COMPARISON OF SMOKING-INDUCED GENE EXPRESSION ON AFFYMETRIX EXON AND 3'-BASED EXPRESSION ARRAYS

XIAOLING ZHANG[1]
zhangxl@bu.edu

GANG LIU[2]
gliu@lung.bumc.bu.edu

MARC E LENBURG[3]
mlenburg@bu.edu

AVRUM SPIRA[2]
aspira@bu.edu

[1]*Bioinformatics Program, Boston University, 24 Cummington Street, Boston, Massachusetts 02215, USA*
[2]*The Pulmonary Center, Boston University Medical Center, 715 Albany Street, Boston, Massachusetts 02118, USA*
[3]*Department of Genetics and Genomics, Boston University, 715 Albany Street, Boston, Massachusetts 02118, USA*

Correspondence should be addressed to Avrum Spira aspira@bu.edu

Cigarette smoke is the major cause of lung cancer and chronic obstructive pulmonary disease in the United States. We have previously defined the impact of tobacco smoke on intrathoracic airway gene expression among healthy nonsmokers and smokers using standard 3'-based expression U133A arrays [12]. In this report, we compared the performance of the Affymetrix GeneChip Human Exon 1.0 ST array with the HG-U133A array for detecting smoking-related gene expression changes in large airway epithelium obtained at bronchoscopy. RNA obtained from the same bronchial airway epithelial cell samples of four current smokers and three never smokers was hybridized to both arrays. Out of 22,215 probe sets on HG-U133A, 14,741 RefSeq transcripts were mapped to 17,800 core transcripts on the Exon array and the 2 platforms were compared for this overlapping transcript set. While the reproducibility of both platforms was high, the Exon array had a slightly stronger correlation for technical replicates. A majority of the genes with the largest smoking-related fold changes were tightly correlated between platforms, but there were a number of smoking-related changes in gene expression that were detected only on the Exon arrays. Furthermore, while the HG-U133A study did not have enough power to detect any differentially expressed genes between the 4 current vs. 3 never smokers at a False Discovery Rate (FDR) < 0.05, seventy differential expressed genes were detected at FDR < 0.05 in the same set of samples using the Exon platform. These findings suggest that the all-Exon array is a more robust platform for measuring airway epithelial gene expression and can serve as an effective tool for exploring host response to and damage from cigarette smoke.

Keywords: cigarette smoke, Affymetrix exon arrays, gene expression, airway epithelium.

1. Introduction

Although cigarette smoking is well recognized as the major cause of lung cancer and chronic obstructive pulmonary disease (COPD) [7], only 10-20% of smokers actually develop these diseases [10]. It is unclear why some smokers remain healthy while others remain at high risk decades even after they have quit [5]. Unfortunately, there are

currently no effective tools for identifying smokers at highest risk for developing tobacco-related lung disease.

Based on the concept that cigarette smoke creates a “field of injury” in epithelial cells throughout the respiratory tract, we have previously measured genome-wide gene expression in large airway epithelium obtained at bronchoscopy in order to gain insights into host response to and damage from smoking. Using the HG-U133A GeneChip array (Affymetrix, Santa Clara, CA), we have defined the impact of tobacco smoke on intrathoracic airway gene expression among healthy nonsmokers and smokers and have demonstrated that a subset of smoking-induced changes persist years after subjects stop smoking [12]. Further, we recently developed a profile of airway gene expression that can distinguish smokers with and without lung cancer and serve as an early diagnostic biomarker for disease [13]. The above studies, however, were limited in terms of their widespread clinical application due to the amounts and quality of RNA needed for the U133A array. That platform requires 4-8ug of total RNA as starting material (a challenge for clinical brushings and biopsy material) and does not perform well in setting of partially degraded RNA. As a result, approximately 10-20% of samples obtained at the time of bronchoscopy in our prior studies were of insufficient quality and/or quantity for microarray studies.

