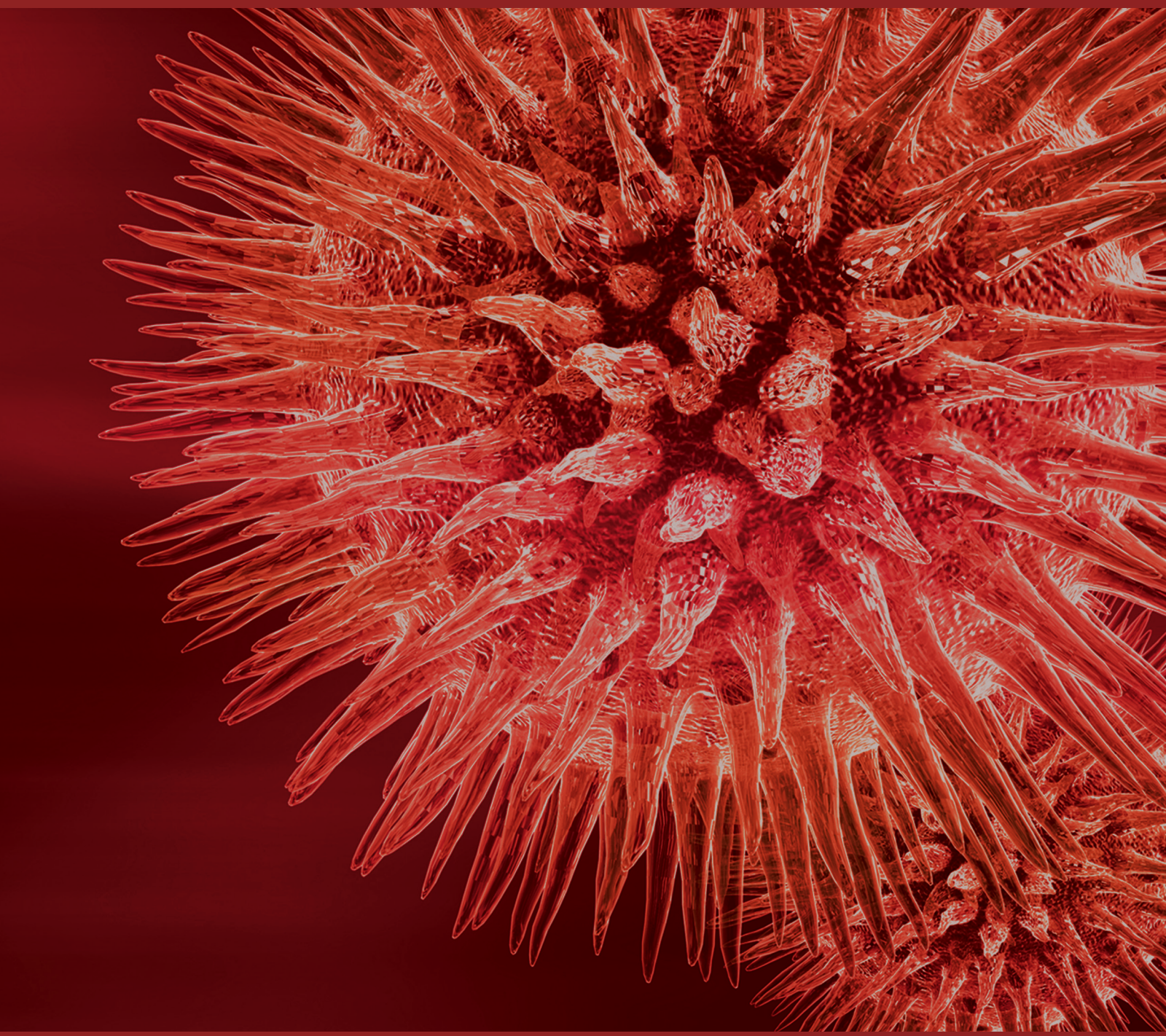
BioMed Research International

Novel Computational Approaches and Applications in Cancer Research

Guest Editors: Min Zhang, Lin Hua, Weiwei Zhai, and Yichuan Zhao



Novel Computational Approaches and Applications in Cancer Research

BioMed Research International

Novel Computational Approaches and Applications in Cancer Research

Guest Editors: Min Zhang, Lin Hua, Weiwei Zhai, and Yichuan Zhao

Contents

Hindawi
BioMed Research International
Volume 2017, Article ID 9509280, 2 pages
http://dx.doi.org/10.1155/2017/9509280

Editorial

Novel Computational Approaches and Applications in Cancer Research

Min Zhang,[1] Lin Hua,[2] Weiwei Zhai,[3] and Yichuan Zhao[4]

[1]*Department of Statistics, Purdue University, West Lafayette, IN, USA*
[2]*School of Biomedical Engineering, Capital Medical University, Beijing 100069, China*
[3]*Genome Institute of Singapore, Singapore*
[4]*Department of Mathematics and Statistics, Georgia State University, Atlanta, GA, USA*

Correspondence should be addressed to Min Zhang; minzhang@purdue.edu

Received 17 November 2016; Accepted 17 November 2016; Published 21 March 2017

Cancer remains one of the leading causes of global morbidity and mortality. As an enormous health burden worldwide, cancer touches every geographic region and is growing at an alarming pace. It is projected that 21.7 million new cases and 13.0 million deaths will occur in 2030 alone. To tackle this vicious disease effectively, a concerted effort by both research and healthcare communities is required to yield significant advances in cancer research and therapy. With the recent developments of high-throughput biotechnologies for genomes, proteomes, and transcriptomes, it is essential to develop innovative computational methods for comprehensive data analysis to improve our understanding of cancer initiation, progression, and metastasis. Therefore, novel computational and statistical methods are needed to analyze each type of omics data, to integrate multiple types of data across platforms, and to discover potential cancer-related biomarkers that can shed light on early detection, monitor disease progression, and eventually facilitate the development of personalized therapy of cancer.

The articles contained in the present issue include basic scientific studies focused on novel computational approaches and tools to analyze high-throughput multiplatform cancer data. In addition, image-based biomarkers were also discussed for early detection of the disease.

Diagnosis of tumor and definition of tumor borders intraoperatively are primarily based on the visualization modalities. However, intraoperative fast histopathology is often not sufficient. The contribution by A. Kamen et al. in "Automatic Tissue Differentiation Based on Confocal Endomicroscopic Images for Intraoperative Guidance in Neurosurgery" proposes an automated endomicroscopic tissue differentiation algorithm based on the machine learning theory. This algorithm offers a useful component to an intraoperative pathology system for guiding the resection procedure based on cellular level information.

Breast cancer is one of the most commonly diagnosed cancers in women all over the world. Osteopontin (OPN) is overexpressed in breast cancers, while its clinical and prognostic significance remain unclear. The contribution by C. Hao et al. in "Prognostic Value of Osteopontin Splice Variant-c Expression in Breast Cancers: A Meta-Analysis" proposes assessing the prognostic value of OPN, especially its splice variants, in breast cancers from eligible studies concerning the OPN and OPN-c expression. It concludes that the high level of OPN-c is suggested to be more reliably associated with poor survival in breast cancer patients.

Apolipoprotein E *(ApoE)* $\varepsilon 4$ allele has been proved to be a risk gene of late-onset Alzheimer's disease. It is very important to look for sensitive and reliable biomarkers in earliest stages. In the paper by Y. Liang "Frequency Specific Effects of *ApoE* $\varepsilon 4$ Allele on Resting-State Networks in Nondemented Elders," the authors applied resting-state functional magnetic resonance imaging to examine the *ApoE* $\varepsilon 4$ allele effects on functional connectivity of the default mode network and the salience network (SN). They conclude a frequency dependent effect of resting-state signals when investigating resting-state networks functional connectivity.

With the accumulation of large scale omic data, finding genomic features (variables) associated with clinical outcomes is an important topic for precision medicine. Variable selection provides a powerful tool for this practical need. In the manuscript entitled "High Dimensional Variable Selection with Error Control," S. Kim and S. Halabi present a sequential method based on false discovery rate (FDR) and iterative sure independence screening (ISIS). On the basis of both simulation study and real data analyses, the authors demonstrate the utility and statistical properties of the new method.

Finding correlation between epigenetic changes and certain phenotypes of interest is an important and hotly debated topic. In the article entitled "An Efficient Approach to Screening Epigenome-Wide Data," M. A. Ray et al. propose an improved and more efficient screening method to look for (filter) informative DNA methylation sites. By incorporating surrogate variable analysis (SVA) and identifying unknown latent variables, the proposed method shows superior performance in identifying epigenetic changes associated with maternal smoking. The developed method has been implemented into an efficient and user-friendly R package.

We hope that the methods proposed in this special issue could help discover potential cancer-related biomarkers or therapy targets and facilitate the biological experiments and biological technology development for cancer-related research.

Acknowledgments

We would like to express our sincere thanks and appreciation to all the authors for their contributions and the reviewers for their valuable inputs as well as constructive critiques to make this special issue possible.

Min Zhang
Lin Hua
Weiwei Zhai
Yichuan Zhao

Hindawi
BioMed Research International
Volume 2017, Article ID 9823501, 8 pages
http://dx.doi.org/10.1155/2017/9823501

Research Article

Frequency Specific Effects of *ApoE* ε4 Allele on Resting-State Networks in Nondemented Elders

Ying Liang,[1,2] Zhenzhen Li,[1,2] Jing Wei,[1,2] Chunlin Li,[1,2] Xu Zhang,[1,2] and Alzheimer's Disease Neuroimaging Initiative

[1]*School of Biomedical Engineering, Capital Medical University, Beijing 100069, China*
[2]*Beijing Key Laboratory of Fundamental Research on Biomechanics in Clinical Application, Capital Medical University, Beijing 100069, China*

Correspondence should be addressed to Xu Zhang; zhangxu@ccmu.edu.cn

Received 23 June 2016; Revised 17 October 2016; Accepted 7 November 2016

Academic Editor: Yichuan Zhao

We applied resting-state functional magnetic resonance imaging (fMRI) to examine the Apolipoprotein E (*ApoE*) ε4 allele effects on functional connectivity of the default mode network (DMN) and the salience network (SN). Considering the frequency specific effects of functional connectivity, we decomposed the brain network time courses into two bands: 0.01–0.027 Hz and 0.027–0.08 Hz. All scans were acquired by the Alzheimer's Disease Neuroscience Initiative (ADNI). Thirty-two nondemented subjects were divided into two groups based on the presence ($n = 16$) or absence ($n = 16$) of the *ApoE* ε4 allele. We explored the frequency specific effects of *ApoE* ε4 allele on the default mode network (DMN) and the salience network (SN) functional connectivity. Compared to ε4 noncarriers, the DMN functional connectivity of ε4 carriers was significantly decreased while the SN functional connectivity of ε4 carriers was significantly increased. Many functional connectivities showed significant differences at the lower frequency band of 0.01–0.027 Hz or the higher frequency band of 0.027–0.08 Hz instead of the typical range of 0.01–0.08 Hz. The results indicated a frequency dependent effect of resting-state signals when investigating RSNs functional connectivity.

1. Introduction

Apolipoprotein E (*ApoE*) ε4 allele has been proved to be a risk gene of late-onset Alzheimer's Disease (AD) [1]. It may cause a variety of functional and structural changes in human brain [2, 3] and is associated with greater amyloid-β (Aβ) accumulation and neurofibrillary tangles than ε3 [4]. It already shows a subtle decline in episodic memory many years before the development of dementia [5]. As there is no effective treatment of AD now, looking for sensitive and reliable biomarkers in earliest stages is very important.

Imaging technologies like functional magnetic resonance imaging (fMRI) offered an opportunity to detect the effects of gene on the brain function by blood oxygenation level dependent (BOLD). Nowadays, emerging computational tools made it possible to study brain networks instead of single brain region in a stereo vision. Specific cortical regions which are spatially separated from brain functional networks to complete cognitive task [6] and neurodegenerative diseases such as AD target specific large-scale brain networks [7]. It has been suggested that two resting-state networks (RSNs), salience network (SN) and the default mode network (DMN) [8], where the atrophy caused by dementia is largely concentrated, play an essential role in AD. A few previous studies have suggested ApoE ε4 allele may affect the activity of DMN and SN [9–11].

Previous studies mainly examined resting-state fMRI activities in the frequency between 0 and 0.08 Hz as this frequency band was thought to be associated with neuronal fluctuations [12], but more and more researchers suggested that functional connectivity may be frequency specific [13]. Complex activities in human brain produce different neuronal firing rate; different frequency may correspond to different cognitive courses [14]. A fMRI combined electroencephalography (EEG) study suggested that each resting-state network corresponds to a specific frequency rhythm [15].

Table 1: Demographic and neuropsychological characteristics of *ApoE* ε4 carriers and noncarriers.

	ApoE ε4 carriers ($n = 16$)	*ApoE* ε4 noncarriers ($n = 16$)	T-value (χ^2)	P
Age (years)	64.13 ± 6.59	63.13 ± 4.86	0.49	0.629
Education (years)	10.75 ± 3.15	10.63 ± 2.45	0.13	0.901
Gender (M/F)	7/9	8/8	0.12	0.723
MMSE	26.50 ± 2.03	28.75 ± 1.00	−3.97	$<0.001^{***}$
RAVLT	3.19 ± 2.08	7.00 ± 2.31	−4.98	$<0.001^{***}$

Values are mean ± standard deviation or numbers of participants. The differences in demographics and neuropsychological scores between the two groups were tested for significance with two-sample t-tests. The P value for gender distribution in the two groups was obtained using a Chi-square test. $^{***}P < 0.01$. MMSE, Mini-Mental Status Examination; RAVLT, Ray Auditory Verbal Learning Test.

Although abnormal resting-state networks were observed in both normal aging and *ApoE* ε4 carriers [9, 16–18], whether the abnormalities are frequency specific is still unknown. Han et al. used 0.027 Hz as dividing level of the low-frequency fluctuations (ALFF) and found that the ALFF abnormalities showed disparate spatial patterns in each frequency band [19]. Many studies also have reported the fMRI signals at both lower and higher frequencies contain important physiological significance; they may lose information if they are considered as a whole [20, 21].

In the present study, we aimed to utilize resting-state fMRI to examine the *ApoE* ε4 allele effects on functional connectivity of DMN and SN. Considering the frequency specific effects of functional connectivity, we decomposed the brain network time courses into two bands: 0.01–0.027 Hz and 0.027–0.08 Hz. We sought to determine (1) whether the *ApoE* ε4 carriers show abnormal resting-state network functional connectivity; (2) whether the functional connectivity abnormalities are frequency specific.

2. Materials and Methods

2.1. Subjects. All data used in this current study were obtained from the ADNI database (http://adni.loni.usc.edu). For the current study we randomly included a total of 16 *ApoE* ε4 carriers (genotypes ε4/ε4 and ε4/ε3) and 16 age- and gender-matched noncarriers (genotype ε3/ε3) from ADNI. Individuals with the ε2 allele were excluded due to its possible protective effects [22]. Only the baseline 3T scans of each subject was utilized. Inclusion criteria of all subjects in this study were aged between 55 and 80, with a Mini Mental State Examination (MMSE) score ≥ 24, lack of MCI or dementia, and a Clinical Dementia Rating-Sum of Boxes (CDR-SB) score of 0. The study was approved by the Institutional Review Boards of all of the participating institutions of ADNI, and informed written consent was written from all participants.

2.2. Neuropsychological Tests and ApoE Genotyping. The cognitive function scores used in this study were downloaded from the ADNI database. In this study, we focused on the results of general cognitive ability tests using Mini Mental State Examination (MMSE) and episodic memory using Ray Auditory Verbal Learning Test (RAVLT) (Table 1). *ApoE* genotyping was analysed from DNA samples of each participant's blood cells, applying an *ApoE* genotyping kit.

2.3. Data Acquisition. Subjects were scanned using a 3.0-Tesla Philips MRI scanner. Resting-state fMRI images were obtained using an echo-planar imaging (EPI) sequence (repetition time (TR) = 3000 ms, echo time (TE) = 30 ms, flip angle = 80°, number of slices = 48, voxel size = 3 mm × 3 mm × 3 mm, slice thickness = 3.3 mm, and voxel matrix = 64 × 64).

2.4. Imaging Preprocessing. Image preprocessing and analysis were performed using Statistical Parametric Mapping (SPM8, http://www.fil.ion.ucl.ac.uk/spm), Resting-State fMRI Data Analysis Toolkit (REST; Song et al., http://restfmri.net), and the Data Processing Assistant for resting-state fMRI (DPARSF, Yan and Zang; http://restfmri.net/forum/DPARSF). The first 10 volumes of the rest scans of each subject were removed for the signal equilibrium and for subject's adaptation to the scanning noise. The functional images left were preprocessed including slice timing, motion correction (exclusion threshold was set as 3 mm for the linear translation), spatial normalization to template in Montreal Neurological Institute (MNI) coordinate space, and resampling with $3 \times 3 \times 3\,mm^3$. Then all images were smoothed with a 4 mm full-width half-maximum (FWHM) Gaussian kernel. Furthermore, the resting-state fMRI data were linearly detrended and processed with regression correction for several nuisance covariates including six motion parameters, white matter signal, global mean signal, and cerebrospinal fluid signal. After preprocessing, data of 2 subjects (1 *ApoE* ε4 carrier; 1 *ApoE* ε4 noncarrier) were excluded from the following analyses due to excessive motion.

2.5. Frequency Division and Functional Connectivity Calculation. After preprocessing, we used low-pass/band-pass filters in REST to generate different data sets including three specific frequency bands: 0.01–0.027 Hz, 0.027–0.08 Hz, and the typical range of 0.01–0.08 Hz. We followed the literature [23] to define seed regions of interest (ROI) of the DMN and SN; 15 ROIs were derived as a 6 mm sphere around the coordinates in MNI space (Table 2). Within the DMN, nine regions were investigated: medial prefrontal cortex (mPFC), left lateral parietal (lLP), right lateral parietal (rLP), posterior cingulate cortex (PCC), left inferior temporal (liTmp), right inferior temporal (riTmp), medial thalamus (mdThal), left posterior cerebellum (lpCBLM), and right posterior cerebellum (rpCBLM). In the SN, six regions were investigated: left

Table 2: Regions and MNI coordinates of the DMN and SN.

ROI	MNI coordinates
Default mode network	
Posterior cingulate cortex (PCC)	(0, −51, 29)
Medial prefrontal cortex (mPFC)	(0, 61, 22)
Left lateral parietal (lLP)	(−48, −66, 34)
Right lateral parietal (rLP)	(53, −61, 35)
Left inferior temporal (liTmp)	(−65, −22, −9)
Right inferior temporal (riTmp)	(61, −21, −12)
Medial thalamus (mdThal)	(0, −9, 7)
Left posterior cerebellum (lpCBLM)	(−28, −82, −32)
Right posterior cerebellum (rpCBLM)	(26, −89, −34)
Salience network	
Right anterior cingulate cortex (rPG-ACC)	(12, 32, 30)
Left anterior cingulate cortex (lPG-ACC)	(−13, 34, 16)
Right ventral anterior cingulate cortex (rSG-ACC)	(10, 34, −6)
Putamen (Put)	(−19, 3, 9)
Left insula (lIns)	(−42, 6, 4)
Right insula (rIns)	(43, 7, 2)

anterior cingulate cortex (lPG-ACC), right anterior cingulate cortex (rPG-ACC), right ventral anterior cingulate cortex (rSG-ACC), putamen (Put), left insula (lIns), and right insula (rIns). Then functional connectivity matrix of each network was produced by averaging the Blood-Oxygen-Level-Dependent (BOLD) signal across all voxels within these ROIs and computing Pearson's correlation coefficients. The correlation coefficients were then transformed into z scores using Fisher r-to-z transformation.

2.6. Statistical Methods. Data were analysed using statistical software (SPSS, version 22.0). Two-sample t-tests were used to examine the significance of group differences in age, education, and neuropsychological scores, as well as the group differences of RSNs functional connectivity matrix in different frequency bands (0.01–0.027 Hz, 0.027–0.08 Hz, and 0.01–0.08 Hz). Gender data were calculated using χ^2-test. Pearson's correlation analyses were performed between the RSNs functional connectivity and the neuropsychological scores to explore whether the functional connectivity in different frequency bands is associated with the cognitive function.

3. Results

3.1. Demographic and Neuropsychological Measurements. The demographic characteristics and neuropsychological test scores of the ε4 carrier and noncarrier groups are shown in Table 1. There is no difference in age, gender, or education between the two groups of carriers and noncarriers. Scores on the MMSE and RAVLT were found to be significantly decreased in *ApoE* ε4 group compared with noncarrier group (Table 1).

3.2. Functional Connectivity. Figure 1 shows the RSNs ROIs and the two-sample t-test results of the connectivity maps of the two networks in each frequency band, respectively. We found some functional connectivities were sensitive to specific frequency band. In the DMN, the functional connectivity between the PCC and lLP, as well as the liTmp and rpCBLM connectivity, showed no significant group differences in the typical frequency band of 0.01–0.08 Hz but significant differences in the lower frequency band of 0.01–0.027 Hz (Table 3). In the SN, the connectivity between lPG-ACC and rSG-ACC, the connectivity between left insula and right insula, the connectivity between putamen and right insula all showed no significant differences in the typical frequency band of 0.01–0.08 Hz. But the first two connectivity group differences become significant in the lower frequency band of 0.01–0.027 Hz while the latter was significant in the frequency band of 0.027–0.08 Hz (Table 4).

In all the significant group differences, the DMN functional connectivities were decreased in the *ApoE* ε4 carriers compared with noncarriers, while the SN connectivities were increased in the *ApoE* ε4 carriers compared with noncarriers (Figure 1).

3.3. Correlation. In the DMN, at the frequency band of 0.01–0.027 Hz, we found significant positive correlations between the PCC and rLP connectivity ($r = 0.40$, $P = 0.028$) and liTmp-rp and rpCBLM connectivity ($r = 0.46$, $P = 0.010$) with RAVLT scores. At the frequency band of 0.027–0.08 Hz, the rLP and rpCBLM connectivity ($r = 0.43$, $P = 0.017$), the liTmp-rp and mdThal connectivity ($r = 0.39$, $P = 0.035$), and rLP and mdThal connectivity ($r = 0.53$, $P = 0.003$) showed significantly positive correlations with the RAVLT scores (Figure 2).

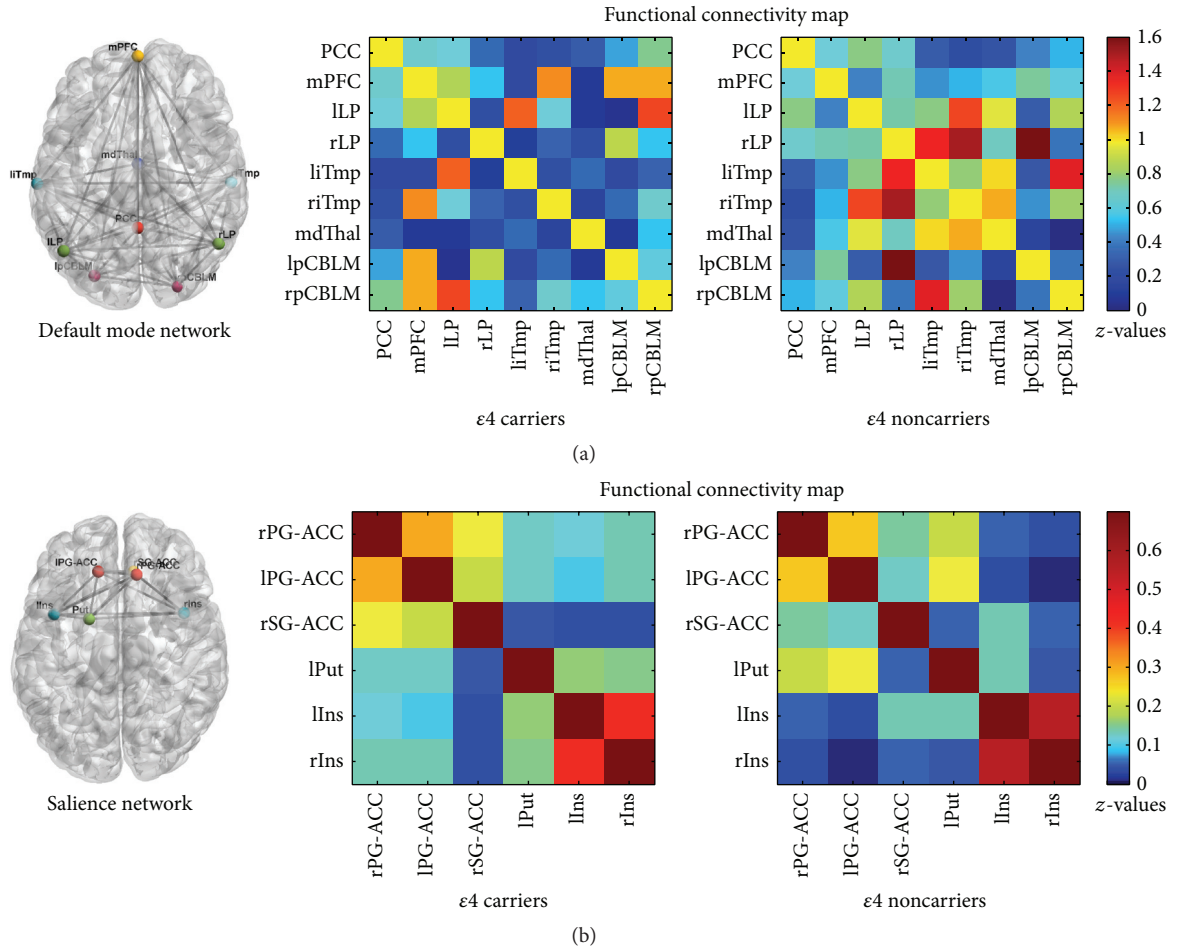


Figure 1: The default mode network and salience network nodes and functional connectivity map (0.01–0.08 Hz) of the ε4 carriers and noncarriers. Correlation matrix of all ROI pairs in each network.

Table 3: Group differences of default mode network functional connectivity at specific frequency bands.

Functional connectivity	0.01–0.027 Hz		0.027–0.08 Hz		0.01–0.08 Hz	
	T-value	P value	T-value	P value	T-value	P value
PCC & lLP	−2.334	0.027*	−1.187	NS	−1.683	NS
PCC & rLP	−2.418	0.022*	−1.696	NS	−2.186	0.037*
liTmp & rpCBLM	−2.083	0.047*	−1.142	NS	−2.005	NS
lLP & mdThal	−1.390	NS	−2.212	0.035*	−2.223	0.034*
lLP & rpCBLM	0.719	NS	−2.907	0.007**	−2.495	0.019*
rLP & rpCBLM	0.210	NS	−2.331	0.027*	−2.357	0.026*
liTmp & mdThal	−0.653	NS	−2.402	0.023*	−3.015	0.005**
rLP & mdThal	−1.042	NS	−2.002	NS	−2.153	0.040*

* $P < 0.05$; ** $P < 0.01$; NS, not significant. $T < 0$ represented the functional connectivity in *ApoE* ε4 carriers was lower than the noncarriers. $T > 0$ represented the functional connectivity in *ApoE* ε4 carriers was higher than the noncarriers.
PCC, posterior cingulate cortex; lLP, left lateral parietal; rLP, right lateral parietal; liTmp, left inferior temporal; rpCBLM, right posterior cerebellum; mdThal, medial thalamus.

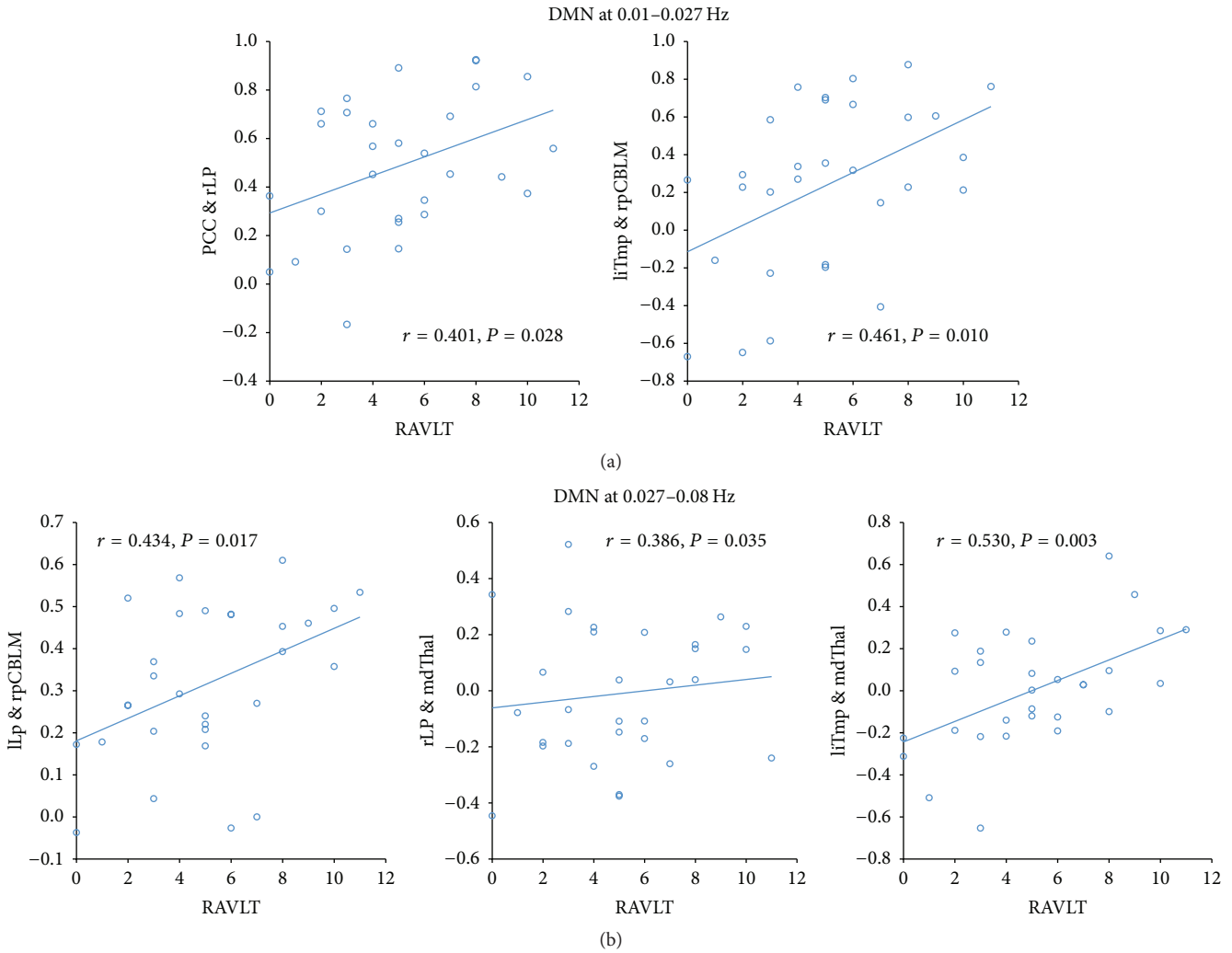


Figure 2: Relationship between altered connectivity and cognition at specific frequency bands in default mode network (0.01–0.027 Hz and 0.027–0.08 Hz).

