Published in final edited form as:
Cancer Prev Res (Phila). 2008 July ; 1(2): 112–118. doi:10.1158/1940-6207.CAPR-07-0017.

Impact of smoking cessation on global gene expression in the bronchial epithelium of chronic smokers

Li Zhang[1], **Jack Lee**[1], **Hongli Tang**[2], **You-Hong Fan**[2], **Lianchun Xiao**[1], **Hening Ren**[2], **Jonathan Kurie**[2], **Rodolfo C Morice**[2], **Waun Ki Hong**[2], and **Li Mao**[2]

[1]Department of Bioinformatics and Computational Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030.

[2]Department of Thoracic/Head and Neck Medical Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030.

Abstract

Cigarette smoke is the major cause of lung cancer and can interact in complex ways with drugs for lung cancer prevention or therapy. Molecular genetic research promises to elucidate the biologic mechanisms underlying divergent drug effects in smokers versus non-smokers and to help in developing new approaches for controlling lung cancer. The present study compared global gene expression profiles (determined via Affymetrix microarray measurements in bronchial epithelial cells) between chronic smokers, former smokers, and never smokers. Smoking effects on global gene expression were determined from a combined analysis of three independent datasets. Differential expression between current and never smokers occurred in 591 of the 13,902 genes measured on the microarrays ($P < 0.01$ and >2 fold change; pooled data)—a profound effect. In contrast, differential expression between current and former smokers occurred in only 145 of the measured genes ($P < 0.01$ and >2 fold change; pooled data). Nine of these 145 genes showed consistent and significant changes in each of the three datasets ($P < 0.01$ and >2 fold change), with 8 being down-regulated in former smokers. Seven of the 8 down-regulated genes, including CYP1B1 and 3 AKR genes, influence the metabolism of carcinogens and/or therapeutic/chemopreventive agents. Our data comparing former and current smokers allowed us to pinpoint the genes involved in smoking–drug interactions in lung cancer prevention and therapy. These findings have important implications for developing new targeted and dosing approaches for prevention and therapy in the lung and other sites, highlighting the importance of monitoring smoking status in patients receiving oncologic drug interventions.

INTRODUCTION

Chronic cigarette smoking is the major cause of lung cancer and remains so for years even after smoking cessation (1,2). Therefore, the development of agents for controlling lung cancer generally target, virtually by default, current and former heavy smokers. Smoking status, however, appears to influence response to various chemopreventive and chemotherapeutic agents and clinical outcomes of their use (3,4). Three large randomized

Correspondence to: Li Zhang, Ph.D., Lzhangli@mdanderson.org, or Li Mao, M.D., lmao@mdanderson.org.

clinical trials to prevent lung cancer--the Alpha-Tocopherol, Beta-Carotene (ATBC) Prevention Study (5), Carotene and Retinol Efficacy Trial (CARET) (6), and Lung Intergroup Trial (LIT) (7)--demonstrated that current heavy smokers had harmful interactions (higher lung cancer mortality, incidence and recurrence) with preventive agents (versus control arms); agent effects in former smokers were generally neutral and were not readily interpretable in never smokers because of the exclusion or very limited number of these patients in these trials. Certain lung cancer therapy regimens have been shown to be less effective in current smokers than in former and never smokers (8,9). Smoking can stimulate the metabolic clearance of targeted anticancer therapies, undoubtedly diminishing therapeutic benefit (9,10). These data highlight the importance of understanding the biologic impact of chronic smoking on lung tissue.

To understand why smokers and former smokers have differential responses to agents for preventing or treating lung cancer, we analyzed and compared global gene expression profiles in three independent cancer-free cohorts comprising current, former and never smokers.

MATERIALS AND METHODS

Study population

This study included current smokers, former smokers, and never smokers with no evidence of cancer and collected from separate, independent studies conducted at The University of Texas M. D. Anderson Cancer Center (MDACC) (2 studies) and the Boston Medical Center (BMC) (one study). The three datasets associated with the three studies are called MDACC-1, MDACC-2 and BMC throughout this manuscript. Former smoking was defined as having quit smoking for at least 12 months before study entry. Participants included in the MDACC-1 and MDACC-2 datasets came from the placebo arm of an ongoing chemoprevention trial performed at M. D. Anderson Cancer Center. All MDACC subjects were clinically free of cancer at enrollment and underwent a bronchoscopy at baseline. Bronchial brushes were performed at 6 predetermined sites including the entry area at each of the 5 main lobes and the carina, as described previously (11). The study was approved by the MDACC Institutional Review Board, and all MDACC participants gave signed informed consent. BMC dataset participants had a bronchoscopy at the BMC and were analyzed in a previously reported study (12) as well as in the current study. Potential subjects for the MDACC or BMC datasets in the current study were excluded if their specimen images (produced as discussed below in “cRNA preparation and microarray hybridization”) had defects, evidence of blood contamination, or other problems that did not meet image quality criteria applied consistently across all three datasets. All MDACC patients had smoking history, with average pack years of 40.6 (±13). Their average age is 58 (±8). 61% of them are male and 78% of them white. More details about the demographic data of MDACC datasets can be found in the Supplementary data Table S2.

Bronchial brush processing and RNA extraction

For samples in MDACC-1 and MDACC-2, brushes were placed on bronchoscopy in 3 ml of plain DMEM culture (Life Technologies, Inc., Gaithersburg, MD) in sterile tissue culture

tubes and stored at 4°C for processing the same day. The tubes were vortexed lightly to detach cells from the brushes. After removal of the brush from the tube, the cell suspension was centrifuged at 2,500 rpm for 5 min. Cell pellets were then washed with 2 ml of PBS twice, and an aliquot of material was saved at −80(C until RNA extraction. For the microarray analysis, cells from the 6 brushing sites of the same individual were pooled together for RNA extraction. We used TRIzol reagent (Invitrogen, Carlsbad, CA) for total RNA extraction according to the manufacturer's protocol, with a yield of 1 ~ 4 μg of total RNA per sample. Integrity of the RNA was confirmed by running it on a RNA 6000 Neno LabChip (Aginent Technologies, Palo Alto, CA). The samples in BMC dataset were processed similarly as described (12), except that a single-round amplification protocol was used.

cRNA preparation and microarray hybridization

The first and second cDNA strands were synthesized as described previously (13). The first reverse transcription was performed in the absence of biotin-labeled ribonucleotides, resulting in unlabeled cRNA, which was then used as starting material for the second cycle. In the second cycle, the first and second cDNA strands were synthesized. The second transcription was performed in the presence of biotin-labeled-ribonucleotides, resulting in labeled cRNA. The cRNA was fragmented and checked by gel electrophoresis, as reported earlier (13). The Affymetrix GeneChip system was used for hybridization, staining and imaging of the probe arrays. Hybridization cocktails of 300 μl each containing 15 μg of cRNA and exogenous hybridization controls were prepared as described previously and hybridized to U133A or U133A plus GeneChips (Affymetrix, Santa Clara, CA) overnight at 42°C. Hybridized fragments were detected using streptavidin-linked to phycoerythrin (Molecular Probes, Eugene, OR). GeneChips were scanned and imaged using Affymetrix Microarray Analysis Suite (MAS), version 5.0.