The recent availability of a new platform, the GeneChip Human Exon 1.0 ST array (Affymetrix, Santa Clara, CA), provides us with an opportunity to get more comprehensive and more reliable genome-wide measurements requiring only 1 μg of starting RNA. The Exon array contains >5million 25mer probes, forming 1.4 million probe sets that are together used to separately interrogate 1 million known and predicted exons [16]. This novel platform offers several key advantages over standard versions of GeneChip arrays: 1) It provides more robust measurements at the transcript level due to more probes per transcript; there are roughly 30-40 probes for each RefSeq transcript, as compared to 11 probes mostly at the 3' end in the U133 array [6]. 2) It has the potential to provide more accurate measurements in the setting of partially degraded mRNA as the probesets are distributed along the entire length of the gene (as opposed to the 3’ end); 3) It has the potential to distinguish between different isoforms of a gene at the level of individual exons, and thus allows the opportunity to identify alternative splicing events that may play important role in host response to smoking.

There is very limited data comparing the robustness and reproducibility of this novel platform and the traditional U133 arrays, particularly in the setting of clinical samples. Okoniewski *et al* [8] reported that irrespective of the mapping methodology applied, Exon 1.0 ST and HG-U133 Plus2 arrays show a high degree of correspondence by hybridizing RNA samples from two human cell lines in triplicate (technical replicates) to both arrays. Gardina *et al* [6] have also demonstrated a reasonable correlation in signals for genes that are significantly differentially expressed between tissue types utilizing a subset of data (3 replicates each tissue) from a panel of 11 normal tissues that was assayed in parallel on both Exon arrays and HG-U133 Plus2 arrays. In this study, we performed a systematic comparison of gene signal estimations from the Exon 1.0 ST and

the U133A arrays by hybridizing the same bronchial airway epithelial RNA obtained from smokers and nonsmokers to both arrays. Our data suggests that while both platforms show a high degree of correlation for detecting smoking-related changes, the all-exon array is a more robust platform for the genome-wide study of smoking-related gene expression changes that occur in airway epithelium.

2. Material and Methods

2.1. *Study Population and Sample Collection.*

In our previous study [12], we recruited nonsmoking and smoking subjects (n = 93) to undergo fiberoptic bronchoscopy at Boston Medical Center (between November 2001 and June 2003). We obtained a sufficient quantity of good-quality RNA for U133A studies from 85 of the 93 subjects recruited into the study. 10 out of 85 samples were excluded based on a quality control filter, and 18 of the remaining 75 samples were collected from former smokers. As a result, there were 34 current smokers and 23 never smokers in our previous analysis [12]. Seven of these 57 subjects (4 smokers and 3 nonsmokers) had sufficient amount of leftover RNA (1ug) to be used in our current study comparing the U133A and the Exon arrays. Of note, there were no significant differences ($P > 0.05$) in age (mean age 40.7±12.7 vs. 39±13.3), race (3 African American (AFA), 2 Caucasian (CAU), 2 Others vs. 24 AFA, 19 CAU, 14 Others), gender (5 male, 2 female vs. 44 male, 13 female) and cumulative tobacco exposure (mean pack-years 39.4± 14 vs. 21.8 ± 20.6) between these 7 subjects and the other 50 subjects not included in this study. The Institutional Review Board at Boston Medical Center approved this study, and all participants provided written informed consent.

Bronchial airway epithelial cells were obtained from brushings of the right mainstem bronchus taken during fiberoptic bronchoscopy with an endoscopic cytobrush (Cellebrity Endoscopic Cytology Brush, Boston Scientific, Boston). After removal from the bronchoscope, the brushes were immediately placed in TRIzol (Invitrogen) and kept at -80° C until RNA isolation was performed. RNA was extracted from the brush using TRIzol reagent (Invitrogen) as per the manufacturer protocol. Integrity of the RNA was confirmed by denaturing gel electrophoresis. Epithelial cell content of representative bronchial brushing samples was quantitated by cytocentrifugation (ThermoShandon Cytospin, Pittsburgh, PA) of the cell pellet and staining with a cytokeratin antibody (Signet, Dedham MA). Using this protocol, we were able to obtain 8-15 μg of total RNA from the intrathoracic airway of these 7 subjects.

