Nucleic Acids Research, 2005, Vol. 33, Database issue ***D573–D579***
doi:10.1093/nar/gki035

SIEGE: Smoking Induced Epithelial Gene Expression Database

Vishal Shah[1,*], Sriram Sridhar[1], Jennifer Beane[1], Jerome S. Brody[2] and Avrum Spira[1,2]

[1]Bioinformatics Program, College of Engineering, Boston University, 44 Cummington Street, Boston, MA 02215, USA and [2]Pulmonary Center and Department of Medicine, Boston University School of Medicine, 715 Albany Street, Boston, MA 02118, USA

Received August 15, 2004; Revised and Accepted September 23, 2004

ABSTRACT

The SIEGE (Smoking Induced Epithelial Gene Expression) database is a clinical resource for compiling and analyzing gene expression data from epithelial cells of the human intra-thoracic airway. This database supports a translational research study whose goal is to profile the changes in airway gene expression that are induced by cigarette smoke. RNA is isolated from airway epithelium obtained at bronchoscopy from current-, former- and never-smoker subjects, and hybridized to Affymetrix HG-U133A Genechips, which measure the level of expression of ~22 500 human transcripts. The microarray data generated along with relevant patient information is uploaded to SIEGE by study administrators using the database's web interface, found at http://pulm.bumc.bu.edu/siegeDB. PERL-coded scripts integrated with SIEGE perform various quality control functions including the processing, filtering and formatting of stored data. The R statistical package is used to import database expression values and execute a number of statistical analyses including *t*-tests, correlation coefficients and hierarchical clustering. Values from all statistical analyses can be queried through CGI-based tools and web forms found on the 'Search' section of the database website. Query results are embedded with graphical capabilities as well as with links to other databases containing valuable gene resources, including Entrez Gene, GO, Biocarta, GeneCards, dbSNP and the NCBI Map Viewer.

INTRODUCTION

Cigarette smoking remains the second largest preventable cause of mortality and morbidity worldwide (1). This year approximately 5 million deaths were attributed directly to tobacco use and if current trends persist, this number could double by the year 2025 to as many as 10 million deaths annually (1). A significant number of these smoking related deaths are due to chronic obstructive pulmonary disease (COPD) and lung cancer. The molecular processes by which these two diseases develop have not been definitively elucidated, nor is it clear why only a subset (10–15%) of smokers actually go on to develop COPD or lung cancer (2). Furthermore, there are no tools currently available for identifying the degree of smoking induced damage in a given smoker or the subset of smokers at highest risk for developing these diseases.

A number of studies have shown that cigarette smoking produces a 'field defect' in airway epithelial cells such that genetic damage occurs throughout the lung and its intra- and extra-pulmonary airways (3,4). This damage is characterized by genetic alterations that can lead to aberrant gene expression and protein production in pulmonary epithelial cells. Based on this concept, we have initiated a DNA microarray-based study to profile the gene expression behavior of epithelial cells in the large airway, i.e. the airway transcriptome (5). The primary goals of this study are to define the normal airway transcriptome, determine the effects of cigarette smoke on that transcriptome and establish how clinical parameters such as age, sex and race modulate the effect of cigarette smoke on the airway transcriptome.

In order to support and facilitate these studies, the SIEGE (Smoking Induced Epithelial Gene Expression) relational database and accompanying web interface were created (http://pulm.bumc.bu.edu/siegeDB). SIEGE is implemented using the MySQL database management system (http://www.mysql.com) and is housed at the Pulmonary Center of the Boston University Medical Center in Boston, MA. All users, including the database administrators, interact with the database using web forms. These forms control user access to all statistical and clinical data and permit the import and export of database information. They also allow for displaying and formatting of returned results by directing a set of CGI and PERL scripts. A full workflow diagram of the SIEGE database

*To whom correspondence should be addressed. Tel: +1 617 638 4860; Fax: +1 617 536 8093; Email: vshah@bu.edu