Microarray data normalization

There were two array types used in this study: U133A and U133 Plus 2. The U133A array contains around 500,000 distinct probe features interrogating 18,400 human transcripts and variants, including 13,902 well characterized genes. The U133 plus 2.0 array contains all probe features that are on the U133A array. In addition, there are 9,921 new probe sets representing 6,500 new genes. To facilitate straight forward comparison of the data, we used only the probes that are common to both array types. We also ignored data of MM probes. We used PM probes common to U133A and U133 Plus 2 arrays to perform quantile normalization on the probe level data (14). The procedure is performed so that the distributions of the probe signal intensities of a sample are identical for all samples within a dataset. Then, we used PDNN model (15) to quantify the gene expression values from the normalized probe signal intensity data. We then applied median -centering normalization on the probeset level data, so that the median of expression values of a sample was made to be the same for all the samples in all of the datasets.

Identification of differentially expressed genes

Differential expression was identified similar to that described in Wang et al (16). We used Z values (defined below) to assess differentially expressed genes between current and never

or former smokers, in whom the magnitude of the Z values is assumed to represent the effect of smoking cessation. For a given dataset containing n_A samples and n_B in groups A and B, respectively, we compute the following test statistic Z for each probeset:

$$Z=D/\sigma \quad \text{(Eq.1)}$$

where D is average difference between the log expression values between A and B groups. σ is the estimated standard deviation of D:

$$\sigma=\sqrt{\sigma_A^2/n_A+\sigma_B^2/n_B} \quad \text{(Eq.2)}$$

where σ^2_A and σ^2_B are estimated variances of log expression values in groups A and B, respectively. These variances were estimated using loess fit between the mean log expression values and the standard deviation of the log expression values. The underlying assumption is that the mean and the standard deviation are related by a smooth function, which allows the analysis method to treat the standard deviation as if it were known.

Combining Z values from different datasets

The Z values obtained from the three datasets can be combined using the following formula:

$$Z=(Z_1/\sigma_1^2+Z_2/\sigma_2^2+Z_3/\sigma_3^2)/\omega^2 \quad \text{(Eq.3)}$$

$$\omega^2=1/\sigma_1^2+1/\sigma_2^2+1/\sigma_3^2 \quad \text{(Eq.4)}$$

where Z_1, Z_2, Z_3 were calculated using Equation 2 from MDACC-1, MDACC-2 and BMC datasets respectively; σ_1, σ_2, σ_3 were calculated using Equation 2 from MDACC-1, MDACC-2 and BMC datasets respectively.

The test statistic Z is supposed to form a T distribution if the log expression values are normally distributed. However, the observed data slightly deviate from the normal distribution for they contain more extreme values. Consequently, the significance of Z can be over estimated.

To alleviate the bias due to the assumption of normal distribution, we used permuted data to compute Z*. The expression values are randomly permuted for each probeset within each dataset. The permutation was performed 10 times to construct an empirical cumulative distribution function of Z*. This distribution was assumed to be the distribution of Z values under null hypothesis *(i.e.*, no differential expression). And it was used to estimate the p-values and the False Discovery Rate (FDR) associated with Z values. The permutation was performed within each of the datasets, but never across the datasets. Note that other than the permutation step, our method is the same as that described by Wang et al (16).

RESULTS

Our overall study population numbered 99 individuals comprising 56 current, 24 former, and 19 never smokers from three independent datasets (Table 1). The MDACC-1 and

MDACC-2 datasets included 41 chronic smokers (26 current, 15 former) enrolled in an ongoing chemoprevention trial at M. D. Anderson Cancer Center. All 41 of these subjects had at least a 20-packyear smoking history. The BMC dataset comprised 75 current, former and never smokers. Never smokers with significant environmental cigarette exposure and subjects with respiratory symptoms or regular use of inhaled medications were excluded. We selected 58 members of the BMC cohort for the present analysis (Table 1) and excluded 17. Exclusions from either the BMC or MDACC datasets were based on image-quality criteria applied consistently across all three datasets.

First, we determined Z values (defined in Methods) in the three datasets separately. Then we compared the Z values from each dataset, and as shown in Fig. 1, we found that the genes with the most significant differential expressions (shown in red) are similar among the three datasets. The largest Z values mostly are located in the first and third quadrants in the scatter plots of Fig. 1, indicating that these changes in gene expression are consistent among the three datasets.

To further assess the statistical significance of the changes in gene expression between former and current smokers, we used Q-Q plots (Fig. 2) of Z values and BUM plots (Fig. 3) to evaluate the distribution of p-values. Fig. 2a compares the quantiles of Z values calculated from combining all three microarray datasets and the quantiles of Z_p (Z values calculated from permuted data). With permuted data, Z values are bounded between -10 and 10. The Z values from observed data contain clear outliers larger than 10. Without differential expression, the data points in Fig. 2a should be close to the diagonal line (shown in red). Ideally, if the gene expression data obey normal distributions and are independent from each other, we would expect values of Z_p to form a standard normal distribution. However, Fig. 2b shows that Z_ps have wider ranges than that from standard normal distribution. Consequently, we used the distribution of Z_p as that from the null hypothesis (no differential expression between former and current smokers) to compute the p-values of Z instead of using the standard normal distribution.