2.2. *Microarray Data Acquisition and Preprocessing*

2.2.1. *Affymetrix HG-U133A GeneChips*

6-8 μg of total RNA was processed, labeled, and hybridized to Affymetrix HG-U133A GeneChips containing 22,215 probesets (for detailed protocol, see [12]). Eight CEL files of 4 current smokers and 3 never smokers profiled on U133A arrays as part of our previous study [12] were retrieved, which includes 1 technical replicate from one of the 3 never smokers.

The Robust Multichip Average (RMA) algorithm [2] was used for background adjustment, normalization, and probe-level summarization of the microarray samples. RMA expression measures were computed using the R statistical package and the GCRMA function in the 'affy' Bioconductor package.

2.2.2. *Affymetrix GeneChip® Human Exon 1.0 ST array*

1 μg of residual RNA from bronchial airway samples for the same subjects (4 current smokers and 3 never smokers) were processed, labeled and hybridized to Affymetrix Human Exon ST 1.0 arrays as described below. Technical replicates were run on one of the current and one of the never smoker samples.

Using a random hexamer incorporating a T7 promoter, double-stranded cDNA was synthesized from 500 ng total RNA in which the majority of the ribosomal RNA had been removed using a RiboMinus Human/Mouse Transcriptome Isolation Kit (Invitrogen, Carlsbad, CA). cRNA was generated from the double-stranded cDNA template though an in-vitro transcription reaction and purified using the Affymetrix sample cleanup module. cDNA was regenerated through a random-primed reverse transcription using a dNTP mix containing dUTP. The RNA was hydrolyzed with RNase H and the cDNA purified. The cDNA was then fragmented by incubation with a mixture of UDG and APE 1 restriction endonucleases; and end-labeled via a terminal transferase reaction incorporating a biotinylated dideoxynucleotide. 5.5 μg of the fragmented, biotinylated cDNA was added to a hybridization cocktail, loaded on a Human Exon 1.0 ST GeneChip and hybridized for 16 hours at 45 ºC and 60 rpm. Following hybridization, the array was washed and stained according to the Affymetrix protocol. The stained array was scanned at 532 nm using an Affymetrix GeneChip Scanner 3000, generating CEL files for each array.

Exon-level expression values were derived from the CEL file probe-level hybridization intensities using the model-based RMA algorithm as implemented in the Affymetrix Expression Console™ software. RMA performs normalization, background correction and data summarization. Whether an exon probeset is reliably detected over background in each sample is determined from the percentile rank of the hybridization intensity of the probes relative to the hybridization intensities observed for 1000 pooled "antigenomic" probes with the same GC content as each perfect match probe (since the Exon array does not include a paired mismatch probe for each perfect match probe) [17].

Each probe's percentile rank is combined using the Fisher Test to determine an overall probability that each probeset is expressed above background. This analysis is performed using the DABG algorithm as implemented in Affymetrix Expression Console, and a p-value threshold of $p < 10^{-5}$ is used as the criterion for expression over background. Transcript-level expression values are also derived using RMA algorithm as implemented in Affymetrix Expression Console.

2.3. *Microarray Data Analysis*

Exon arrays were designed by selecting a diverse set of genomic annotations, including empirically supported and predicted transcripts. By doing so, it enables the discovery of novel transcriptional events. On the other hand, probes that may target incorrectly predicted exons (genes) could increase the exon identification errors and transcript misassignments, which could potentially decrease power to detect differential expression. As a result, in our current study, we focused our analyses on the approximately 230,000 "core" exon probesets that have been mapped to approximately 17,800 empirically supported core transcripts (RefSeq and full-length GenBank mRNAs [6]) with a high degree of confidence.