Table 4: Group differences of salience network functional connectivity at specific frequency bands.

Functional connectivity	0.01–0.027 Hz		0.027–0.08 Hz		0.01–0.08 Hz	
	T-value	P value	T-value	P value	T-value	P value
lPG-ACC & rSG-ACC	2.751	0.010*	−0.100	NS	1.079	NS
rPG-ACC & rIns	2.054	0.049*	2.191	0.037*	2.495	0.019*
lIns & rIns	−2.187	0.037*	−1.354	NS	−1.983	NS
rSG-ACC & lIns	0.698	NS	3.244	0.003**	2.550	0.017*
lPG-ACC & rIns	1.565	NS	2.117	0.043*	2.281	0.030*
Put & rIns	0.819	NS	2.398	0.023*	1.670	NS

* $P < 0.05$; ** $P < 0.01$; NS, not significant. $T < 0$ represented the functional connectivity in *ApoE* ε4 carriers was lower than the noncarriers. $T > 0$ represented the functional connectivity in *ApoE* ε4 carriers was higher than the noncarriers.
lPG-ACC, left anterior cingulate cortex; rSG-ACC, right ventral anterior cingulate cortex; rPG-ACC, right anterior cingulate cortex; rIns, right insula; lIns, left insula; Put, putamen.

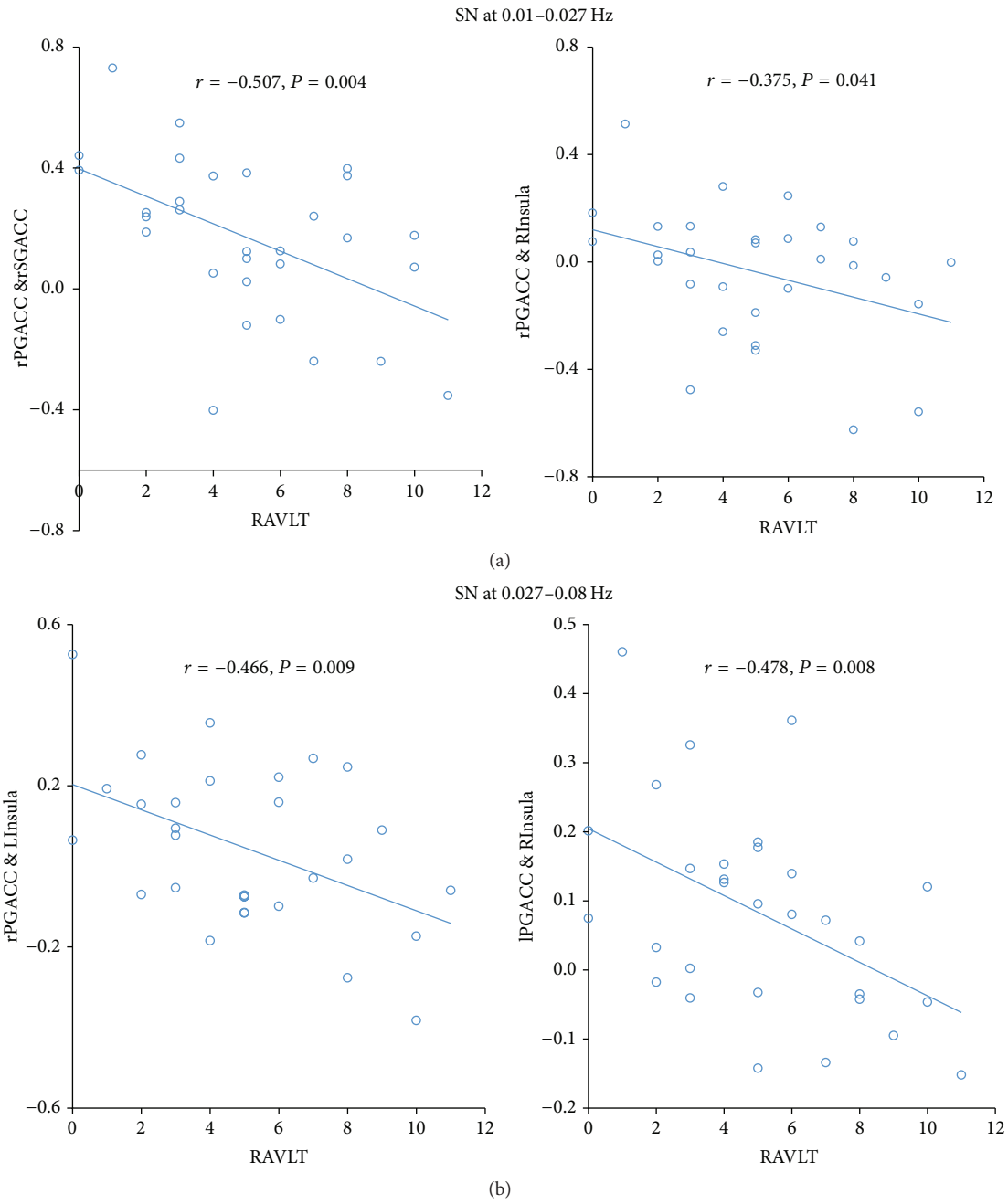


Figure 3: Relationship between altered connectivity and cognition at specific frequency bands in salience network (0.01–0.027 Hz and 0.027–0.08 Hz).

In the SN, at the frequency band of 0.01–0.027 Hz, we found significant negative correlations between the rPG-ACC and rSG-ACC connectivity ($r = -0.51$, $P = 0.004$) and rPG-ACC and rIns connectivity ($r = -0.38$, $P = 0.041$) with RAVLT scores. At the frequency band of 0.027–0.08 Hz, the rPG-ACC and lIns connectivity ($r = -0.47$, $P = 0.009$) and lPG-ACC and rIns connectivity ($r = -0.48$, $P = 0.008$) showed significantly negative correlations with the RAVLT scores (Figure 3).

4. Discussion

In the current study, we examined the functional connectivity changes in DMN and SN in the nondemented *ApoE* ε4 carriers at three different frequency bands (the typical range of 0.01–0.08 Hz, 0.01–0.027, and 0.027–0.08 Hz). Generally, the DMN functional connectivities were decreased while the SN connectivities were increased in the *ApoE* ε4 carriers compared with noncarriers. Importantly, we found that many

functional connectivities showed significant differences at the lower frequency band of 0.01–0.027 Hz and the higher frequency band of 0.027–0.08 Hz instead of the typical range of 0.01–0.08 Hz. The results indicated a frequency dependent effect of resting-state signals when investigating RSNs functional connectivity.

Possession of $\varepsilon 4$ allele disrupts the cognition, especially episodic memory at an early time [24]. We also found a significant decline of the general cognitive ability and episodic memory in $\varepsilon 4$ carriers, which illustrated that the $\varepsilon 4$ allele may affect the cognition long before the conversion to AD. More and more studies proved that brain cognitive functions were performed on the base of specific resting-state networks such as DMN and SN [25]. RSNs and their functional connectivity patterns have already shown a potential power of disease diagnosis and prediction [10, 26].

DMN and SN abnormalities were widely found in AD studies, implying that early detection of RSNs changes can offer opportunities to distinguish AD patients from healthy people in early stage [27, 28]. Comparing to the $\varepsilon 4$ noncarriers, our results showed that several functional connectivities in DMN, such as PCC-lLP connectivity, were significantly reduced in the $\varepsilon 4$ carriers. Meanwhile, the SN showed widely increased functional connectivity between regions like lPG-ACC and rSG-ACC. These changing trends of functional connectivity we found in the $\varepsilon 4$ carriers were consistent with the results of early AD [8, 29]. Functional connectivity describes the degree of the dynamic and synchronized oscillations between brain regions. These imaging biomarkers may help a better understanding of early disease pathogenesis and measuring genetic effects on brain, indicated by the significant correlations between the frequency specific functional connectivity and RAVLT scores in our results.

Exploring the high risk genotypes of diseases like AD may reveal the early disease-causing effects. With the development of novel algorithms and theories, nowadays we can correlate the neuroimaging data with the genetic data and discover how genotype affecting brain function and connectivity just as the phonotype, and to identify risks for neurological and psychiatric diseases [30]. The frequency characteristics of functional connectivity may be distinct for different brain networks [13]. The typical range of 0.01–0.08 Hz has its limitations. More and more researches indicate that the fMRI signals contain important physiological significance at specific frequency bands [21, 31, 32]. Moreover, these studies suggested that different frequency bands may play different role in the low-frequency fluctuations [33]. It is important to note that each frequency bands of neuronal oscillations are produced by different oscillators with distinct physiological functions and properties [6].

In the current study, we took advantages of the frequency characteristics of the resting-state fMRI signals to investigate the $\varepsilon 4$ allele effects on the DMN and SN. We found that some functional connectivities were not sensitive to the typical frequency band of 0.01–0.08 Hz, but when we segmented the frequency band into 0.01–0.027 Hz and 0.027–0.08 Hz group differences emerged. The results indicated that the abnormalities of brain functional connectivity in *ApoE* $\varepsilon 4$ carriers are associated with the choice of specific frequency bands. Our result supported that the RSNs functional connectivity can be modulated by frequency band. To the best of our knowledge, this is the first study on the frequency specific effect of the *ApoE* $\varepsilon 4$ allele on the RSNs functional connectivity.

There are also some limitations in our study. Although we found a frequency specific effect on the RSNs functional connectivity, the nature of these effects remains unclear. Future studies are necessary to investigate the physiological significance of the frequency specific effects.

5. Conclusions

In conclusion, we used genetic neuroimaging methods and found alterations of both DMN and SN functional connectivity in *ApoE* $\varepsilon 4$ allele carriers. Our results supported that the RSNs functional connectivity can be modulated by frequency band and emphasized the importance of considering frequency specific effects when investigating the genotypical effect on the brain function.

Disclosure

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

Data used in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

References

[1] E. H. Corder, A. M. Saunders, W. J. Strittmatter et al., "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families," *Science*, vol. 261, no. 5123, pp. 921–923, 1993.

[2] E. M. Reiman, R. J. Caselli, L. S. Yun et al., "Preclinical evidence of Alzheimer's disease in persons homozygous for the $\varepsilon 4$ allele for apolipoprotein E," *New England Journal of Medicine*, vol. 334, no. 12, pp. 752–758, 1996.

[3] A. J. Trachtenberg, N. Filippini, and C. E. Mackay, "The effects of APOE-$\varepsilon 4$ on the BOLD response," *Neurobiology of Aging*, vol. 33, no. 2, pp. 323–334, 2012.

[4] E. M. Reiman, K. Chen, X. Liu et al., "Fibrillar amyloid-β burden in cognitively normal people at 3 levels of genetic risk for Alzheimer's disease," *Proceedings of the National Academy*

of Sciences of the United States of America, vol. 106, no. 16, pp. 6820–6825, 2009.

[5] K. L. Lange, M. W. Bondi, D. P. Salmon et al., "Decline in verbal memory during preclinical Alzheimer's disease: examination of the effect of APOE genotype," *Journal of the International Neuropsychological Society*, vol. 8, no. 7, pp. 943–955, 2002.

[6] G. Buzsáki and A. Draguhn, "Neuronal olscillations in cortical networks," *Science*, vol. 304, no. 5679, pp. 1926–1929, 2004.

[7] W. W. Seeley, R. K. Crawford, J. Zhou, B. L. Miller, and M. D. Greicius, "Neurodegenerative diseases target large-scale human brain networks," *Neuron*, vol. 62, no. 1, pp. 42–52, 2009.

[8] J. Zhou, M. D. Greicius, E. D. Gennatas et al., "Divergent network connectivity changes in behavioural variant frontotemporal dementia and Alzheimer's disease," *Brain*, vol. 133, no. 5, pp. 1352–1367, 2010.

[9] E. T. Westlye, A. Lundervold, H. Rootwelt, A. J. Lundervold, and L. T. Westlye, "Increased hippocampal default mode synchronization during rest in middle-aged and elderly APOE $\varepsilon 4$ carriers: relationships with memory performance," *Journal of Neuroscience*, vol. 31, no. 21, pp. 7775–7783, 2011.

[10] W. Koch, S. Teipel, S. Mueller et al., "Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease," *Neurobiology of Aging*, vol. 33, no. 3, pp. 466–478, 2012.

[11] M. M. Machulda, D. T. Jones, P. Vemuri et al., "Effect of *APOE* $\varepsilon 4$ status on intrinsic network connectivity in cognitively normal elderly subjects," *Archives of Neurology*, vol. 68, no. 9, pp. 1131–1136, 2011.

[12] B. Biswal, F. Zerrin Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo−planar MRI," *Magnetic Resonance in Medicine*, vol. 34, no. 4, pp. 537–541, 1995.

[13] C. W. Wu, H. Gu, H. Lu, E. A. Stein, J.-H. Chen, and Y. Yang, "Frequency specificity of functional connectivity in brain networks," *NeuroImage*, vol. 42, no. 3, pp. 1047–1055, 2008.

[14] H. Lu, Y. Zuo, H. Gu et al., "Synchronized delta oscillations correlate with the resting-state functional MRI signal," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 46, pp. 18265–18269, 2007.

[15] D. Mantini, M. G. Perrucci, C. Del Gratta, G. L. Romani, and M. Corbetta, "Electrophysiological signatures of resting state networks in the human brain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 32, pp. 13170–13175, 2007.

[16] N. Filippini, B. J. MacIntosh, M. G. Hough et al., "Distinct patterns of brain activity in young carriers of the APOE-$\varepsilon 4$ allele," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 17, pp. 7209–7214, 2009.

[17] H.-J. Li, X.-H. Hou, H.-H. Liu, C.-L. Yue, Y. He, and X.-N. Zuo, "Toward systems neuroscience in mild cognitive impairment and Alzheimer's disease: a meta-analysis of 75 fMRI studies," *Human Brain Mapping*, vol. 36, no. 3, pp. 1217–1232, 2015.

[18] H.-J. Li, X.-H. Hou, H.-H. Liu, C.-L. Yue, G.-M. Lu, and X.-N. Zuo, "Putting age-related task activation into large-scale brain networks: a meta-analysis of 114 fMRI studies on healthy aging," *Neuroscience and Biobehavioral Reviews*, vol. 57, pp. 156–174, 2015.

[19] Y. Han, J. Wang, Z. Zhao et al., "Frequency-dependent changes in the amplitude of low-frequency fluctuations in amnestic mild cognitive impairment: a resting-state fMRI study," *NeuroImage*, vol. 55, no. 1, pp. 287–295, 2011.

[20] X.-N. Zuo, A. Di Martino, C. Kelly et al., "The oscillating brain: complex and reliable," *NeuroImage*, vol. 49, no. 2, pp. 1432–1445, 2010.

[21] L. Wei, X. Duan, C. Zheng et al., "Specific frequency bands of amplitude low-frequency oscillation encodes personality," *Human Brain Mapping*, vol. 35, no. 1, pp. 331–339, 2014.

[22] S. Suri, V. Heise, A. J. Trachtenberg, and C. E. Mackay, "The forgotten APOE allele: a review of the evidence and suggested mechanisms for the protective effect of APOE $\varepsilon 2$," *Neuroscience and Biobehavioral Reviews*, vol. 37, no. 10, pp. 2878–2886, 2013.

[23] M. R. Brier, J. B. Thomas, A. Z. Snyder et al., "Loss of intranetwork and internetwork resting state functional connections with Alzheimer's disease progression," *Journal of Neuroscience*, vol. 32, no. 26, pp. 8890–8899, 2012.

[24] L.-G. Nilsson, R. Adolfsson, L. Bäckman et al., "The influence of APOE status on episodic and semantic memory: data from a population-based study," *Neuropsychology*, vol. 20, no. 6, pp. 645–657, 2006.

[25] S. L. Bressler and V. Menon, "Large-scale brain networks in cognition: emerging methods and principles," *Trends in Cognitive Sciences*, vol. 14, no. 6, pp. 277–290, 2010.

[26] E. Challis, P. Hurley, L. Serra, M. Bozzali, S. Oliver, and M. Cercignani, "Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI," *NeuroImage*, vol. 112, pp. 232–243, 2015.

[27] M. D. Greicius, G. Srivastava, A. L. Reiss, and V. Menon, "Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 13, pp. 4637–4642, 2004.

[28] M. L. F. Balthazar, B. M. de Campos, A. R. Franco, B. P. Damasceno, and F. Cendes, "Whole cortical and default mode network mean functional connectivity as potential biomarkers for mild Alzheimer's disease," *Psychiatry Research—Neuroimaging*, vol. 221, no. 1, pp. 37–42, 2014.

[29] F. Agosta, M. Pievani, C. Geroldi, M. Copetti, G. B. Frisoni, and M. Filippi, "Resting state fMRI in Alzheimer's disease: beyond the default mode network," *Neurobiology of Aging*, vol. 33, no. 8, pp. 1564–1578, 2012.

[30] N. Jahanshad, D. P. Hibar, A. Ryles et al., "Discovery of genes that affect human brain connectivity: a genome-wide analysis of the connectome," in *Proceedings of the 9th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI '12)*, May 2012.

[31] X. Song, Y. Zhang, and Y. Liu, "Frequency specificity of regional homogeneity in the resting-state human brain," *PLOS ONE*, vol. 9, no. 1, Article ID e86818, 2014.

[32] R. Yu, Y.-L. Chien, H.-L. S. Wang et al., "Frequency-specific alternations in the amplitude of low-frequency fluctuations in schizophrenia," *Human Brain Mapping*, vol. 35, no. 2, pp. 627–637, 2014.

[33] A. T. Baria, M. N. Baliki, T. Parrish, and A. V. Apkarian, "Anatomical and functional assemblies of brain BOLD oscillations," *Journal of Neuroscience*, vol. 31, no. 21, pp. 7910–7919, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 8209453, 11 pages
http://dx.doi.org/10.1155/2016/8209453

Research Article

High Dimensional Variable Selection with Error Control

Sangjin Kim and Susan Halabi

Department of Biostatistics and Bioinformatics, Duke University Medical Center, Box 2717, Durham, NC 27710, USA

Correspondence should be addressed to Susan Halabi; susan.halabi@duke.edu

Received 3 April 2016; Accepted 25 May 2016

Academic Editor: Weiwei Zhai

Background. The iterative sure independence screening (ISIS) is a popular method in selecting important variables while maintaining most of the informative variables relevant to the outcome in high throughput data. However, it not only is computationally intensive but also may cause high false discovery rate (FDR). We propose to use the FDR as a screening method to reduce the high dimension to a lower dimension as well as controlling the FDR with three popular variable selection methods: LASSO, SCAD, and MCP. *Method.* The three methods with the proposed screenings were applied to prostate cancer data with presence of metastasis as the outcome. *Results.* Simulations showed that the three variable selection methods with the proposed screenings controlled the predefined FDR and produced high area under the receiver operating characteristic curve (AUROC) scores. In applying these methods to the prostate cancer example, LASSO and MCP selected 12 and 8 genes and produced AUROC scores of 0.746 and 0.764, respectively. *Conclusions*. We demonstrated that the variable selection methods with the sequential use of FDR and ISIS not only controlled the predefined FDR in the final models but also had relatively high AUROC scores.

1. Introduction

Prognosis will continue to play a critical role in patient management and decision making in 21st century medicine. Advanced technologies for genomic profiling are now available and they include millions of sets of molecular data in these assays. A critical element of personalized medicine is utilizing and implementing validated diagnostic signatures (or classifiers) for diagnosing or treating cancer patients. These signatures are built and validated utilizing common statistical methods and machine learning tools. For example, the Decipher signature has been developed as a prognostic model to predict metastasis after radical prostatectomy in patients with prostate cancer [1]. The Decipher score is a 22-feature genomic classifier that has been used to predict metastasis and has been independently validated for prediction of prostate metastasis [2–5]. Another example is oncotypeDx that has been used to stratify randomization and guide treatment in women with breast cancer [6].

A vital step in model building is data reduction. It is assumed that there are several variables that are associated with the clinical outcome in the large dimensional data. The main purpose of the variable selection is to detect only those variables related to the response. Variable selection is composed of two steps: screening and model building. The screening step is to reduce the large number of variables into moderate size while maintaining most of the informative variables relevant to the clinical response. In contrast, in the model building step, investigators develop a single best model utilizing a proper evaluation criterion.

Penalized variable selection methods have played a key role in identifying important prognostic models in several areas in oncology [7–9]. Many articles focused on the development of methodologies related to "small N and large P" with the advent of high throughput technology in cancer. The sure independence screening (SIS) was introduced to reduce the high dimension to below the sample size to efficiently select the best subset of variables to predict clinical responses [10]. Although this approach is popular, it does not perform well under some situations. First, unimportant variables that are heavily correlated with important variables are more highly likely to be selected than important variables that are weakly associated with the response. Second, important variables that are not marginally significantly related to the response are screened out. Finally, there may be collinearity between variables that may impact the calculations of the individual predictors.

The iterative sure independence screening (ISIS) was proposed to overcome the above issues. The procedure is to apply iteratively high dimensional variable screening followed by the proper scale of variable selection until the best subset of variables with high predictive accuracy is obtained. ISIS screening, however, is also computationally intensive and leads to high false discovery rate (FDR) in ultra-high dimensional setting ($P \gg 1$ mils).

The oncology literature is rich in articles related to the use of validated signatures. Despite their abundance, comparisons and the performance of these various methods have not been studied. We propose to use the false discovery rate (FDR) of the multiple testing correction methods as a screening method to reduce the high dimension to lower dimension as well as controlling the false discovery rate in the final model. We investigate the feasibility of the sequential use of FDR screening method with the ISIS and utilize three popular variable selection methods: LASSO [11], SCAD [12, 13], and MCP [14], through the extensive simulation studies. To the best of our knowledge, this is the first paper that thoroughly analyzes and compares the performance of the variable selection methods with the sequential use of FDR and ISIS screening methods. We use a prostate cancer signature as an example [1] where the number of probes is around 1.4 million and the clinical outcome is binary in nature: presence of metastasis (presence of metastasis = 1, no metastasis = 0) by fitting models based on the simulation results.

In addition, we provide a broad review of the existing penalized variable selection methods with screening methods. The remainder of this paper is organized as follows. In Section 2, we provide general details of the screening methods of FDR [15] and ISIS [10] and the variable selection methods with the penalized logistic regression. In Section 3, we describe the simulation studies and in Section 4, we summarize the results of the simulations. We then apply the best screening methods from the simulation studies to the real data in Section 5. Finally in Section 6, we discuss our findings.

2. Methods

We divide this section into several subsections describing the methods used in our paper. The screening section briefly discusses commonly used methods that reduce high dimensionality: false discovery rate (FDR) and iterative sure independence screening (ISIS). We then describe the methods needed to assess variable selection models. The final section considers three existing popular variable selection methods with the logistic regression. All simulations and calculations were carried out using glmnet and ISIS packages in the R library, and the code is available at https://www.duke.edu/halab001/FDR.

2.1. Benjamini and Hochberg False Discovery Rate (FDR). The false discovery rate is defined as the expected proportion of incorrectly rejected null hypotheses. That is,

$$E\left(\frac{V}{R} \mid R > 0\right), \tag{1}$$

where V is the number of falsely rejected hypotheses and R is the total number of rejected hypotheses. We focus on the Benjamini and Hochberg FDR [15] method as a screening method in the simulation studies and application. Briefly, the procedure works as follows. Let q denote the FDR, where $q \in (0, 1)$.

(1) Let $p_1, \ldots, p_m$ be the p values of the m hypothesis tests and sort them from smallest to largest. Denote these ordered p values by $p_{(1)}, \ldots, p_{(m)}$.

(2) Let $\widehat{k} = \max\{k : p_{(k)} \leq (k \times q)/m\}$, $k = 1, 2, \ldots, m$.

If $\widehat{k} > 1$, then reject $p_{(1)}, \ldots, p_{(k)}$ and if $\widehat{k} = 0$, then there is no rejection of the m hypothesis.

2.2. Iterative Sure Independence Screening (ISIS). The ISIS method was proposed to overcome the difficulties caused by the sure independence screening [16]. Briefly, the algorithm works in the following way:

(1) The likelihood of marginal logistic regression (LMLR) is computed for every $j \in S = \{1, 2, \ldots, p\}$. Then d which is $N/4\log(N)$ of the top ranked variables of the descending order list of the LMLR is selected to obtain the index set $\widehat{I}_1$.

(2) Apply those variables in $\widehat{I}_1$ to the penalized logistic models to obtain a subset of indices $\widehat{M}_1$.

(3) For every variable $j \in \{S - \widehat{M}_1\}$, the likelihood of the marginal logistic regression condition on the variables in $\widehat{M}_1$ is solved. Then the likelihood estimators are sorted in descending order and then the d top ranked variables are selected to get the index set $\widehat{I}_2$.

(4) Apply those variables in $\widehat{I}_2 \cup \widehat{M}_1$ to the penalized logistic models to obtain a new index set $\widehat{M}_2$.

(5) Steps (3) and (4) are repeated until $\widehat{M}_l = d$ or $\widehat{M}_l = \widehat{M}_{l-1}$.

2.3. Regularizing Methods with Penalized Logistic Regression. The logistic regression is one of the most commonly used methods for assessing the relationship between a binary outcome and a set of covariates and building prognostic models of clinical outcomes. In addition, it is widely used in the classification of two classes such as the development of metastasis in prostate cancer [1]. The purpose of variable selection with the logistic regression model in high dimensional setting is to select the optimal subset of variables that will improve the prediction accuracy [17]. Variable selection in high dimensional setting is composed of two components: a likelihood function and a penalty function in order to obtain better estimates for prediction.

Let the covariates of ith individual be denoted as $x_i = (x_{i1}, \ldots, x_{ip})^T$ for $i = 1, \ldots, N$ and p is the total number of covariates. The penalized logistic regression is as follows:

$$-\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(p_i\right) + \left(1 - y_i\right)\log\left(1 - p_i\right)\right) + p\left(\beta\right), \quad i = 1, 2, \ldots, N, \tag{2}$$

where $p(\beta)$, a penalty, is function and y_i is 1 for cases and 0 for controls. The probability that ith individual is a case based on covariates' information is expressed as

$$p_i = \frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}, \quad i = 1, 2, \ldots, N. \tag{3}$$

The regression coefficients are obtained by minimizing the objective function (2).

One of the most popular penalty functions is the least absolute shrinkage and selection operator (LASSO) [11]. It forces the coefficients of unimportant variables to be set to 0 and then the LASSO has sparsity property. The LASSO estimates are obtained by minimizing the above penalized logistic regression form (2). It has a satisfactory performance in identifying a small number of representative variables. Though LASSO is widely used in most applications [18–21], its robustness is open to question as it has the tendency to randomly select one of the variables with high correlation and exclude the rest of the predictors [22]. Another disadvantage of LASSO is that it always chooses at most N (sample size) number of predictors even though there are more than N variables with true nonzero coefficients [23]. The coefficients estimates are obtained by minimizing the following objective function based on the likelihood function of logistic regression:

$$\widehat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ -\frac{1}{N} \cdot \sum_{i=1}^{N} \left(y_i \log\left(p_i\right) + \left(1-y_i\right)\log\left(1-p_i\right)\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right\}. \tag{4}$$

Another method commonly employed is the smoothly clipped absolute deviation (SCAD) with a concave penalty function that overcomes some of the limitations of the LASSO [12]. The coefficients from SCAD are solved by minimizing the following objective function:

$$\widehat{\beta}^{\text{scad}} = \underset{\beta}{\operatorname{argmin}} \left\{ -\frac{1}{N} \cdot \sum_{i=1}^{N} \left(y_i \log\left(p_i\right) + \left(1-y_i\right)\log\left(1-p_i\right)\right) + \sum_{i=1}^{p} f_{\lambda,\gamma}\left(\beta_j\right) \right\}. \tag{5}$$

The SCAD penalty function, $f_{\lambda,\gamma}(\beta_j)$, is defined by

$$f_{\lambda,\gamma}\left(\beta_j\right) = \lambda\gamma\left|\beta_j\right| \mathbf{I}_{(0\le|\beta_j|\le\lambda)} + \left(\frac{\lambda\gamma\left(\left|\beta_j\right|-\lambda\right) - \left(\left|\beta_j\right|^2 - \lambda^2\right)/2}{(a-1)} + \lambda^2 \right) \cdot \mathbf{I}_{(\lambda<|\beta_j|\le\lambda\gamma)} + \left(\frac{(\gamma+1)\lambda^2}{2} \right) \tag{6}$$

with $\lambda \ge 0$ and $\gamma > 2$.

