[Frontiers in Bioscience 7, c63-67, May 1, 2002]

TREE-BASED ANALYSIS OF MICROARRAY DATA FOR CLASSIFYING BREAST CANCER

Heping Zhang and Chang-Yung Yu

Department of Epidemiology and Public Health , Yale University School of Medicine, New Haven, CT

TABLE OF CONTENTS

1. ABSTRACT

DNA microarray data have provided us with the opportunity to assess the expression levels for thousands of genes simultaneously. One of the uses of this information is to classify cancer tumors. A noted challenge in using microarray information is analytical. Following the work of Zhang *et al.* (1), we further pursue the use of recursive partitioning in analyses of microarray data for cancer classification. Not only does the recursive partitioning technique create intuitive classification rules, but also it is most flexible as to the handling of a massive number of genes, missing expressions, and multi-class tissues. Using a published data set (2), we demonstrate that the recursive partitioning technique creates a more precise and simpler classification rule than other commonly used approaches. In particular, we introduce the concept of A-tree and propose a procedure to assess a large number of A-trees. One of the identified genes (ERBB2) is in the close region of BRCA1 (17q21.1) and has been shown by others to have altered expression levels in breast cancer. Nonetheless, our identified genes warrant further investigation as to whether they play a role in the etiology of breast cancer.

2. INTRODUCTION

Following the work of Zhang *et al.* (1), we further pursue the use of classification trees in analyses of microarray data. Based on a recursive partitioning technique, classification trees (3) have become one of the premier nonparametric classification methods. The applications and merits of this method have been discussed in detail by Zhang and Singer (4), particularly in health related applications. One of the key advantages of the

method is its ability to extract and select useful information for classification purpose from a large number of possibly correlated, discrete or continuous variables or features. As a result of the Human Genome Project (5,6), the ability to accommodate tens or hundreds of thousands of genes simultaneously is critically important and challenging.

With gene chip technology, DNA microarray data have provided us with the opportunity to assess the expression levels for thousands of genes simultaneously and to explore the gene function and pathways. The use of DNA microarrays holds a great promise to advance biological and medical research, for example, in cancer research (7) and drug discovery (8). In particular, many studies have employed microarrays to analyze gene expression in tumors of the colon, breast and other organs (2, 9, 10). For cancer diagnosis and treatment, gene expression profiles may lead to an alternative to morphology-based tumor classification systems with abundant, insightful information. The technology is being rapidly improved and the cost is becoming cheaper and cheaper. The greatest challenge is, however, analytical (11). One of the analytic issues is classification.

In machine learning literature, classification is a supervised learning process. We are given a training set of observations (also referred to as learning sample) that contain vectors of gene expressions as well as the labeled (normal or tumor) tissues. These observations are used to induce a classification scheme with the intent to accurately predict the class label (normal or tumor) for a tissue sample that may or may not be a part of the training data. In the context of analyzing microarray data, Golub *et al.* (12) and Xiong *et al.* (13) among others used discriminant analysis to produce decision rules, Brown *et al.* (14) applied support vector machine, Moler *et al.* (15) adopted a Bayesian approach, and Zhang *et al.* (1) advocated classification trees. Dudoit *et al.* (16) provided an interesting comparison of several commonly used classification methods. In this work, we exploit the use of classification trees or recursive partitioning to improve cancer diagnosis based on microarray data.

Not only does the recursive partitioning technique create intuitive classification rules, but also it is most flexible as to the handling of a massive number of genes, missing expressions, and multi-class tissues. Using a published data set (2), we demonstrate that the recursive partitioning technique gives rise to a more precise classification rule and uses fewer gene profiles than other commonly used approaches, as conducted by Hedenfalk *et al.* (2).