All statistical analyses below were performed with R 2.4.0 (available at http://r-project.org). The gene annotations used for each probe set were from the October 2003 NetAffx HG-U133A, and HuEx-1_0-st-transcript-annot.affy.csv.

Technical replicates were obtained from selected subjects. Pearson correlations were calculated for technical replicate samples from the same individual on both Exon 1.0 ST and U133A arrays. To compare the fold change of smoking-induced gene changes on both arrays, we considered the RefSeq mappings [3] as a nonredundant and relatively complete database of transcripts [9]. By mapping 22,215 U133A probe sets to 17,800 core transcripts on the Exon array, 14,741 RefSeq transcripts are found on both platforms. Fold changes between the log2 mean values for the smokers and nonsmoker replicates were calculated independently for Exon and U133A arrays.

Using a number of statistical methods, we also identified differentially expressed genes (DEGs) on the U133A and Exon arrays from the same subjects independently. For the U133A array, 3 common methods for identifying differentially expressed genes were applied: significance analysis of microarrays (SAM) [14], Limma [11], and Student's *t*-test. The estimated false discovery rate (FDR) for each of these analyses was calculated using the Benjamini and Hochberg approach [1] in order to correct for multiple comparisons. Due to the Exon array's unique ability to support exon-level data in which each transcript has several measurements, we were able to utilize an ANOVA model that has a term for smoking status and another that accounts for differences in exon-to-exon expression level in order to detect differential transcript-level expression on this platform. We could not apply the ANOVA model to the U133A data since only one summarization signal is available for each transcript on that platform.

For the Exon array, we leveraged the power of this platform to detect differential transcript-level expression using traditional linear model approaches to predict the observed expression level of each exon that include terms for experimental variables, terms to account for differences in exon-to-exon expression level, and a term to account for multiple measurements of transcript abundance being made from each sample. We assess the significance of the effect of experimental variables using ANOVA and correct for multiple comparisons using the Benjamini and Hochberg approach [1]. This type of modeling approach is similar to what has been proposed by others for analyzing exon array gene-expression data [15]. Additionally, we removed exon probesets that exhibited invariant expression. Measurements from exon probesets that are not significantly expressed over background were first removed by excluding probesets that are below background in > 90% of samples. Exon probesets that are expressed over background but display low variance across all samples were further removed by comparing the variance of each exon to the average variance of all exons in that transcript using a Chi-Square test.

In order to identify functional categories that were overrepresented within the genes differentially expressed between current vs. never smokers, DAVID software [4] was used to functionally classify these genes by the molecular function categories within gene ontology using total 30, 000 human genes as population background. Fisher Exact P-value was used to measure the gene-enrichment in annotation terms (for details, see [4]).

To further characterize the smoking-related genes, 2D hierarchical clustering of all never and current smokers using the differentially expressed genes was performed. Hierarchical clustering of the genes and samples was performed by using log-transformed *z*-score normalized data with a Pearson correlation (uncentered) similarity metric and average linkage clustering with CLUSTER and TREEVIEW software obtained at http://rana.lbl.gov/EisenSoftware.htm

3. Results and Discussion

3.1. *Technical Replicates comparison on the U133A and Exon arrays*

Technical replicates were run on 1 current and 1 never smoker sample. Overall transcript abundance estimates from the Exon array demonstrated a slightly higher level of reproducibility (r = 0.989) as compared to transcript abundance estimates from U133A technical replicates (r = 0.984) (see Fig. 1). However, due to the limited number of technical replicates, we were unable to get confidence intervals for the correlation coefficients in order to assess the statistical significance of this difference.