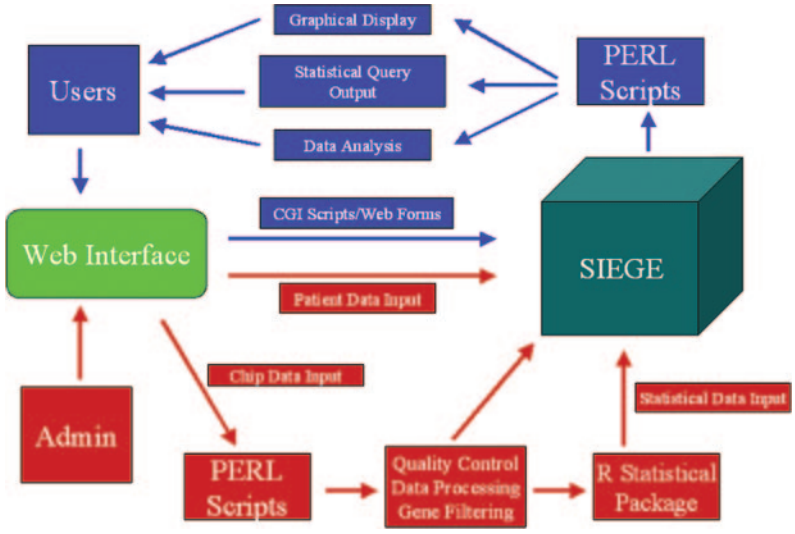


Figure 1. SIEGE Database workflow chart. This figure displays the flow of information processing that occurs when either Users or Administrators of the site access SIEGE through the web Interface. Red arrows indicate processing steps for administrative functions and blue arrows indicate steps in returning results to users.

is shown in Figure 1. Curation of the database is performed automatically through integrated software scripts encoded in PERL, and updates of newly added clinical, patient or microarray data are made available on a weekly basis. The database format provides several distinct benefits for our study. First, MySQL offers robust data storage capabilities and provides efficient query and display functionality for the millions of data points contained in our study. The SIEGE database structure allows for the seamless and accurate linkage of multiple categories of information. This includes linking a patients' clinical and sample data with the corresponding set of individual gene expression values and respective gene annotations. Second, the output from MySQL queries is easily retrieved and processed to facilitate the calculation of a wide variety of statistical tests in a dynamic manner. Furthermore, this output is in a suitable format for export to various third party statistical and visualization software. Finally, the SQL database query language enables users to write and execute user-defined database searches to power their own customized mining of our dataset.

DATABASE STRUCTURE AND CONTENT

The framework of the SIEGE database is structured around three entities: the patient, the microarray sample and the statistical values derived from the gene expression data. Each of these entities is a critical component of the study, and therefore each is described using multiple tables, which are linked together by keys of unique identifiers. The database schema portraying the entire list of table characteristics recorded for each entity and the full set of relationships between database tables is shown on its own dedicated web page (http://pulm.bumc.bu.edu/siegeDB/schema.html). SIEGE's schema conforms to the minimal information about microarray experiment (MIAME) guidelines (6).

Patient/clinical information

The database contains clinical and demographic information from our study on approximately 100 patients who are classified into three groups based on smoking histories (current-, former- and never-smokers). Upon recruitment to the study, each subject provides our study coordinator with detailed information regarding their demographic, clinical and smoking histories. Furthermore, each patient undergoes a diagnostic workup, which may include a chest X-ray, CT scan of the chest and a complete set of pulmonary function tests. This study has been approved by the Institution Review Board of the Boston Medical Center, and all subjects are coded with identification numbers in order to protect patient privacy.