The BUM plot (17) presented a histogram of the p-values. Under the null hypothesis, the p-values should form a uniform distribution. The sharp spike at the left side of Fig. 3 represents the effects of differential expression contradicting the null hypothesis. The uniform part of the histogram is indicated by the red line in Fig. 3. The peak on the left side of the figure is composed of 1.2×10^3 probesets (transcripts), which is our estimated number of genes that are differentially expressed between the former smokers and current smokers. Only a subset of these genes is identifiable, however. According to the BUM method (17), we found 345 probesets that were differentially expressed at a P-value of < 0.01, for which the false discovery rate is approximately 32%. Of the 345 probesets, 176 have greater than a 2-fold difference in expression (details of these176 probesets are shown in supplementary Table S1). These 176 probesets represent 145 non-redundant significantly differentially expressed genes (> 2 fold change, $P < 0.01$). These 145 genes include 9 genes (Table 2) with consistent and significant changes in each of the three datasets ($P < 0.01$, >2 fold change). Eight of the 9 genes are down-regulated after smoking cessation, one is up-regulated. To confirm that our measurements using microarrays are accurate, we selected a panel of genes to testing using RT-PCR and found that microarray and RT-PCR measurements are highly

correlated. For gene ALDH3A1, the correlation is 96% (Figure S1 in supplementary material).

For comparison, we also examined differential gene expression between current and never smokers (Fig. 4a). Similar to that in Fig. 3, the peak volume above the red line represents the number of differentially expressed genes, which is ~11,000 probesets. This number is more than 9 times greater than the number detected in the comparison between former and current smokers (Fig. 3). We found 591 non-redundant genes with statistically significant changes (fold change >2 and P < 0.01) in pooled data of the three datasets, a group that is over 4 times larger than the group of such differentially expressed genes detected in the comparison between current and former smokers. Among the 145 genes significantly changed between current and former smokers, 77 are consistent with the 591 genes significantly changed between current and never smokers whereas other 68 genes are not, suggesting some of the changes between current and former smokers are results of re-alignment of gene expression profiles due to the remaining genetic and gene expression abnormalities in the airway cells. The 77 genes have been highlighted in Table S1 for easy reference (response to comment 6 of Reviewer 3). Similar to Figure 4a, Figure 4b, showed the comparison between former smokers and never smokers.

One possible reason that there is much more differential expression in Figure 4a and than Figure 3 is that the sample sizes are bigger in the Figure 3, which led to more statistical power. To address this problem, we performed principal component analysis and presented in the results in Figure 5. The gene expression profile of each patient is represented by its two main principal components. Two clusters emerge from the analysis. The cluster on the left (Comp1 < -10) contained mostly the never smokers while the cluster on the right contained a mixture of all kinds of smoking status. The former smokers (in red) and current smokers (in black) were intermingled. This results suggest that among the current, former and never smokers, the never smokers are the most distinct from the three groups. Some of the never smokers appear to resemble smokers. It is possible that these never smokers may have suffered from environmental hazards other than smoking. It is also possible that these never smokers are simply artifacts of principal component analysis since lesser components were ignore in Figure 5.

DISCUSSION

In probing 13,902 genes, we found that 591 were expressed differently in current versus never smokers and that only 145 (25%) of these 591 also were expressed differently in current versus former smokers. Among these 145 genes, 9 were significantly differentially expressed (8 overexpressed, 1 underexpressed; Table 2) by > 2 fold in current versus former smokers in each ($P < 0.01$) and in the pooled data ($P < 0.0001$) of the three datasets (2 MDACC, 1 BMC) included in this study. Therefore, our present study pinpoints and validates 9 differentially expressed genes in former versus current smokers, the first such validation reported to date.

Seven of the 8 validated genes overexpressed in current smokers--CYP1B1, 4 AKRs, ALDH3A1, and NQO1 (Table 2)--are involved in drug and/or carcinogen metabolism (18–

27). Polycyclic aromatic hydrocarbons (PAHs) in tobacco smoke are known to bind to and activate the aryl hydrocarbon receptor (AhR) and thus induce CYP1B1 (10). CYP1B1 expression is of special interest because it may contribute both to increased drug metabolism and to carcinogenesis of the aerodigestive tract (1,18–20). The metabolic clearance of docetaxel, tamoxifen, gefitinib, erlotinib, and other cancer prevention and therapy drugs is enhanced by CYP1B1 (9,21–23). Up-regulation of CYP1B1 and the 6 other validated overexpressed metabolizing genes by smoking likely is involved in the adverse interactions between smoking and drugs for lung cancer prevention and therapy; smoking cessation downregulates these gene expressions and thus may reduce or eliminate the adverse drug interactions.

AKR1C1, AKR1C3, and AKR1B10 were among the significantly upregulated genes in oral epithelial cells treated with cigarette smoke condensate (Nagaraji, Toxicol. Letters, 2006), suggesting a role of the enzymes in preventing or repairing carcinogen-induced DNA damages. However, AKR1C1, AKR1C2, and AKR1B10 have been shown to overexpress in NSCLC (Hsu, Cancer Res, 2001; Woenckhaus, J Pathology, 2006 ref 26 in our previous version; Fukomoto, Clin Cancer Res, 2005, Ref 24 in our previous version) and the overexpression was associated with a poor clinical outcome (Hsu, Cancer Research, 2001). It is possible that in the early carcinogenic process, ARKs plays important role in detoxification and DNA damage repair whereas in the later stage of the tumorigenesis, the same enzymes may act as survival factors to prevent transformed cells from extensive DNA damage which will lead death of the transformed cells. In fact, several studies have shown that overexpression of AKR1C1, AKR1C2, or AKR1C3 is a potential contributing mechanism for resistance to platin-based chemotherapy to various tumor types including lung cancer (Deng, JBC, 2002; Deng, Cancer Chemother. Pharmacol, 2004; Chen, Cancer Chemother. Pharmacol., 2007). AKR1B10 expression has also been implicated in retinoic acid signaling (reference 25 of previous version but was not cited previously) which may be important in retinoic acid resistance in both chemoprevention and chemotherapy settings. Another enzyme in the list is ALDH3A1 which was linked to responses to cyclophosphamide-based chemotherapy in breast cancer (reference 27 of the previous version but was not cited previously), suggesting ALDH3A1 may serve as a biomarker in individualized application of oxazaphosphorine-based cancer chemotherapeutic regimens (for Reviewer 4).

