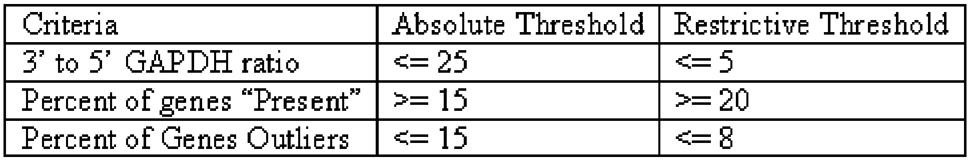
Supplementary Material

Quality Control Procedure:

The first step of the quality control procedure involves visually reviewing each scanned microarray sample image to ensure that the chip has no visible defects such as bubbles, smears, or scratches. Secondly, PERL-encoded scripts that are integrated with SIEGE evaluate three separate quality parameters for each chip: 3'/5' GAPDH ratio, Percent Present and Percent Outliers. The 3'/5' GAPDH ratio is a measure of RNA degradation and the Percent Present evaluates the percent of genes on the microarray that have been accurately measured as per the detection p-value (P value <= 0.05). Percent Outliers calculates the number of genes per sample whose expression levels are greater than two standard deviations above/below the average across all samples. We established quality cut-off thresholds for each of the above parameters as shown in the table below:

Table 1: Quality Parameter Thresholds



Criteria	Absolute Threshold	Restrictive Threshold
3' to 5' GAPDH ratio	<= 25	<= 5
Percent of genes "Present"	>= 15	>= 20
Percent of Genes Outliers	<= 15	<= 8

For a sample to be deemed acceptable it must satisfy ALL three of the absolute thresholds and at least 2 out of 3 of the Restrictive Thresholds.

The second quality filter we employ detects hybridization of RNA contaminated from non-epithelial cells. Through review of the literature, we have established a list of present and absent control genes, which are shown in Table 2 and 3. 80% of the present control genes need to be present (P value of detection <= 0.05) in good quality samples while at least 80% of absent control genes need to be absent (P value of detection > 0.05).

Table 2: Genes present in Bronchial Epithelial Cells

Affymetrix ID/Gene

207847_s_at = muc 1 transmembrane
204895_x_at = muc 4
213693_s_at = muc 1 transmembrane
214303_x_at = muc 5 AC
214385_s_at = muc 5 B
217109_at = muc 4
217110_s_at = muc 4
205725_at = clara cell
209008_x_at = cytokeratin 8

Table 3: Genes absent in Bronchial Epithelial Cells

Affymetrix ID/Gene

205049_s_at = cd-79 alpha (B cell)
205297_s_at = cd-79 beta (B cell)
205264_at = cd-3 epsilon associated protein (T cell)
210031_at = cd-3 zeta (T cell)
206804_at = cd-3 gamma (T cell)
205982_x_at = surfactant associated protein C (distal alveolar epithelium)
211735_x_at = surfactant associated protein C(distal alveolar epithelium)
214199_at = surfactant associated protein D(distal alveolar epithelium)
218835_at = surfactant associated protein A2(distal alveolar epithelium)
215454_x_at = surfactant associated protein C(distal alveolar epithelium)
203948_s_at = MPO (myeloid/monocyte)
203949_at = MPO (myeloid/monocyte)
206120_at = CD33(myeloid specific)
210184_at = CD11c(myeloid specific)
204625_s_at = CD61 (megakaryocyte)
211579_at = CD61 (megakaryocyte)

Multiple Comparison Correction:

One of the central statistical analyses we performed in our study was a Two-sample unequal variance Student's T-test. Due to the large number of comparisons between expression values of each microarray transcript across two sample sets, there is an increased chance that the Student's T-test will identify genes as being differentially expressed when no real difference exists. This is known as the Multiple Comparison problem. We have employed two statistical methods to correct for the multiple comparison problem in our study: 1) Q-value correction and 2) Permutation based correction. The Q-value score is a correction that was performed by using the Q-VALUE software program, which can be downloaded from http://faculty.washington.edu/~jstorey/qvalue (1). The Q-value measures the proportion of false positives incurred (called the false discovery rate) for a particular P-value threshold, and it is estimated using a histogram analysis of an entire list of P-Values resulting from the simultaneous testing of many hypotheses. The Permutation based approach employed a PERL script we developed in-house to strictly control the probability of having even one non-differentially expressed gene pass our significance threshold by chance alone. After calculating the actual Student T-Test P-value for each gene, we randomly assigned samples to the two different groups being compared and evaluated this randomized Student T-Test P-value 1000 separate times. The Permutation corrected P-value for each gene is the number of times out of 1000 that a randomly generated P-value better than the actual P-value for that gene.

Grubbs Outlier Test:

The Grubbs Test, which is also known as the maximum normed residual test, is used to identify outlier values in a univariate set of data. The test assumes normality and so care must be taken to ensure that the values in the dataset being tested reasonably approximate a normal distribution. The Grubbs Test works to identify outliers in an iterative manner by first evaluating the dataset value that is furthest away from the dataset mean. It first calculates the Grubbs statistic which is: $G = \frac{\max|A_i - \overline{A}|}{std}$ where A_i is the value being tested and $\bar{A}$ the dataset mean. The hypothesis of no outliers is rejected if:

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/(2N),N-2)}}{N-2+t^2_{(\alpha\ /\ (2N),\ N\ -\ 2)}}}$$ where N = number of values in dataset

and $t^2_{(\alpha\ /\ (2N),\ N\ -\ 2)}$ is the critical value of the t-distribution with N-2 degrees of freedom and significance level of α/2N. If an outlier is detected, this value is eliminated and another round of Grubbs testing begins with the remaining dataset. This continues until no new outliers are detected. It is not recommended to use the Grubbs Test on dataset with N less than 6.

Definition of the Normal Airway Transcriptome:

Our method of defining the normal airway transcriptome relies on identifying genes with low detection P values (P value <= 0.05) in a majority if not all microarray samples. One of the limitations of microarray technology is that it does not accurately measure expression levels of lowly expressed genes. This results in a relatively high detection P value for genes that are expressed at low levels in vivo. Therefore, our approach will result in a transcriptome gene list that will be biased against lowly expressed genes and weighted towards abundantly expressed genes.

Additional Sample Collection Sites:

In addition to collecting bronchial airway epithelial cells from subjects at Boston University Medical Center, we have also recruited patients from 2 additional medical centers into our study:

Lahey Clinic:

41 Mall Road
Burlington, MA 01805

St James's Hospital:

James's Street, Dublin 8
Ireland

References:

1. Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci.*, **100**, 9440-9445
2. Grubbs, F. (1969) Procedures for Detecting Outlying Observations in Samples. *Technometrics*, **11(1)**, 1-21.