Sample information

Large airway epithelial cells are obtained from brushings of the right mainstem bronchus at the time of bronchoscopy from non-smoking and smoking subjects. RNA is extracted from the brush using TRIzol reagent (Invitrogen) according to manufacturer's instructions. An aliquot of 6–8 μg of total RNA is processed, labeled and hybridized to Affymetrix HG-U133A Genechips (containing ~22 500 human transcripts) as described previously (5). A single weighted mean expression level for each gene is derived using Microarray Suite 5.0 software, along with a detection *P*-value which indicates whether the transcript is reliably detected. This information is stored in the CHIP_DATA table, linking each gene in each sample with its corresponding signal intensity and *P*-value of detection (constituting approximately three million rows of data).

Data pre-processing

After new data are entered into the database using the 'Admin' section of our site, PERL encoded scripts that we have developed in-house perform supplementary sample quality checks. These scripts evaluate three specific quality control metrics: the 3′/5′ GAPDH ratio, the percentage of genes detected as present and the percentage of outlier genes for each sample processed. Acceptable threshold values have been set for each of these parameters, and a sample must pass the quality threshold for at least two out of the three criteria to be deemed

acceptable (see Supplementary Material for more information on Quality Control procedures). Although cytological examination of selected specimens reveals that ~90% of nucleated cells obtained from the airway are epithelial, we have developed an additional gene filter to exclude specimens potentially contaminated with other cell types. A group of genes on the U133A array have been identified that should be expressed in bronchial epithelial cells and a list of genes identified that should not be expressed in airway epithelium and are specific for various lineages of white blood cells and distal alveolar epithelial cells (7). At least 80% of the present control genes should have a detection *P*-value ≤ 0.05 in good quality samples, just as at least 80% of the absent control genes must have a *P*-value detection >0.05. A PERL script automatically stores the results for all these quality metrics in the database once they are computed. This same script will immediately tag a sample that does not satisfy our checklist of quality criteria so that the database recognizes it is not to be included in further analyses.

Statistical information

One of the strengths of our database lies in the statistical data that it contains. These data have been generated by leveraging the functionality of the R statistical package (http://www.r-project.org), which is run off the same server that houses SIEGE. Embedded PERL scripts retrieve gene expression and clinical information from the database tables and feed them directly to R to perform a battery of statistical tests. Comparative statistics (*t*-test and Wilcoxon Rank Sum) evaluate the difference in the expression of each gene between the various smoker subgroups (current, former and never). Correlation measures (Pearson and Spearman) examine the potential relationship between gene expression and a number of continuous clinical variables. The R package is also utilized as a gene filter in order to remove genes that are not reliably detected. Only genes whose 20% percentile detection *P*-value is <0.05 across all samples are evaluated for the above-mentioned statistical tests.

Once gene expression data for a new sample have been added, its patient class is determined and all statistical values pertaining to the class are automatically recalculated and uploaded. This allows for the dynamic updating of all statistical test values so that users are always evaluating the most current information. Complementing the statistical data on each gene are tables providing annotation information on each given probeset including gene descriptions as well as unique Entrez Gene, Gene Ontology (GO), HUGO and GeneCards identifiers.

The construction of the statistical portion of our database on a robust package such as R has given us the power and flexibility to develop the various data-mining tools discussed in the following paragraphs.