The minimum concave penalty (MCP) is also a recognized method with SCAD, where the coefficients are estimated via minimization of the following objective function:

$$\widehat{\beta}^{\text{mcp}} = \underset{\beta}{\operatorname{argmin}} \left\{ -\frac{1}{N} \cdot \sum_{i=1}^{N} \left(y_i \log\left(p_i\right) + \left(1-y_i\right)\log\left(1-p_i\right)\right) + \sum_{i=1}^{p} f_{\lambda,\gamma}\left(\beta_j\right) \right\}. \tag{7}$$

The MCP penalty function, $f_{\lambda,\gamma}(\beta)$, is defined by

$$f_{\lambda,\gamma}\left(\beta\right) = \left(\frac{2\lambda\gamma\left|\beta_j\right| - {\beta_j}^2}{2\gamma} \right) \mathbf{I}_{(|\beta_j|\le\lambda\gamma)} + \left(\frac{\lambda^2\gamma}{2} \right) \mathbf{I}_{(|\beta_j|>\lambda\gamma)}, \tag{8}$$

for $\lambda \ge 0$ and $\gamma > 1$.

3. Simulation Studies

3.1. Simulation Setup. We performed extensive simulation studies to explore the performance of three popular variable selection methods: LASSO, SCAD, and MCP in high dimensional setting. We employed 10-fold cross validation to tune the regularization parameter for the methods. Figure 1 describes the schema of the simulation procedures.

Based on the logistic regression model, we generated the binary outcome and covariates for each simulation as follows. First, we generated $z_1, z_2, \ldots, z_p$ independently from $N(0, 1)$, and each of z_i is an $N \times 1$ vector. We defined $x_1 = z_1$, $x_i = \rho x_{i-1} + \sqrt{(1-\rho^2)}z_i$, where $i = 2, \ldots, p$, so that correlation of x_k and x_l was $\rho^{|k-l|}$ for some $\rho \in [0, 1)$. That is, the covariates were generated with serialized correlation structure (AR (1)). Next, we specified the true regression coefficients β. We fixed all of β's except the first 25 β's to be 0. The true nonzero β's were sampled independently from uniform distribution $[-1.5, 2]$. We considered 25 true effects of the regression coefficients since several classifiers including the Decipher score had selected 20–25 genes [1, 2] and because that number predicted reasonably well the outcome. The number of variables was fixed at $P = 100{,}000$, and the

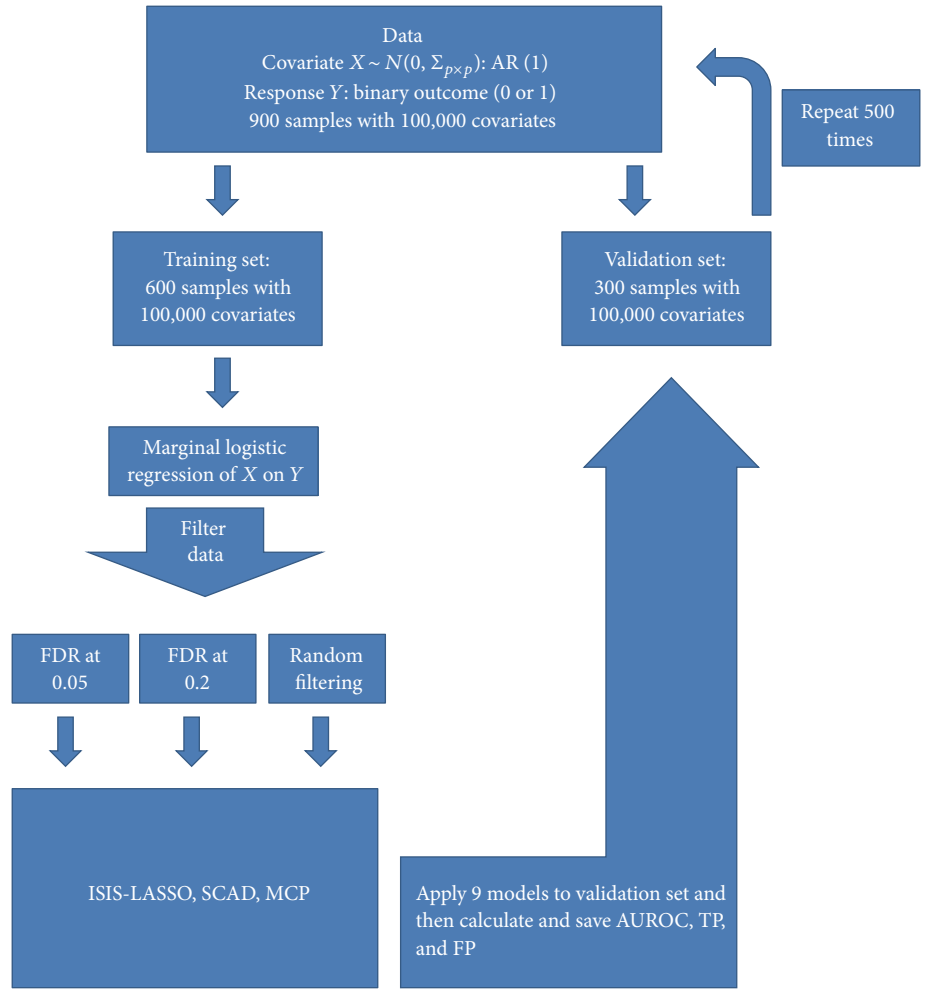


Figure 1: Diagram showing simulation procedures.

sample size was set at $N = 900$. Finally, the corresponding binary response y_i was simulated based on the Bernoulli distribution with the following:

$$y_i \sim \text{Bern}\left(p\left(x_i\right)\right), \quad p\left(x_i\right) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}}, \tag{9}$$

where $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$. Covariates were generated until the target number of 450 cases and 450 controls was reached.

We considered different simulation scenarios for the correlation matrix $\Sigma_{P\times P}$, $\rho = \{0, 0.1, 0.4\}$ among variables. Each simulation scenario was composed of the nine different models with the combination of the FDR, the ISIS, and the random filtering (RF_{1000}). The RF_{1000} selected 1,000 variables with the smallest unadjusted p values obtained from the marginal logistic regression with the three variable selection methods (LASSO, SCAD, and MCP). The reason we used RF_{1000} from the 100,000 potential variables was that the number of false discovery rates is low relative to the other random filtering (such as 2,000 or higher). Therefore, we considered the top 1,000 variables to be a reasonable number of variables screened as reference to be compared with our proposed methods.

We then simulated the data 500 times because of computational intensity. In each simulation, we randomly divided the data into two parts: the training set ($N = 600$) for model selection and the testing set ($N = 300$) for validation.

3.2. Metrics of Performance. We calculated the true positive rate (TP), the false positive rate (FP), the false discovery rate (FDR), the average number of false positives in the final model, the average model size, the average of the area under receiver operating characteristic (AUROC), and the number of screened true important variables through the FDR and the RF_{1000} to assess the impact of the FDR-ISIS screening method with the three variable selection methods.

The true positive rate (TP), also called sensitivity, is the proportion of positives that are identified correctly given true positives:

$$\text{True Positive rate (TP)} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \tag{10}$$

where TP is the number of the true positives and FN is the number of false negatives. The false positive rate is the

Table 1: The true positive rate (TP), the false positive rate (FP), and the false discovery rate (FDR), the average number of false positives (ANFP) in the final models, the final model size (size), the area under the curve (AUROC), and the number of filtered truly important variables from FDR and RF (# filter) under the low correlation coefficients, $\rho = 0.1$ among variables.

Screening	Method	TP	FP	FDR	ANFP	Size	AUROC	# filter
$FDR_{.05}$-ISIS	LASSO	0.22336 (0.00203)	$2.2e-06$ (0)	0.03086 (0.00303)	0.216 (0.02199)	5.8 (0.05816)	0.80285 (0.00157)	5.826 (0.05985)
	SCAD	0.22336 (0.00203)	$2.1e-06$ (0)	0.03071 (0.003)	0.214 (0.02157)	5.798 (0.05791)	0.80286 (0.00157)	
	MCP	0.22336 (0.00203)	$2.1e-06$ (0)	0.03071 (0.003)	0.214 (0.02157)	5.798 (0.05791)	0.80286 (0.00157)	
$FDR_{.20}$-ISIS	LASSO	0.2676 (0.00229)	$1.17e-05$ (0)	0.12618 (0.0055)	1.17 (0.05914)	7.86 (0.09344)	0.81965 (0.00154)	7.904 (0.09499)
	SCAD	0.2676 (0.00229)	$1.16e-05$ (0)	0.12522 (0.00547)	1.158 (0.05866)	7.848 (0.09288)	0.81964 (0.00154)	
	MCP	0.2676 (0.00229)	$1.13e-05$ (0)	0.12355 (0.00541)	1.134 (0.05698)	7.824 (0.09096)	0.81967 (0.00154)	
RF_{1000}-ISIS	LASSO	0.42112 (0.00236)	0.0001237 (0)	0.53987 (0.00265)	12.364 (0.06298)	22.892 (0.01389)	0.83244 (0.00149)	13.286 (0.06807)
	SCAD	0.42336 (0.00243)	0.0001226 (0)	0.53635 (0.00276)	12.26 (0.06622)	22.844 (0.01624)	0.83196 (0.00147)	
	MCP	0.42656 (0.00249)	0.0001217 (0)	0.53261 (0.00283)	12.17 (0.06797)	22.834 (0.01666)	0.83191 (0.00147)	

( ): standard deviation.

proportion of incorrect identification as a true positive given true negatives. That is,

$$\text{False Positive rate (FP)} = \frac{\text{FP}}{(\text{TN} + \text{FP})}, \tag{11}$$

where the FP is the number of false positives and the TN is the number of true negatives. In addition, the average number of false positives (ANFP) was computed as the number of false positives that were selected in the final model out of 500 simulations. Furthermore, the average model size was computed as the number of variables selected in the final model out of 500 simulations.

Finally, the AUROC was utilized as a measure of the performance of the logistic regression and is the proportion of the time which a model predicts correctly given observations of a random positive and negative. A perfect model produces an AUROC = 1 whereas a random model has an AUROC = 0.5.

4. Simulation Results

We summarized the simulation results for $\rho = 0.1$, one of three correlation structures in Table 1, where all 25 important covariates were assumed to have linear effects. Table 1 presents the performance of the nine different models with the FDR, ISIS, and random filtering based on 500 simulations. The average true positive rates (TP) were 0.223 and 0.268 for the three variable selection methods using $FDR_{.05}$ − ISIS and $FDR_{.20}$ − ISIS. The average true positive rates (TP) of the LASSO, SCAD, and MCP with RF_{1000} − ISIS were 0.46013, 0.46365, and 0.46739, respectively. These values were much higher than the two FDR screening methods which were below 0.30. On the other hand, the three variable selection methods with RF_{1000} − ISIS selected several of the false positive variables that consequently increased the false positive rate (FP). LASSO, SCAD, and MCP with RF_{1000} − ISIS included a higher average number of the false positives of 12.364, 12.260, and 12.170, respectively. Although the FDR filtering method did not select a higher number of true important variables, this screening method reduced the false positive rates below the predefined target α.

The average numbers of the false positives in the final models with the FDR − ISIS methods were much smaller than that of using RF_{1000} − ISIS (Table 1). Specifically, the average numbers of the false positives in the final models with the LASSO, SCAD, and MCP with $FDR_{.05}$ − ISIS were 0.216, 0.214, and 0.214 with the corresponding standard deviations 0.0219, 0.0215, and 0.0215. As expected, the three variable selection methods with RF_{1000} − ISIS had selected a higher average model size of 22.8 than the FDR methods. Similar results were observed for $FDR_{.20}$ − ISIS. We also calculated the false discovery rate. The variable selection models with the FDR at the target $\alpha = 0.05$ and $\alpha = 0.20$ controlled the false discovery rate below α whereas over 40% of the finally selected variables were incorrectly selected using the random filtering methods. The average AUROC scores with RF_{1000} − ISIS were relatively higher than the $FDR_{.05}$ − ISIS and $FDR_{.20}$ − ISIS. Similar results were noted for independent and moderate correlation $\rho = \{0, 0.4\}$ as presented in Tables S1 and S2 in Supplementary Material available online at http://dx.doi.org/10.1155/2016/8209453.

Figure 2 presents the selection frequency for the 25 important variables under the three different screening methods. The x-axis denotes the variable name and the y-axis represents the frequency of selection out of 500 simulations. The variables not depicted on the x-axis in Figure 2 did not

Figure 2: Selection frequencies of each of the 25 variables across the LASSO, the SCAD, and the MCP during 500 simulations with $\rho = 0.1$. The x-axis depicts the names of the variables, and the y-axis is the frequency of variables selected out of 500 simulations. The variables not depicted on the x-axis in Figure 2 did not have any counts. Each of the three methods is identified by color in legend.

have any counts and thus were not selected in the simulation. The variables with the highest selection frequencies had true regression coefficients that were strongly associated with the clinical response. These variables were g07, g03, g19, g23, g01, g24, g12, and g06 and were selected over 100 times out of 500 simulations with average corresponding regression coefficients of 1.65, −1.45, 1.57, −1.17, 1.12, 0.968, 0.963, and −1.01 (see Table S3 in Supplementary File). The coefficients of the eight variables were ranked the highest among the 25 absolute values of the true regression coefficients which had

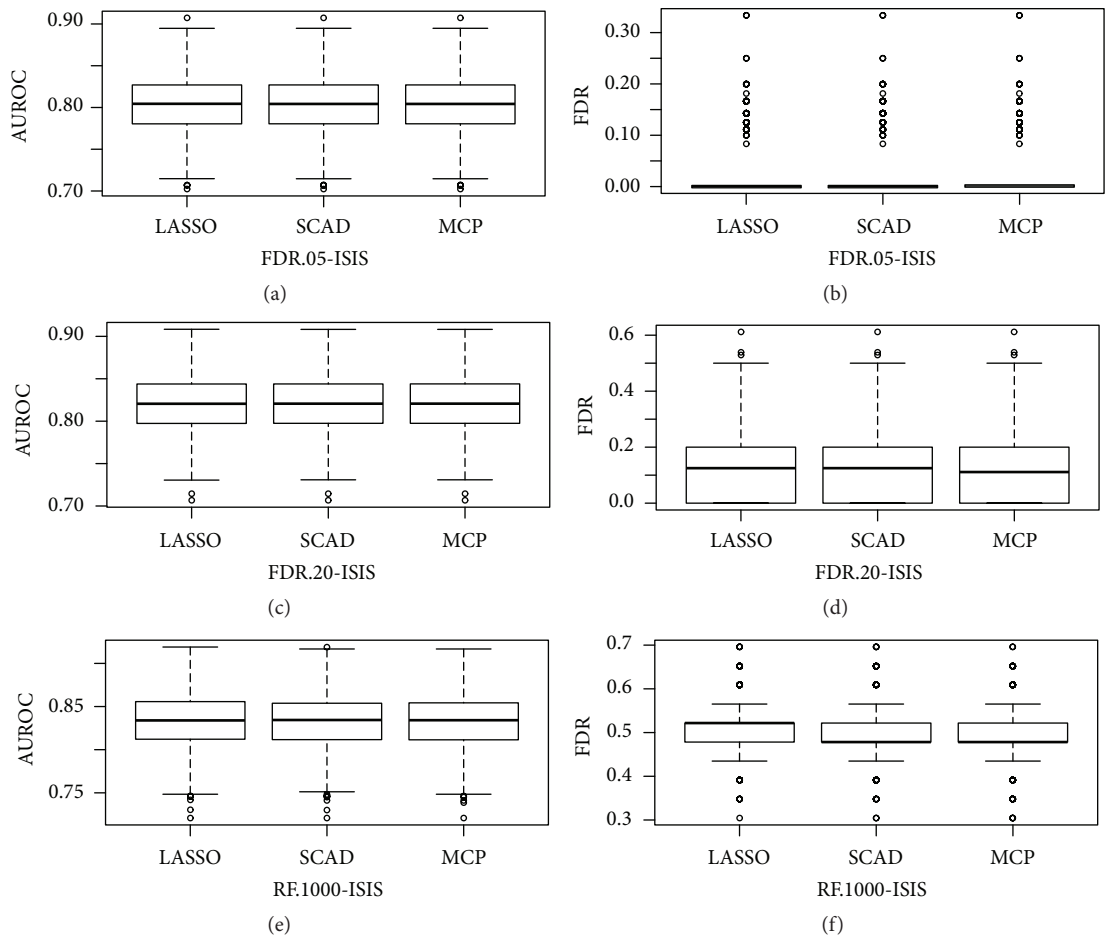


Figure 3: (a, c, e) The AUROC scores under $\rho = 0.1$. The x-axis is the name of methods and y-axis is AUROC scores. (b, d, f) The corresponding mean proportion of falsely selected variables in the model. The x-axis is the name of methods and the y-axis is the false discovery rate.

strong effects on the response. There were no differences in selecting the important variables by the variable selection methods (LASSO, SCAD, and MCP). In addition, similar patterns of the selection frequencies were observed for both $\text{FDR}_{.05} - \text{ISIS}$ and $\text{FDR}_{.20} - \text{ISIS}$ as shown in Figures 2(a) and 2(b) while Figure 2(c) showed a little variation with $\text{RF}_{1000} - \text{ISIS}$. The results were similar for independent and moderate correlation $\rho = \{0, 0.4\}$ (Figures S1 and S2 in Supplementary File).

To gain more insights into the comparisons of the methods, we present the plots of the AUROC scores and the corresponding false discovery rate under $\rho = 0.1$ in Figure 3. (a), (c), and (e) in Figure 3 represent the AUROC scores whereas (b), (d), and (f) represent the false discovery rates using three different screening methods. The variable selection methods with random filtering screening had relatively higher AUROC scores compared to the FDR methods. However, there were a number of false positive in the final models as seen in Figure 3(f). It is noteworthy that the variable selection methods using the FDR not only controlled the FDR below the target $\alpha = 0.05$ and $\alpha = 0.20$ but also had AUROC scores that were relatively high (Figures 3(d) and 3(e)). Similar patterns were observed for independent and moderate correlation $\rho = \{0, 0.4\}$ (Figures S3 and S4 in Supplementary File).

Therefore, the FDR − ISIS screening method is preferred to $\text{RF}_{1000} - \text{ISIS}$ since it allowed the variable selection methods to obtain the proper AUROC scores while controlling the false discovery rate at the nominal level of α. As a result of the simulation studies, we applied the three variable selection methods with the FDR and ISIS screening to the high dimensional data of the prostate cancer in the following section.

5. Real Data Analysis

We analyzed the prostate cancer data from the public domain (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46691: GSE46691). The dataset has 1.4 million probes and the primary outcome is presence of metastasis (yes or no) by fitting the LASSO, SCAD, and MCP methods using the FDR and ISIS screenings suggested from the simulation studies with the sequential filtering of both FDR and ISIS. In the prostate cancer application, we considered the false discovery rate (FDR) at $\alpha = 0.01$ as the screening method. Figure 4 describes the schema of the prognostic model

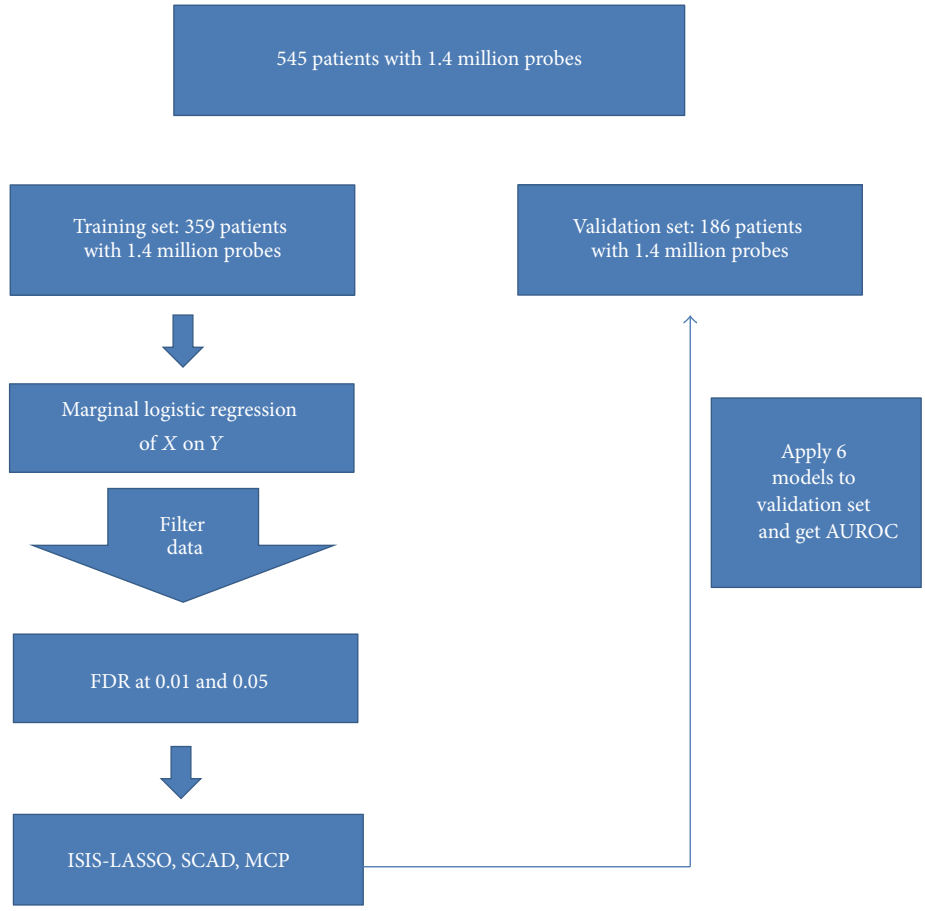
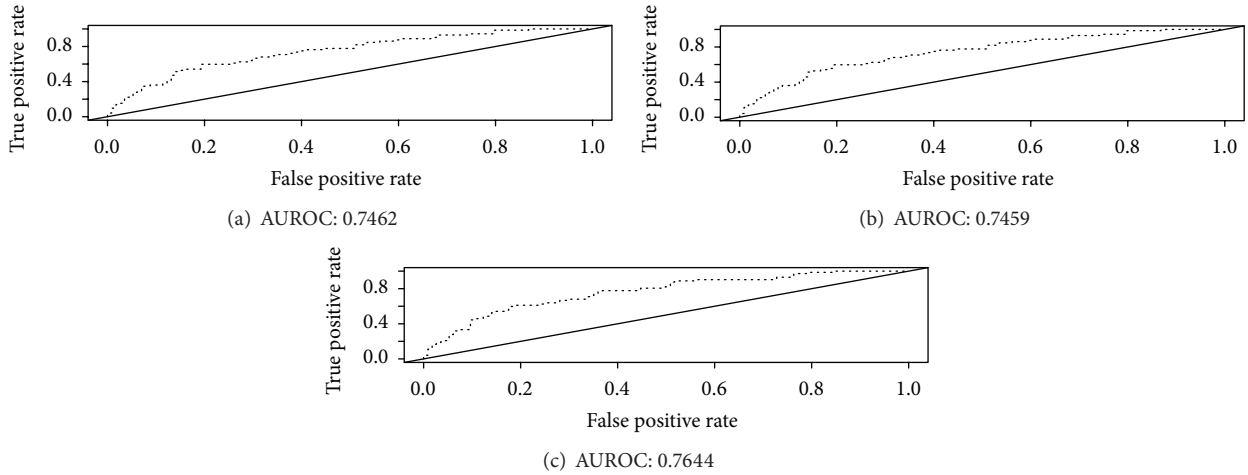


Figure 4: The schema of prognostic model building for the prostate cancer.

building for the prostate cancer. We utilized the training set that was obtained from the random split and was composed of 359 individuals (140 cases and 219 controls) with 1.4 million probes to build each of the three models. We then estimated the AUROC scores with the validation set with 186 individuals (72 cases and 114 controls). We used 10-fold cross validation for each of the variable selection models to tune the parameters after the screening. We obtained 39 variables with FDR at $\alpha = 0.01$. We repeated each of the three models 100 times to improve the AUROC with those screened variables.

Figure 5 shows the AUROC plots of the three models. Based on FDR at $\alpha = 0.01$, the LASSO, SCAD, and MCP identified 12 genes (CAMK2N1, AN07, RPL7A, MALAT1, MYBPC1, TMP0, UBE2C, DID01, RAB25, LOC728875, FTH1, and MKI67), 11 genes (CAMK2N1, AN07, RPL7A, MALAT1, MYBPC1, TMP0, UBE2C, DID01, RAB25, LOC728875, and FTH1), and 8 genes (CAMK2N1, AN07, RPL7A, MALAT1, MYBPC1, TMP0, UBE2C, and DID01) gene models out of 39 potential variables with AUROC scores of 0.746 (95% CI = 0.675–0.818), 0.746 (95% CI = 0.674–0.817), and 0.764 (95% CI = 0.695–0.834), respectively (refer to Table S4 for more details in Supplementary File). It is noteworthy to note that MCP selected the same set of genes as SCAD and LASSO and the 95% confidence intervals were overlapping. On the other hand, using the FDR at $\alpha = 0.05$, LASSO, SCAD, and MCP selected 15, 13, and 15 gene models out of 565 potential genes with corresponding AUROC scores of 0.697 (95% CI = 0.619–0.775), 0.714 (95% CI = 0.637–0.791), and 0.683 (95% CI = 0.603–0.763), respectively. It is worthwhile to note that MCP had the highest AUROC score (FDR-ISIS at $\alpha = 0.01$ and AUROC = 0.764) followed by LASSO (FDR-ISIS at $\alpha = 0.01$ and AUROC = 0.746) although the results were not consistent with the FDR at $\alpha = 0.05$. This could be due to the larger number of potential variables (565 variables) when using FDR at a higher level. Nevertheless because our interest was to use FDR at $\alpha = 0.01$, MCP and LASSO methods were used for the variable selection in our real example.

Table 2 presents the selected probes and their corresponding genes from the two models that had two highest AUROC scores among the six models. LASSO and MCP identified each of the 12 and 8 genes that were associated with developing prostate cancer metastasis. The four genes (ANO7, UBE2C, MYBPC1, and CAM2KN1) associated with developing prostate cancer metastasis were detected in both models. These four genes were a subset of the 22 biomarkers for the Decipher PCa classifier [1]. MYBPC1 (Myosin-Binding Protein C) on chromosome 12 and ANO7 (Anoctamin 7) on chromosome 2 were only



(a) AUROC: 0.7462

(b) AUROC: 0.7459

(c) AUROC: 0.7644

Figure 5: AUROC plots of using the LASSO, SCAD, and MCP with the screenings of both FDR at 0.01 and ISIS. (a), (b), and (c) are for the LASSO, the SCAD, and the MCP variable selection methods, respectively.

Table 2: Probes and corresponding genes identified by the LASSO and MCP methods with FDR at $\alpha = 0.01$ for association with the prostate cancer metastases. The Adj. p is based on the marginally adjusted p values by the BH-FDR method.

Gene	Probe ID	Ch	Start	Stop	Adj. p	LASSO	MCP
RAB25	2361272	chr1	156041891	156042035	0.003122859	*	
CAMK2N1	2400181	chr1	20810150	20810212	0.003122859	*	*
LOC728875	2432120	chr1	143692898	143692956	0.007211879	*	
AN07	2536262	chr2	242163962	242164581	0.003122859	*	*
FTH1	2590344	chr2	181551038	181551091	0.009379258	*	
RPL7A	3284321	chr10	33483529	33483624	0.007451575	*	*
MKI67	3312502	chr10	129899547	129899701	0.003122859	*	
MALAT1	3377635	chr11	65206468	65206658	0.009379258	*	*
MYBPC1	3428626	chr12	102030464	102030494	0.009379258	*	*
TMP0	3467302	chr12	98943231	98943926	0.008090076	*	*
UBE2C	3887068	chr20	44445472	44445507	0.001041641	*	*
DID01	3913561	chr20	156041891	156042035	0.003122859	*	*

*Each gene is identified by the variable selection method.

downregulated genes whereas the other 10 genes including UBE2C (Ubiquitin-Conjugating Enzyme E2C) on chromosome 20 and CAMK2N1 (Calcium/Calmodulin-Dependent Protein Kinase II Inhibitor 1) on chromosome 1 were the top upregulated genes as presented in Figure S5.

6. Discussion

This paper explored the feasibility of using the false discovery rate (FDR) followed by ISIS as screening methods in conjunction with three popular variable selection methods in ultra-high dimensional data for the purpose of controlling FDR and improving AUROC scores.

Our simulation studies demonstrated that the variable selection methods with FDR − ISIS not only controlled the false discovery rate below the target α, but also produced high AUROC scores. Furthermore, the results showed that the false discovery rate was controlled conservatively even with the increased correlation structures. As demonstrated in the simulation studies, if the truly prominent variables have not passed through the screening, they would lose the opportunity to be selected to the final model. Thus, the prediction accuracy may be relatively reduced. Currently, most multiple testing correction methods underscore the priority of identifying prominent variables. Therefore, effective filtering techniques are ultimately needed for the situation when there are weak effects among the important variables.