3.2. *Gene-level fold change comparison on U133A and Exon arrays*

The Exon array can serve as a gene-level expression array by summarizing multiple probes signals on different exons into an expression level of all transcripts from one gene. Since RefSeq is known to be a nonredundant and relatively complete database of high

confident transcripts [9], we compared the tobacco-smoke induced gene level changes for 14,741 RefSeq transcripts found on both platforms. This approach allowed us to compare the fold change of gene level signals from the same 4 current smoker vs. 3 never smoker samples assayed on both arrays (see Fig. 2). In this scatter plot, each point corresponds to a pair of probe sets for which a successful cross-chip mapping could be found. Ideally, fold changes would be identical, and all points would fall on the major diagonal of the scatter plot. While the majority of gene-expression differences detected correlate tightly between platforms (r=.62; p-value < $2*10^{-16}$), there were a number of smoking-related gene expression changes that were detected only on the Exon array (see colored circles along the horizontal plane in Fig. 2). This suggests that the U133A array may lack probesets to measure expression of these particular transcripts. Gardina *et al*. [6] also observed a shift of apparently low expressing genes on the U133 Plus 2 array to higher signals on the Exon array. In order to evaluate whether these results extended to the other U133A samples collected as part of our original study [12], we compared smoking-related fold changes in all 57 samples from that study to those detected on the exon array in seven subjects. These results reinforced the strong correlation between the 2 platforms (r=.4; p-value < $2*10^{-16}$) even in the setting of unmatched samples and once again demonstrated that there are a number of smoking-related changes only detected by the exon array (data not shown).

3.3. *Gene-level differential expression changes on U133A and Exon arrays*

To examine the effect of smoking on the U133A array, three statistical methods (SAM, Limma and a two-sample Student's *t*-test) were used to test for genes differentially expressed between current (n = 4) and never smokers (n = 3). After correcting for multiple testing using the Benjamini and Hochberg FDR [1], there were no significant genes found at FDR < 0.05. We also randomly sampled, 1000 times, 4 current vs. 3 never smokers from the complete set of 57 subjects in our previous U133A study[12] and were unable to find any genes that passed the FDR cutoff in this analysis.

As mentioned in *material and methods*, we focused our analyses on the approximately 230,000 "core" exon probesets that have been mapped to approximately 17,800 core transcripts on the Exon array. With the Exon arrays, we are able to use an exon-level mixed-model ANOVA to determine whether a gene is differentially expressed at the transcript level. As a result, when comparing 4 current vs. 3 never smokers analyzed on the Exon array, we found 70 differentially expressed genes at p<0.001 (FDR<0.05), while analysis of the same samples on the U133A failed to yield any significantly differentially expressed genes. An example of how the exon-level measurements aid in the identification of differentially expressed genes is shown in Fig. 3. These results suggest that the Exon array is a more robust gene expression platform, yielding more power for detecting differential expression for a given sample size.

To further characterize the genes identified as differentially expressed on the Exon array, hierarchical clustering of all 7 subjects was performed using exon gene expression measurements for these 70 DEGs (see Fig. 4). Current and never smokers grouped with their appropriate classes according to the expression of these 70 genes.

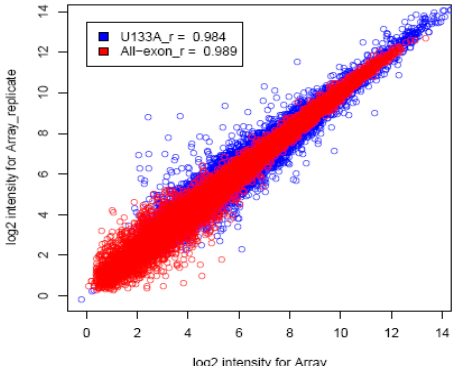

Fig. 1. Correlation between technical replicates of all-exon and U133A arrays demonstrates that all-exon arrays produce slightly more accurate estimates of transcript abundance.