DATABASE FEATURES

Search pages

The central database mining capability of our site is found on the Compsearch page. Within this section, it is possible to query the different statistical tables to return lists of biologically relevant genes. The *t*-test or Wilcoxon search returns genes that exhibit differences in gene expression between any comparison of smoker groups (current, former and never) that are statistically significant (user-defined; see below). A race comparison evaluating differences in expression between Caucasian and African American smokers is also available. Similarly, the Pearson and Spearman search options retrieve genes whose level of expression in a specific smoker subgroup can be correlated to clinical variables such as degree of tobacco exposure (measured in pack-years), age and lung function [i.e. forced expiratory volume (FEV1%) and forced vital capacity (FVC%)]. Significance thresholds for these tests are defined by the user, and in the case of the *t*-Test search option, the thresholds can be applied to filter based on the *P*-value, *Q*-value or permutation *P*-value. The *Q*-value and permutation *P*-value are calculated to correct for the multiple comparison problem inherent in any microarray analysis. The *Q*-value measures the proportion of false positives incurred (called the false discovery rate) for a particular *P*-value threshold (8). The permutation *P*-value is determined based on repeated random trials, which measure the likelihood of obtaining significant *t*-test *P*-values by chance alone (for additional details see Supplementary Material). Further streamlining of the search results can be achieved by restricting the output using a specific GO keyword. Sorting of the results is possible by Affymetrix ID, *P*-value, correlation score, gene expression fold change or chromosomal location of the gene. Furthermore, genes returned from statistical queries are linked to sites on external databases that contain additional gene-specific information including Entrez Gene, GO (9), Biocarta, GeneCards (10), dbSNP (11) and NCBI's Map Viewer (see Figure 2).

Information on a particular set of genes of interest to the user can be found on the Gene Reference page, which provides 'gene-centric' tools to find the relevant gene annotation, expression and statistical data. This includes tools to display all statistical results for a list of given Affymetrix identifiers as well as the ability to output the specific gene's expression behavior across any patient subgroup.

The Advanced Search page and accompanying Schema page aim to provide users with the ability to interactively access and mine data from SIEGE. With a basic knowledge of SQL, users are able to customize their own personal searches and thus are not limited to the tools/searches we have developed and made available. Output from Advanced Search queries can be easily downloaded in tab-delimited text file format.

Visualization

The ability to visualize the data graphically is made possible through display tools integrated within the various search options. Clicking on the Affymetrix IDs of genes returned from any *t*-test or Wilcoxon based search will produce a color-coded bar graph showing the gene expression behavior of samples from all three patient groups combined (see Figure 2). Similarly, clicking the Affymetrix ID of a gene returned from a Pearson- or Spearman-based search will produce a scatterplot of gene expression versus the variable being correlated for all patients in the specified patient group. On the 'Graph' section of our website, the user can correlate and produce scatterplots