3. MATERIALS AND METHOD

3.1. Data

Many cases of hereditary breast cancer are due to mutations in either the BRCA1 (MIM: 113705) or BRCA2 (MIM: 600185) gene. The histopathological changes in these cancers are often characteristic of the mutant gene. These germ-line mutations account for a substantial proportion of inherited breast and ovarian cancers, but it is likely that additional susceptibility genes will be discovered. Hedenfalk *et al.* (2) gave a detailed comparison between BRCA1 and BRCA2 mutations. They collected and analyzed biopsy specimens of primary breast cancer tumors from patients with germ-line mutations of BRCA1 (7 patients) and BRCA2 (8 patients). In addition, seven patients with sporadic cases of primary breast cancer whose family history was unknown were also identified. They obtained cDNA microarrays from 5361 unique genes, of which 2905 are known genes and 2456 are unknown. The data set can be downloaded from http://www.nhgri.nih.gov/DIR/Microarray/NEJM_Supplement. Using a variety of analytic techniques including a modified F- and t-test and a mutual-information scoring, Hedenfalk *et al.* (2) selected nine differentially expressed genes to classify BRCA1-mutation-positive and negative tumors and then 11 genes for BRCA2-mutation-positive and negative tumors.

3.2. Tree Model

Suppose we have data from n units of observations. Each unit contains a vector of feature measurements or covariates (gene expression profiles from a tissue) and a class label (normal or tumor). The recursive partitioning technique extracts homogeneous strata from the data and constructs tree-based classification rules. In essence, the classification tree is constructed through a recursive partitioning process that divides the study sample into smaller and smaller samples (every sub-sample is called a node, the starting sample being termed the root node) according to whether a particular selected predictor is above a chosen cut-off value or belongs to a subset of the discrete levels. The choices of the selected predictor and its corresponding split are decided to purify the distribution of the response; namely, separating the normal tissues from the cancer tissues in the present context. After an initial (usually over grown) tree structure results from the recursive partitioning process, a pruning step usually follows so that a balance is reached between the tree size and its apparent misclassification rate based on the learning sample. This procedure becomes more apparent as we present the analysis from the tree-based analysis.

3.3. Recursive Partitioning

Using the method described in Zhang and Singer (4) and Zhang *et al.* (1) and the RTREE program developed by Heping Zhang (http://peace.med.yale.edu), we have identified classification trees (rules) that are very accurate, as determined by the cross-validation procedures for classifying these three types of tumors. Using only one classification rule (instead of two as in [2]) on the basis of three genes (instead of 20 as in [2]) we can achieve high accuracy.

Figure 1 is a classification tree that divides 22 samples into 4 groups generated by RTREE automatically. It first uses the expression from HV16A12 to divide the original 22 tumor samples (7 BRCA1, 8 BRCA2, and 7 sporadic, respectively arranged at the top, middle, and bottom in the top circle of the tree) into two (right and left circles) sets of samples according to whether the expression level is higher than 0.835. The selection of HV16A12 and

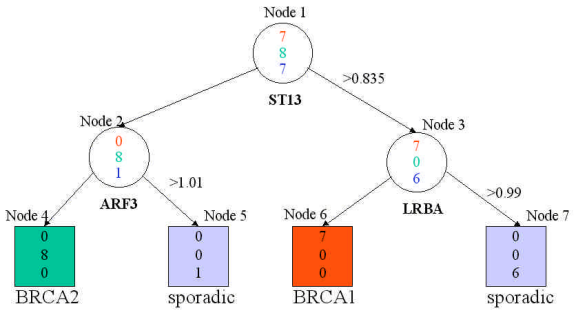

Figure 1. An Automatically Produced Tree Structure. Inside each circle and box (both are called nodes) are the numbers of BRAC1, BRCA2, and sporadic tissue samples (from top to bottom). Under the circles are the genes whose changes in expressions are used to classify the tumor types. The threshold levels of the expressions on displayed on the right arrows of the circles.