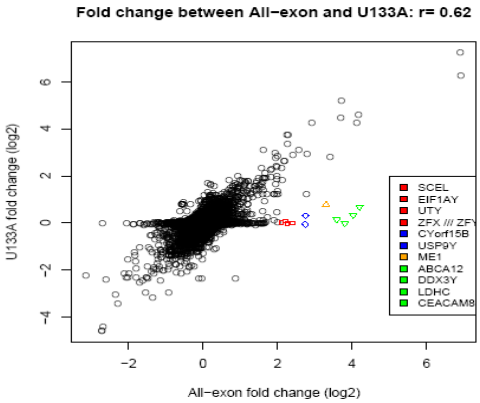

Fig. 2. Correlation between smoking induced gene-expression differences detected by all-exon (x-axis) and U133A (y-axis) arrays demonstrates that the all-exon arrays detect gene expression differences that are not detected on the U133A array. The highlighted gene expression differences (colored circles) are greater than 4 fold and detected only on the all-exon array.

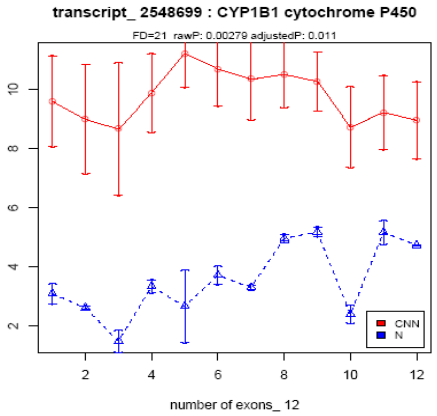

Fig. 3. Exon-level expression estimates of a gene that is detected as significantly differentially expressed on the all-exon array, but does not reach statistical significance when measured on the U133A array in the same 4 current (red: CNN) vs. 3 never smokers (blue: N).

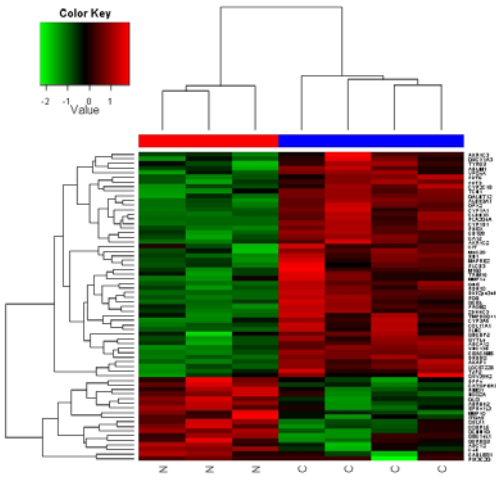

Fig. 4. Hierarchical clustering of 4 current smokers and 3 never smokers according to the expression of the 70 genes differentially expressed between current and never smokers on the Exon array. Current and never smokers are separated into their appropriate classes.

Table 1 depicts the DAVID functional annotation [4] of the 70 DEGs detected on the Exon arrays. Genes involved in oxidoreductase activity (P = 5.17E-07) and xenobiotic

functions (P = 1.32E-05) are significantly enriched among the smoking-related DEG including CYP1B1, AKR1C3, PGD, NOS2A, ALDH3A1, and GPX2. We observed very similar biological findings, both at a gene level and a functional annotation level, in our previous study by analyzing a larger U133A data set (34 current vs. 23 never smokers) [12].

Table 1. Functional categories significantly enriched (by Fisher Exact test) among the genes differentially expressed between current and never smokers on the exon array with Benjamini and Hochberg adjusted p value < 0.05.

Category	Term	Count	PValue
SP_PIR_KEYWORDS	oxidoreductase	12	5.17E-07
GOTERM_MF_ALL	catalytic activity	40	1.33E-06
INTERPRO_NAME	IPR000008:C2	7	1.13E-05
KEGG_PATHWAY	HSA00980:METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	7	1.32E-05
GOTERM_MF_ALL	oxidoreductase activity	13	1.84E-05
SP_PIR_KEYWORDS	metal-binding	20	2.10E-05
GOTERM_MF_ALL	transition metal ion binding	23	2.31E-05
SP_PIR_KEYWORDS	membrane	26	3.44E-05
GOTERM_MF_ALL	tetrapyrrole binding	6	3.93E-05
GOTERM_MF_ALL	heme binding	6	3.93E-05
GOTERM_MF_ALL	cation binding	27	5.04E-05
GOTERM_MF_ALL	ion binding	28	5.75E-05
GOTERM_MF_ALL	metal ion binding	28	5.75E-05
GOTERM_MF_ALL	monooxygenase activity	6	5.91E-05
SP_PIR_KEYWORDS	chromoprotein	5	6.13E-05
SMART_NAME	SM00239:C2	6	6.89E-05
GOTERM_MF_ALL	unspecific monooxygenase activity	4	8.19E-05