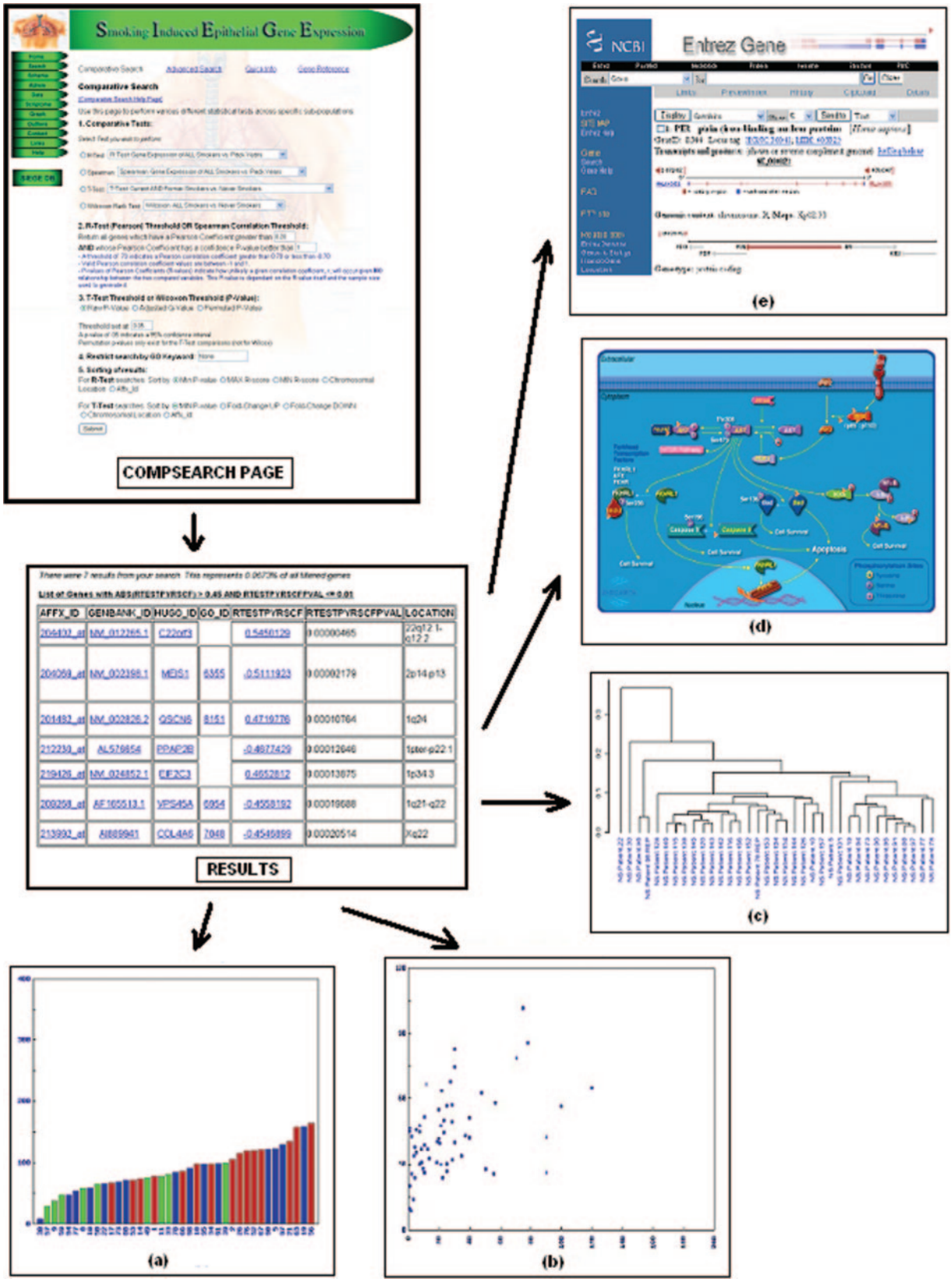


Figure 2. Comparative search function options. Statistical query results obtained by using the Compsearch section of the database website can be linked directly to (**a**) expression bar graph for given gene, with expression level on the *y*-axis and Patient ID number on the *x*-axis. Subjects are color coded for smoking status (blue, never-smokers; green, former-smokers; and red, current-smokers); (**b**) scatter plot for correlation analysis, with expression level on one axis and the associated continuous variable on the other; (**c**) hierarchical cluster of never-smoker samples based on Pearson correlation of expression levels of a set of genes; (**d**) Biocarta Pathway diagram for a given gene; and (**e**) Entrez Gene entry for a given gene.

of all gene expression values between two different samples. This ability is important for two reasons: (i) it permits comparison of gene expression behavior between two patients within the same patient class or across classes and (ii) it allows assessment of the reproducibility of technical and biological replicate samples, an important quality control feature in a microarray experiment. There are three types of replicates in our study: technical replicates (RNA divided into two and processed separately), spatial replicates (RNA obtained from two slightly different locations in the bronchi) and temporal replicates (RNA obtained from the same patient three months apart). Finally, samples can be compared visually on the Graph page using Hierarchical Clustering. All samples or a subset class of patient samples can be clustered using any of a number of clustering methods and distance metrics provided. Similarly, the gene set used for clustering can be established using expression variability criteria such as standard deviation/mean expression ratios, Max/Min expression ratios and others. Alternatively, the entire list of filtered genes or a user supplied gene list can be clustered. The final clustering dendogram generated is in the PDF format and can also be downloaded.

Data download

In order to allow users the ability to completely reanalyze our results, all data from our study has been made downloadable through the 'Data' section of our site. All expression data files for each sample, including the raw .DAT files generated prior to analysis by Affymetrix Microarray Suite 5.0 software, are available in zipped format. Demographic and clinical data on each patient recruited into the study is also provided. Up-to-date data dumps of all four statistical tables (*t*-test, Wilcoxon, Spearman and R-test) are downloadable in tab-delimited format. In addition, quality control data that is stored in the database is presented for user verification.