Although RF_{1000} − ISIS produced the highest AUROC through the simulation studies, it also had the highest false discovery rate. It is reasonable to expect that if one variable is not selected during the screening step, then the other variables that were correlated with the unselected variable have a tendency not to be chosen in the final model. As expected, the true positive rates of RF_{1000} − ISIS were relatively higher than those of using FDR − ISIS. This is because random filtering had more opportunity to select the true important variables. Due to the relatively high number of true important variables selected, the number of unimportant variables highly correlated with the true important variables was also high. This may explain why the average AUROC

scores were highest in our simulation studies for RF_{1000}−ISIS.

There are some caveats to the sequential use of FDR with ISIS. First, the total number of true important variables was restricted to 25 variables in the simulation studies. This may explain why the three variable selection methods, LASSO, SCAD, and MCP, performed similarly in the simulations. Second, the computational time for 500 simulations with ISIS alone was 52,500 minutes (around 36.45 days). Although ISIS is computationally intensive, FDR with ISIS performed very well and it took 9,625 minutes to perform the 500 simulations (7 days).

Turning back to our motivation example of prostate cancer, LASSO and MCP under the $FDR_{0.01}$ − ISIS screening methods produced the best AUROC scores. We also present the results for other LASSO and MCP methods selected 12 and 8 genes out of 39 screened probes using the FDR at the target $\alpha = 0.01$ and had the AUROC scores of 0.7462 and 0.7644, respectively. The AUROC score of the MCP was 0.144 points higher than what was reported by Erho et al. ([1]; AUC = 0.75). Although the authors did not report a 95% confidence interval for the AUROC scores, it is most likely that the 95% confidence intervals for the AUROC scores of Erho et al. and the MCP were overlapping.

In summary, based on our extensive simulations, FDR with ISIS seems to be superior to random filtering in terms of error control and is less computationally intensive compared with ISIS only. We also showed that the classifier based on 8 genes detected by the MCP had similar performance to the prognosis for early clinical metastasis prostate cancer model. To our knowledge, this is the first paper that systematically compared the performance of high dimensional methods with screening methods. Based on the extensive simulation studies, effective screening procedures with penalized logistic regression methods would not only lead to controlling the FDR but also produce high area under receiver operating characteristic curve.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This research was supported in part by National Institutes of Health Grants R01CA155296 and U01CA157703 and United States Army Medical Research W81XWH-15-1-0467. The authors thank Lira Pi for her help and would like to acknowledge the contribution of the Duke University Compact for Open-Access Publishing Equity fund for its support of this paper.

References

[1] N. Erho, A. Crisan, I. A. Vergara et al., "Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy," *PLoS ONE*, vol. 8, no. 6, article e66855, 2013.

[2] R. J. Karnes, E. J. Bergstralh, E. Davicioni et al., "Validation of a genomic classifier that predicts metastasis following radical prostatectomy in an at risk patient population," *The Journal of Urology*, vol. 190, no. 6, pp. 2047–2053, 2013.

[3] A. E. Ross, F. Y. Feng, M. Ghadessi et al., "A genomic classifier predicting metastatic disease progression in men with biochemical recurrence after prostatectomy," *Prostate Cancer and Prostatic Disease*, vol. 17, no. 1, pp. 64–69, 2014.

[4] A. E. Ross, K. Yousefi, E. Davicioni et al., "Utility of risk models in decision making after radical prostatectomy: lessons from a natural history cohort of intermediate and high-risk men," *European Urology*, vol. 69, no. 3, pp. 496–504, 2015.

[5] E. A. Klein, K. Yousefi, Z. Haddad et al., "A genomic classifier improves prediction of metastatic disease within 5 years after surgery in node-negative high-risk prostate cancer patients managed by radical prostatectomy without adjuvant therapy," *European Urology*, vol. 67, no. 4, pp. 778–786, 2015.

[6] J. A. Sparano, R. J. Gray, D. F. Makower et al., "Prospective validation of a 21-gene expression assay in breast cancer," *The New England Journal of Medicine*, vol. 373, no. 21, pp. 2005–2014, 2015.

[7] S. Halabi, A. J. Armstrong, O. Sartor et al., "Prostate-specific antigen changes as surrogate for overall survival in men with metastatic castration-resistant prostate cancer treated with second-line chemotherapy," *Journal of Clinical Oncology*, vol. 31, no. 31, pp. 3944–3950, 2013.

[8] S. Halabi, C.-Y. Lin, W. Kevin Kelly et al., "Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer," *Journal of Clinical Oncology*, vol. 32, no. 7, pp. 671–677, 2014.

[9] S. Lee, J. Rahnenführer, M. Lang et al., "Robust selection of cancer survival signatures from high-throughput genomic data using two-fold subsampling," *PLoS ONE*, vol. 9, no. 10, Article ID e108818, 2014.

[10] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 70, no. 5, pp. 849–911, 2008.

[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B: Methodological*, vol. 58, no. 1, pp. 267–288, 1996.

[12] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[13] J. Fan and H. Peng, "Nonconcave penalized likelihood with a diverging number of parameters," *Annals of Statistics*, vol. 32, no. 3, pp. 928–961, 2004.

[14] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

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

[16] J. Fan, R. Samworth, and Y. Wu, "Ultrahigh dimensional variable selection: beyond the linear model," *Journal of Machine Learning Research*, vol. 10, pp. 1829–1853, 2009.

[17] M. Pourahmadi, *High-Dimensional Covariance Estimation: with High-Dimensional Data*, John Wiley & Sons, 2013.

[18] N. Städler, P. Bühlmann, and S. van de Geer, "ℓ_1-penalization for mixture regression models," *TEST*, vol. 19, no. 2, pp. 209–256, 2010.

[19] H. Li, "Censored data regression in high-dimension and low sample size settings for genomic applications," in *Statistical Advances in Biomedical Sciences: State of Art and Future Directions*, A. Biswas, S. Datta, J. Fine, and M. Segal, Eds., John Wiley & Sons, Hoboken, NJ, USA, 1st edition, 2008.

[20] T. Cai, J. Huang, and L. Tian, "Regularized estimation for the accelerated failure time model," *Biometrics. Journal of the International Biometric Society*, vol. 65, no. 2, pp. 394–404, 2009.

[21] B. A. Johnson, "Rank-based estimation in the ℓ_1-regularized partly linear model for censored outcomes with application to integrated analyses of clinical predictors and gene expression data," *Biostatistics*, vol. 10, no. 4, pp. 659–666, 2009.

[22] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[23] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 7310694, 8 pages
http://dx.doi.org/10.1155/2016/7310694

Research Article

Prognostic Value of Osteopontin Splice Variant-c Expression in Breast Cancers: A Meta-Analysis

Chengcheng Hao,[1,2] Zhiyan Wang,[1,2] Yanan Gu,[1,2] Wen G. Jiang,[1,2,3] and Shan Cheng[1,2]

[1]*Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Capital Medical University, Beijing 100069, China*
[2]*Beijing Key Laboratory of Cancer & Metastasis Research, Capital Medical University, Beijing 100069, China*
[3]*Cardiff China Medical Research Collaborative, Cardiff University School of Medicine, Heath Park, Cardiff CF14 4XN, UK*

Correspondence should be addressed to Shan Cheng; chengs@ccmu.edu.cn

Received 20 April 2016; Accepted 8 June 2016

Academic Editor: Yichuan Zhao

Objectives. Osteopontin (OPN) is overexpressed in breast cancers, while its clinical and prognostic significance remained unclear. This study aimed to assess the prognostic value of OPN, especially its splice variants, in breast cancers. *Methods.* Data were extracted from eligible studies concerning the OPN and OPN-c expression in breast cancer patients and were used to calculate the association between OPN/OPN-c and survival. Two reviewer teams independently screened the literatures according to the inclusion and exclusion criteria based on quality evaluation. Following the processes of data extraction, assessment, and transformation, meta-analysis was carried out *via* RevMan 5.3 software. *Results.* A total of ten studies involving 1,567 patients were included. The results demonstrated that high level OPN indicated a poor outcome in the OS (HR = 2.22, 95% CI: 1.23–4.00, and $P = 0.008$; random-effects model) with heterogeneity ($I^2 = 62\%$) of breast cancer patients. High level OPN-c appeared to be more significantly associated with poor survival (HR = 2.14, 95% CI: 1.51–3.04, and $P < 0.0001$; fixed-effects model) with undetected heterogeneity ($I^2 = 0\%$). *Conclusions.* Our analyses indicated that both OPN and OPN-c could be considered as prognostic markers for breast cancers. The high level of OPN-c was suggested to be more reliably associated with poor survival in breast cancer patients.

1. Introduction

Breast cancer is one of the most commonly diagnosed cancers in women worldwide and it is the leading cause of cancer death. According to the statistics provided by the American Cancer Society, approximately 231,840 new cases of invasive breast cancer and 40,290 breast cancer deaths occurred among US women in 2015 [1]. It was estimated by GLOBOCAN 2008 that breast cancer was the most frequent cancer in Chinese women, with an age standardized rate (ASR) of 21.6 cases per 100,000 individuals [2]. Although multiple treatment methods including surgery, chemotherapy, radiotherapy, and targeted therapy have been applied over the past few decades, the prognosis of breast cancer remains unsatisfactory. Due to the limitations in prognostic values of the conventional predicting factors, such as TNM stage, age, sex, and histological type, it is necessary to explore novel biomarkers to better predict the outcome and assist in the clinical management of breast cancer patients.

Osteopontin (OPN), a phosphorylated glycoprotein secreted by various tissues and cells, has been implicated with important roles in physiological and pathological processes [13–15]. In recent years, accumulating evidences have shown that aberrant OPN expression was closely associated with tumorigenesis and metastasis in several types of tumors, including breast cancer [16, 17]. Overexpression of OPN was found in multiple malignant breast cancer cell lines, and the transfection of OPN into benign breast epithelial cells induced invasive behavior [18]. In addition, several studies investigated the association between OPN expression and the clinical outcome and prognosis of breast cancer patients, but the results failed to demonstrate

a consented conclusion regarding the ability of OPN to predict cancer progression [7, 19–21].

The biological functions of metastasis-associated gene products are often mediated by different splicing isoforms. Alternative splicing OPN secreted by various cells has diverse structural characteristics. Tumor-derived OPN forms are smaller than OPN secreted by nontransformed cells. Full length OPN (OPN-a) consists of 7 exons while OPN-b and OPN-c lack exon 4 and exon 5, respectively [22, 23]. Alternative splicing occurred at the upstream of the central integrin binding domain and the C-terminal CD44 binding domain [24–26]. Clinical data suggested that the shortest splice variant OPN-c could be a selective diagnostic and prognostic candidate for human breast cancer [27–29]. However, the clinical function of OPN-c in breast cancer remains poorly defined.

Therefore, to clarify the prognostic significance of OPN, as well as its splicing variants, in breast cancer, it is necessary to investigate the association of OPNs expression with patient survivals. In the present study, we pooled data from the available reports and analyzed the association between the expression of OPN and the prognostic measures of breast cancer patients.

2. Materials and Methods

2.1. Study Selection. A comprehensive literature search (dated to December 2015) was conducted through the PubMed database. The query strategy was the combinations of the following terms: "Osteopontin", "OPN", "OPN-c", "Osteopontin-c", "breast cancer", "prognosis", "prognostic" and "survival" without restrictions on regions, languages, and publication types. All eligible studies were retrieved.

2.2. Inclusion and Exclusion Criteria. The criteria for reference inclusion were (1) data associated with OPN levels (negative/positive or low/high expression), as measured by reverse transcription-polymerase chain reaction (RT-PCR), immunohistochemistry (IHC), or enzyme linked immunosorbent assay (ELISA); (2) confirmed diagnosis of breast cancer with pathological or histological evidence; (3) hazard ratio (HR) and their 95% confidence intervals (CIs) that could be directly extracted or disease-free survival or overall survival with sufficient data that was provided; (4) some data supplied by authors upon our requests for the calculation of the HR or 95% CIs values.

The studies were excluded for (1) meeting abstracts, letters, case reports, and reviews; (2) being animal studies; (3) missing necessary data; and (4) using redundant or overlapped database records.

2.3. Data Extraction and Quality Assessment. To reduce bias and improve reliability, two investigator groups independently reviewed potentially relevant studies. The data of the number of patients, follow-up, detection method, and cut-off value, together with name of the first author, name of the journal, year of publication, and ethnicity, were extracted and then analyzed against OPN/OPN-c expression-related survival. Conflicting interpretation problems were resolved through consensus with the third investigator. The outcome assessment focused on survival curves in patients with different OPN/OPN-c expression. The GetData Graph Digitizer 2.24 software (http://getdata-graph-digitizer.com/) and HR digitizer Engauge 4.1 software (http://engauge-digitizer.software.informer.com/) were used to digitize and extract the data from the Kaplan-Meier curves, in cases when only survival curves were provided. The quality of each included study was based on two independent assessments by Newcastle-Ottawa Scale (NOS) scoring. We allocated a score of 0–9 to each included study, and those with a NOS score $\geq$ 6 were assigned as high-quality studies.

2.4. Statistical Analysis. The association between OPN/OPN-c expression and the survival outcome of breast cancer was estimated according to the HR and 95% CI directly or indirectly extracted from each eligible study. The heterogeneity among studies was detected using Cochran's Q test and Higgins I-squared statistic. Severe heterogeneity was taken into account based on measures of $I^2 > 50\%$. The fixed-effects model was used when $I^2 < 50\%$ or $P \geq 0.10$ in the Q test, while random-effects model was conducted otherwise. The potential publication bias was estimated by Egger's and Begg's tests with significance of $P < 0.05$. All the statistical analyses were performed using Review Manager version 5.3.

3. Results

3.1. Characteristics of Included Studies. Following the search strategy given in Section 2, a total of 214 potentially relevant citations were retrieved (Figure 1). After reviewing the titles and abstracts, 188 studies failed to meet our selection criteria and were excluded. The remaining 26 were subjected to full-text screening, of which 16 publications were excluded because of the lack in survival or follow-up data associated with OPN/OPN-c. Eventually, a total of 10 studies including 1,567 patients were qualified for further analysis, 7 of which investigated the impact of OPN expression on survival and 4 were about OPN-c. Three of the 10 studies used ELISA, 5 used IHC, and 2 used qRT-PCR to detect serum- or tissue-derived OPN/OPN-c expression. Detailed values from the included publications were summarized in Table 1.

3.2. Association of OPN on OS of Breast Cancer Patients. Seven included studies provided data for calculating the association between OPN expression and survival. We pooled all the available data into our meta-analysis and found that poor OS was significantly associated with a positive status for OPN (HR = 2.22, 95% CI: 1.23–4.00, and $P = 0.008$) (Figure 2), with heterogeneity in the data ($I^2 = 62\%$, $P = 0.01$).

To determine whether the heterogeneity in OS was caused by data bias, we performed sensitivity analysis to assess the stability of the results. The results showed that, after exclusion of the study by Liao et al., the heterogeneity for OS did not significantly decrease ($I^2 = 56\%$, $P = 0.0007$) with a combined HR of 2.64 (95% CI: 1.51–4.63) (Figure 3). The fact

Table 1: General characteristics of included studies.

	First authors	Journal	Year	Country	Number of patients	Follow-up (months)	Index	Methods	Cut-off point (high/low)	OPN+/OPN−	HR estimation	Quality assessment (0–9)
1	Anborgh [3]	Am J Transl Res	2015	UK	53	76	OPN	ELISA	Median	31/22	HR + 95% CI	7
2	Martinetti [4]	Endocrine-Related Cancer	2004	Italy	34	60	OPN	ELISA	Median	12/22	Survival curves	8
3	Singhal [5]	Clin Cancer Res	1997	Canada	70	19	OPN	ELISA	>203	24/46	Survival curves	7
4	Tuck [6]	Int. J. Cancer	1998	Canada	154	195.6	OPN	IHC	Score > 4	11/143	Survival curves	5
5	Rudland [7]	Cancer Res	2002	UK	333	228	OPN	IHC	>5%	221/112	RR + 95% CI	8
6	Liao [8]	Modern Oncology	2010	China	70	110	OPN	IHC	Score ≥ 2	51/19	Survival curves	5
7	Ortiz-Martínez [9]	Human Pathology	2014	Spain	309	302	OPN/OPN-c	qRT-PCR	FC > 32	76/231 70/204	DFS + OS	7
8	Pang [10]	Cancer Epidemiology	2013	China	170	92	OPN-c	IHC	Score ≥ 2	119/51	HR + 95% CI	6
9	Zduniak [11]	British Journal of Cancer	2015	USA	291	60	OPN-c	IHC	Score ≥ 2	181/110	Survival curves	7
10	Patani [12]	Anticancer Research	2008	UK	83	150	OPN-c	qRT-PCR	NPI	21/62	OS	6

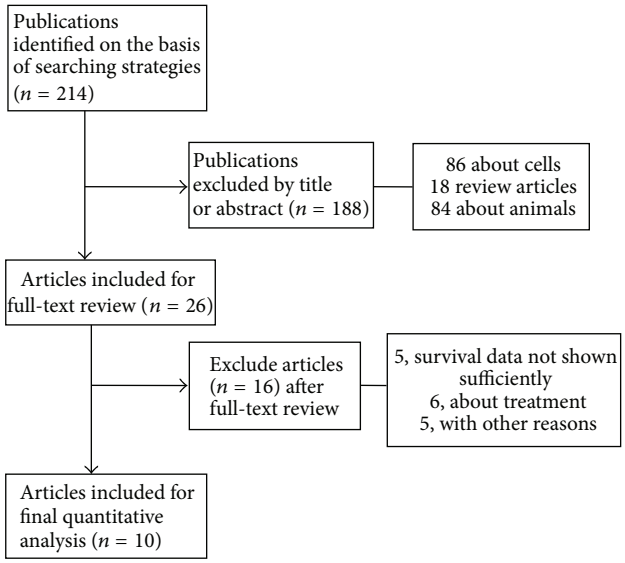


Figure 1: Flowchart of the selection process of studies for inclusion in this meta-analysis.

that the HR and I^2 values were not significantly altered with the exclusion of distracted dataset indicated that the method was appropriate and the results were credible.

The subgroup meta-analyses were performed to exclude the potential influence from sample heterogeneity. Four studies (three using IHC, one by qPCR) were included to evaluate the connection between OPN expression in breast tumor tissues and survival. Positive OPN expression was associated with poor OS (HR = 2.10, 95% CI: 0.81–5.43) with heterogeneity (I^2 = 80%, P = 0.002), but no significant difference (P = 0.13) was found (Figure 4). Three studies about the association between plasma OPN level and survival were also investigated. Results showed that reduced survival was significantly associated with a positive status for plasma OPN (HR = 2.46, 95% CI: 1.31–4.60, and P = 0.005) (Figure 5), with diminished heterogeneity in the data (I^2 = 0%, P = 0.76).

3.3. Association of OPN-c on Survival of Breast Cancer Patients. Four studies (two OS, one on DFS, and one with five-year survival) including 853 patients were investigated. The statistical results once again indicated that high OPN-c expression in tissues was significantly associated with poor survival (HR = 2.14, 95% CI: 1.51–3.04, and $P < 0.0001$) with diminished heterogeneity (I^2 = 0%, P = 0.72) (Figure 6).

3.4. Publication Bias. In order to assess the publication bias of the included studies, Begg's funnel plot was deployed. As shown in Figures 7 and 8, nonasymmetric funnel plots for the synthesis of the HRs for survival were obtained. Hence, no publication bias was detected among all the comparisons.

4. Discussion

As a popular and effective approach for systematic reviews, meta-analysis has been successfully applied to evaluate prognostic indicators in patients with varied diseases. It has been indicated in recent research that the expression of biomarkers was promisingly associated with tumor progression, including breast cancers. Besides, its elevated levels often suggested high tumor grade, metastasis, and poor prognosis. Several markers have been used in breast cancer diagnosis, including estrogen receptor (ER) and progesterone receptor (PR). However, these markers still lack the adequate sensitivity and specificity for detecting prognosis of breast cancer patients. It was reported that OPN might be a more potential clinical marker candidate for predicting the survival of breast cancer patients than ER and PR [6, 7, 9]. A meta-analysis was further performed in our group to clarify the above ambiguous conclusion and to investigate the association of OPN with prognostic factors in breast cancer.

From a number of reports, OPN was overexpressed in primary tumors and maintained a high secreted level in the blood of breast cancer patients, which correlated with a poor prognosis [3–6, 8, 9, 30]. The results of our present meta-analysis study basically supported the association between increased OPNs with poor OS in patients with breast cancer (P = 0.008). However, there could be a concern with the heterogeneity (I^2 = 62%). Through the sensitivity analysis, the resulting heterogeneity value (I^2 = 56%) without substantial changes indicated an intrinsic property and would not compromise the robustness and prognostic values of OPN in breast cancer. The heterogeneity problem was also reported in a recent similar meta-analysis conducted by Xu et al., which explored the prognostic value of OPN by OS and DFS in breast cancer patients. Notably, the further subgroup analysis showed that the increased OPN was not correlated with clinicopathological parameters such as tumor grade, tumor stage, PR status, ER status, and p53 status, suggesting that the source of heterogeneity could be rather complicated [31]. The prognostic value of OPN has been reported in systematic reviews and meta-analyses on lung [32], colorectal [33], and pancreatic [34] cancers with a vast range in heterogeneities and the positive association of increased OPN with the survival of cancer patients appeared to be consistent, despite the fact that certain additional factors were indeed to be considered for the statistical analysis of heterogeneity and its origins.

Alternative RNA splicing of human OPN results in three transcriptive variants: OPN-a (full length), OPN-b (lacking exon 5), and OPN-c (lacking exon 4). Recent studies have shown that cell-type specific expression of OPN splice variants exerted different functions in malignant tumors. Compared to full length OPN, OPN-c was specifically expressed in breast cancer cells. Hence, it was likely to be a better prognostic marker for human breast cancer [9–12]. Unfortunately, studies have not conducted separate measures on OPN splice variants. The results were actually the total OPN (including OPN-a, OPN-b, and OPN-c) expression in plasma or tumor of breast cancer patients. The only study that can be found reported that the survival for breast cancer patients with high levels of OPN-b mRNA expression was found to differ significantly from that of their low level counterparts [12]. The available data of OPN-c for cancer

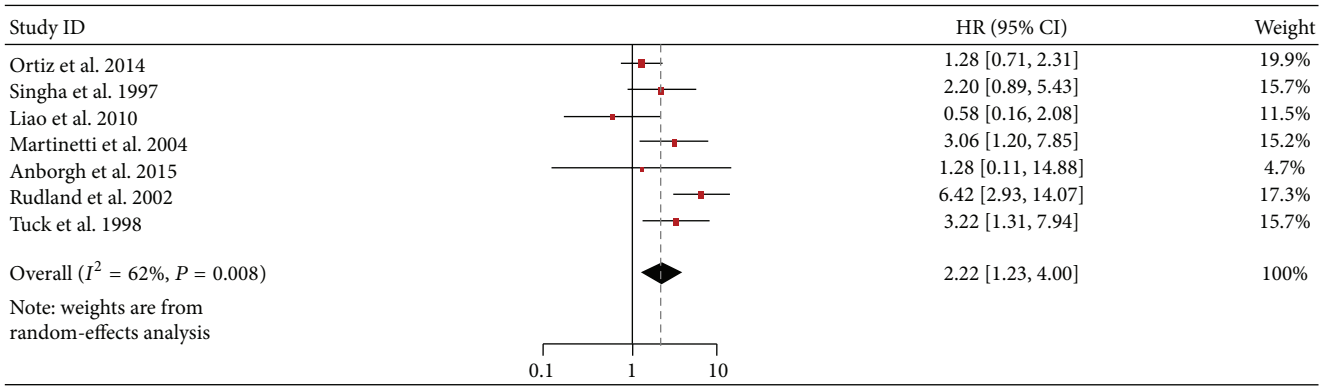


Figure 2: Forest plot demonstrating the association of OPN expression with overall survival in breast cancer. HR, hazard ratio; CI, confidence interval.

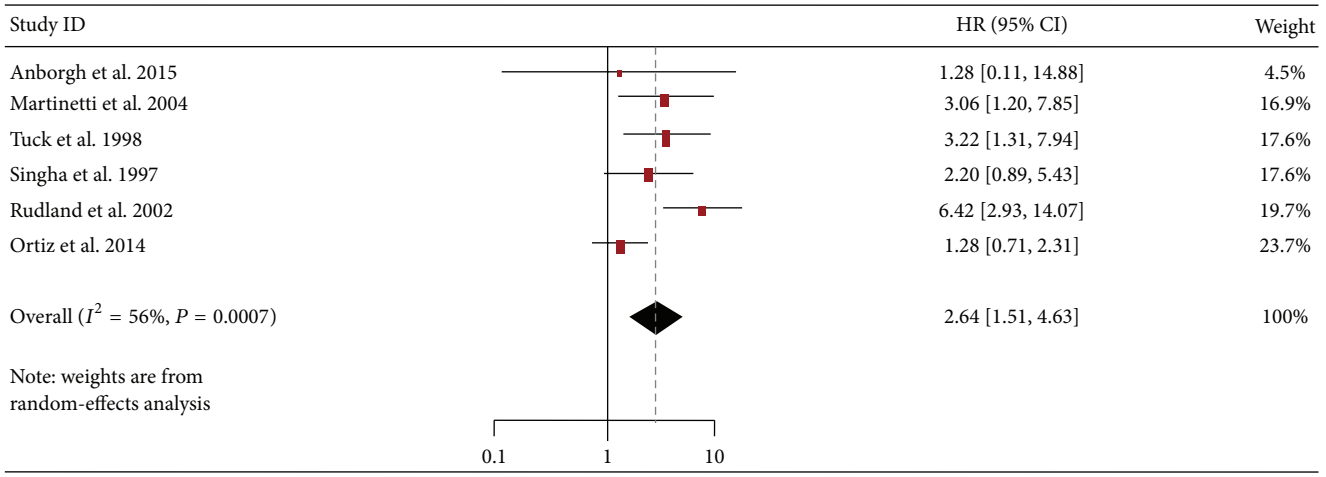


Figure 3: Forest plot on the association between OPN expression and overall survival in breast cancer by sensitivity analysis. HR, hazard ratio; CI, confidence interval.

survival only allowed primitive comprehensive analysis. In the present study, we carried out a meta-analysis and noted that the high level of OPN-c was associated with poor survival with better statistical significance ($P < 0.0001$) and, most of all, with a drastically reduced heterogeneity ($I^2 = 0\%$). Thus, OPN-c could be by far a most significant predictor of the poor prognosis of breast cancers, which could be further investigated in studies involving more breast cancer patients or perhaps in other cancer types as well.

Interestingly, in the subgroup meta-analysis, positive OPN expression in breast tumor tissues was associated with poor OS, but no significant difference ($P = 0.13$) was found (Figure 4), while reduced survival was significantly associated with a positive status for OPN in plasma ($P = 0.005$) (Figure 5). These results seemed to be contrary; however, they were still considered to confirm the results that the total tumor OPN expression was very heterogeneous. As a more significant marker candidate, OPN-c is likely to be of particular utility as a prognostic marker and should be included in further validation studies. The significant association between plasma OPN and survival suggested that the plasma detection could be a powerful supplement to the analysis of tumor tissues. However, there were no reports about the association between plasma OPN-c and the prognosis of breast cancers till now. Therefore, it is very important and necessary to perform the corresponding studies.

We have recognized that there are several limitations existing in this meta-analysis. First, the literatures covering this topic are limited in numbers. The database documenting their related data is small in scale, especially those about OPN-c. Second, the investigation regarding the correlations between OPN/OPN-c expression and clinical features, such as tumor stage, lymph node metastasis, and tumor size, is not performed, because most of the primary studies have not provided sufficient information for this analysis. Third, different methods are adopted and inconsistent cut-off values are used for determining the expression of OPNs in different reports, and therefore it is a challenge for the normalization of the results. Fourth, in dealing with the reports which do not contain the HRs in the full text, we extrapolated the values from the survival curves. Although this approach is

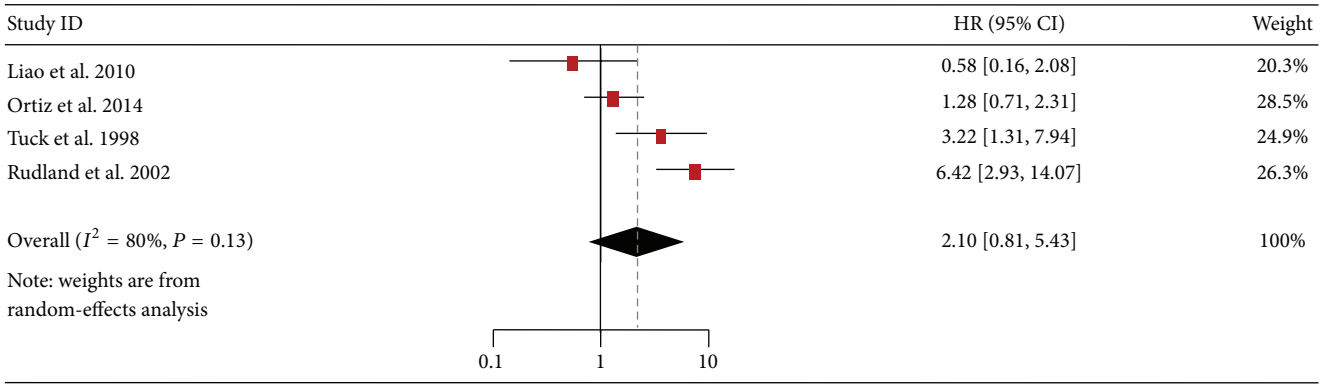


Figure 4: Forest plot demonstrating the association of OPN expression with overall survival in tissues of breast cancer patients. HR, hazard ratio; CI, confidence interval.