Various biases can produce inconsistencies between similar datasets. These biases can stem from differences in age, race, sex, smoking history, and sample processing. Regarding sample processing for example, MDACC-1 and MDACC-2 samples involved two rounds of RNA amplification versus a single round in the BMC set. Two rounds of amplification are known to cause loss of signals for probes that target far away from the 3’ end of mRNA sequences. The consistent changes in smoking cessation-related genes in all three independent data sets support the robustness of our present findings.

Gene expression profiling in bronchoscopy specimens offers a direct assessment of the effects of cigarette smoking in the lungs. Spira and colleagues have reported the analyses of global genomic profiling in bronchoscopy specimens (12, and Genome Biology, 2007) of which we have used the data to enhance our study. Gene expression patterns vary greatly

between individuals (because of genetic variations and different environmental influences), and the Spira report, which was published during the course of our study, provided us with the opportunity to increase the robustness of our MDACC cohort data with the addition of the BMC dataset. Our present findings complement, confirm and extend the previous findings, and both are unique in the literature. Whereas the Spira study highlighted important genomic profiling differences between current and never smokers and a relatively broad array of differentially expressed metabolizing and antioxidant genes in current versus former smokers, our current results hone in on the specific drug-metabolizing genes involved in smoking–drug interactions. The precision we achieved is due to the increased power and cross validation provided by adding the BMC dataset to our MDACC-1 and -2 datasets from the extensive program of tobacco-related carcinogenesis research conducted at MDACC. Without access to a validating independent dataset, the important earlier study could not pinpoint over-expressed genes in current versus former smokers. Another previous report used Serial Analysis of Gene Expression (SAGE) for global gene expression profiles of bronchial epithelial cells from current, former, and never smokers with relatively small sample size (BMC Genomics, 2007). This study used a single cohort and but the genes found differentially expressed between current and former smokers are quite different as those identified in our study except MUC5AC. It is possible that the two different methodology platforms (SAGE vs. DNA microarray) have different preference for certain gene signals. Therefore, comparative analyses of the two technologies using same RNA samples may be needed to address this issue (all marked here are for comment 3 of Reviewer 2).

Our results also show that the scope of genetic changes following smoking cessation is much smaller than that associated with chronic smoking (Figs. 2 and 4), possibly explaining the persistent high lung cancer risk in former smokers (28). Surprisingly at the time (about 10 years ago), we and others previously found in assessments limited to specific genetic alterations that smoking-related genetic changes persisted after smoking cessation in a population similar to those of MDACC-1, -2 and BMC (29,30). Showing similar genetic alterations in current and former smokers, results of the more sophisticated global genomic profiling approach of our present study are consistent with the earlier findings.

Our findings underscore the importance of smoking status in clinical trials, demonstrating that smoking effects on metabolizing genes potentially can interfere with drugs in standard or investigational chemoprevention or therapy not only in the lung but in other sites as well. Future research directions should include (a) increased monitoring of smoking status and increased smoking-cessation efforts in any trial setting because of adverse smoking effects on drug uptake and metabolism and (b) the development of new dosing and targeted approaches to counteract adverse smoking–drug interactions in the lung. New targeted approaches should consider the signaling pathways of drug-metabolizing genes that were validated in this study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants from the National Cancer Institute (P01 CA91844) and Department of Defense (W81XWH-04-1-0142).

References

1. Hecht SS. Tobacco carcinogens, their biomarkers and tobacco-induced cancer. Nature Rev Cancer. 2003; 3:733–744. [PubMed: 14570033]
2. Cancer Facts and Figures 2003. Am Cancer Soc. 2003:1–52.
3. Mayne ST, Lippman SM. Cigarettes: a smoking gun in cancer chemoprevention. J Natl Cancer Inst. 2005; 97:1319–1321. [PubMed: 16174848]
4. Gritz ER, Dresler C, Sarna L. Smoking, the missing drug interaction in clinical trials: ignoring the obvious. Cancer Epidemiol Biomarkers Prev. 2005; 14:2287–2293. [PubMed: 16214906]
5. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. N Engl J Med. 1994; 330:1029–1035. [PubMed: 8127329]
6. Omenn GS, Goodman GE, Thornquist MD, et al. Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. N Engl J Med. 1996; 334:1150–1155. [PubMed: 8602180]
7. Lippman SM, Lee JJ, Karp DD, et al. Randomized phase III intergroup trial of isotretinoin to prevent second primary tumors in stage I non-small-cell lung cancer. J Natl Cancer Inst. 2001; 93:605–618. [PubMed: 11309437]
8. Zhang Z, Xu F, Wang S, et al. Influence of smoking on histologic type and the efficacy of adjuvant chemotherapy in resected non-small cell lung cancer. Lung Cancer. (In press).
9. Hamilton M, Wolf JL, Rusk J, et al. Effects of smoking on the pharmacokinetics of erlotinib. Clin Cancer Res. 2006; 12:2166–2171. [PubMed: 16609030]
10. Port JL, Yamaguchi K, Du B, et al. Tobacco smoke induces CYP1B1 in the aerodigestive tract. Carcinogenesis. 2004; 25:2275–2281. [PubMed: 15297370]
11. Lee JS, Lippman SM, Benner SE, et al. Randomized placebo-controlled trial of isotretinoin in chemoprevention of bronchial squamous metaplasia. J Clin Oncol. 1994; 12:937–945. [PubMed: 8164045]
12. Spira A, Beane J, Shah V, et al. Effects of cigarette smoke on the human airway epithelial cell transcriptome. PNAS. 2004; 101:10143–10148. [PubMed: 15210990]
13. Gold D, Coombes K, Medhane D, et al. A comparative analysis of data generated using two different target preparation methods for hybridization to high-density oligonucleotide microarrays. BMC Genomics. 2004; 5:2. [PubMed: 14709180]
14. Bolstad BM, Irizzary RA, Astrand M, et al. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics. 2003; 19:185–193. [PubMed: 12538238]
15. Zhang L, Miles MF, Aldape KD. A model of molecular interactions on short oligonucleotide microarrays. Nat Biotechnol. 2003; 21:818–821. [PubMed: 12794640]
16. Wang J, Coombes KR, Highsmith WE, et al. Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. Bioinformatics. 2004; 20:3166–3178. [PubMed: 15231529]
17. Pounds S, Morris SW. Estimating the occurrence of false positives and false negatives in microarray studies b approximating and partitioning the empirical distribution of p-values. Bioinformatics. 2003; 19:1236–1242. [PubMed: 12835267]
18. Mahadevan B, Luch A, Atkin J, et al. Inhibition of human cytochrome P450 1B1 further clarifies its role in the activation of dibenzo[*a*,*l*]pyrene in cells in culture. J Biochem Mol Toxicol. 2007; 21:101–109. [PubMed: 17623886]
19. Roos PH, Bolt HM. Cytochrome P450 interactions in human cancers: new aspects considering CYP1B1. Expert Opin Drug Metab Toxicol. 2005; 1:187–202. [PubMed: 16922636]