Transcriptome

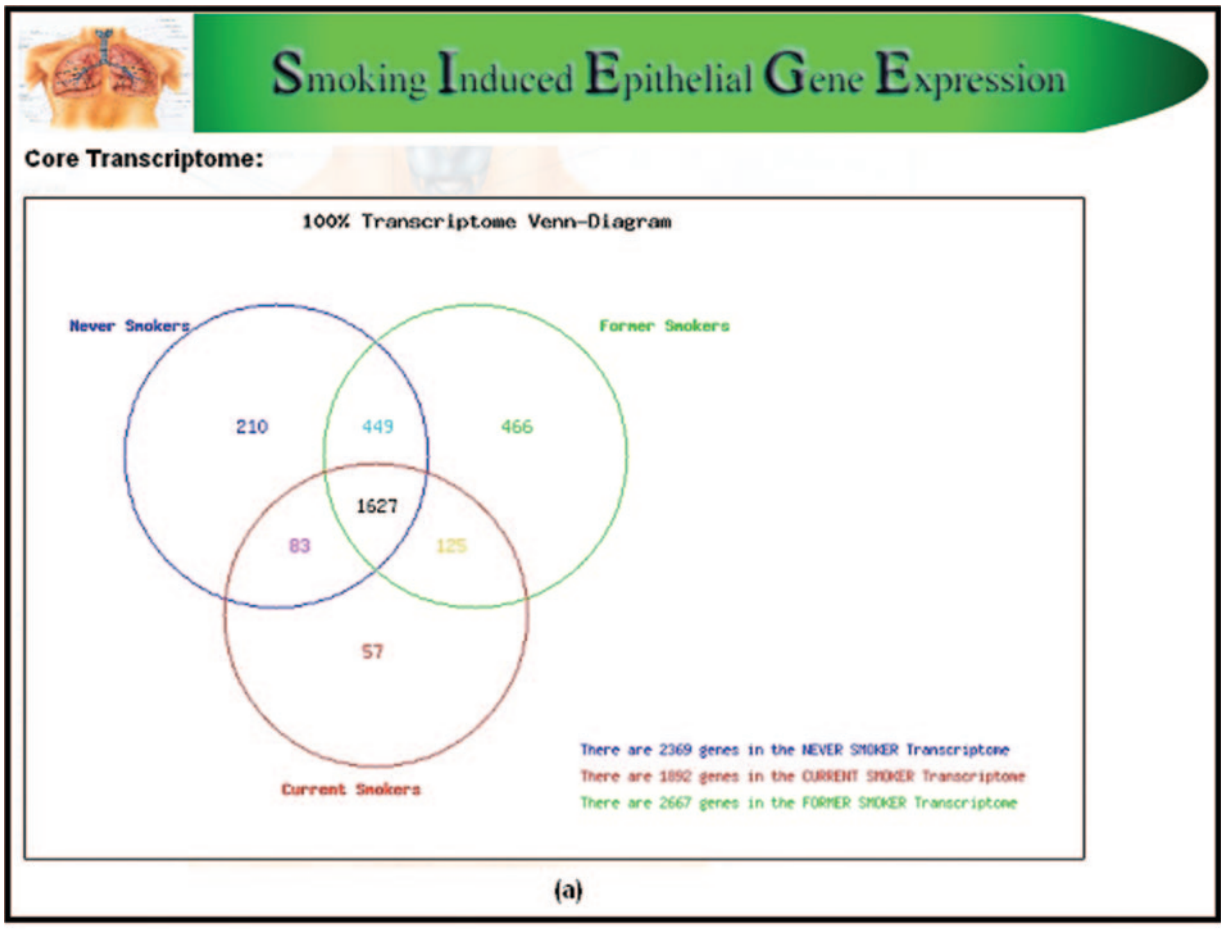
One of the primary objectives of this translational research project is the definition of the normal airway transcriptome, and these data can be accessed on the 'Scriptome' page of the website. In our study, the airway transcriptome is defined as the set of genes that are expressed in airway epithelium at levels that are accurately measured with microarray technology (for further details see Supplementary Material). In order for a gene to qualify as 'expressed' in our study, the *P*-value of detection for that gene in that particular sample has to be <0.05. Transcriptomes are defined using respective patient class, and thus the normal airway transcriptome is generated from the never-smoker samples. The 100% transcriptome for any patient class is the set of genes that are shown to be expressed in EVERY sample of that class. Similarly, the 50% transcriptome is made up of genes that were expressed in at least 50% of the samples of that patient class. Each of the transcriptomes is searchable through the 'Scriptome' page of the website. Selecting any of the links on the first section of the Scriptome page will generate a Venn diagram showing the number of genes that fall into each transcriptome, as well as those that intersect across transcriptomes. A text file output of genes in each cell of the Venn diagram is available for download. Genes within a given transcriptome can be further filtered using any of four different expression variability criteria or through GO keywords. Finally, the Scriptome page also presents results from an analysis of functional categories represented by the genes in a transcriptome. This analysis was performed using the GOMINER software, which classifies the genes into the biological categories outlined by the GO consortium (9) and assesses whether any of these categories are statistically over- or under-represented in a given transcriptome (12). GOMINER output is shown on our site in two forms: (i) a diacyclic graph (DAG), which is an interactive graphical file showing the hierarchical structure of GO categories color coded for over- and under-represented categories. (ii) A tab-delimited file listing each GO category, the number of times it was represented in a given transcriptome and a Fisher's exact *P*-value relating the statistical significance of that particular categories over- or under-representation when compared to a background set of genes (see Figure 3).