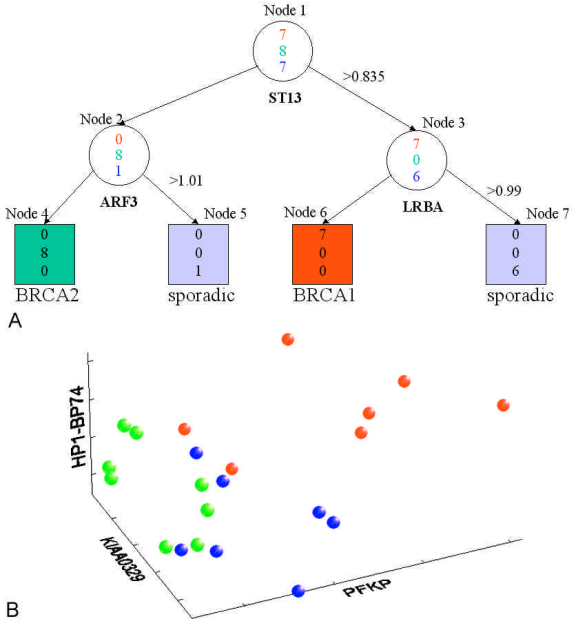

Figure 2. (A) One of the Best A-Trees in Table 1. See Figure 1 for the Structure. (B) A 3-D View of Figure 2A.

its corresponding cut-off level was determined after comparing it with all other possible dichotomous splits that were derived from all available gene expressions.

Then the resulting two sub-samples are further divided by expressions from genes HK1A2 and HV8D4 as the levels specified in the figure. Again, the selections of genes HK1A2 and HV8D4 and their corresponding splitting levels were determined through extensive comparisons and searches considering all possible binary splits.

3.4. A-Tree

In essence, Figure 1 illustrates a process that can continue as long as the data permit. This process is called recursive partitioning. In general, this process can produce a large tree structure, and requires a pruning procedure to remove the over-grown nodes of the tree. However, gene expression profiles present us a rather unusual form of data, and nearly diminish the need of pruning. As recognized by Zhang *et al.* (1), the correlation among many gene profiles makes it likely that there exist other competing tree structures that could have similar predictive precision. Indeed, we have observed consistently that there are many trees that are small and classify the labels accurately for the learning data. For this reason, we introduce the concept of A-tree as follows:

Definition: A classification tree is called an A-tree if it has the same structure as shown in Figure 1. Sometimes, we extend this structure into one more layer, namely, the terminal nodes in Figure 1 are allowed to have two offspring nodes.

3.5. Sorting A-Trees

For the present data, we have identified over one hundred thousands A-trees that make perfect classifications for the learning data. Because we have such a large number of apparently pure A-trees, we began our examination with those A-trees. By examining these A-trees, not only can we identify more reliable classification trees, but also reveal alternative biological mechanisms and pathways.

Not surprisingly, the first split was most critical to the A-tree structures. Thus, we categorized the tree structures according to the gene chosen as the first split variable. There were 29 such genes that ultimately lead to pure A-trees. It turns out that, even within a category, there can be many pure A-trees. To sort them out, we employed a local leave-one-out cross validation (LLOOCV) procedure as described in Zhang *et al.* (1). The cross validation was applied "locally" in order to keep the tree structures intact. Specifically, for a given A-tree such as the one in Figure 1, we left out one of the 22 samples. While maintaining the same splitting genes and their order, e.g., HV16A12, HK1A2 and HV8D4 in Figure 1, the corresponding splitting values were re-determined based on the remaining 21 samples. After this determination, the newly formed tree was used to classify the left-out sample and the number of errors is recorded. This process was repeated 22 times until every sample was left out once. After the completion of the LLOOCV, a ranking was performed among the A-trees in the same category, and those resulting in the fewest LLOOCV errors were kept for further consideration. For example, in one category, there may not be a perfect A-tree according to LLOOCV, but more than one A-tree that make one LLOOCV error. All A-trees with one LLOOCV error were evaluated further.

3.6. Stratified Cross Validation

To further examine the performance of the selected A-trees, we also carried out a stratified cross validation and considered leaving out one sample among samples of the same label. The procedure was repeated 100 times. In each category, we chose the A-tree with the lowest stratified cross validation error rate.