20. Purnapatre K, Khattar SK, Saini KS. Cytochrome P450s in the development of target-based anticancer drugs. Cancer Letters. (In press).
21. Sissung TM, Price DK, Sparreboom A, et al. Pharmacogenetics and regulation of human cytochrome *P*450 1B1: implications in hormone-mediated tumor metabolism and a novel target for therapeutic intervention. Mol Cancer Res. 2006; 4:135–150. [PubMed: 16547151]
22. Li J, Zhao M, He P, et al. Differential metabolism of gefitinib and erlotinib by human cytochrome P450 enzymes. Clin Cancer Res. 2007; 13:3731–3737. [PubMed: 17575239]
23. Rochat B, Morsman JM, Murray GI, et al. Human CYP1B1 and anticancer agent metabolism: mechanism for tumor-specific drug inactivation? JPET. 2001; 296:537–541.
24. Fukumoto S, Yamauchi N, Moriguchi H, et al. Overexpression of the aldo-keto reductase family protein AKR1B10 is highly correlated with smokers' non-small cell lung carcinomas. Clin Cancer Res. 2005; 11:1776–1785. [PubMed: 15755999]
25. Penning TM. AKR1B10: a new diagnostic marker of non-small cell lung carcinoma in smokers. Clin Cancer Res. 2005; 11:1687–1690. [PubMed: 15755988]
26. Woenckhaus M, Klein-Hitpass L, Grepmeier U, et al. Smoking and cancer-related gene expression in bronchial epithelium and non-small-cell lung cancers. J Pathol. (In press).
27. Sladek NE, Kollander R, Sreerama L, et al. Cellular levels of aldehyde dehydrogenases (ALDH1A1 and ALDH3A1) as predictors of therapeutic responses to cyclophosphamide-based chemotherapy of breast cancer: a retrospective study. Rational individualization of oxazaphosphorine-based cancer chemotherapeutic regimens. Cancer Chemother Pharmacol. 2002; 49:309–321. [PubMed: 11914911]
28. Tong L, Spitz MR, Fueger JJ, et al. Lung carcinoma in former smokers. Cancer. 1996; 78:1004–1010. [PubMed: 8780538]
29. Mao L, Lee JS, Kurie JM, et al. Clonal genetic alterations in the lung of current and former smokers. J Natl Cancer Inst. 1997; 89:857–862. [PubMed: 9196251]
30. Wistuba II, Lam S, Behrens C, et al. Molecular damage in the bronchial epithelium of current and former smokers. J Natl Cancer Inst. 1997; 89:1366–1373. [PubMed: 9308707]

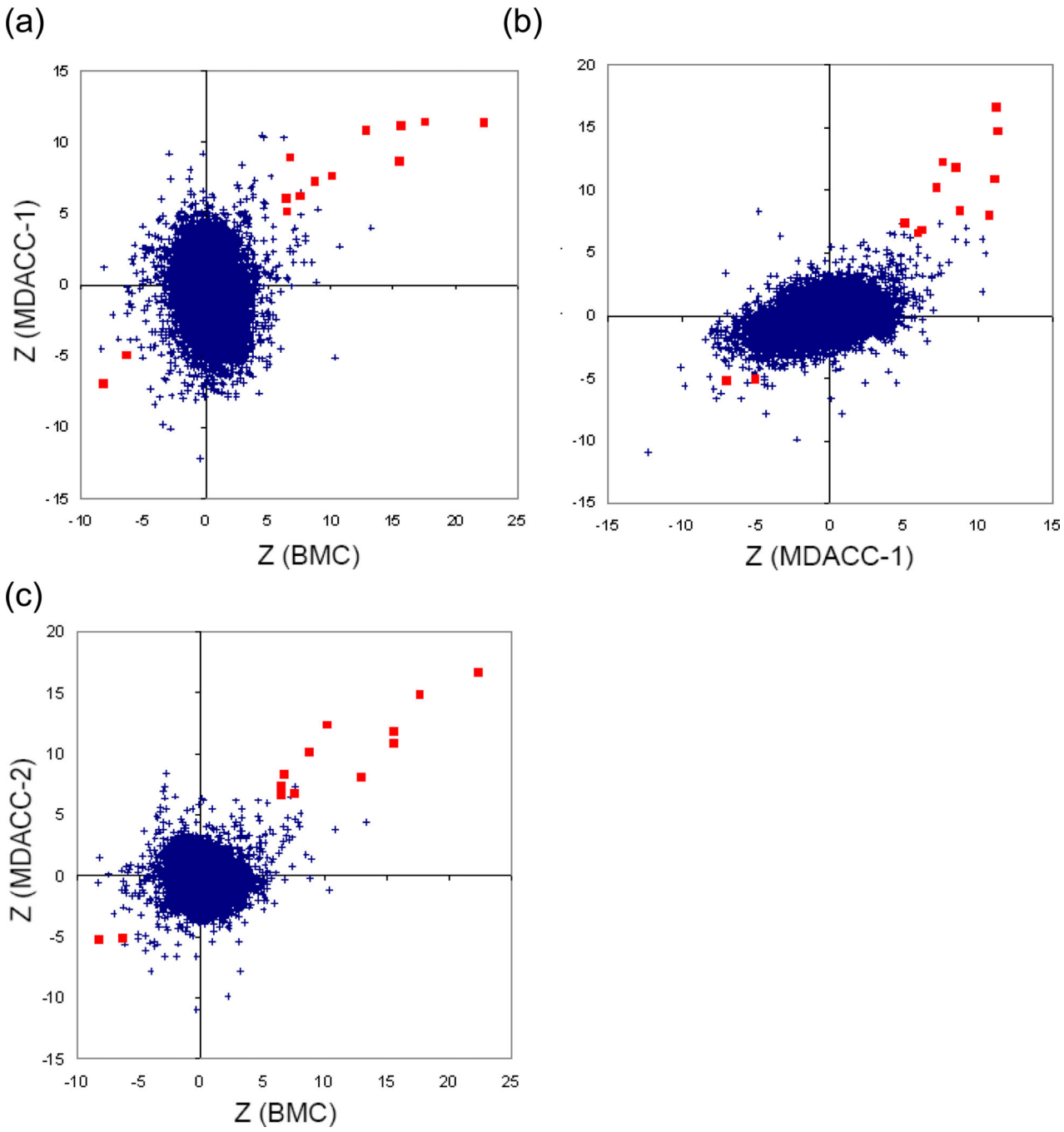

Fig. 1.