4. Conclusions

We compared the performance of the Affymetrix GeneChip Human Exon 1.0 ST array with the HG-U133A array for detecting smoking-related gene expression changes in airway epithelium obtained at bronchoscopy. The Exon array appears to be a reproducible platform capable of working with smaller amounts (1ug) of RNA obtained from the airway epithelium. In a gene-level fold change comparison, we found a strong correlation between the 2 platforms for smoking-related changes in gene expression, although a number of gene expression changes in airway epithelium were detected only on the Exon array. Furthermore, the exon array appears to be a more robust for detecting

differentially expressed genes in the setting of a relatively small sample size. These findings suggest that the Exon array can serve as a clinically-relevant gene expression platform for measuring host response to and damage from tobacco smoke.

Acknowledgments

This work was supported by the Doris Duke Charitable Foundation (A.S.), NIH/NCI R21CA10650 (A.S.) and NIH/NCI R01CA124640 (A.S. and M.E.L)

References

[1] Benjamini, Y. and Hochberg, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society Series B*, (57):289–300, 1995.

[2] Bolstad, B.M., *et al.*, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics*, 19(2):185-193, 2003.

[3] Dai, M., *et al.*, Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data, *Nucleic Acids Res.*, 33(20):e175, 2005.

[4] Dennis, G., Jr., *et al.*, DAVID: Database for Annotation, Visualization, and Integrated Discovery, *Genome Biol.*, 4(5):P3, 2003.

[5] Ebbert, J.O., *et al.*, Lung cancer risk reduction after smoking cessation: observations from a prospective cohort of women, *J. Clin. Oncol.*, 21(5):921-926, 2003.

[6] Gardina, P.J., *et al.*, Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array, *BMC Genomics*, **7**:325, 2006.

[7] Greenlee, R.T., *et al.*, Cancer statistics, 2001., *CA Cancer J. Clin.*, 51(1):15-36, 2001.

[8] Okoniewski, M.J., *et al.*, High correspondence between Affymetrix exon and standard expression arrays, *Biotechniques*, 42(2):181-185, 2007.

[9] Pruitt, K.D., Tatusova, T., and Maglott, D.R., NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res.*, 35(Database issue):D61-65, 2007.

[10] Shields, P.G., Molecular epidemiology of lung cancer, *Ann. Oncol.*, 10 Suppl 5:S7-11, 1999.

[11] Smyth, G.K., Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Stat. Appl. Genet. Mol. Biol.*, 3:Article3, 2004.

[12] Spira, A., *et al.*, Effects of cigarette smoke on the human airway epithelial cell transcriptome, *Proc, Natl. Acad. Sci. USA*, 101(27):10143-10148, 2004.

[13] Spira, A., *et al.*, Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer, *Nat. Med.*, 13(3):361-366, 2007.

[14] Tusher, V.G., Tibshirani, R., and Chu, G., Significance analysis of microarrays applied to the ionizing radiation response, *Proc. Natl. Acad. Sci. USA*, 98(9):5116-5121, 2001.

[15] Yoshida, R., *et al*., A statistical framework for genome-wide discovery of biomarker splice variations with GeneChip Human Exon 1.0 ST Arrays, *Genome Informatics*, 17(1):88-99, 2006.
[16] Affymetrix. 2005. Exon probe set annotations and transcript cluster groupings. Affymetrix, Santa Clara, CA.
[17] Affymetrix_Inc. Affymetrix White Papers: http://www.affymetrix.com/support/technical/whitepapers.affx