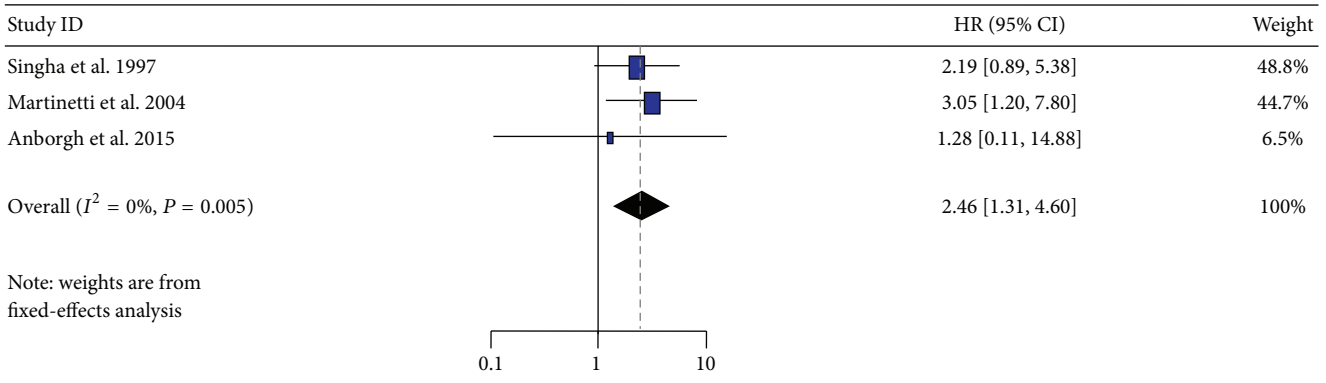


Figure 5: Forest plot demonstrating the association of OPN expression with overall survival in blood of breast cancer patients. HR, hazard ratio; CI, confidence interval.

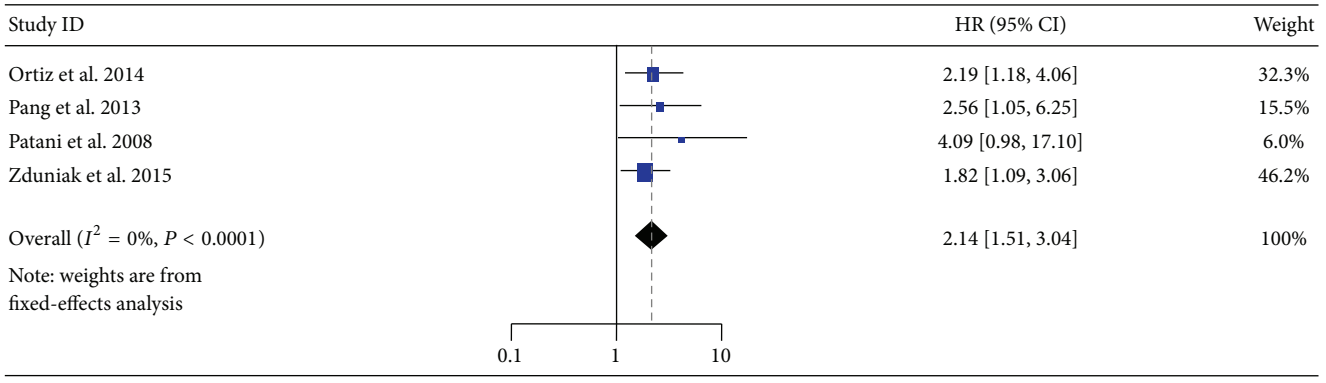


Figure 6: Forest plot demonstrating the association of OPN-c expression with overall survival in breast cancer. HR, hazard ratio; CI, confidence interval.

a common practice, we cannot exclude the possible introduced errors. Finally, to avoid the potential publication bias, factors, such as sex, age, and geographical origins, are not considered in this analysis.

In summary, both OPN and OPN-c could be considered as markers for breast cancer prognosis. The finding of better performance of OPN-c in statistics suggested that the stratification of different isoforms or variants of gene expression could be a valid approach for searching better biomarkers in cancer patients. Based on the findings from this study, we recommend a prospective study to confirm the prognostic value of OPN-c in breast cancer patients.

Competing Interests

The authors declare that they have no competing interests.

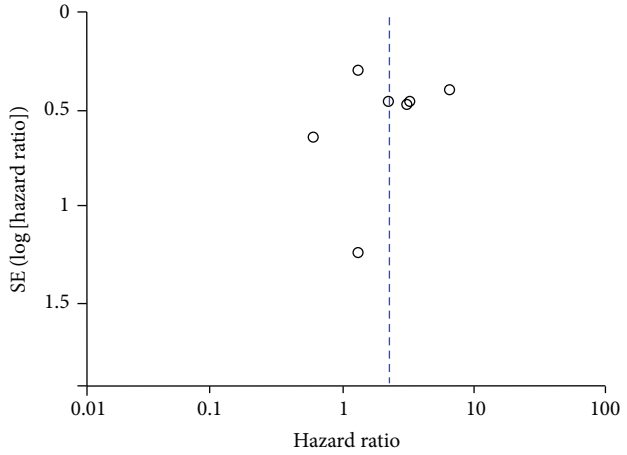


Figure 7: Begg's funnel plot estimating the publication bias of the included studies about OPN. HR, hazard ratio.

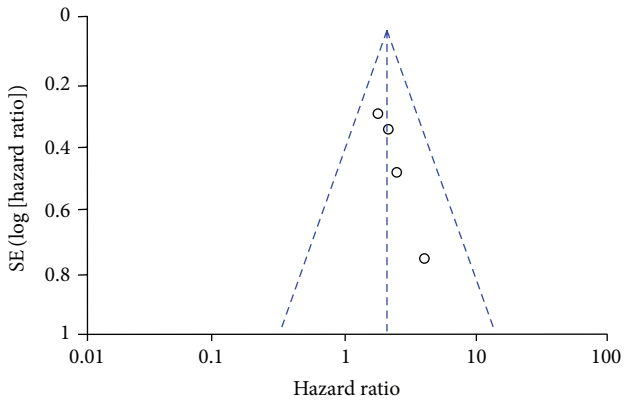


Figure 8: Begg's funnel plot estimating the publication bias of the included studies about OPN-c. HR, hazard ratio.

References

[1] C. E. DeSantis, S. A. Fedewa, A. G. Sauer, J. L. Kramer, R. A. Smith, and A. Jemal, "Breast cancer statistics, 2015: convergence of incidence rates between black and white women," *CA: A Cancer Journal for Clinicians*, vol. 66, no. 1, pp. 31–42, 2016.

[2] L. Fan, K. Strasser-Weippl, J.-J. Li et al., "Breast cancer in China," *The Lancet Oncology*, vol. 15, no. 7, pp. e279–e289, 2014.

[3] P. H. Anborgh, L. B. R. Caria, A. F. Chambers, A. B. Tuck, L. W. Stitt, and M. Brackstone, "Role of plasma osteopontin as a biomarker in locally advanced breast cancer," *American Journal of Translational Research*, vol. 7, no. 4, pp. 723–732, 2015.

[4] A. Martinetti, E. Bajetta, L. Ferrari et al., "Osteoprotegerin and osteopontin serum values in postmenopausal advanced breast cancer patients treated with anastrozole," *Endocrine-Related Cancer*, vol. 11, no. 4, pp. 771–779, 2004.

[5] H. Singhal, D. S. Bautista, K. S. Tonkin et al., "Elevated plasma osteopontin in metastatic breast cancer associated with increased tumor burden and decreased survival," *Clinical Cancer Research*, vol. 3, no. 4, pp. 605–611, 1997.

[6] A. B. Tuck, F. P. O'Malley, H. Singhal et al., "Osteopontin expression in a group of lymph node negative breast cancer patients," *International Journal of Cancer*, vol. 79, no. 5, pp. 502–508, 1998.

[7] P. S. Rudland, A. Platt-Higgins, M. El-Tanani et al., "Prognostic significance of the metastasis-associated protein osteopontin in human breast cancer," *Cancer Research*, vol. 62, no. 12, pp. 3417–3427, 2002.

[8] N. Liao, L. L. Tang, Z. Q. Wang, Z. G. Liao, and C. J. Zhang, "Expression of OPN in recurrent and metastatic breast cancer," *Modern Oncology*, vol. 4, pp. 692–695, 2010.

[9] F. Ortiz-Martínez, A. Perez-Balaguer, D. Ciprián et al., "Association of increased osteopontin and splice variant-c mRNA expression with HER2 and triple-negative/basal-like breast carcinomas subtypes and recurrence," *Human Pathology*, vol. 45, no. 3, pp. 504–512, 2014.

[10] H. Pang, H. Lu, H. Song et al., "Prognostic values of osteopontin-c, E-cadherin and β-catenin in breast cancer," *Cancer Epidemiology*, vol. 37, no. 6, pp. 985–992, 2013.

[11] K. Zduniak, P. Ziolkowski, C. Ahlin et al., "Nuclear osteopontin-c is a prognostic breast cancer marker," *British Journal of Cancer*, vol. 112, no. 4, pp. 729–738, 2015.

[12] N. Patani, F. Jouhra, W. Jiang, and K. Mokbel, "Osteopontin expression profiles predict pathological and clinical outcome in breast cancer," *Anticancer Research*, vol. 28, no. 6, pp. 4105–4110, 2008.

[13] D. T. Denhardt and M. Noda, "Osteopontin expression and function: role in bone remodeling," *Journal of Cellular Biochemistry*, vol. 30, pp. 92–102, 1998.

[14] C. M. Giachelli and S. Steitz, "Osteopontin: a versatile regulator of inflammation and biomineralization," *Matrix Biology*, vol. 19, no. 7, pp. 615–622, 2000.

[15] K. X. Wang and D. T. Denhardt, "Osteopontin: role in immune regulation and stress responses," *Cytokine & Growth Factor Reviews*, vol. 19, no. 5-6, pp. 333–345, 2008.

[16] S. R. Rittling and A. F. Chambers, "Role of osteopontin in tumour progression," *British Journal of Cancer*, vol. 90, no. 10, pp. 1877–1881, 2004.

[17] D. Coppola, M. Szabo, D. Boulware et al., "Correlation of osteopontin protein expression and pathological stage across a wide variety of tumor histologies," *Clinical Cancer Research*, vol. 10, no. 1, pp. 184–190, 2004.

[18] G. Chakraborty, S. Jain, and G. C. Kundu, "Osteopontin promotes vascular endothelial growth factor-dependent breast tumor growth and angiogenesis via autocrine and paracrine mechanisms," *Cancer Research*, vol. 68, no. 1, pp. 152–161, 2008.

[19] A. B. Tuck and A. F. Chambers, "The role of osteopontin in breast cancer: clinical and experimental studies," *Journal of Mammary Gland Biology and Neoplasia*, vol. 6, no. 4, pp. 419–429, 2001.

[20] V. H. C. Bramwell, G. S. Doig, A. B. Tuck et al., "Serial plasma osteopontin levels have prognostic value in metastatic breast cancer," *Clinical Cancer Research*, vol. 12, no. 11, pp. 3337–3343, 2006.

[21] L. R. Rodrigues, J. A. Teixeira, F. L. Schmitt, M. Paulsson, and H. Lindmark-Mänsson, "The role of osteopontin in tumor progression and metastasis in breast cancer," *Cancer Epidemiology Biomarkers & Prevention*, vol. 16, no. 6, pp. 1087–1097, 2007.

[22] M. F. Young, J. M. Kerr, J. D. Termine et al., "cDNA cloning, mRNA distribution and heterogeneity, chromosomal location, and RFLP analysis of human osteopontin (OPN)," *Genomics*, vol. 7, no. 4, pp. 491–502, 1990.

[23] N. Hijiya, M. Setoguchi, K. Matsuura, Y. Higuchi, S. Akizuki, and S. Yamamoto, "Cloning and characterization of the human osteopontin gene and its promoter," *Biochemical Journal*, vol. 303, no. 1, pp. 255–262, 1994.

[24] H. Rangaswami, A. Bulbule, and G. C. Kundu, "Osteopontin: role in cell signaling and cancer progression," *Trends in Cell Biology*, vol. 16, no. 2, pp. 79–87, 2006.

[25] D. S. Bautista, J.-W. Xuan, C. Hota, A. F. Chambers, and J. F. Harris, "Inhibition of Arg-Gly-Asp (RGD)-mediated cell adhesion to osteopontin by a monoclonal antibody against osteopontin," *The Journal of Biological Chemistry*, vol. 269, no. 37, pp. 23280–23285, 1994.

[26] Y. Yokosaki, N. Matsuura, T. Sasaki et al., "The integrin $\alpha_9\beta_1$ binds to a novel recognition sequence (SVVYGLR) in the thrombin-cleaved amino-terminal fragment of osteopontin," *The Journal of Biological Chemistry*, vol. 274, no. 51, pp. 36328–36334, 1999.

[27] B. He, M. Mirza, and G. F. Weber, "An osteopontin splice variant induces anchorage independence in human breast cancer cells," *Oncogene*, vol. 25, no. 15, pp. 2192–2202, 2006.

[28] M. Mirza, E. Shaughnessy, J. K. Hurley et al., "Osteopontin-c is a selective marker of breast cancer," *International Journal of Cancer*, vol. 122, no. 4, pp. 889–897, 2008.

[29] J. Sun, A. Feng, S. Chen et al., "Osteopontin splice variants expressed by breast tumors regulate monocyte activation via MCP-1 and TGF-β1," *Cellular & Molecular Immunology*, vol. 10, no. 2, pp. 176–182, 2013.

[30] P. S. Rudland, A. Platt-Higgins, M. El-Tanani et al., "Prognostic significance of the metastasis-associated protein osteopontin in human breast cancer," *Cancer Research*, vol. 62, no. 12, pp. 3417–3427, 2002.

[31] Y. Y. Xu, Y. Y. Zhang, W. F. Lu, Y. J. Mi, and Y. Q. Chen, "Prognostic value of osteopontin expression in breast cancer: a meta-analysis," *Molecular and Clinical Oncology*, vol. 3, pp. 357–362, 2015.

[32] Y. Liu, X. Gu, Q. Lin et al., "Prognostic significance of osteopontin in patients with non-small cell lung cancer: results from a meta-analysis," *International Journal of Clinical and Experimental Medicine*, vol. 8, no. 8, pp. 12765–12773, 2015.

[33] M. Zhao, F. Liang, B. Zhang, W. Yan, and J. Zhang, "The impact of osteopontin on prognosis and clinicopathology of colorectal cancer patients: a systematic meta-analysis," *Scientific Reports*, vol. 10, pp. 1038–1045, 2015.

[34] J.-J. Li, H.-Y. Li, and F. Gu, "Diagnostic significance of serum osteopontin level for pancreatic cancer: a meta-analysis," *Genetic Testing and Molecular Biomarkers*, vol. 18, no. 8, pp. 580–586, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 6183218, 8 pages
http://dx.doi.org/10.1155/2016/6183218

Research Article

Automatic Tissue Differentiation Based on Confocal Endomicroscopic Images for Intraoperative Guidance in Neurosurgery

Ali Kamen,[1] Shanhui Sun,[1] Shaohua Wan,[1] Stefan Kluckner,[1] Terrence Chen,[1] Alexander M. Gigler,[2] Elfriede Simon,[2] Maximilian Fleischer,[2] Mehreen Javed,[3,4] Samira Daali,[3,4] Alhadi Igressa,[3] and Patra Charalampaki[3,4]

[1]*Siemens Healthcare, Technology Center, Princeton, NJ 08540, USA*
[2]*Siemens Corporate Technology, 81739 Munich, Germany*
[3]*Department of Neurosurgery, Hospital Merheim, Cologne Medical Center, 51109 Cologne, Germany*
[4]*Department of Neurosurgery, Heinrich Heine University Düsseldorf, 40255 Düsseldorf, Germany*

Correspondence should be addressed to Mehreen Javed; mehreen.javed2@gmail.com

Received 26 November 2015; Revised 15 January 2016; Accepted 18 January 2016

Academic Editor: Lin Hua

Diagnosis of tumor and definition of tumor borders intraoperatively using fast histopathology is often not sufficiently informative primarily due to tissue architecture alteration during sample preparation step. Confocal laser microscopy (CLE) provides microscopic information of tissue in real-time on cellular and subcellular levels, where tissue characterization is possible. One major challenge is to categorize these images reliably during the surgery as quickly as possible. To address this, we propose an automated tissue differentiation algorithm based on the machine learning concept. During a training phase, a large number of image frames with known tissue types are analyzed and the most discriminant image-based signatures for various tissue types are identified. During the procedure, the algorithm uses the learnt image features to assign a proper tissue type to the acquired image frame. We have verified this method on the example of two types of brain tumors: glioblastoma and meningioma. The algorithm was trained using 117 image sequences containing over 27 thousand images captured from more than 20 patients. We achieved an average cross validation accuracy of better than 83%. We believe this algorithm could be a useful component to an intraoperative pathology system for guiding the resection procedure based on cellular level information.

1. Introduction

Therapy of choice in most malignant and benign tumors in brain is to attempt a total resection with preservation of normal functional tissue, followed by radiochemotherapy. An incomplete resection of a tumor with remaining infiltrative growing cells increases the risk of recurrence with adjacent therapies, decreasing the quality of life and shortening lifetime.

The determination of tumors' borders during the operation is primarily based on the surgeon visual inspection on the tissue through microscope or a histopathologic examination of a limited number of biopsy specimens. Surgical pathology is focused on arriving at a definitive diagnosis of disease in the excised sample, involving a combination of gross, histologic, or even evaluation of molecular properties through immunohistochemistry. Unfortunately, intraoperative histopathology is often not sufficiently informative. Often the biopsies are nondiagnostic due to many reasons. These include sampling errors, which account for possible fact that the biopsies may not originate from the most aggressive part of the tumor. Furthermore, the tissue architecture of the tumor can be altered during the specimen preparation (e.g., frozen section [1]). Other disadvantages are the lack

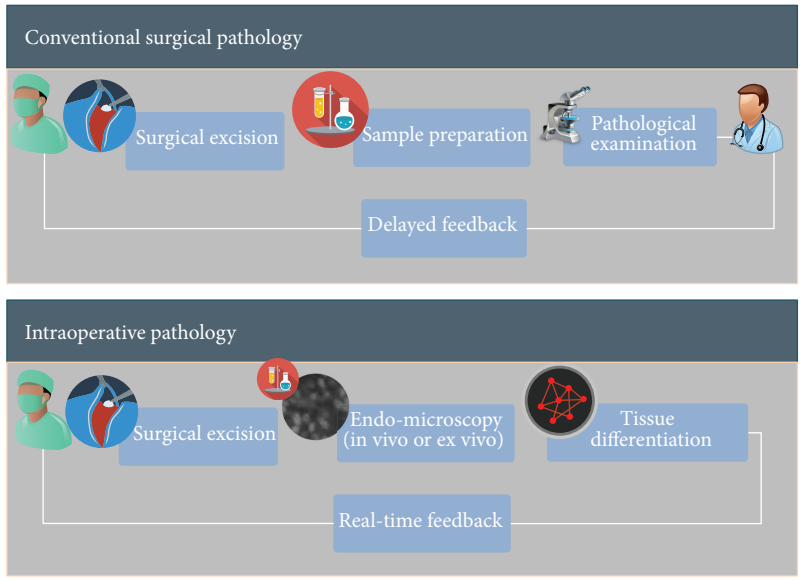


Figure 1: Conventional intraoperative pathology versus proposed intraoperative pathology.

of interactivity and a waiting time of about 30–45 minutes for the result. The impact of "lower quality" slides on the diagnosis and surgical management is investigated in [1]. In summary, optimal surgical therapy, which is the combination of maximal near total resection and minimal injury of the normal tissue can only be achieved if the assessment of cellular, vascular, and connective structures to differentiate tumor from normal functional tissue is done accurately and during the course of operation with minimal time delay (preferably real-time).

A number of recently introduced optical imaging technologies have started to be utilized in the clinical setting both macroscopically and microscopically during surgeries. Endomicroscopy is a technique for obtaining histology-like images from inside the human body in real-time [2], process known as "optical biopsy" [3]. It generally refers to fluorescence confocal microscopy, although multiphoton microscopy and optical coherence tomography have also been adapted for endoscopic use [4]. These images provide abundant information regarding cellular, vascular, and connective tissue structures and specific descriptors which could be used to differentiate various tissue types [5]. Our aim is to be able to computerize these analyses and assist surgeons in delineating tissue boundaries by analyzing real-time streams of images as quickly as possible (see Figures 1 and 2).

To this end, in this paper we propose a method for classifying cellular images and videos, which consist of a multistage image processing pipeline by extracting pixel level features and descriptors. We apply a learnt codebook to quantize each descriptor and generate specific code for each. An image or sequences of images are then represented by pooling result of generated codes into histograms. The pooling result is then used for classifying the image(s) contents into multiple types. The major novelty of our system includes (1) a processing pipeline, where a variety of the feature extraction methods, codebook generations, and classification algorithms can easily be tested and validated, (2) an advanced feature coding method that takes into account both locality and sparsity for computing feature codes, (3) a fast approximate optimization algorithm that significantly improves the computation efficiency with little compromise on ultimate system performance, (4) an image entropy-based data pruning function that increases system robustness to outliers, and (5) a majority voting based classification scheme that boosts the recognition performance for video stream based input data. Experiments demonstrate that the proposed system and method perform well on cellular image classification tasks.

2. Material and Methods

We use a commercially available clinical endomicroscope on the market from Mauna Kea Technologies, Paris, France, called Cellvizio. The main applications for such system are currently in imaging of the gastrointestinal tract, particularly for the diagnosis and characterization of Barrett's Esophagus, pancreatic cysts, and colorectal lesions. We use this system for imaging excised brain tissue during the operation for the identification of the tumor type. The image processing and classification pipeline is developed to automate the process of assigning labels to tumor types.

2.1. System Description. We use a commercially available clinical endomicroscope on the market Cellvizio (Mauna Kea Technologies, Paris, France). The main applications are currently in imaging of the gastrointestinal tract, particularly for the diagnosis and characterization of Barrett's Esophagus, pancreatic cysts, and colorectal lesions. Cellvizio is a probe-based CLE system. It consists of a laser scanning unit,

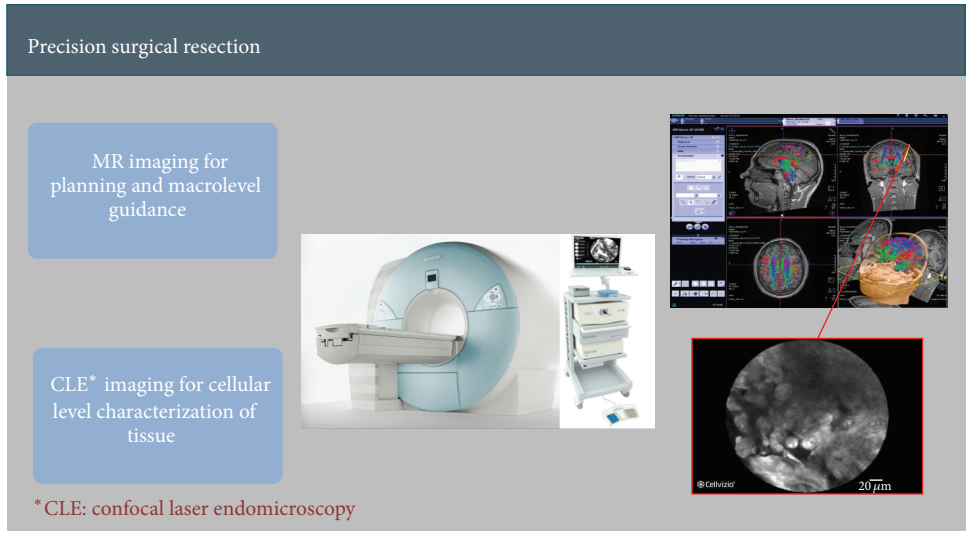


Figure 2: Potential application of intraoperative pathology and surgical guidance within a hybrid OR (this concept is an investigational tool and not approved for clinical use).

proprietary software, a flat-panel display, and miniaturized fiber optic probes. The device is intended for imaging the internal microstructure of tissues in the anatomical tract (gastrointestinal or respiratory) that are accessed by an endoscope. Endomicroscopy enables subsurface analysis of the gut mucosa and in vivo histology during ongoing endoscopy in full resolution by point scanning laser fluorescence analysis [6]. Cellular, vascular, and connective structures can be seen in detail. The new detailed images seen with confocal laser endomicroscopy are unequivocally the beginning of a new era, where this optical development will allow a unique look onto cellular structures and functions at and below the surface of the gut. We consider the application of this technology for brain surgery where identification of malignant (glioblastoma) and benign (meningioma) tumors from normal tissues is clinically important.

2.2. Method Description. Specimens were collected from patients undergoing neurosurgical operations at the Department of Neurosurgery of the Hospital Merheim, Cologne Medical Center, in Cologne, Germany. All studies on human subjects were performed according to the requirements of the local ethic committee and in agreement with the Declaration of Helsinki. Tissue samples were excised from the tumor bed after the resection of the tumor. First, 1-2 drops of 0.01 mg/mL acriflavine hydrochloride AF from Sigma Pharmaceuticals, Victoria, Australia, dissolved in saline were administered topically to the excised tissue sample. This is primarily to stain the nuclei and to a minor extent the cell membrane and the extracellular matrix. The excess dye was washed off with saline. In this study, we used probes from Mauna Kea Technologies, Paris, France, to examine tissue samples. The tip of the probe was placed gently on the tissue and a sequence of images was taken. After imaging, human tissue samples were stored in 4% formalin and transferred for histopathology. Preliminary tests showed that neither the fixation process nor the age of the sample or the preoperative administration of 5-ALA has an effect on CLE examination after topical application of AF. Figure 3 shows examples of typical brain tumor imaging by endomicroscopy technology.

2.3. Intraoperative Endomicroscopic Image Classification. Our classification pipeline includes three parts: offline unsupervised codebook learning, offline supervised classifier training, and online image and video classification. The online classification system is shown in Figure 4. The core components are local feature extraction, feature coding, feature pooling, and classification. Local feature points are detected on the input image and descriptors such as "SIFT" [7] and "HOG" [8] are extracted from each feature point. To encode local features, codebooks are learned offline. A codebook with m entries is applied to quantize each descriptor and generate the "code" layer. As a preferred embodiment, hierarchical K-means clustering method is utilized. For the supervised classification, each image is then converted into an m-dimensional code represented as a histogram, where each bin encodes the occurrence of a quantized feature descriptor. Finally, a classifier is trained using the coded features. As one preferred embodiment, support vector machine (SVM) [9] is utilized. Note that our system is not limited to SVM classifier. For example, as another embodiment, random forest classifier [10] can be utilized alternatively. Two variations of our system are considered. (1) If input images are considered as video streams, our system is able to incorporate the visual cues from adjacent (prior) image frames. This significantly improves the performance of our recognition system. (2) If input images are low-contrast and contain little categorical information, our system can automatically discard those images from further processing. These two variations increase the robustness of the overall system.

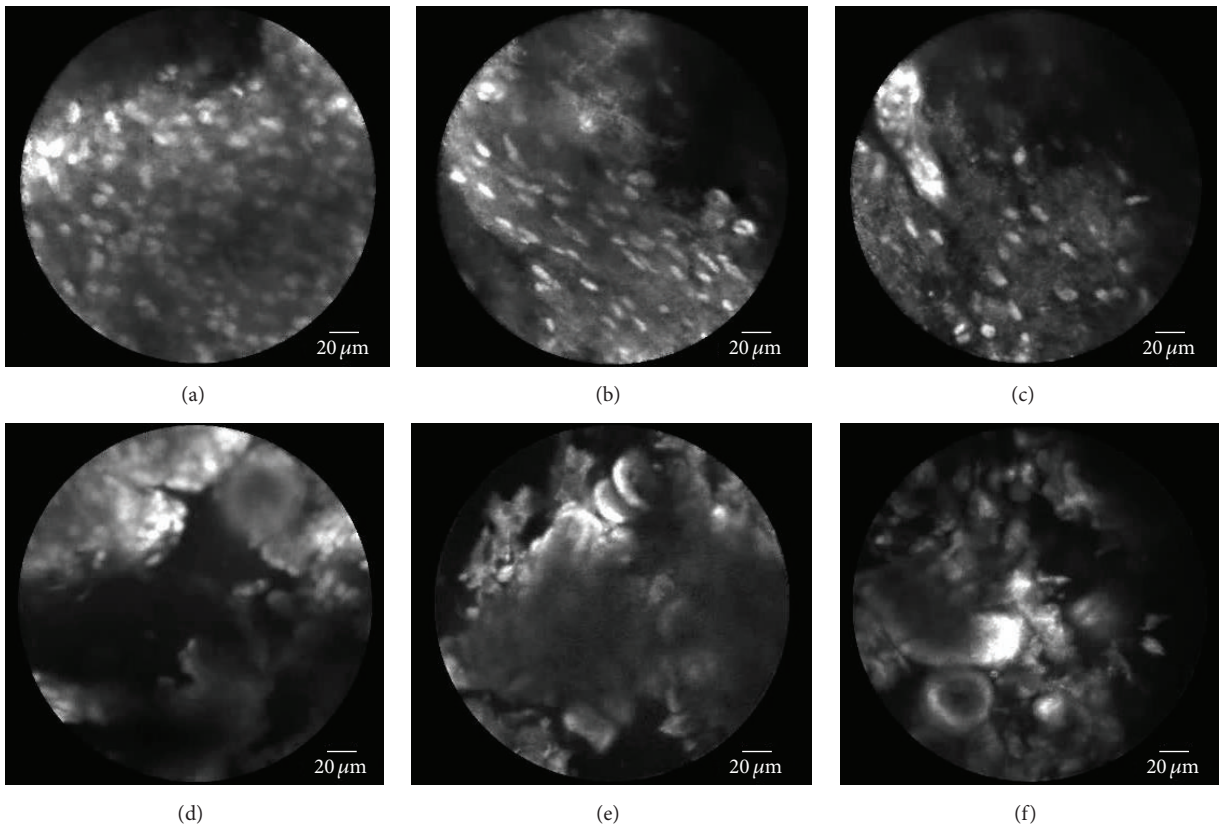


FIGURE 3: Examples of brain tumor imaging by endomicroscopy technology. (a)–(c) Glioblastoma (tumor) images. (d)–(f) Meningioma (tumor) images.