Comparison of Z values obtained from the three datasets (BMC, MDACC-1, MDACC-2). Each point in these scatter plots represents a probeset. The probesets with an absolute Z values greater than 5 in all three datasets are shown in red. Detailed data on these probesets are in Table 2.

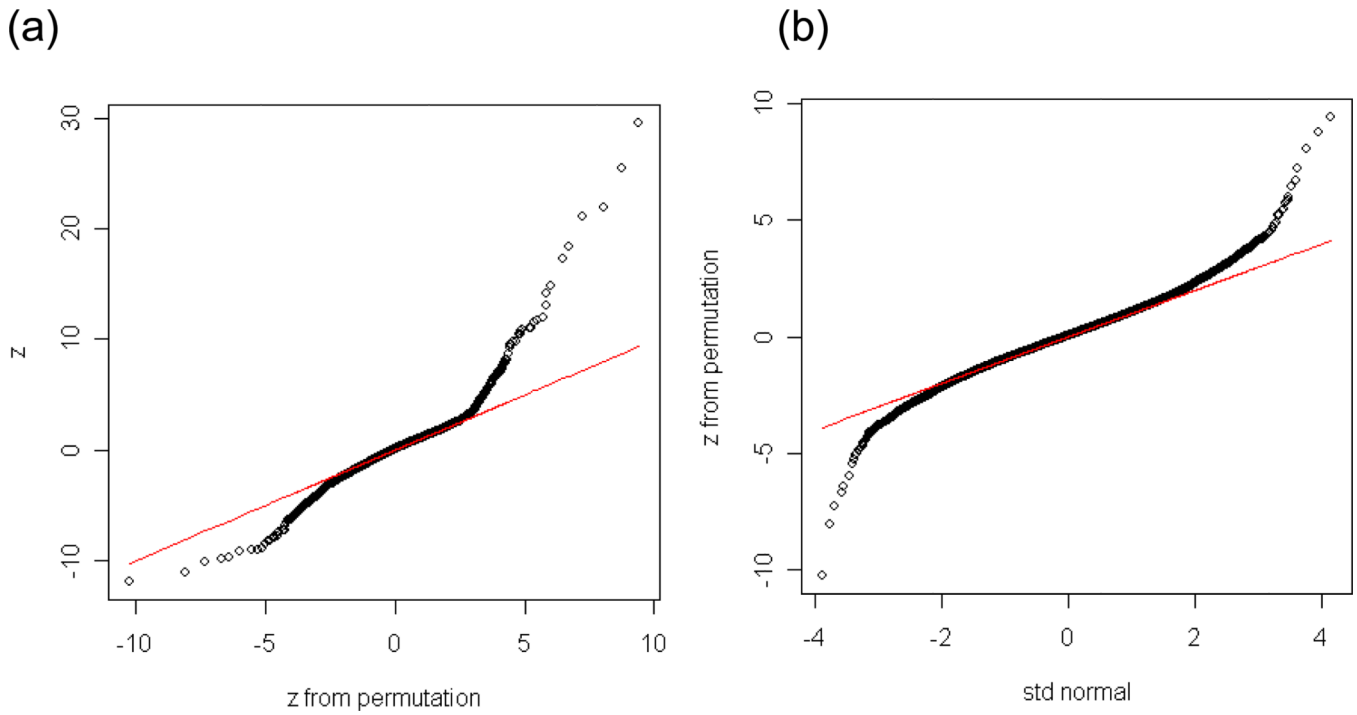

Fig. 2.

Quantile-Quantile plots of Z values. (a) Quantile of Z values *vs.* quantile of Z values obtained from permuted data. (b) Quantile of Z values from permutation data *vs.* quantile values of standard normal distribution.