In Table 1, we present the best A-tree from every category on the basis of the performance in the LLOOCV and the stratified cross validation. Each row represents a

Table 1. The Performance of the Best A-trees

A-trees in order of impurity			LLOOCV	Avg. # errors in SCV	# of errors in 100 SCV		
1st split	2nd split	3rd split			BRCA1	BRCA2	Sporadic
ST13	ARF4L	HP1-BP74	2	1.83	0	83	100
ACTR1A	HP1-BP74	SNRP70	2	2.00	0	100	100
PPP1CB	HNRPA1	P84	1	1.04	0	4	100
GMPS	SEMA4D	PECAM1	1	1.01	0	1	100
TP53BP2	SCNN1A	GLUD1	2	2.00	100	0	100
ILF2	SEMA4D	PFKP	1	1.09	0	8	101
MSH2	112158	HADHB	3	2.76	87	89	100
NIFU	ZNF161	PFKP	3	2.97	98	177	22
HP1-BP74	ST13	HMGCL	1	1.00	0	100	0
SEMA4D	ESTs	RANBP1	1	1.25	3	86	36
LRP1	HP1-BP74	PTPN9	1	1.16	0	98	18
PFKP	KIAA0329	HP1-BP74	0	0.04	0	4	0
D123	TGM2	P84	1	1.13	0	104	9
CBX3	MYD88	TMF1	2	1.54	69	71	14
MTMR4	ESTs	EIF4A2	0	0.34	4	0	30
G22P1	MSN	HP1-BP74	1	1.34	16	106	12
PCNA	CTNND1	VASP	3	2.94	98	100	96
ERBB2	OSBPL3	SEMA4D	0	0.08	0	8	0
ATP6F	HP1-BP74	PRNP	1	1.68	15	105	48
GART	CRADD	DCTD	2	1.89	91	98	0
FLJ12442	NAGA	PXN	0	0.78	72	6	0
ZNF146	IGFBP2	CCT6A	1	1.10	0	10	100
ZNF161	FLJ12442	ADSL	1	0.99	0	99	0
GTF2I	TNFAIP1	KIAA0090	2	1.94	94	100	0
GCSH	SEMA4D	KIAA0090	1	1.02	97	5	0
LPIN1	KIAA0329	LRBA	1	0.89	0	89	0
FOXM1	C14orf2	GNAI3	0	0.07	0	7	0
NSEP1	SEMA4D	TFAP2C	2	1.46	100	32	14
PTPRU	RANBP1	CTPS	2	1.82	0	92	90

Note: Each row represents the best A-tree within a category of A-trees using the same gene as the first split as listed in the first column. The 2nd and 3rd columns display the genes used in the subsequent splits. The fourth column is the number of the LLOOCV errors for the respectively best A-tree. The fifth column is the average number of errors made during the stratified cross validation (SCV). The last three columns dissect the total numbers of errors during the 100 repetitions of the stratified cross validation for each label of the samples.

category. The first three columns display the genes used in the three splits. The fourth column is the number of the LLOOCV errors for the corresponding best A-tree. Because we have 22 samples, the maximum number of errors is 22. It is clear from Table 1 that the selected A-trees have high classification precisions based on the LLOOCV. The fifth column is the average number of errors made during the 7-fold stratified cross validation. The maximum number of errors is again 22. The last three columns dissect the total numbers of errors during the 100 repetitions of the stratified cross validation for each label of the samples. For BRCA1 (the 6th column), BRCA2 (the 7th column), and sporadic (the last column), the maximums are 700 (7 samples by 100 repetitions), 800 (8 samples by 100 repetitions), and 700 (7 samples by 100 repetitions), respectively.

From Table 1, we see that there are three exceptionally precise trees that would not have been identified by the existing automated recursive partitioning techniques. These A-trees are presented in Figures 2A, 3A and 4A. All of the seven BRCA1 tumors and the seven sporadic tumors are correctly classified by these three trees. At the same time, the misclassification rate for the eight BRCA2 tumors is also low. Figures 2B, 3B, and 4B also display three-dimensional views of the results with the gene expressions as the three coordinates. They provide an intuitive view of how perfect classifications for three types of tumors are made.

