Kubo A, Tsukita S, Tsukita S, Amagai M: Cloning of the gene gob-4, which is required for mouse pulmonary goblet cell differentiation and regulates a network of genes associated with their potential roles for different-tissue processes during airway remodeling. Curr Stem Cell Res Ther 2008, 3:237-246.

77. Stephens RE, Lemieux NA. Tektins as structural determinants in basal bodies. Cell Motil Cytoskeleton 1998, 40:379-392.

76. Galgano MT, Hampton GM, Frierson HF Jr, Reynolds PR, Mucenski ML, Whitsett JA: Expression of the polymeric immunoglobulin receptor: bridging innate and adaptive immune responses at mucosal surfaces. Immunol Rev 2005, 206:83-99.

75. Chan YR, Liu JS, Pociask DA, Zheng M, Mietzner TA, Berger T, Mak TW, Hubbard RC, Crystal RG: Presence of a putative tumor-initiating progenitor cell population predicts poor prognosis in smokers with non-small cell lung cancer. Cancer Res 2010, 70:6659-6664.

74. Pequeux C, Breton C, Hendrick JC, Hagelstein MT, Mertens H, Winkler R, Geenen V, Legros JJ: Oxytocin synthesis and oxytocin receptor expression by cell lines of human small cell carcinoma of the lung stimulate tumor growth through autocrine/paracrine signaling. Cancer Res 2008, 68:4243-4250.

73. Galbally M, Lewis AJ, Ijzendoorn M, Permezel M: Lipocalin 2 is required for the formation and function of secretory leukoprotease inhibitor. Exp Lung Res 2003, 27:237-246.

72. Vaerman JP, Decramer M, Sibille Y: Presence of a putative tumor-initiating progenitor cell population predicts poor prognosis in smokers with non-small cell lung cancer. Cancer Res 2010, 70:6659-6664.

71. Taggart CC, Oyan SA, Weldon S, Gibbons A, Greene CM, Kelly E, Low TB, O'Brien SJ, McEvany NG: Secretory leukoprotease inhibitor binds to NF-kappaB binding sites in monocyttes and inhibits p65 binding. J Exp Med 2005, 202:1659-1668.

70. Vogelmeier C, Hubbard RC, Fells GA, Schnebi HP, Thompson RC, Fritz H, Crystal RG: Anti-neutrophil elastase defense of the normal human respiratory epithelial surface provided by the secretory leukoprotease inhibitor. J Clin Invest 1991, 87:482-488.

69. Tektins as structural determinants in basal TFF (trefoil factor family) peptides and their potential roles for different-tissue processes during airway remodeling. Curr Stem Cell Res Ther 2008, 3:237-246.

68. Jost D, Rousseau K, McGuckin MA: Role of oxytocin in mother-infant relations: a systematic review of human studies. Harv Rev Psychiatry 2011, 19:1-14.

67. Thornton DJ, Rousseau K, McGuckin MA: Structure and function of the cell surface glycoproteins in health and disease. Physiol Rev 2006, 86:245-278.

66. D’Agostino J, Zhang X, Wu H, Ling G, Wang S, Zhang QY, Liu F, Ding X: Characterization of CYP2A13*2, a variant cytochrome P450 allele previously found to be associated with decreased incidences of lung adenocarcinoma in smokers. Drug Metab Dispos 2008, 36:2316-2323.

65. Bourdin A, Kotsimbos T, Nguyen K, Vachier I, Mainprice B, Farce M, Paganin F, Decalvi D, Caninao GI: Clara cell 16 protein in COPD sputum: a marker of small airways damage? Respir Med 2007, 101:2119-2124.

64. Lacewell P, Goddard K, Viss M, Delahoz M, Verheugen E, Deaksteiner C, De PK, Vaerman JP, Decramer M, Sibille Y: Reduced epithelial expression of secretory component in small airways correlates with airflow obstruction in chronic obstructive pulmonary disease. J Am Resp Crit Care Med 2008, 160:185-194.

63. Linnola R, Szabo B, DelMato F, Witschi A, Sabouni C, Malkinson A: The role of CC10 in pulmonary carcinogenesis: from a marker to tumor suppressor. Ann N Y Acad Sci 2003, 993:249-267.

62. Ratschek C, Folli C, Descalzi D, Canonica GW: Evaluation of serum CC-16 as a biomarker for COPD in the ECLIPSE cohort. Thorax 2008, 63:1053-1058.

61. Bourdin A, Kotsimbos T, Nguyen K, Vachier I, Mainprice B, Farce M, Paganin F, Marty-Ane C, Vernhet H, Godard P, Chanez P: Aquaporin-3 expression in human fetal airway epithelial progenitor cells. Stem Cells 2007, 25:1583-1590.

60. Hoffmann T, Woff TF (trefoil factor family) peptides and their potential roles for different-tissue processes during airway remodeling. Curr Stem Cell Res Ther 2007, 11:2914-2924.

59. Thornton DJ, Bourdin A, Kotsimbos T, Nguyen K, Vachier I, Mainprice B, Farce M: Beta microseminoprotein is not a prostate-specific protein. Histochem Cell Biol 2010, 133:505-515.

58. Weiber H, Andersson C, Murne A, Rannevik G, Lindstrom C, LiJa H, Fernlund P: Beta microseminoprotein is not a prostate-specific protein.
Its identification in mucous glands and secretions. Am J Pathol 1990, 137:593-603.