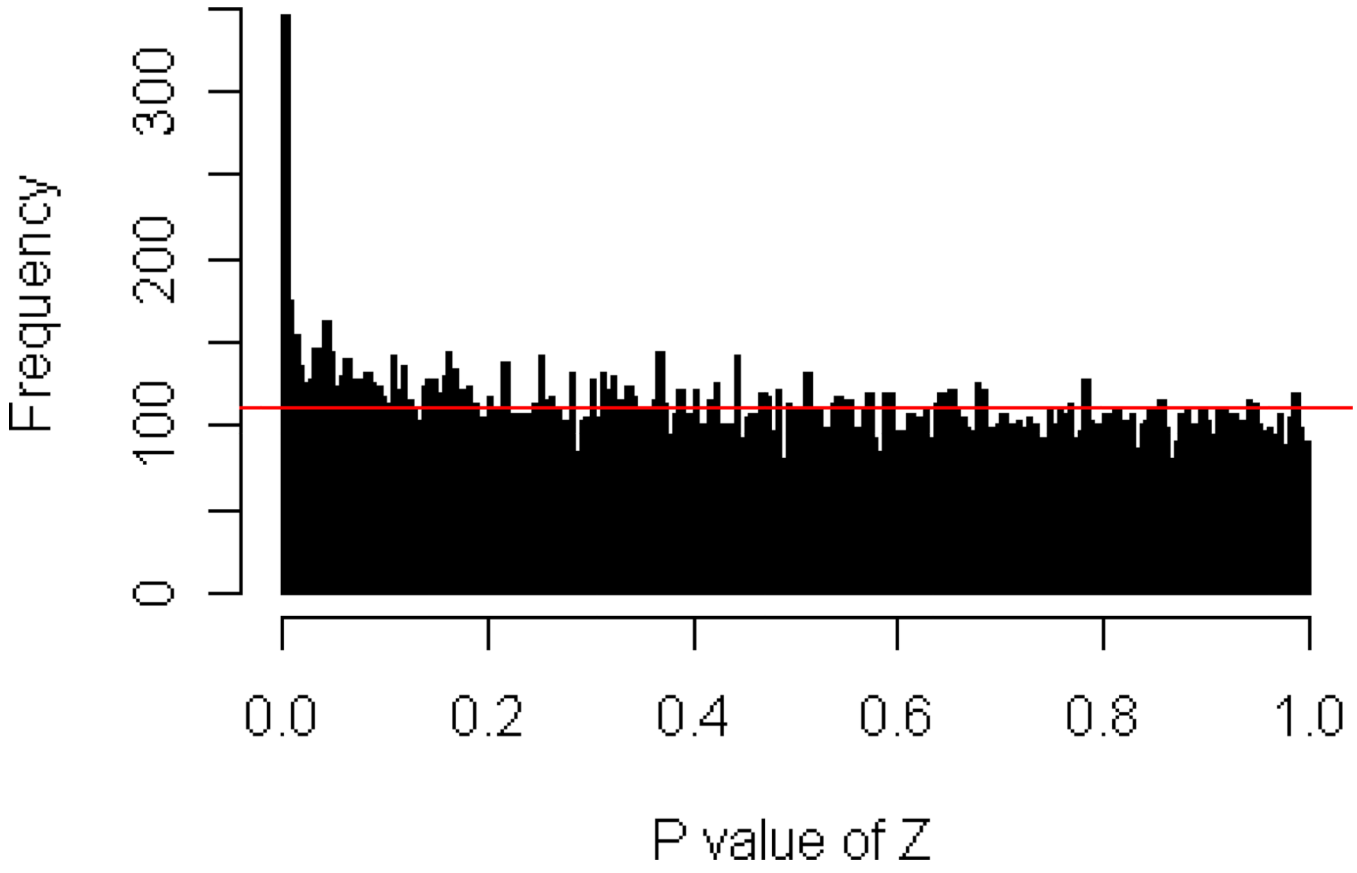

Fig. 3.

Histogram of P-values in search of differential expression between current and former smokers. Based on BUM estimate, 345 probesets were identified as differentially expressed with a false discovery rate of 32%. 176 Of the 345 probesets have a fold change >2. Detailed gene information on the 176 probesets (145 genes) is provided in Supplementary Table S1. The P-values were evaluated on the basis of the combined Z values from the three datasets and the combined Z values from the permuted data.

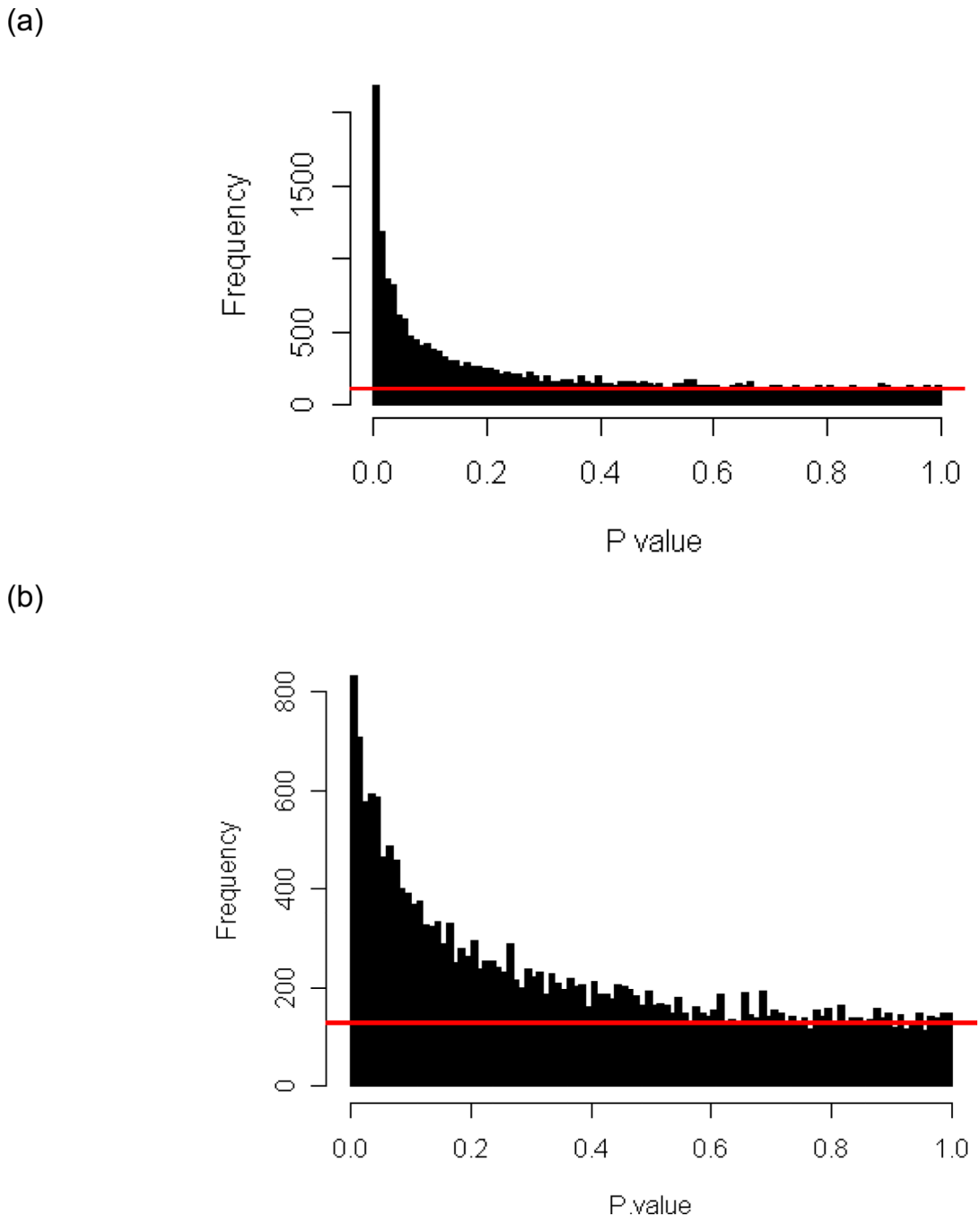

Figure 4.