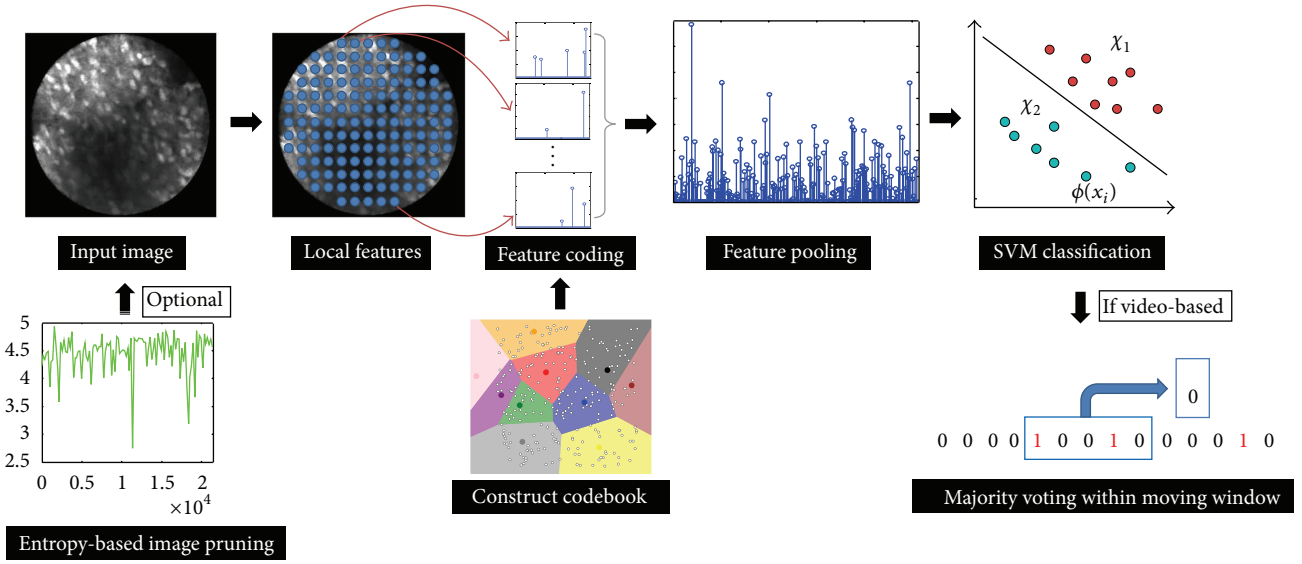


FIGURE 4: Illustration of the image recognition system for tissue classification.

2.4. Entropy-Based Image Data Pruning. Image frames with low image texture information are not clinically interesting or not discriminative for image classification task. Image entropy [11] is a quantity which is used to describe the "informativeness" of an image region, that is, the amount of information contained in a region. On the one hand, low-entropy images have very little contrast where large numbers of pixels have the same or similar intensity values. On the other hand high entropy images have a great deal of contrast from one pixel to the next.

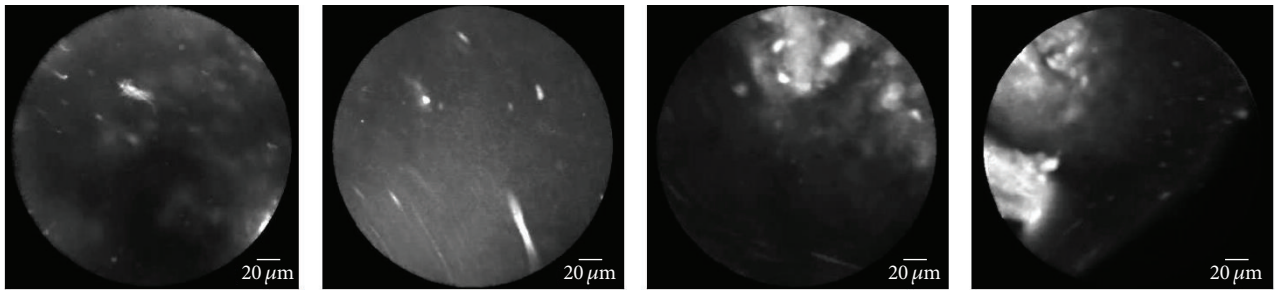
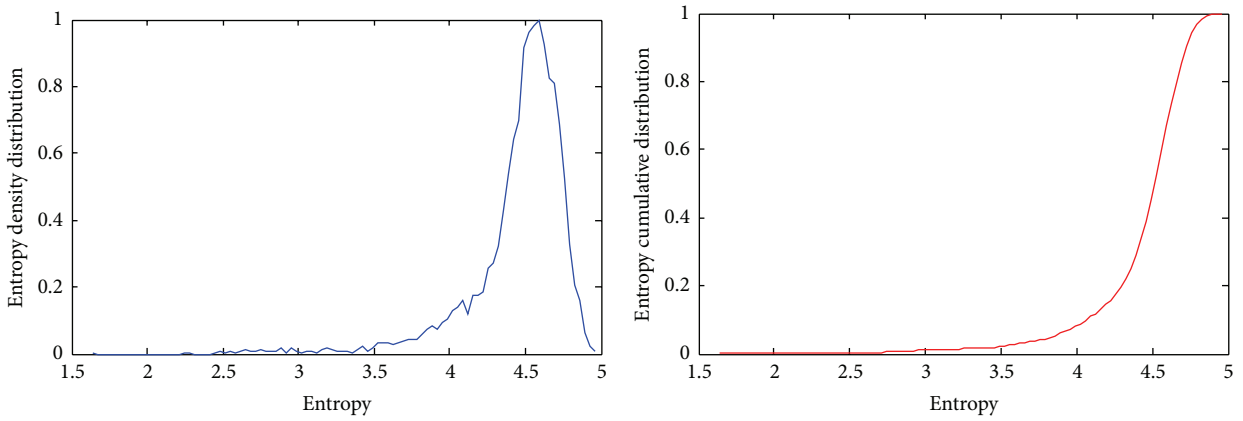


Figure 5: Example of excluded images due to low entropy.



(a) Density distribution of image entropy

(b) Cumulative distribution of image entropy

Figure 6: Image entropy distribution for brain tumor dataset.

See examples of images with low information content in Figure 5.

To estimate a proper entropy threshold, we first calculate the distribution of the image entropy throughout the available dataset. Figure 6 shows the image entropy distribution for brain tumor dataset. As it can be seen, there are a relatively large number of images with entropy significantly lower than that of the rest of the images within a sequence. We therefore set the entropy threshold such that 10% of images will be discarded from later stages of our system (e.g., 4.05 for brain tumor dataset).

3. Theory and Calculations

3.1. Feature Descriptor. Scale Invariant Feature Transform (SIFT) [12] is a local feature descriptor that has been used for a large number of applications in computer vision. It is invariant to translations, rotations, and scaling transformations in the image domain and robust to moderate perspective transformations and illumination variations. Experimentally, the SIFT descriptor has been proven very useful in practice for image matching and object recognition under real-world conditions.

In our system, we utilize dense SIFT descriptors of 20×20 pixel patches computed over a grid with spacing of 10 pixels. Dense image descriptors are necessary to capture uniform regions in cellular structures such as low-contrast regions in case of meningioma.

3.2. Codebook Generation. We perform hierarchical K-means clustering on a random subset of 100,000 local features, extracted from the training set to form a visual vocabulary. We utilized Euclidean distance based exhaustive nearest neighbor search to obtain the feature clusters. As another embodiment, one can also utilize a vocabulary tree structure (binary search tree) to obtain the clusters. For *Bag of Words* (BoW, described in the following section), we preferred the vocabulary tree structure with tree depth of 8. For *sparse coding* (SC, described in the following section) [13], and *locality-constrained sparse coding* (LLC, described in the following section) [14], we prefer K-means of Euclidean distance based exhaustive nearest neighbor search. The utilized vocabulary size of the codebook for our experiments is m and that is set to be 256. Codebook represents the most discriminative set of features, which can supposedly be used for classification. In addition, the codebook is represented within a high dimensional space defined by the number of histogram entries, that is, Bin (e.g., 256 bins). Due to high dimensional space the discrimination between the two classes may not be readily visualized.

3.3. Feature Coding. Let $\mathbf{X}$ be a set of d-dimensional local descriptors extracted from an image; that is, $\mathbf{X} = [x_1, \ldots, x_n] \in \mathbf{R}^{d \times n}$. Given a codebook with m entries, $\mathbf{B} = [b_1, \ldots, b_m] \in \mathbf{R}^{d \times m}$, various coding schemes have been developed to convert each descriptor into an m-dimensional

code $\mathbf{c}_i = [c_{i1}, \ldots, c_{im}] \in \mathbf{R}^m$. In this section, we first review three existing coding schemes (Bag of Words, sparse coding, and locality-constrained linear coding). As embodiments of our classification pipeline, these three coding schemes can be utilized in our proposed pipeline. Toward the end of this section, we will describe the proposed novel locality-constrained sparse coding method and the solution to this problem with more details.

Bag of Words (BoW). As one embodiment, BoW based approach can be utilized. For a local feature x_i, there is one and only one nonzero coding coefficient. The nonzero coding coefficient corresponds to the nearest visual word subject to a predefined distance. When we adopt the Euclidean distance, the code c_i is calculated as

$$c_{ij} = \begin{cases} 1, & \text{if } j = \arg\min_{j=1,\ldots,n} \left\| x_i - b_j \right\|_2^2, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Sparse Coding. As one embodiment, sparse coding [13] based approach can be utilized. The local feature x_i is represented by a linear combination of a sparse set of basis vectors in the codebook. The coefficient vector c_i is obtained by solving an l_1-norm regularized problem:

$$c_i = \arg\min \quad \left\| x_i - \mathbf{B}c_i \right\|_2^2 + \lambda \left\| c_i \right\|_1 \tag{2a}$$

$$\text{s.t.} \quad 1^T c_i = 1, \quad \forall i, \tag{2b}$$

where $\|.\|_1$ denotes the l_1-norm of the a vector. The constraint $1^T c_i = 1$ follows the shift-invariant requirements of the sparse code [15].

Locality-Constrained Linear Coding (LLC). Unlike sparse coding, LLC [14] enforces codebook locality instead of sparsity. This leads to smaller coefficients for basis vectors farther away from x_i. The code c_i is computed by solving the following regularized least squares error:

$$c_i = \arg\min \quad \left\| x_i - \mathbf{B}c_i \right\|_2^2 + \lambda \left\| d_i \odot c_i \right\|_2^2 \tag{3a}$$

$$\text{s.t.} \quad 1^T c_i = 1, \quad \forall i, \tag{3b}$$

where $\odot$ denotes the element-wise multiplication and $d_i \in R^m$ is the locality adaptor that gives different freedom for each basis vector proportional to its similarity to the input descriptor x_i. Specifically,

$$d_i = \exp \frac{\operatorname{dist}(x_i, \mathbf{B})}{\sigma}, \tag{4}$$

where $\operatorname{dist}(x_i, B) = [\operatorname{dist}(x_i, b_1), \ldots, \operatorname{dist}(x_i, b_m)]^T$, and $\operatorname{dist}(x_i, b_1)$ is the Euclidean distance between x_i and b_i. σ is used for adjusting the weight decay speed for local adaptation.

Locality-Constrained Sparse Coding (LSC). The proposed LSC feature coding method compares favorably to state-of-the-art methods in that it not only enforces code sparsity for better discriminative power but also preserves code locality in the sense that each descriptor is best coded within its local-coordinate system. Specifically, the LSC code can be formulated as

$$c_i = \arg\min \quad \left\| x_i - \mathbf{B}c_i \right\|_2^2 + \lambda \left\| d_i \odot c_i \right\|_1 \tag{5a}$$

$$\text{s.t.} \quad 1^T c_i = 1, \quad \forall i. \tag{5b}$$

To the best of our knowledge, this is the first feature coding method that takes into account both locality and sparsity in converting feature descriptor into feature code. Although various algorithms exist for solving the conventional sparse coding problem, it becomes a significantly challenging optimization problem once we incorporate the locality weight vector d_i as in (5a) and (5b). In the following, we present an algorithm for solving (5a) and (5b) based on Alternating Direction Method of Multipliers (ADMM) [16]. A fast approximate solver to (5a) and (5b) is also described based on k-NN search. We kindly point out that this is the first time such an algorithm (ADMM and its k-NN based fast solver) is used to solve the LSC problem.

Optimization Algorithm. We employ the Alternating Direction Method of Multipliers (ADMM) [16] to solve (5a) and (5b). Let us first introduce a dummy variable $y_i \in R^m$ and reformulate (5a) and (5b) as

$$c_i = \arg\min \quad \left\| x_i - \mathbf{B}c_i \right\|_2^2 + \lambda \left\| d_i \odot y_i \right\|_1 \tag{6a}$$

$$\text{s.t.} \quad 1^T y_i = 1, \quad \forall i \tag{6b}$$

$$c_i = y_i. \tag{6c}$$

Then, we can form the augmented Lagrangian of the above objective, which becomes

$$\begin{aligned} \min L(c_i, y_i) = {} & \left\| x_i - \mathbf{B}y_i \right\|_2^2 + \lambda \left\| d_i \odot c_i \right\|_1 + \mu \left\| c_i - y_i \right\|_2^2 \\ & + \rho^T (c_i - y_i) + \mu \left\| 1^T - y_i \right\|_2^2 \\ & + \gamma \left(1^T y_i - 1 \right). \end{aligned} \tag{7}$$

The ADMM consists of three iterations

$$y_i^{t+1} = \arg\min_{y_i} L\left(y_i, c_i^t, \rho^t, \gamma^t \right), \tag{8a}$$

$$c_i^{t+1} = \arg\min_{c_i} L\left(y_i^{t+1}, c_i, \rho^t, \gamma^t \right), \tag{8b}$$

$$\begin{aligned} \rho^{t+1} &= \rho^t + \mu \left(c_i - y_i \right), \\ \gamma^{t+1} &= \gamma^t + \mu \left(1^T y_i - 1 \right), \end{aligned} \tag{8c}$$

which allows us to break our original problem into a sequence of subproblems. In subproblem (8a), we are minimizing $L(y_i, c_i^t, \rho^t, \gamma^t)$ with respect to only y_i, and the l_1-penalty $\|d_i \odot c_i\|$ disappears from the objective making it a very efficient and simple least squares regression problem. In subproblem (8b), we are minimizing $L(y_i^{t+1}, c_i, \rho^t, \gamma^t)$ with

respect to only c_i, and the term $\|x_i - \mathbf{B}y_i\|_2^2 + \mu\|1^T y_i - 1\|_2^2 + \gamma(1^T y_i - 1)$ disappears allowing for c_i to be solved independently across each element. This now allows us to use soft-thresholding efficiently. The current estimates of y_i and c_i are then combined in subproblem (8c) to update our current estimate of the Lagrangian multipliers ρ and γ. Note that ρ and γ play a special role here, as they allow us to employ an imperfect estimate of ρ and γ when solving for both y_i and c_i. For convenience, we introduce the following soft-thresholding (shrinkage) operator:

$$S_\epsilon[x] = \begin{cases} x - \epsilon, & \text{if } x > \epsilon, \\ x + \epsilon, & \text{if } x < -\epsilon, \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

Fast Approximate Coding. The size of the codebook B has a direct effect on the time complexity of algorithm described above. To develop a fast approximate solution to LSC, we can simply use the K ($K < n$) nearest neighbors of x_i as the local bases B_i and solve a much smaller sparse reconstruction system to get the codes:

$$\widehat{c}_i = \arg\min \quad \left\|x_i - \mathbf{B}\widehat{c}_i\right\|_2^2 + \lambda \left\|\widehat{d}_i \odot \widehat{c}_i\right\|_1 \quad (10a)$$

$$\text{s.t.} \quad 1^T \widehat{c}_i = 1, \quad \forall i. \quad (10b)$$

As K is usually very small, solving (10a) and (10b) is very fast. For searching K-nearest neighbors, we can apply a simple but efficient hierarchical K-NN search strategy. In this way, a much larger codebook can be used to improve the modelling capacity, while the computation in LSC remains fast and efficient.

Majority Voting. With an endomicroscopy imaging device, the clinician may obtain a real-time stream of images from the surgical field. This section describes an advanced feature of our image recognition system that can potentially reduce the classification error rate by smoothing out the classification results using a sliding time window (Figure 4). The idea is to assign class labels to the current image frame using the majority voting result of the images within a fixed length time window surrounding the current frame. To the best of our knowledge, this is the first time the majority voting scheme is used to reduce the error rate in a cellular image classification system. We experimentally verify this idea on the brain tumor dataset, and relevant results are presented in the following section.

4. Results

The brain tumor dataset consists of videos of two tumor categories, that is, glioblastoma and meningioma. The equipment is used to collect 86 short videos, each from a unique patient suffering from glioblastoma and 29 relatively longer videos from patients with meningioma. All videos are captured at 24 frames per second, under a resolution of 464 × 336. The collection of videos is hereafter being referred to as the brain tumor dataset. Due to the limited imaging capability of CLE devices or intrinsic properties of brain tumor tissues, the resultant images often contain little categorical information and are not useful for recognition algorithms. Image entropy has been constantly used in the literature to quantitatively determine the information content of an image. Specifically, low-entropy images have very little contrast and large runs of pixels with the same or similar values.

Uninformative video frames are discarded using entropy-based thresholding. The threshold is determined to be 4.05 after computing gray-level entropies of more than 34000 frames in our dataset. This simple thresholding scheme allows us to select more than 16 thousand frames containing glioblastoma and more than 10 thousand frames containing meningioma cases, respectively. We select 80% of these, evenly distributed over both classes, for training and the remaining 20% for testing. All experiments report average classification accuracy on 5 such training/testing splits, ensuring no frame from a training video ending up in the validation split.

We verify majority voting idea using the brain tumor dataset, with the same experimental setup as described in the previous sections. We set the sliding time window to be T in length and derive the class label for the current frame using the majority voting result of the frames within the sliding time window. The recognition performance with respect to the time window length T is given in Figure 7. As it can be seen the optimal performance is achieved at $T = 5$. It is quite likely that higher recognition accuracy can be achieved using much longer time window. In practice, however, one has to balance the relative importance between recognition speed and accuracy. Table 1 depicts the results from three coding approaches including BoW, LLC, and LSC. The results include the classification metrics as well as the processing time per frame. As it can be seen form

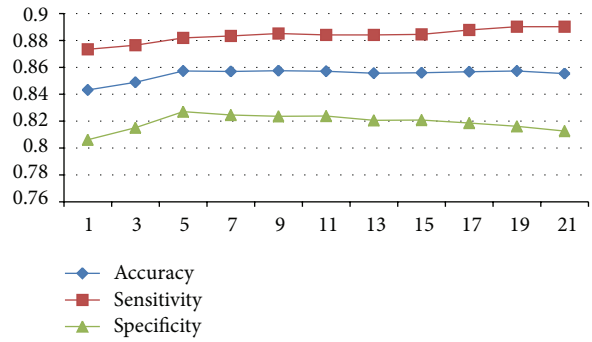


Figure 7: The performance of the majority voting based classification with respect to time window size.

Table 1: The recognition accuracy and speed of different methods on the brain tumor dataset.

	Accuracy	Sensitivity	Specificity	Time (s)
BoW	0.838735	0.893285	0.771834	0.5935
LLC	0.840300	0.877257	0.794974	0.9154
LSC	0.843205	0.873402	0.806171	5.413

the table, the LSC method is the most accurate. However, it does come with a heavy computational cost. The future work includes improving the quality of the classification and also improving the performance of the algorithm to achieve real-time response.

5. Discussion and Conclusion

In this paper we have described a novel system for classifying cellular images and videos. We utilized the state-of-the-art feature coding scheme in our proposed pipeline and demonstrate that they work well on cellular image datasets. We also propose a novel feature coding method (LSC) and a novel fast solution for its implementation. As one embodiment local features are extracted densely to represent the appearance of the local image patch. However, the local feature descriptors are not limited to, for example, SIFT; various types of local feature descriptors, for example, Local Binary Pattern (LBP) [17], Histogram of Oriented Gradient (HOG) [8], and Gabor features [18], can be plugged into our pipeline easily. Future work would consist of investigating the usefulness of feature learning techniques such as Convolutional Neural Networks (CNN) [19]. Local features obtained using machine-learned filters are arguably better than hand-designed features. We expect to significantly improve the performance of our system by exploiting feature learning techniques.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Alhadi Igressa and Patra Charalampaki contributed equally to this paper.

Acknowledgments

This work was supported by Siemens AG, Germany. The authors thank Mauna Kea Technologies, Paris, France, for the technical assistance.

References

[1] E. Mahe, S. Ara, M. Bishara et al., "Intraoperative pathology consultation: error, cause and impact," *Canadian Journal of Surgery*, vol. 56, no. 3, pp. E13–E18, 2013.

[2] P. E. Paull, B. J. Hyatt, W. Wassef, and A. H. Fischer, "Confocal laser endomicroscopy: a primer for pathologists," *Archives of Pathology & Laboratory Medicine*, vol. 135, no. 10, pp. 1343–1348, 2011.

[3] R. C. Newton, S. V. Kemp, P. L. Shah et al., "Progress toward optical biopsy: bringing the microscope to the patient," *Lung*, vol. 189, no. 2, pp. 111–119, 2011.

[4] G. J. Tearney, M. E. Brezinski, B. E. Bouma et al., "In vivo endoscopic optical biopsy with optical coherence tomography," *Science*, vol. 276, no. 5321, pp. 2037–2039, 1997.

[5] P. Charalampaki, M. Javed, S. Daali, H. Heiroth, A. Igressa, and F. Weber, "Confocal laser endomicroscopy for real-time histomorphological diagnosis," *Neurosurgery*, vol. 62, pp. 171–176, 2015.

[6] R. Kiesslich and M. F. Neurath, "Endoscopic confocal imaging," *Clinical Gastroenterology and Hepatology*, vol. 3, no. 7, pp. S58–S60, 2005.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 886–893, IEEE, San Diego, Calif, USA, June 2005.

[9] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[10] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 2, pp. 2161–2168, New York, NY, USA, June 2006.

[11] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB*, Prentice-Hall, Upper Saddle River, NJ, USA, 2003.

[12] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV '99)*, pp. 1150–1157, IEEE, September 1999.

[13] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1794–1801, IEEE, Miami, Fla, USA, June 2009.

[14] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 3360–3367, San Francisco, Calif, USA, June 2010.

[15] L. K. Saul and S. T. Roweis, "An introduction to locally linear embedding," http://www.cs.toronto.edu/~roweis/lle/publications.html.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[17] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[18] H. G. Feichtinger and T. Strohmer, *Gabor Analysis and Algorithms: Theory and Applications*, Springer, 2012.

[19] V. Jain and H. S. Seung, "Natural image denoising with convolutional networks," in *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS '08)*, pp. 769–776, December 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 2615348, 16 pages
http://dx.doi.org/10.1155/2016/2615348

Research Article

An Efficient Approach to Screening Epigenome-Wide Data

Meredith A. Ray,[1] Xin Tong,[2] Gabrielle A. Lockett,[3] Hongmei Zhang,[1] and Wilfried J. J. Karmaus[1]

[1]*Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Zach Curlin Street, Memphis, TN 38152, USA*
[2]*Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, 915 Green Street, Columbia, SC 29208, USA*
[3]*Human Development and Health, Faculty of Medicine, University of Southampton, 801 South Academic Block Tremona Road, Southampton SO16 6YD, UK*

Correspondence should be addressed to Hongmei Zhang; hzhang6@memphis.edu

Received 9 January 2016; Accepted 11 February 2016

Academic Editor: Weiwei Zhai

Screening cytosine-phosphate-guanine dinucleotide (CpG) DNA methylation sites in association with some covariate(s) is desired due to high dimensionality. We incorporate surrogate variable analyses (SVAs) into (ordinary or robust) linear regressions and utilize training and testing samples for nested validation to screen CpG sites. SVA is to account for variations in the methylation not explained by the specified covariate(s) and adjust for confounding effects. To make it easier to users, this screening method is built into a user-friendly R package, *ttScreening*, with efficient algorithms implemented. Various simulations were implemented to examine the robustness and sensitivity of the method compared to the classical approaches controlling for multiple testing: the false discovery rates-based (FDR-based) and the Bonferroni-based methods. The proposed approach in general performs better and has the potential to control both types I and II errors. We applied *ttScreening* to 383,998 CpG sites in association with maternal smoking, one of the leading factors for cancer risk.

1. Background

Due to its high throughput, accuracy, small sample requirement, and acceptable cost, the Illumina Infinium HumanMethylation450 BeadChip has been widely used to analyze deoxyribonucleic acid (DNA) methylation profiles in epigenetic studies that target various types of cancer. In particular, the illumina infinium assay utilizes a pair of probes (a methylated probe and an unmethylated probe) to measure the intensities of methylated and unmethylated alleles at the interrogated cytosine-phosphate-guanine dinucleotide (CpG) sites [1]. Two measures of DNA methylation are usually used: beta-values and M-values. A beta-value is the ratio of signal from the methylated probe relative to the sum of both methylated and unmethylated probes. Beta-values are in the range of $(0, 1)$ with 0 being completely unmethylated and 1 being fully methylated. M-values are $\log 2$ ratio of intensities for methylated and unmethylated probes and range from $(-\infty, +\infty)$ [2, 3]. M-values are used more often in appreciation of its wide data range and variance homogeneity compared to beta-values.

Given the feature of high dimensionality of high-throughput methylation data, when performing designed and possibly complicated statistical analyses, it is wise to target potentially important CpG sites, for instance, CpG sites potentially associated with single-nucleotide polymorphisms (SNPs) and/or other covariate(s) of interest. Otherwise, the statistical power will be substantially lost. There is evidence that methylation is affected by genetic and some known factors such as smoking [4, 5]. Screening CpG sites have become overwhelmingly important across multiple health fields of study such as cancer research, genetic diseases, and epigenetic research.

Common methods for screening CpG sites assume some relationship of the M-values in association with SNPs or some other genetic or environmental factors conditional on

the assumption of linearity with some post hoc adjustment for multiple comparisons. The advantage to this method is the flexibility of incorporating additional covariates and their interactions. The primary limitation lies in controlling for multiple testing. Two popular adjustment methods are the Bonferroni-based method [6, 7] and the Benjamini-Hochberg method for controlling the false discovery rate (FDR) [8, 9]. These methods alter the p-value or critical value to control for type I error. Bonferroni correction is the most conservative by dividing the linear regression p-value, respective of the regression term of interest, by the total number of comparisons (m) or CpG sites in this case such that those adjusted p-values above the significance level are rejected. The FDR method first orders the p-values, $P(k)$ for $k \in 1 \cdots m$, such that lower ordered p-values that are less than or equal to $k/m \times \alpha$ are rejected [9]. It follows that the conservative Bonferroni-based method cannot control for type II error while the FDR-based method cannot control for type I error.

There are other potential issues that arise when dealing with DNA methylation. It is possible that the variation in methylation cannot be fully explained by the known covariates and there exist latent factors that confound with these known covariates [5]. To improve the screening quality, it is thus important to account for variations introduced by other unknown factors. Furthermore, CpG sites screened from one data set may not be consistent with those from another data set which directly affects the type I error rate and leads to a loss of power. It is thus equally important to improve the reproducibility of the selected CpG sites.

In this paper, we propose a novel collaboration of existing statistical techniques to screen genome-wide methylation. It takes unknown factor effects into account and achieves better reproducibility. The method has the ability to control for both types I and II error while adjusting for covariates as well as latent variables. The proposed screening method incorporates surrogate variable analysis [5], which identifies unknown latent variables, in conjunction with a training and testing approach [10] across CpG methylation sites linearly associated with covariates of interest, including the identified surrogate variables. Independently, each method is well established for different purposes. We mingle these methods and form, compared to existing methods, an improved and more efficient process to screen (filter) informative DNA methylation sites. In addition, this proposed method has been built into an efficient and user-friendly R package. In the following sections, further description and details of the proposed method can be found in Section 2, simulation studies and a real data application are included in Section 3, and we summarize the approach in Section 4.