112. Bingle L, Cross SS, High AS, Wallace WA, Rasul D, Yuan G, Hellstrom L, Campos MA, Bingle CD. WDFC2 (HE4): a potential role in the innate immunity of the oral cavity and respiratory tract and the development of adenocarcinomas of the lung. Respir Res 2006, 7:6.

113. Taupin D, Pedersen J, Familiari M, Cook G, Yeomans N, Giraud AS. Augmented intestinal trefoil factor (TFF3) and loss of p52 (TFF1) expression precedes metaplastic differentiation of gastric epithelium. Lab Invest 2001, 81:397-408.

114. Nakata K, Nagai E, Ohuchida K, Hayashi A, Miyaoka Y, Aishima S, Oda Y, Mizumoto K, Tanaka M, Tsuneyoshi M. S100P is a novel marker to identify intraductal papillary mucinous neoplasms. Hum Pathol 2010, 41:824-831.

115. Diederichs S, Bulk E, Steffen B, Ji P, Tickenbrock L, Lang K, Zanker KS, Metzger R, Schneider PM, Gerke V, Thomas M, Berdel WE, Serve H, Muller-Tidow C. S100 family members and trepsinogens are predictors of distant metastasis and survival in early-stage non-small cell lung cancer. Cancer Res 2004, 64:5564-5569.

116. Patel M, Lu L, Zander DS, Seeramana L, Coco D, Moreb JS. ALDH1A1 and ALDH3A1 expression in lung cancers: correlation with histologic type and potential precursors. Lung Cancer 2008, 59:340-349.

117. Gemignani F, Landi S, Szeszenia-Dabrowska N, Zaridze D, Lissowska J, Rudnai P, Fa-bianova E, Mates D, Foretova L, Janout V, Bencko V, Gaborieau V, Gioia-Patricola L, Bellini I, Barale R, Canzian F, Hall J, Roffetta P, Hung RJ, Brennan P. Development of lung cancer before the age of 50: the role of xenobiotic metabolizing genes. Carcinogenesis 2007, 28:1287-1293.

118. Bartek J, Bartkova J, Taylor-Papadimitriou J. Keratin 19 expression in the adult and developing human mammary gland. Histochem J 1990, 22:537-544.

119. Yoshise H, Puddicombe SM, Wilson SJ, Haitch HM, Powell RM, Wilson DI, Pandit A, Berger AE, Davies DE, Holgate ST, Holloway JW. Characterization of ciliated bronchial epithelium 1, a ciliated cell-associated gene induced during mucociliary differentiation. Am J Respir Cell Mol Biol 2004, 31:491-500.

120. Norford D, Koo JS, Gray T, Alder K, Netteymeier P. Expression of nitric oxide synthase isoforms in normal human tracheobronchial epithelial cells in vitro: dependence on retinoic acid and the state of differentiation. Exp Lung Res 1998, 24:355-366.

121. Ibanez-Tallon I, Miwa JM, Wang HL, Adams NC, Crabtree GIW, Sine SM, Heintz N. Novel modulation of neuronal nicotinic acetylcholine receptors by association with the endogenous protoxin lynx1. Neuron 2002, 33:893-903.

122. Ardell MD, Aragon I, Oliveira L, Porche GE, Burke E, Pittler SJ. The beta subunit of human rod photoreceptor cGMP-gated cation channel is generated from a complex transcription unit. FEBS Lett 1996, 389:213-218.

123. Michalske S, Resiert J, Geiger H, Wetzl C, Zong X, Bradley J, Spehr M, Huettl S, Gerstner A, Pfeifer A, Hart H, Yau KW, Bell M. Loss of CNGR1 protein leads to olfactory dysfunction and subciliary cyclic nucleotide-gated channel trapping. J Biol Chem 2006, 281:35156-35166.

124. Shah AS, Ben-Shahar Y, Moninger TO, Klein JN, Walsh MA. Motile cilia of human airway epithelia are chemosensory. Science 2009, 325:1131-1134.

125. Chen RS, Deng TC, Garcia T, Sellers JM, Best PM. Calcium channel gamma subunits: a functionally diverse protein family. Cell Biochem Biophys 2007, 47:178-186.

126. Merger S, Drost A, Rechstein WO, Neuhaus P, Wiedenmann B. Cal(2+) channel properties in neuroendocrine tumor cell cultures investigated by whole-cell patch-clamp technique. Ann N Y Acad Sci 2004, 1014:137-139.

127. Xi L, Feber A, Gupta V, Wu M, Bergemann AD, Landreneau RJ, Little VR, Pennathur A, Luketch JD, Godfrey TE. Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. Nucleic Acids Res 2008, 36:6523-6547.

128. Misquitta-Alu CM, Cheng E, O’Hanlon D, Liu N, McGlade CJ, Tiao MS, Blencowe BJ. Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer. Mol Cell Biol 2011, 31:138-150.

129. Butler MW, Fuku T, Saltit J, Shaykhiev R, Mezey J, Hackett NR, Crystal RG. Modulation of Cystatin A expression in the human Small Airway Epithelium by Genotype, Smoking and COPD. Cancer Res 2011, 71:2572-2581.

130. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009, 10:R25.

Cite this article as: Hackett et al. RNA-Seq quantification of the human small airway epithelium transcriptome. BMC Genomics 2012, 13:82.