Histograms if p-values in search of differential expression (a) between Never smokers and current smokers; (b) between former smokers and never smokers. Only data from BMC dataset were used the plots. FDRs were estimated to be 5% and 16% in for p-value <0.01 in (a) and (b), respectively.

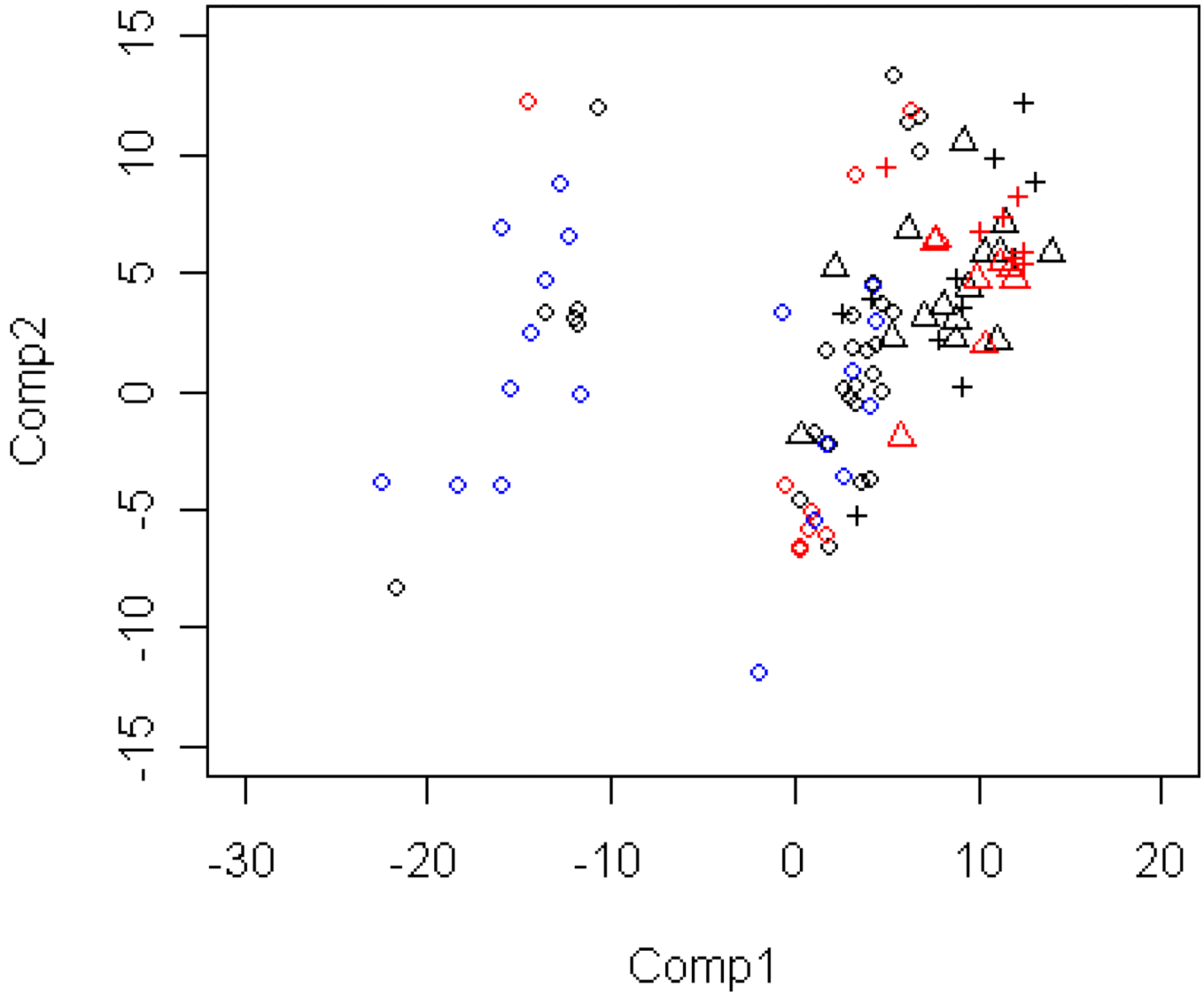

Figure 5.

Principal component analysis. The two main principal components were used to visualize the relationships amongst patients with different smoking status. Each point represents a patient. Current smokers were shown in black; former smoker in red; and never smoker in blue. Data from BMC were shown in circles; MDACC1 in pluses; and MDACC2 in triangles.

Table 1

Sample sizes and array types of the microarray datasets.*

Dataset	FS	CS	NS	ArrayType
MDACC-1	7	11	0	U133A
MDACC-2	8	15	0	U133Plus2
BMC	9	30	19	U133A

*FS, former smoker; CS, current smoker; NS, never smoker.

Table 2

Genes with consistent fold changes > 2 in each ($P < 0.01$) and across ($P \leq 0.0001$) the three data sets.*

Gene	Fold Changes				P-value (comb.)	Full Name	Refseq	Probeset
	BMC	MDACC-1	MDACC-1	Comb.				
ALDH3A1	6.9	9.4	4.0	6.2	0.0000	aldehyde dehydrogenase 3 family, memberA1	NM_000691	205623_at
CYP1B1	4.2	5.7	6.7	4.9	0.0000	cytochrome P450, member 1B1	NM_000104	202436_s_at
MUC5AC	2.2	9.6	3.0	3.5	0.0000	mucin 5AC, oligomeric mucus/gel-forming	XM_001130382	214385_s_at
AKR1C2	3.3	4.2	3.5	3.5	0.0000	aldo-keto reductase family 1, member C2	NM_001354	209699_x_at
AKR1B10	3.2	4.2	3.8	3.5	0.0000	aldo-keto reductase family 1, member B10	NM_020299	206561_s_at
AKR1C1	2.8	4.0	3.3	3.2	0.0000	aldo-keto reductase family 1, member C1	NM_001353	204151_x_at
NQO1	2.8	2.3	2.4	2.6	0.0001	NAD(P)H dehydrogenase, quinone 1	NM_000903	210519_s_at
AKR1C3	2.5	2.1	3.1	2.5	0.0000	aldo-keto reductase family 1, member C3	NM_003739	209160_at
SCGB1A1	−2.0	−2.4	−2.6	−2.4	0.0001	secretoglobin, family 1A, member 1 (uteroglobin)	NM_003357	205725_at