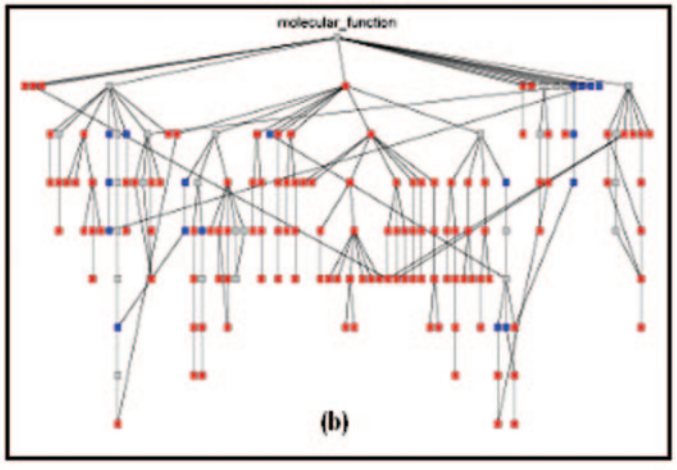
Outliers

Genetic alterations induced by cigarette smoke are not uniform across all smokers (13). Subgroups of smokers may be inherently more susceptible to tobacco-induced genetic damage leading to aberrations in gene expression behavior. These aberrations may over time lead to diseases such as COPD or lung cancer (14). In order to identify smokers who may not be responding to tobacco appropriately, we have begun to catalog genes that exhibit outlier expression for different subsets of the 43 current-smokers currently represented in our database. Outlier expression is determined by using the Grubbs test for Outliers (see Supplementary Material). With this page, users are able to search for and view the number of outlier samples/patients as well as outlier expression level for a given gene of interest. Alternatively, the user can retrieve a list of all genes that have more than the specified number of outliers. More advanced tools for the Outliers section are under development.