3.7. Significance Level

We have identified trees that have a simple structure and make use only three genes while having remarkable precision. An important question to be addressed is whether this is a chance occurrence or whether it is a result of a small sample being subject to extensive searches. To this end, we performed a permutation test. We generated 10,000 data sets by permuting the response. Less

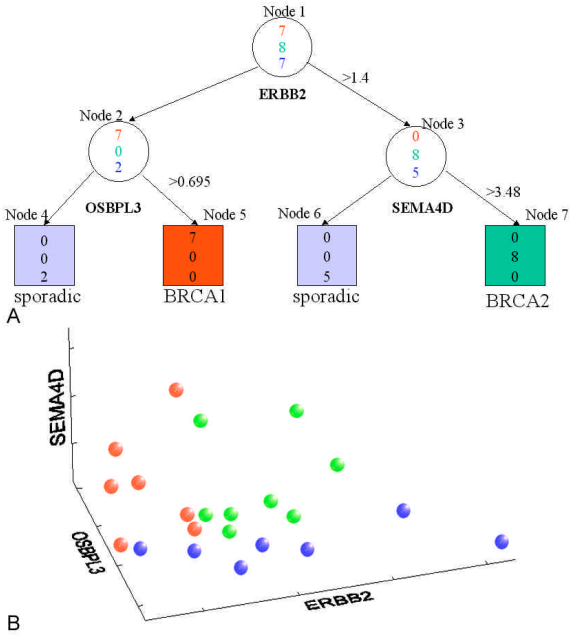

Figure 3. (A) One of the Best A-Trees in Table 1. See Figure 1 for the Structure. (B) A 3-D View of Figure 3A.

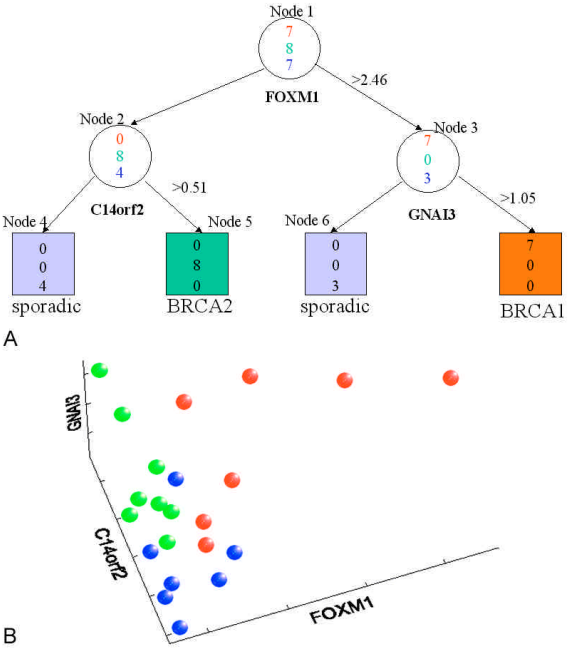

Figure 4. (A) One of the Best A-Trees in Table 1. See Figure 1 for the Structure. (B) A 3-D View of Figure 4A.

than 0.01% of the data sets in which the LLOOCV are applied resulted in perfect classification. Thus, the chance of having perfect LLOOCV trees as presented in Figures 2 and 3 is only 0.0001.

4. DISCUSSION

We have introduced the concept of the A-tree construction and established a procedure to assess a large of number of A-trees. Using a published data set, we have demonstrated the usefulness of this procedure in identifying simple and accurate tree-structures. Unlike many other classification schemes, our analysis shows that this procedure performs very well with the multi-class classification.

Most of the genes appearing in Figures 2, 3, and 4 have been isolated (PFKP[MIM:171840], ERBB2[MIM: 164870], SEMA4D[MIM: 601866], FOXM1[MIM: 602341], C14ORF2[MIM: 604573], GNAI3[MIM: 139370]). However, their roles in the etiology of breast cancer are largely unknown, with a few exceptions. Kroll *et al.* (17) analyzed the gene expression patterns of four breast cancer cell lines: MCF-7, SK-BR-3, T-47D, and BT-474, and reported unique high levels of expressions in the receptor tyrosine kinase ERBB2. In addition, Yu *et al.* (18) evaluated the mechanisms by which FK228 mediates apoptosis in non-small-cell lung cancer cells. Using Western blot and immunoprecipitation techniques, they analyzed expression of signaling-related proteins ERBB2 among others and found that FK228-treated cells were also depleted of ErbB2. It is also interesting to note that ERBB2 (17q21.1) is mapped in the same region of BRCA1.