2. Materials and Methods

The proposed screening procedure is built for analyzing the associations between methylation data (M-values), or some high-dimensional data, and covariates of interest and their potential interactions. It consists of two consecutive components and surrogate variable analysis followed by a series of regressions while controlling for multiple testing. Surrogate variable analysis (SVA) aims to identify and estimate latent factors or surrogate variables (SVs) that potentially affect the association between known factors and the response variable, for example, SNPs (known factor) and DNA methylation (response variable) [5]. Including the estimated surrogate variables into the screening process has the potential to reduce unexplained variations, adjust for confounding effects, and consequently improve the accuracy of screening in terms of important variable identification [5]. In the context of DNA methylation, inclusion of the surrogate variables explains the variation in DNA methylation not explained by the covariates currently under consideration. This implies that the identified surrogate variables can be further used to identify important factors (markers) showing large contribution to the variation in the response variable explained by the surrogate variables [11]. These surrogate variables along with other variable(s) of interest can be included in regression analysis as independent variables. After SVA, we then begin the screening process with regressions and adjust for multiple comparisons.

As mentioned in the introduction, several methods exist to adjust for multiple testing but lack the ability to control for both types I and II errors. We have elected to implement a method that will control for both while simultaneously helping the issue of reproducibility. We let randomly chosen training and testing samples estimate and test the effects of the primary covariate(s), termed the TT method. General ideas of this approach are discussed in Dobbin and Simon and Faraggi and Simon [10, 12]. This method follows the concept of cross-validation. It has been shown that the implementation of the training and testing technique can provide a better control of type I error rate [10, 12]. In the next two subsections, we provide detailed steps and options available to both the surrogate variable analysis component and the TT method.

2.1. Identifying Surrogate Variables. Surrogate variables are inferred prior to screening using an algorithmic method developed by Leek and Storey (2007) called surrogate variable analysis [5]. Following the descriptions in Leek and Storey (2007), these SVs are developed by removing the amount of methylation or signal due to the variable(s) of interest and then decomposing the remaining residuals to identify an orthogonal basis of singular vectors that can be reproduced. These vectors are further examined for significant variation to form surrogate variables. Leek and Storey built an R package to perform the surrogate variable analyses (SVAs). The first step in SVA is to identify the number of surrogate variables based on the data using one of two methods, "be" or "leek" as noted in the package. The "be" method, being the default choice according Leek and Storey's R package, is based on a permutation procedure originally proposed by Buja and Eyuboglu in 1992 [13], while the "leek" method provides an interface to the asymptotic approach proposed by Leek in 2011 [14]. Once the number of surrogate variables is calculated, they are then estimated using one of three algorithms, the iteratively reweighted ("irw"), supervised

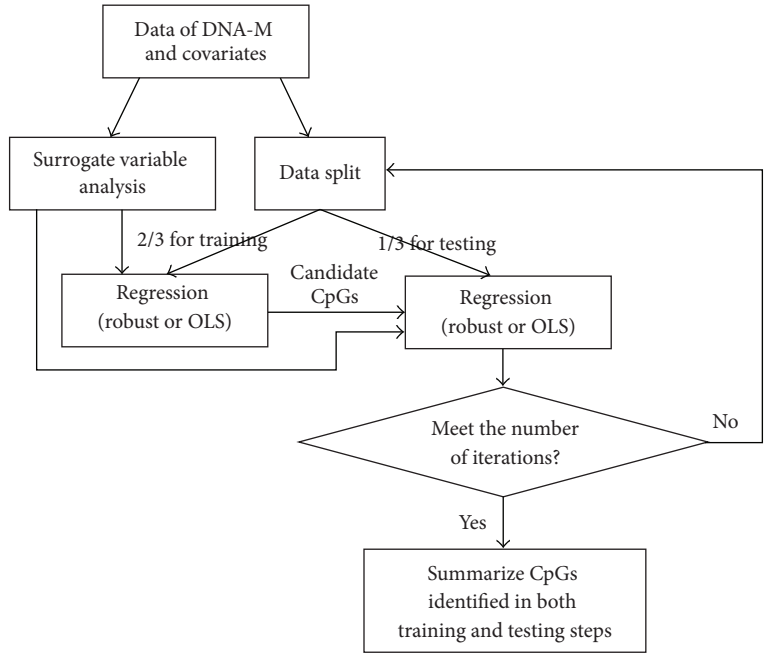


Figure 1: A conceptual representation of the training and testing algorithm.

("supervised"), and the two-step ("two-step") method. Iteratively reweighted method is for empirical estimation of control probes, supervised method is for when control probes are known, and the two-step method is for general estimation of surrogate variables [15]. We elected to implement the "two-step" method following Leek and Storey [5]. Conditional on the data, a number of latent unknown variables have been identified, estimated, and will now be incorporated into the regression in association with DNA methylation as additional covariates.

2.2. The Training and Testing (TT) Screening Method. After surrogate variable analysis is complete, the TT screening method then begins as an iterative process of randomly sampling the training and testing data by a specified proportion. By default, 2/3 of the data will be included in the training data set, which is suggested in Dobbin and Simon [10] to maximize statistical power. Linear regressions are first applied to the training data to calculate the p values for the association between the CpG site and the covariates, including SVs, using either ordinary least squares (this is a default choice in our R package) or robust regression. Robust regression is a type of linear regression that allows for more relaxed assumptions about normality and presence of outliers in the data. A CpG site is included as a candidate if the covariate(s) of interest is statistically significant according to a prespecified significance level, for example, 0.05. In our designed R package, we give user the flexibility to define which term (covariate) is used to decide the selection of CpG sites. For example, suppose the defined right-hand side of the regression is $x_1 + x_2 + x_1 \times x_2$, where $x_1 \times x_2$ denotes the interaction of x_1 and x_2. If the decision of selecting a CpG site is based on one single term, for example, the significance of the interaction effect, the p value for the interaction term would be used to test statistical significance against the prespecified significance level of 0.05.

The process continues with these candidate CpG sites being further tested using the remaining subjects (testing data set) with linear regressions. For one pair of training and testing data sets, a candidate CpG site is deemed as being important if the significance still holds in the testing data. The significance level for the testing data by default is set at the same level as for the training data, 0.05. This screening process will be repeated i times (i.e., iterations); at each iteration, a training and testing data set will be randomly selected. After one iteration, a pool of candidate CpG sites will be selected. This process is continued for total i iterations. We summarize this screening process in Figure 1. Across all i iterations, CpG sites selected in at least m iterations will be included in the final pool of potentially important CpG sites; that is, the cutoff percentage of selection is $m/i \times 100\%$. Final estimates of the associations and statistical significance for the selected CpG sites are inferred by use of the complete data via the same analytical methods (i.e., linear regression including surrogate variables previously estimated from complete data) as in the training and testing process. A cutoff percentage of 50 ($m = 50$ across 100 total iterations ($i = 100$)) was used to determine the final pool of potentially important CpG sites. Suggestions on the determination of this predefined value, m, are discussed later in Section 3.1.

The user-friendly R package, *ttScreening*, is available from the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/web/packages/ttScreening/index.html, which implements the proposed screening procedure discussed above. This *ttScreening* package also provides access to other screening methods: FDR and Bonferroni methods. Various options, such as type of linear regression and surrogate variable estimation method, are available for

the user to specify while other options are data specific and will need to be defined by the user. However, the package does provide acceptable defaults values for those options that are not data specific. A list of package options along with descriptions are available in the package manual at https://cran.r-project.org/web/packages/ttScreening/ttScreening.pdf.

3. Results and Discussion

Simulations are used to demonstrate and assess the TT screening method in comparison with the FDR- and Bonferroni-based methods using the *ttScreening* R package. These are followed up by an application to a data set of 383,998 CpG sites and their association with maternal smoking.

3.1. Simulations. Simulation Scenarios. In total, 2,000 CpG sites across $n = 600$ subjects were simulated. In these 2,000 CpG sites, k sites were assumed to be important. Different settings of k were considered, $k = 10$, 100, 200, and 400. Among the k important sites, DNA methylation at 90% of the k CpG sites was associated with two variables x_1 and x_2, their interaction, and 5 unobservable independent uncorrelated variables, and the remaining 10% of the k sites were associated with x_1, the interaction of x_1 and x_2, and the 5 unobservable independent uncorrelated variables. Variable x_1 is generated from normal distribution with mean 1 and variance 1, and x_2 is a four-level categorical variable generated from multinomial with parameters n and $\pi = \{0.15, 0.25, 0.25, 0.35\}$. The remaining CpG sites $(2000 - k)$ were only associated with the 5 unobservable variables and were deemed as unimportant ones. Linear regressions were applied to simulate the data. To assess the robustness of each method (TT screening and FDR- and Bonferroni-based methods), we considered two types of random error in the regressions, one following normal distribution with mean 0 and variance $\sigma^2 = 1.5$ and the other χ^2 distribution with a degree freedom of 1. Combining the choices on k and the distributions of random error, in total, we have 8 settings. For each setting, we generated 100 Monte Carlo (MC) replicates. The results presented below include means of the number of incorrect selections (rounded to the nearest integer), estimates of sensitivity, and estimates of specificity across the 100 MC replicates. The number of incorrect selections refers to the number of CpG sites misidentified (refers to both false positive and false negative CpG sites) out of the 2,000 CpG sites.

3.2. Results. Variables x_1 or x_2 would be selected if their interaction effect was statistically significant. All package options were chosen as the default values with training and testing significance levels both set to 0.05. The screening results (Table 1) indicate that, in general, the sensitivity from the FDR-based method is comparable to that from the TT screening method but its specificity is lower when the number of important variables is not sparse; for example, $k = 200$. Compared to the Bonferroni-based method, the TT method in general gave better sensitivity and comparable specificity. These were as expected, as the Bonferroni-based method lacks the ability to control type II error while the FDR-based method cannot control well type I error [16]. The TT screening method, on the other hand, has the potential to control both types I and II errors. This is reflected by the results that the TT screening method overall produced the smallest number of incorrectly identified variables, a statistic incorporating information from both sensitivity and specificity. We also performed the screening using robust regressions and similar results were obtained (not shown). These findings are invariant to the distribution pattern of random errors, normally, or skew distributed.

Recall that the default value of the cutoff percentage required for a CpG site to be treated as an informative site across all iterations was 50%, and the significance levels for both the training and testing data set were at 0.05. To further evaluate how the choices of this cutoff percentage and the significance levels in the training and testing steps influence the screening results, we chose a set of the cutoff percentage values ranging from 30 to 90 and set the training significance level to 0.1 instead of default value 0.05. All other settings were kept the same. As seen in Figure 2, across the different number of important CpG sites and sample sizes, overall taking the cutoff percentage close to 50% works the best with significance level of 0.05 for both the training and testing steps. If a higher significance level is chosen in the testing step, then a higher value for cutoff percentage should be used.

The above examination on cutoff percentage is with sample size $n = 600$. To assess whether and how sample size influences the choice of cutoff percentage, we repeated the above analyses for different sample sizes (and for each sample size, 100 MC replicates were generated), $n = 200$, 400, and 800 with all options in *ttScreening*() set at default values except for cutoff percentage. The results (Figures 3–5) indicated that if the number of important variables is expected to be sparse (e.g., 0.5% of the total number variables), the default cutoff value (50%) is a reasonable choice regardless of the sample size. On the other hand, if important variables are not sparse, for example, at least 5% of the total number of variables, then the cutoff value tends to be influenced by the sample size only if sample size is not large, for example, ≤400. In this case, the smaller the sample size is, the lower the cutoff percentage should be taken. This is as expected; when the sample size is smaller, the probability of true positives will be lower. We thus do not expect a higher proportion of correct selection and consequently a lower cutoff percentage is desired. From our simulations, we recommend 20% cutoff if both the following two conditions are not satisfied: (1) important variables are not expected to be sparse (which rarely happens in high-throughput and high-dimensional data) and (2) we have small samples compared to the number of candidate CpG sites. However, if sample size is large, the default cutoff 50% is still recommended. The selection results for sample sizes $n = 200$, 400, and 800 following these recommendations are included in the appendix (Tables 3, 4, and 5).

To demonstrate the benefit of using the identified surrogate variables in screening, we reanalyzed the simulated data (for $n = 600$) but with the surrogate variable analysis excluded (this option is available in the package). The relative patterns of statistics between different methods were

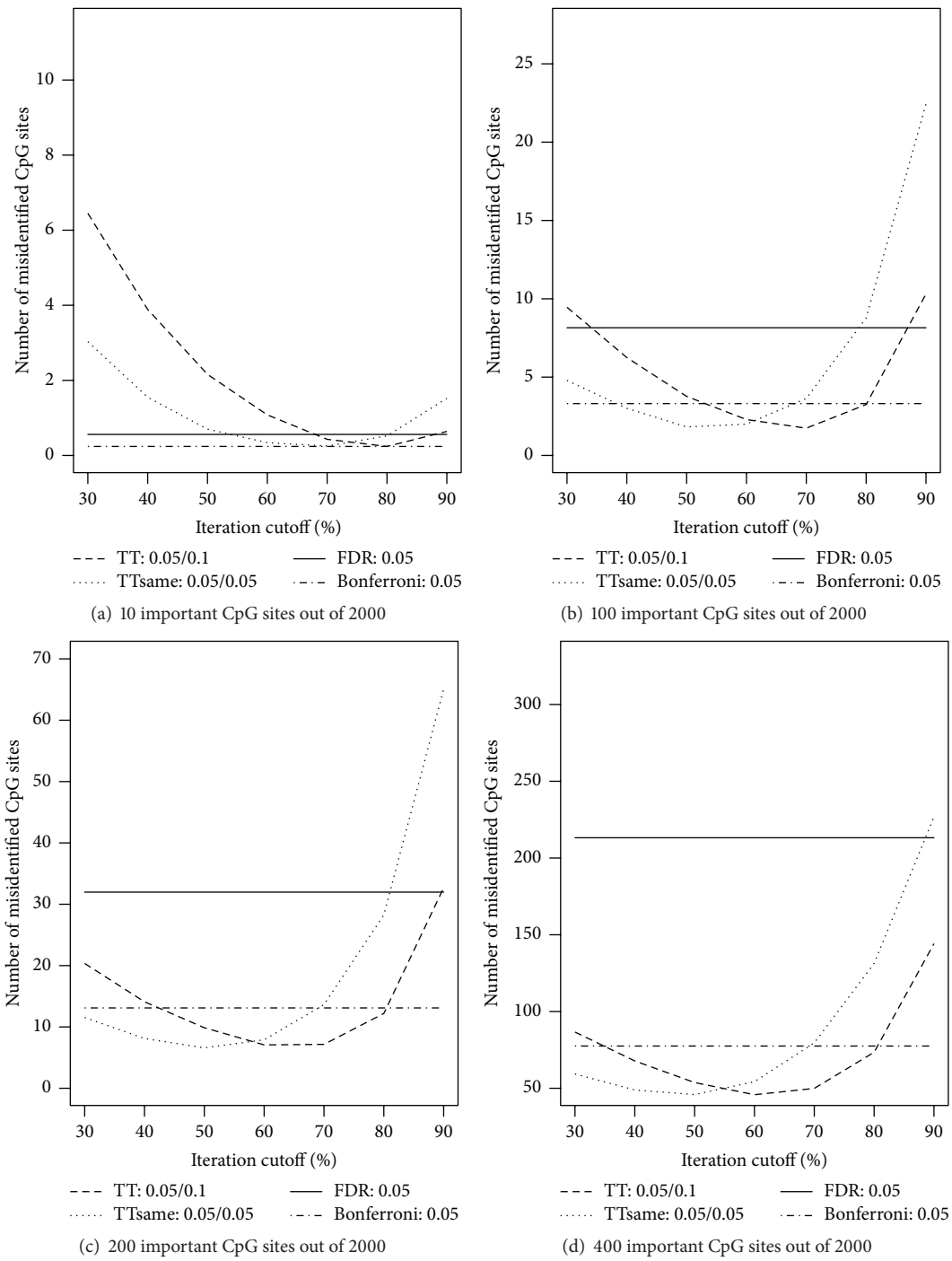


Figure 2: Numbers of misidentified CpG sites versus cutoff frequency for a CpG site being potentially important (based on ordinary least squares regressions). The true numbers of important CpG sites are (a) 10, (b) 100, (c) 200, and (d) 400 out of 2,000 CpG sites. For the TT screening method, two sets of significance levels are considered: (1) 0.05 for training data and 0.1 for testing data; (2) 0.05 for both training and testing data. For the FDR-based and Bonferroni methods, the level was set at 0.05.

intact compared to the results when surrogate variables were included in the screening process. However, when surrogate variables were not considered, a drastic drop in the mean sensitivity was observed across all the settings (Table 2). The sensitivity measures across all the 100 replicates ranged from 0 to 30%, implying that all methods had trouble deciphering CpG sites that are truly important. However, instead of completely excluding surrogate variables, what if we include a set of surrogate variables with each surrogate variable explaining quite a small portion of variation (i.e., less informative)? To examine this, we extracted surrogate variables that are less informative and reanalyzed the data. Similar findings as

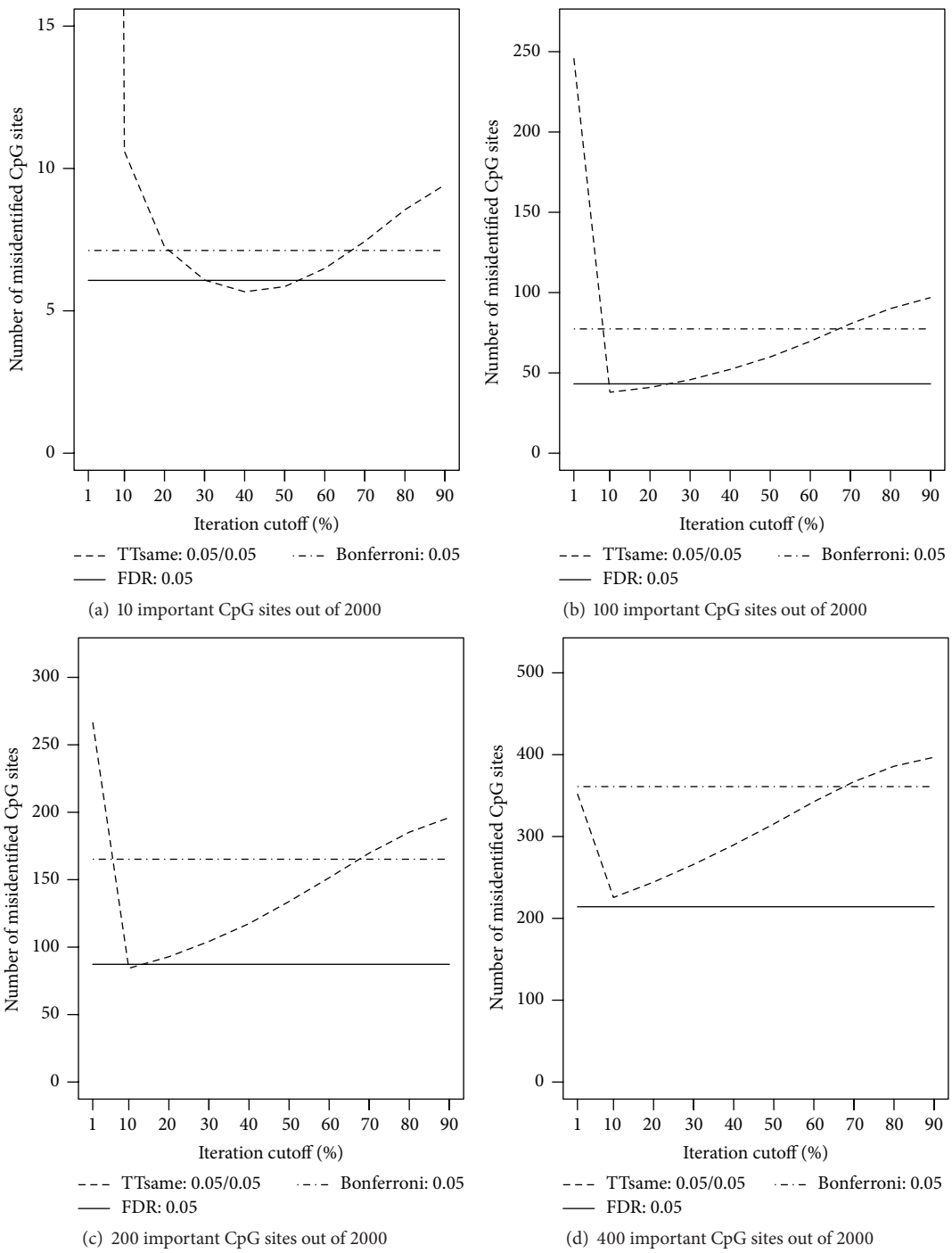


Figure 3: Numbers of misidentified CpG sites versus cutoff frequency for a CpG site being potentially important (based on ordinary least squares regressions). The true numbers of important CpG sites are (a) 10, (b) 100, (c) 200, and (d) 400 out of 2,000 CpG sites across 200 subjects. For the TT screening method, significance levels considered are 0.05 for both training and testing data. For the FDR-based and Bonferroni methods, the level was set at 0.05.

shown in Table 2 were concluded (Table 6), indicating the importance of including informative surrogate variables in order to substantially improve the quality of selection.

We further examined confounding effect of surrogate variables. To achieve this, we followed the same simulation scenarios noted earlier but, instead of assuming independence between all the (observed and unobserved) independent variables, we allow the 5 unobserved variables to be correlated with observed variable x_1 with a correlation of $0.7^{|i-j|}$ where $i, j \in L$ and $L = (0, 1, 2, 3, 4, 5)$ represents an arbitrary

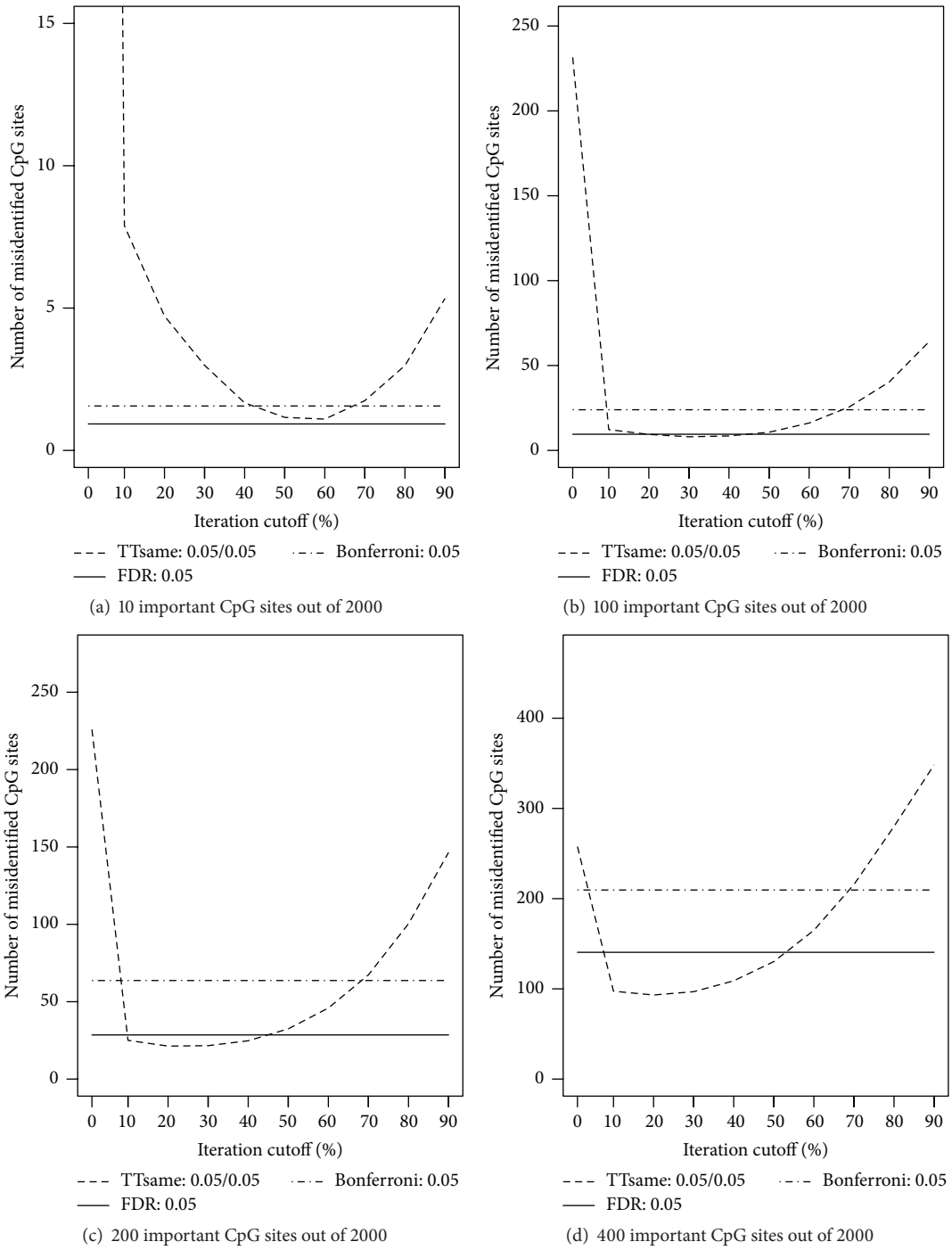


Figure 4: Numbers of misidentified CpG sites versus cutoff frequency for a CpG site being potentially important (based on ordinary least squares regressions). The true numbers of important CpG sites are (a) 10, (b) 100, (c) 200, and (d) 400 out of 2,000 CpG sites across 400 subjects. For the TT screening method, significance levels considered are 0.05 for both training and testing data. For the FDR-based and Bonferroni methods, the level was set at 0.05.

locations of the variables. For example, the observed variable is identified at location 0 and the 5 unobserved variables are locations $(1, 2, 3, 4, 5)$. Then we generated 100 Monte Carlo replicates based on normal distributions with a sample size of $n = 600$. In the simulations, we considered the impact of including and excluding identified surrogate variables. The screening results (results not shown) are consistent with the previous findings when no correlations between unobserved and observed variables were assumed (Tables 1 and 2); that is, including the identified surrogate variable substantially improves the screening statistics (number of incorrectness, sensitivity, and specificity).

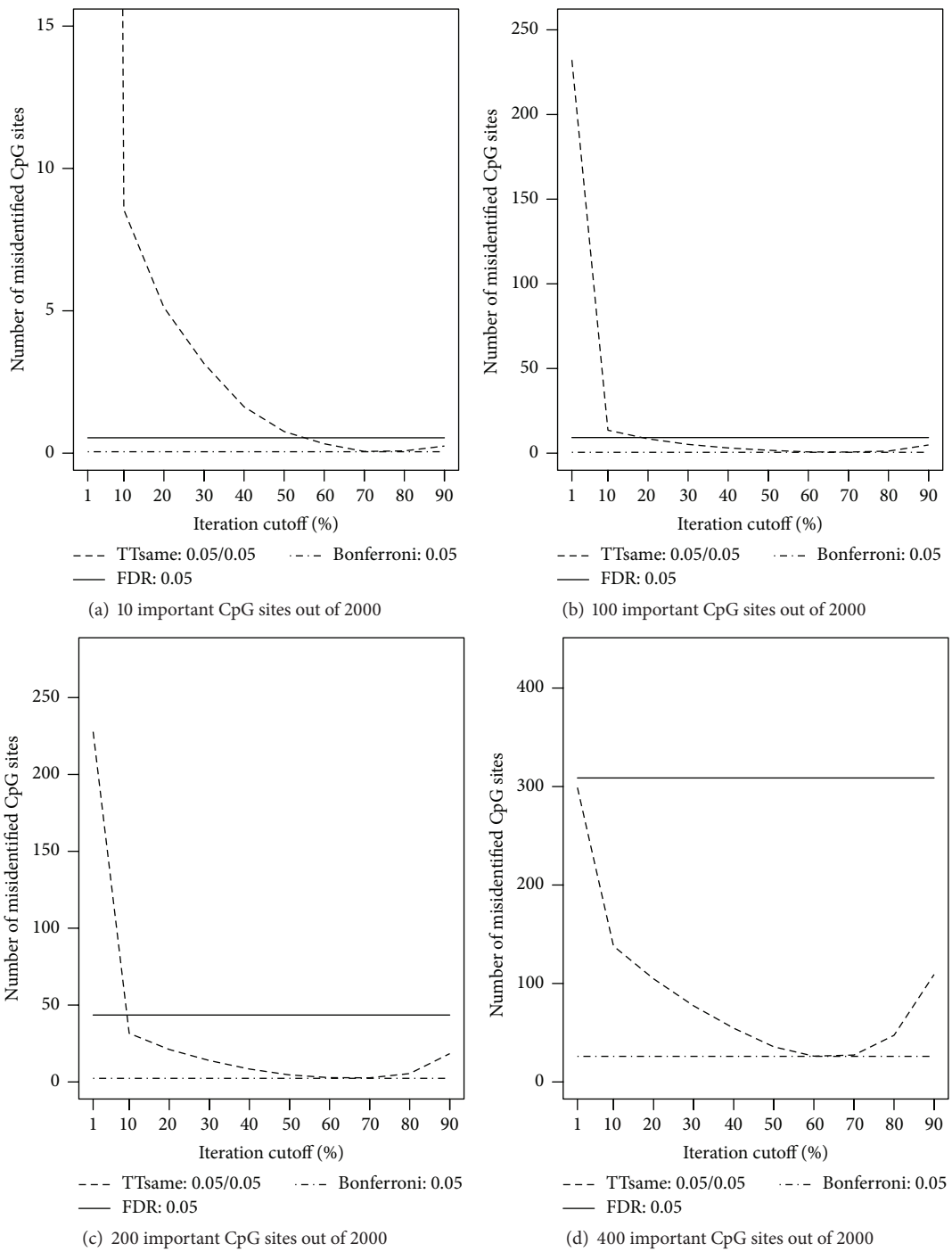


Figure 5: Numbers of misidentified CpG sites versus cutoff frequency for a CpG site being potentially important (based on ordinary least squares regressions). The true numbers of important CpG sites are (a) 10, (b) 100, (c) 200, and (d) 400 out of 2,000 CpG sites across 800 subjects. For the TT screening method, significance levels considered are 0.05 for both training and testing data. For the FDR-based and Bonferroni methods, the level was set at 0.05.