FUTURE DIRECTIONS

SIEGE will continue to add gene expression data through the acquisition of additional patient samples from Boston University Medical Center as well as several additional medical centers (see Supplementary Material). The processing of greater numbers of microarray samples will provide sufficient power to run advanced statistical tests including ANCOVA and other multivariate analyses. These tests will aid in discerning the subtle differences in airway gene expression that are modulated based on race, sex and age. One of the most important future directions for SIEGE will be the inclusion of airway microarray samples from current- and former-smokers with lung cancer. Comparing this new set of airway samples with the previously obtained samples from smokers without lung cancer may allow for the identification of gene expression profiles that could serve as a clinically applicable diagnostic test for distinguishing smokers with and without lung cancer. The long-term goal of our database is to facilitate the





GO_ID	TOTAL	CHANGED	PVALUE	GO_DESCRIPTION
7155	330	45	0.05430	cell adhesion

(c)

Figure 3. Transcriptome analysis: (**a**) Venn diagram of 100% transcriptome displaying number of genes in each smoker class transcriptome. (**b**) DAG of molecular function GO categories represented in 100% transcriptome. (**c**) Statistical *P*-value of over- or under-representation of a given GO category in 100% transcriptome.

development of airway gene expression profiling as a relatively non-invasive tool for assessing the degree of epithelial cell damage in the airway and risk for the development of lung cancer among smokers. Our database will provide researchers with valuable insight into the biological mechanisms/pathways by which cigarette smoke damages airway epithelium and provide clinicians and epidemiologists with novel biomarkers for smoking-induced lung diseases.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Michael Schaffer for his expert advice on the use of appropriate database software and server configuration. This work was supported in part by a Doris Duke Charitable Foundation Clinical Scientist Development Award (A.S) and by the National Institutes for Health Grant HL47049. Affymetrix provided the U133A arrays for these studies.

REFERENCES

1. Proctor,R.N. (2004) The global smoking epidemic: a history and status report. *Clin. Lung Cancer*, **5**, 371–376.
2. Shields,P.G. (1999) Molecular epidemiology of lung cancer. *Ann. Oncol.*, **10** (Suppl. 5), S7–S11.
3. Wistuba,I.I., Lam,S., Behrens,C., Virmani,A.K., Fong,K.M., LeRiche,J., Samet,J.M., Srivastava,S., Minna,J.D. and Gazdar,A.F. (1997) Molecular damage in the bronchial epithelium of current and former smokers. *J. Natl Cancer Inst.*, **89**, 1366–1373.
4. Powell,C.A., Spira,A., Derti,A., DeLisi,C., Liu,G., Borczuk,A., Busch,S., Sahasrabudhe,S., Chen,Y., Sugarbaker,D. *et al.* (2003) Gene expression in lung adenocarcinomas of smokers and nonsmokers. *Am. J. Respir. Cell Mol. Biol.*, **29**, 157–162.
5. Spira,A., Beane,J., Shah,V., Liu,G., Schembri,F., Yang,X., Palma,J. and Brody,J.S. (2004) Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl Acad. Sci. USA*, **101**, 10143–10148.
6. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al.* (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genet.*, **29**, 365–371.
7. Hermans,C. and Bernard,A. (1999) Lung epithelium-specific proteins: characteristics and potential applications as markers. *Am. J. Respir. Crit. Care. Med.*, **159**, 646–678.
8. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
9. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
10. Safran,M., Chalifa-Caspi,V., Shmueli,O., Olender,T., Lapidot,M., Rosen,N., Shmoish,M., Peter,Y., Glusman,G., Feldmesser,E. *et al.* (2003) Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.*, **31**, 142–146.
11. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
12. Zeeberg,B., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.
13. van Oijen,M.G., van de Craats,J.G. and Slootweg,P.J. (1999) p53 overexpression in oral mucosa in relation to smoking. *J. Pathol.*, **187**, 469–474.
14. Anderson,G.P. and Bozinovski,S. (2003) Acquired somatic mutations in the molecular pathogenesis of COPD. *Trends Pharmacol. Sci.*, **24**, 71–76.