We have devised a permutation procedure that supports the tree structures in Figures 2, 3 and 4 beyond the chance. However, the biological implication of our results needs to be validated by further experiments.

As a future project, we will look into the issue of how we can take further advantage of the abundant number of high quality A-trees and appropriately summarize them.

5. ACKNOWLEDGEMENT

This research was supported in part by grants AA-12044 and DA-12468 from the National Institutes of Health, the United States of America. The authors wish to thank Drs. Chin-Pei Tsai and Suddhasatta Acharyya for their careful review and comments on this article.

6. REFERENCES

1. Zhang, H. P., C. Yu, B. Singer, M. Xiong: Recursive partitioning for tumor classification with gene expression microarray data, *Proc. Natl. Acad. Sci. USA* 98, 6730-6735 (2001)
2. Hedenfalk I., D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, Z. Yakhini, A. Ben-Dor, E. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Pittaluga, S. Gruvberger, N. Loman, O. Johannsson, H. Olsson, B. Wilfond, G. Sauter, O.-P. Kallioniemi, A. Borg, J. Trent: Gene-expression profiles in hereditary breast cancer, *N. Engl. J. Med.* 344, 539-48 (2001)
3. Breiman, L., J. Friedman, C. Stone, R. Olshen: Classification and Regression Trees, Wadsworth, California (1984)
4. Zhang, H.P., B. Singer: Recursive Partitioning in the Health Sciences, Springer, New York (1999)
5. The Chipping Forecast, *Nature Genetics* Supplement, 21 (1999)

6. Genome Prospecting (Special Issue), *Science*, 286 (1999)
7. J. G. Hacia: Resequencing and mutational analysis using oligonucleotide microarrays,
Nature Genetics supplement, 21, 42-47 (1999)
8. Debouck, C., P. N. Goodfellow: DNA microarrays in drug discovery and development. *Nature Genetics* supplement, 21, 48-50 (1999)
9. Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA*, 96, 6745-6750 (1999)
10. Perou, C. M., S. S. Jeffrey, M. van de Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. F. Lee, D. Lashkari, D. Shalon, P. O. Brown, D. Botstein: Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, *Proc. Natl. Acad. Sci. USA*, 96, 9212–9217 (1999)
11. Lander, E. S.: Array of hope. *Nature Genetics* Supplement, 21: 3-4 (1999)
12. Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander: Molecular classification of cancer. Class discovery and class prediction by gene expression monitoring, *Science*, 286, 531-537 (1999)
13. Xiong, M. M., L. Jin, W. Li, E. Boerwinkle: Computational methods for gene expression-based tumor classification, *Biotechniques*, 29, 1264-1270 (2000)
14. Brown, M. P. S., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr., D. Haussler: Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proc. Natl. Acad. Sci. USA*, 97, 262-267 (2000)
15. Moler, E. J., M. L. Chow, I. S. Mian: Analysis of molecular profile data using generative and discriminative methods, *Physiol. Genomics*, 4, 109-126 (2000)
16. Dudoit S, J. Fridlyand, T. Speed: Comparison of discrimination methods for the classification of tumors using gene expression data, *JASA*, 97, 77-87 (2002)
17. Kroll, T., L. Odyvanova, H. Clement, C. Platzer, A. Naumann, N. Marr, K. Hoffken, S. Wolfl: Molecular characterization of breast cancer cell lines by expression profiling, *J Cancer Res Clin Oncol*, 128, 125-34 (2002)
18. Yu, X., Z. S. Guo, M. G. Marcu, L. Neckers, D. M. Nguyen, G. A. Chen, D. S. Schrump: Modulation of p53, ErbB1, ErbB2, and Raf-1 Expression in Lung Cancer Cells by Depsipeptide FR901228, *J Natl Cancer Inst*, 94, 504-13 (2002)

Key Words: Classification Trees; Recursive Partitioning; Breast Cancer; BRCA1; BRCA2; Microarray; Gene Expression; Gene Chip

Send all correspondence to: Heping Zhang, Ph.D., Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520-8034, Tel: 203-785-2838, Fax: 203-785-6912. E-mail: heping.zhang@yale.edu