Based on the simulations, it is advised that users follow default setting for the TT screening method: 2/3 of the data for training, "two-step" for SVA analysis as described in the published literature [5], 100 iterations for the total number of TT screenings, 50% as the cutoff proportion of those 100 iterations, and 0.05 significance level for the training and testing data. We recommend 100 iterations as a balance between computing efficiency and adequate resampling of the subjects to decipher true associations. We recommend 50% as the default because in general informative CpGs are sparse compared to the number of candidate CpG sites, in which case, as seen in simulations the 50% cutoff percentage is suitable for small and large sample sizes.

Table 1: Simulation results for selecting k important variables among 2,000 candidates with surrogate variables included.

Statistics	Bon	FDR	TT	Bon	FDR	TT
Random error normally distributed						
		$k = 10$			$k = 100$	
# incorrect	0	1	1	3	8	2
Sensitivity	0.981	0.991	0.994	0.968	0.999	0.991
Specificity	1	1	1	1	0.996	1
		$k = 200$			$k = 400$	
# incorrect	13	32	7	78	213	46
Sensitivity	0.936	0.998	0.982	0.814	0.996	0.931
Specificity	1	0.982	0.998	0.998	0.868	0.989
Random error χ^2 distributed (df = 1)						
		$k = 10$			$k = 100$	
# incorrect	1	1	1	16	9	7
Sensitivity	0.892	0.958	0.963	0.837	0.982	0.941
Specificity	1	1	1	1	0.996	1
		$k = 200$			$k = 400$	
# incorrect	46	29	21	169	158	95
Sensitivity	0.771	0.981	0.906	0.582	0.969	0.790
Specificity	1	0.986	0.999	0.999	0.909	0.993

Bon: Bonferroni, FDR: false discovery rate, and TT: training and testing.

Table 2: Simulation results for selecting k important variables among 2,000 candidates with surrogate variables excluded.

Statistics	Bon	FDR	TT	Bon	FDR	TT
Random error normally distributed						
		$k = 10$			$k = 100$	
# incorrect	10	51	9	99	114	90
Sensitivity	0.011	0.054	0.100	0.014	0.167	0.102
Specificity	1	0.979	1	1	0.984	1
		$k = 200$			$k = 400$	
# incorrect	197	182	180	394	310	359
Sensitivity	0.015	0.232	0.102	0.015	0.288	0.104
Specificity	1	0.984	1	1	0.984	1
Random error χ^2 distributed (df = 1)						
		$k = 10$			$k = 100$	
# incorrect	10	38	9	98	103	91
Sensitivity	0.018	0.053	0.095	0.015	0.158	0.096
Specificity	1	0.986	1	1	0.99	1
		$k = 200$			$k = 400$	
# incorrect	197	176	181	394	314	363
Sensitivity	0.014	0.217	0.094	0.014	0.273	0.095
Specificity	1	0.989	1	1	0.986	1

Bon: Bonferroni, FDR: false discovery rate, and TT: training and testing.

3.3. Real Data Analysis. We applied the *ttScreening* package to 383,998 CpG sites with DNA methylation available for 245 subjects. The data were collected from a study cohort of 18-year-old females on the Isle of Wight (IOW) in the United Kingdom [17]. We examined a single factor that may be potentially associated with DNA methylation, maternal smoking status during pregnancy (0/1). It is thought that maternal smoking in utero increases chances of asthma and wheezing in children and modifies defensive mechanisms such as xenobiotic detoxification systems, antioxidant responses, and damage repair mechanisms [18]. Certain modifications to such systems have been known to increase risk of lung cancer [18].

We applied the TT, FDR-based, and Bonferroni-based methods to identify potentially important CpG sites. In the TT method, 2/3 of the samples were used for training and

Table 3: Simulation results for selecting k important variables among 2,000 candidates including surrogate variables across 200 subjects.

Statistics	Bon	FDR	TT	Bon	FDR	TT
Random error normally distributed						
		$k = 10$			$k = 100^{\dagger}$	
# incorrect	7	6	6	77	43	41
Sensitivity	0.292	0.423	0.475	0.227	0.605	0.641
Specificity	1	1	1	1	0.998	0.997
		$k = 200^{\dagger}$			$k = 400^{\dagger}$	
# incorrect	165	87	93	361	214	244
Sensitivity	0.175	0.616	0.571	0.099	0.565	0.432
Specificity	1	0.994	0.996	1	0.975	0.989
Random error χ^2 distributed (df = 1)						
		$k = 10$			$k = 100^{\dagger}$	
# incorrect	9	8	7	88	62	55
Sensitivity	0.126	0.193	0.307	0.124	0.408	0.498
Specificity	1	1	1	1	0.998	0.997
		$k = 200^{\dagger}$			$k = 400^{\dagger}$	
# incorrect	182	123	119	380	274	288
Sensitivity	0.093	0.422	0.435	0.051	0.377	0.314
Specificity	1	0.996	0.997	1	0.985	0.992

Bon: Bonferroni, FDR: false discovery rate, and TT: training and testing. †Cutoff percentage of 20%.

Table 4: Simulation results for selecting k important variables among 2,000 candidates including surrogate variables across 400 subjects.

Statistics	Bon	FDR	TT	Bon	FDR	TT
Random error normally distributed						
		$k = 10$			$k = 100^{\dagger}$	
# incorrect	2	1	1	24	10	9
Sensitivity	0.851	0.949	0.941	0.761	0.968	0.965
Specificity	1	1	1	1	0.997	0.997
		$k = 200^{\dagger}$			$k = 400^{\dagger}$	
# incorrect	64	29	21	210	141	93
Sensitivity	0.682	0.968	0.947	0.479	0.948	0.864
Specificity	1	0.988	0.994	0.999	0.925	0.976
Random error χ^2 distributed (df = 1)						
		$k = 10$			$k = 100^{\dagger}$	
# incorrect	4	3	3	47	19	18
Sensitivity	0.612	0.767	0.786	0.531	0.878	0.883
Specificity	1	1	1	1	0.997	0.997
		$k = 200^{\dagger}$			$k = 400^{\dagger}$	
# incorrect	111	42	42	289	146	141
Sensitivity	0.445	0.88	0.837	0.281	0.841	0.717
Specificity	1	0.99	0.995	0.999	0.949	0.983

Bon: Bonferroni, FDR: false discovery rate, and TT: training and testing. †Cutoff percentage of 20%.

the remaining for testing. The number of iterations was set at 100 with relative frequency of 50% as the cutoff and the significance level in both training and testing steps was set at 0.05. The significance level for FDR-based and Bonferroni-based methods was set to 0.05. OLS regression was used to estimate associations and p values. Other settings were chosen as the default.

FDR- and Bonferroni-based methods, respectively, identified ten and five CpG sites associated with maternal smoking status. The five CpG sites identified by the Bonferroni-based method were also included in the ten CpG sites identified by FDR. The TT screening method identified 91 CpG sites potentially linked to maternal smoking status. The 10 CpG sites collectively identified by Bonferroni and FDR

Table 5: Simulation results for selecting k important variables among 2,000 candidates including surrogate variables across 800 subjects.

Statistics	Bon	FDR	TT	Bon	FDR	TT
Random error normally distributed						
		$k = 10$			$k = 100$	
# incorrect	2	44	5	26	309	36
Sensitivity	0.991	1	0.998	0.949	1	0.988
Specificity	1	0.976	0.998	0.996	0.807	0.981
		$k = 200$			$k = 400$	
# incorrect	64	29	32	210	141	130
Sensitivity	0.682	0.968	0.847	0.479	0.948	0.696
Specificity	1	0.988	0.999	0.999	0.925	0.995
Random error χ^2 distributed (df = 1)						
		$k = 10$			$k = 100$	
# incorrect	0	1	1	4	9	2
Sensitivity	0.975	0.994	0.993	0.961	0.998	0.99
Specificity	1	1	1	1	0.995	0.999
		$k = 200$			$k = 400$	
# incorrect	15	34	7	82	214	50
Sensitivity	0.928	0.998	0.98	0.803	0.994	0.924
Specificity	1	0.981	0.998	0.998	0.868	0.988

Bon: Bonferroni, FDR: false discovery rate, and TT: training and testing.

were included in the 91 identified by TT. Significant CpG site locations and annotated genes identified are included in Table 7 in Figure 6.

To understand the biological meaning of the identified sets of CpGs, pathway analysis was used. The genes annotated to each significant CpG site were extracted from the 450 K array manifest file (v1.2; available: http://support.illumina.com/downloads/humanmethylation450_15017482_v1-2_product_files.html). Where a CpG site was annotated to more than one gene, all annotated genes were included. The resulting gene lists were analyzed with Ingenuity Pathway Analysis tool (IPA; Qiagen). Statistical significance for each pathway is reported by IPA using p values. The set of 91 CpGs differentially methylated with maternal smoking identified by TT (includes the 5 and 10 CpG sites identified by Bonferroni and FDR, resp.) were mapped to 54 unique genes.

The 91 CpGs differentially methylated with maternal smoking included 18 previously identified top maternal smoking-associated CpG sites in *AHRR*, *CYP1A1*, *GFI1*, *MYO1G*, *CNTNAP2* [4, 19–21], *FRMD4A* [4, 21], *LRRC32*, and intergenic CpGs near *LOC284998* and *PDE10A-SDIM1* [21]. Bonferroni identified five maternal smoking-associated CpGs and FDR identified ten (including all five identified by Bonferroni), all of which have been previously published in association with maternal smoking. However, the TT method identified many more CpGs that have been previously identified as statistically significant in other cohorts, suggesting that these are not simply false positive results, and represent additional truly maternal smoking associated genes worthy of future investigation and validation. Our observation of differential methylation at age 18 in response to maternal smoking during pregnancy also agrees with previous observations that these epigenetic responses are preserved at least into adolescence [20]. The top pathways enriched among genes containing the 91 CpG sites included aryl hydrocarbon receptor signaling ($p = 0.005$) and xenobiotic metabolism signaling ($p = 0.030$), which support a long-lasting effect of maternal smoking on metabolic pathways controlling responses to smoke exposure. Furthermore, these pathways overlap with other pathways including the nicotine degradation II and nicotine degradation III pathways, as well as pathways for the metabolism of cigarette smoke components, and producing known effects of smoke exposure (Figure 6), such as melatonin degradation [22]. The TT method also identified six differentially methylated CpG sites in *HOXA2*, whose mutation causes cleft palate [23], the risk of which is known to be affected by maternal smoking [21].

In conclusion, the CpGs detected by TT in association with maternal smoking were enriched among pathways related to the known biology of those processes. Disruption or modification of some of these pathways results in greater risk of lung cancer [18]. TT also identified CpGs located in genes not previously identified which indicates potentially new findings.

4. Conclusions

We developed a unique screening procedure built into an R package for the purpose of screening important variables. It includes the developed method that involves training and testing steps (the TT screening method), along with another two existing methods, the method controlling for FDR and the other controlling overall significance level through the

Table 6: Simulation results for selecting k important variables among 2,000 candidates including the most and least important surrogate variables across 600 subjects.

	Bonferroni			FDR			TT		
	n.sv = 5	n.sv = 10	n.sv = 15	n.sv = 5	n.sv = 10	n.sv = 15	n.sv = 5	n.sv = 10	n.sv = 15
Most important surrogate variables included									
# incorrect									
k = 10	0	0	0	6	6	6	5	4	3
k = 100	2	3	2	19	19	19	7	5	5
k = 200	6	6	5	26	29	29	6	6	4
k = 400	12	11	11	40	38	39	7	7	7
Sensitivity									
k = 10	1	1	1	1	1	1	1	1	1
k = 100	0.98	0.97	0.98	1	1	1	0.98	0.98	0.98
k = 200	0.97	0.97	0.975	0.995	0.995	0.995	0.985	0.985	0.985
k = 400	0.97	0.973	0.973	1	1	1	0.988	0.988	0.985
Specificity									
k = 10	1	1	1	0.997	0.997	0.997	0.997	0.998	0.998
k = 100	1	1	1	0.99	0.99	0.99	0.997	0.998	0.998
k = 200	1	1	1	0.986	0.984	0.984	0.998	0.998	0.999
k = 400	1	1	1	0.975	0.976	0.976	0.999	0.999	0.999
Most important surrogate variables not included									
# incorrect									
k = 10	10	10	10	10	10	10	10	10	10
k = 100	100	100	100	100	100	100	100	100	100
k = 200	200	200	200	200	200	200	200	200	200
k = 400	400	400	400	400	400	400	400	400	400
Sensitivity									
k = 10	0	0	0	0	0	0	0	0	0
k = 100	0	0	0	0	0	0	0	0	0
k = 200	0	0	0	0	0	0	0	0	0
k = 400	0	0	0	0	0	0	0	0	0
Specificity									
k = 10	1	1	1	1	1	1	1	1	1
k = 100	1	1	1	1	1	1	1	1	1
k = 200	1	1	1	1	1	1	1	1	1
k = 400	1	1	1	1	1	1	1	1	1

FDR: false discovery rate, TT: training and testing, n.sv = number of surrogate variables, and k: the number of truly important CpG sites out of 2,000.

Bonferroni method. Simulations are used to demonstrate and assess the TT screening method.

Overall, the TT screening method produced comparable sensitivity results to that of FDR-based method and comparable specificity results of Bonferroni-based methods. However, the number of misidentified CpG sites from the TT method is in general smaller than those from the other two approaches. The findings on sensitivity and specificity were as expected, because the Bonferroni-based method lacks the ability to control type II error while the FDR-based method cannot control well type I error [16]. The TT screening method, on the other hand, has the potential to control both the types I and II errors, which was supported by the smaller numbers of incorrect detections. It was also noticed that, by incorporating surrogate variable analysis, all three methods produced higher sensitivity measures. In the real data application, the TT method identified a larger number of CpGs compared to the other two methods. The subsequent pathway analyses further support the practical strength of the proposed method. It is worth noting that since the screening process is to identify the CpG sites, to properly assess the functionality of these sites, it is critical to evaluate them jointly, for example, using pathway analyses.

An important contribution of the developed process is its computing efficiency. Although we combined three methods into one package, the computing is not burdensome. The computing time for screening 383,998 CpG sites (the real data analysis) with the *ttScreening*() function using default options only takes 82 minutes on a personal computer with a central processing unit (CPU) of 2.0 gigahertz (GHz) and memory of 4.0 gigabytes (GB) of random-access memory (RAM). The *ttScreening*() function always automatically adjusts for

Table 7: Information for CpG sites associated with maternal smoking identified by the training and testing screening method.

Model 2: maternal smoking only				
CpG site name	Chromosome: map information	Gene name	Promoter region and exon	CpG island location
cg20464068	chr1: 14026054	*PRDM2*	TSS1500	14026481–14027200
cg07951355	chr1: 40123717			
cg14179389	chr1: 92947961	*GFI1*	Body	92945907–92952609
cg00909806	chr1: 212688762			212688031–212688448
cg26143053	chr2: 3718125	*ALLC*	5′UTR	
cg03690080	chr2: 39188006	*LOC100271715, LOC375196*	Body, TSS1500	39186777–39187968
cg19243656	chr2: 73340094	*RAB11FIP5*	5′UTR, 1stExon	73339292–73340733
cg18703066‡	chr2: 105363536			
cg07516970	chr2: 157181359	*NR4A2*	3′UTR	157184389–157184632
cg19273101	chr2: 191734576			
cg14075934	chr2: 200137014	*SATB2*	Body	
cg03158780	chr3: 577964			
cg10663973	chr4: 6642559	*MRFAP1*	1stExon, 5′UTR	6642194–6643322
cg10364374	chr4: 125857940			
cg23743778	chr4: 140357491			140357362–140357610
cg21401642	chr4: 174421114			174421347–174421559
cg17924476†	chr5: 323794	*AHRR*	Body	320788–323010
cg05575921	chr5: 373378	*AHRR*	Body	373842–374426
cg21161138	chr5: 399360	*AHRR*	Body	
cg12287936	chr5: 1800606	*NDUFS6, MRPL36*	TSS1500	1799461–1801905
cg16244648	chr5: 141555043			
cg18349863	chr6: 29912713	*HLA-A*	Body	29910202–29911367
cg11492288	chr6: 30290596	*HCG18*	Body	30294169–30295071
cg04325960	chr6: 147124986	*LOC729176, C6orf103*	TSS200, Body	
cg00004963	chr6: 147124996	*LOC729176, C6orf103*	TSS200, Body	
cg11881038	chr6: 154408701	*OPRM1*	Body, 1stExon, 5′UTR	
cg20418529	chr6: 166260012			
cg00794911	chr6: 166260532			
cg18132363†	chr6: 166260572			
cg08634229	chr6: 169326603			
cg06769202	chr7: 27142535	*HOXA2*	TSS200	27143181–27143479
cg23206851	chr7: 27143046	*HOXA2*	TSS1500	27143181–27143479
cg02225599	chr7: 27143252	*HOXA2*	TSS1500	27143181–27143479
cg10319053	chr7: 27143370	*HOXA2*	TSS1500	27143181–27143479
cg00445443	chr7: 27143478	*HOXA2*	TSS1500	27143181–27143479
cg06401979	chr7: 27143717	*HOXA2*	TSS1500	27143181–27143479
cg11986226	chr7: 40026390	*CDK13*	Body	
cg19089201‡	chr7: 45002287	*MYO1G*	3′UTR	45002111–45002845
cg04180046‡	chr7: 45002736	*MYO1G*	Body	45002111–45002845
cg12803068‡	chr7: 45002919	*MYO1G*	Body	45002111–45002845
cg25949550‡	chr7: 145814306	*CNTNAP2*	Body	145813030–145814084
cg11207515	chr7: 146904205	*CNTNAP2*	Body	
cg21015808	chr7: 149809179			
cg21330896	chr7: 28205705	*ZNF395*	3′UTR	
cg17199018	chr7: 28206278	*ZNF395*	Body	
cg04690729	chr8: 133494328	*KCNQ3*	TSS1500	133492398–133493586
cg15707110	chr8: 144311102			144311708–144311985
cg08126560	chr9: 92291523	*LOC100129066*	Body	92291268–92291524

Table 7: Continued.

Model 2: maternal smoking only				
CpG site name	Chromosome: map information	Gene name	Promoter region and exon	CpG island location
cg13393408	chr9: 132874232	*GPR107*	Body	
cg11813497	chr10: 14372879	*FRMD4A*	TSS200	
cg12490835	chr10: 22623821			22623350–22625875
cg26520012	chr10: 42672589			42672509–42673432
cg05329352	chr10: 112838983	*ADRA2A*	1stExon	112835990–112839303
cg18424850	chr10: 132945786	*TCERG1L*	Body	
cg19494188	chr11: 1466780	*BRSK2*	Body	1466304–1467210
cg26204383	chr11: 2435667	*TRPM5*	Body	2435295–2436651
cg14436038	chr11: 6494706	*TRIM3*	5′UTR	6494725–6495453
cg15627089	chr11: 16625751			16626053–16629180
cg25160605	chr11: 21087846	*NELL1*	Body	
cg17517598	chr11: 61659090	*FADS3*	TSS200	61658569–61659592
cg10788371	chr11: 76381040	*LRRC32*	5′UTR, 1stExon	76381449–76382295
cg11395306	chr11: 98939366	*CNTN5*	5′UTR	
cg01186919	chr11: 111742365	*ALG9*	TSS1500, TSS200	111741953–111742292
cg02820646	chr11: 115398838			
cg05730269	chr11: 118477055	*PHLDB1*	TSS200, TSS1500	118478235–118481896
cg18493761	chr11: 125386885			
cg09932758	chr12: 58022542	*B4GALNT1*	Body	58021294–58022037
cg09644707	chr12: 114885161			114885105–114885418
cg27103591	chr12: 124809023	*NCOR2*	3′UTR	124808972–124809176
cg02032696	chr14: 67982198	*TMEM229B*	TSS200	67981514–67982380
cg21511816	chr14: 76597824			76597648–76597911
cg24874277	chr15: 33211107	*FMN1*	Body	
cg03643241	chr15: 44487910	*FRMD5*	TSS1500	44486741–44487860
cg20596162	chr15: 45408861	*DUOXA2*	Body	45408573–45409528
cg16754378	chr15: 57179394	*LOC145783*	Body	57179277–57179838
cg06899985	chr15: 65689298	*IGDCC4*	Body	65689142–65689362
cg05549655[†]	chr15: 75019143	*CYP1A1*	TSS1500	75018186–75019336
cg17852385	chr15: 75019188	*CYP1A1*	TSS1500	75018186–75019336
cg11924019[†]	chr15: 75019283	*CYP1A1*	TSS1500	75018186–75019336
cg18092474[†]	chr15: 75019302	*CYP1A1*	TSS1500	75018186–75019336
cg01060282	chr16: 17033575			
cg07675285	chr16: 27121267			27121011–27121241
cg11705699	chr16: 87742845	*KLHDC4*	Body	87742556–87743109
cg09554007	chr16: 89627174	*RPL13, SNORD68*	5′UTR, 1stExon, TSS1500	89626644–89627869
cg16483033	chr17: 1090441	*ABR*	1stExon, 5′UTR	
cg08682866	chr17: 13818946			
cg17624073	chr17: 79393583	*BAHCC1*	Body	79393341–79393742
cg13723693	chr17: 80656731	*RAB40B*	TSS200	80655335–80657183
cg18495341	chr17: 80909836	*B3GNTL1*	Body	
cg18449879	chr19: 16045054	*CYP4F11*	1stExon, Body	
cg02543506	chr20: 33876594	*FAM83C*	Body	33879904–33880215

CpG: cytosine-phosphate-guanine dinucleotide, chr: chromosome, locations of each CpG are for v37 of the human genome, [†]CpG sites identified by false discovery rate, and [‡]CpG sites identified by false discovery rate and Bonferroni-based methods.

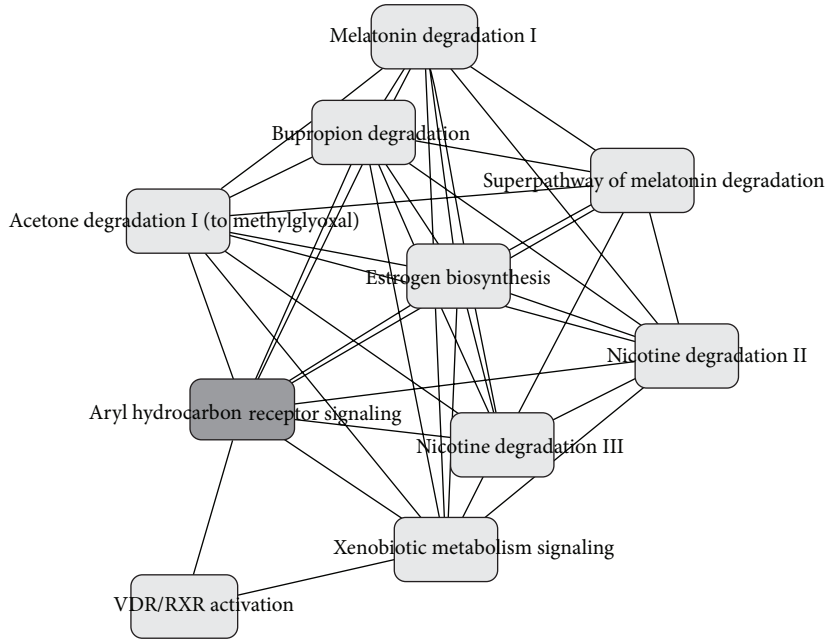


Figure 6: Top pathways enriched among genes identified by training and testing method from the Isle of Wight data include aryl hydrocarbon receptor signaling and xenobiotic metabolism signaling, which overlap with other pathways including the nicotine degradation II and nicotine degradation III pathways, as well as pathways for the metabolism of cigarette smoke components and producing known effects of smoke exposure.

multiple testing using three methods, FDR, Bonferroni, and the training and testing method.

Lastly, the versatility of the proposed screening method allows it to be applied to a variety of scientific fields in which large data or high dimensionality is a computational problem. DNA methylation is of growing interest in all aspects of public health, in cancer and genetics/epigenetics specifically. TT reduces dimensionality in a timely manner while controlling for types I and II errors and adjusting unknown latent variables estimated using the surrogate variable analysis. Finally, the package incorporates a variety of options which allows the user to create very specific settings while maintaining convenient usability.

Appendix

See Tables 3, 4, 5, 6, and 7 and Figures 3, 4, 5, and 6.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

Many thanks to Dr. Nelis Soto-Ramirez for her effort in assisting the authors with the testing of the package. This work is supported by NIH Grant nos. R01AI091905 (Principal Investigator (PI): Karmaus) and R21AI099367 (PI: Zhang).

References

[1] D. Weisenberger, D. Van Den Berg, F. Berg, B. Berman, and P. Laird, *Comprehensive DNA Methylation Analysis on the Illumina Infinium Assay Platform*, Illumina, San Diego, Calif, USA, 2008.

[2] R. A. Irizarry, C. Ladd-Acosta, B. Wen et al., "The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores," *Nature Genetics*, vol. 41, no. 2, pp. 178–186, 2009.

[3] P. Du, X. Zhang, C.-C. Huang et al., "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis," *BMC Bioinformatics*, vol. 11, article 587, 2010.

[4] B. R. Joubert, S. E. Håberg, R. M. Nilsen et al., "450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy," *Environmental Health Perspectives*, vol. 120, no. 10, pp. 1425–1431, 2012.

[5] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genetics*, vol. 3, no. 9, article e161, 2007.

[6] O. J. Dunn, "Estimation of the medians for dependent variables," *The Annals of Mathematical Statistics*, vol. 30, pp. 192–197, 1959.

[7] O. J. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52–64, 1961.

[8] Y. Hochberg and Y. Benjamini, "More powerful procedures for multiple significance testing," *Statistics in Medicine*, vol. 9, no. 7, pp. 811–818, 1990.

[9] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing,"

Journal of the Royal Statistical Society—Series B: Methodological, vol. 57, no. 1, pp. 289–300, 1995.

[10] K. K. Dobbin and R. M. Simon, "Optimally splitting cases for training and testing high dimensional classifiers," *BMC Medical Genomics*, vol. 4, no. 1, article 31, 2011.

[11] X. Ren, Y. Wang, L. Chen, X.-S. Zhang, and Q. Jin, "ellipsoidFN: a tool for identifying a heterogeneous set of cancer biomarkers based on gene expressions," *Nucleic Acids Research*, vol. 41, no. 4, article e53, 2013.

[12] D. Faraggi and R. Simon, "A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis," *Statistics in Medicine*, vol. 15, no. 20, pp. 2203–2213, 1996.

[13] A. Buja and N. Eyuboglu, "Remarks on parallel analysis," *Multivariate Behavioral Research*, vol. 27, no. 4, pp. 509–540, 1992.

[14] J. T. Leek, "Asymptotic conditional singular value decomposition for high-dimensional genomic data," *Biometrics*, vol. 67, no. 2, pp. 344–352, 2011.

[15] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, *sva: Surrogate Variable Analysis*, Package Version 3.8.0, 2014.

[16] J. D. Storey, "The positive false discovery rate: a bayesian interpretation and the *q*-value," *The Annals of Statistics*, vol. 31, no. 6, pp. 2013–2035, 2003.

[17] A. H. Ziyab, W. Karmaus, J. W. Holloway, H. Zhang, S. Ewart, and S. H. Arshad, "DNA methylation of the filaggrin gene adds to the risk of eczema associated with loss-of-function variants," *Journal of the European Academy of Dermatology and Venereology*, vol. 27, no. 3, pp. e420–e423, 2013.

[18] F. D. Gilliland, Y.-F. Li, L. Dubeau et al., "Effects of glutathione S-transferase M1, maternal smoking during pregnancy, and environmental tobacco smoke on asthma and wheezing in children," *American Journal of Respiratory and Critical Care Medicine*, vol. 166, no. 4, pp. 457–463, 2002.

[19] K. W. K. Lee, R. Richmond, P. Hu et al., "Prenatal exposure to maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age," *Environmental Health Perspectives*, vol. 123, no. 2, pp. 193–199, 2015.

[20] R. C. Richmond, A. J. Simpkin, G. Woodward et al., "Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC)," *Human Molecular Genetics*, vol. 24, no. 8, Article ID ddu739, pp. 2201–2217, 2015.

[21] C. A. Markunas, Z. Xu, S. Harlid et al., "Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy," *Environmental Health Perspectives*, vol. 122, no. 10, pp. 1147–1153, 2014.

[22] F. Ozguner, A. Koyu, and G. Cesur, "Active smoking causes oxidative stress and decreases blood melatonin levels," *Toxicology and Industrial Health*, vol. 21, no. 10, pp. 21–26, 2005.

[23] M. Gendron-Maguire, M. Mallo, M. Zhang, and T. Gridley, "Hoxa-2 mutant mice exhibit homeotic transformation of skeletal elements derived from cranial neural crest," *Cell*, vol. 75, no. 7, pp. 1317–1331, 1993.