Med Biol Eng Comput (2010) 48:459–468
DOI 10.1007/s11517-010-0595-0

ORIGINAL ARTICLE

A method for detecting significant genomic regions associated with oral squamous cell carcinoma using aCGH

Ki-Yeol Kim · Jin Kim · Hyung Jun Kim · Woong Nam · In-Ho Cha

Received: 17 November 2009 / Accepted: 26 February 2010 / Published online: 20 March 2010

Abstract Array comparative genomic hybridization (aCGH) provides a genome-wide technique for identifying chromosomal aberrations in human diseases, including cancer. Chromosomal aberrations in cancers are defined as regions that contain an increased or decreased DNA copy number, relative to normal samples. The identification of genomic regions associated with systematic aberrations provides insights into initiation and progression of cancer, and improves diagnosis, prognosis, and therapy strategies. The McNemar test can be used to detect differentially expressed genes after discretization of gene expressions in a microarray experiment for the matched dataset. In this study, we propose a method to detect significantly altered DNA regions, shifted McNemar test, which is based on the standard McNemar test and takes into account changes in copy number variations and the region size throughout the whole genome. In addition, this novel method can be used to detect genomic regions associated with the progress of oral squamous cell carcinoma (OSCC). The performance of the proposed method was evaluated based on the homogeneity within the selected regions and the classification accuracies of the selected regions. This method might be useful for identifying new candidate genes that neighbor known genes based on the whole-genomic variation because it detects significant chromosomal regions, not independent probes.

K.-Y. Kim · J. Kim · I.-H. Cha (✉)
Oral Cancer Research Institute, College of Dentistry, Yonsei University, Seoul 120-752, Republic of Korea
e-mail: cha8764@yuhs.ac

H. J. Kim · W. Nam · I.-H. Cha
Department of Oral and Maxillofacial Surgery, College of Dentistry, Yonsei University, Seoul 120-752, Republic of Korea

Keywords Shifted McNemar · Oral squamous cell carcinoma · aCGH · Genomic variations · Systematicaberrations · Significant chromosomal region

1 Introduction

Chromosomal aberrations such as deletions, amplifications, and structural rearrangements are hallmark of cancer [10, 14, 23]. Therefore, identifying genomic regions associated with systematic aberrations provides insights into the initiation and progression of cancer, and improves the diagnosis, prognosis, and treatment strategies [15]. For understanding genome-wide genetic aberrations, comparative genomic hybridization (CGH) and array technology combined as an array comparative genomic hybridization (aCGH) have been used. Since aCGH data includes high throughput genetic information, these different analytic methods should be applied in conjunction with microarray techniques to analyze this type of data. While the purpose of microarray data analysis is significant gene selection, the main issue in aCGH analysis is to segment the sequence of log ratios along the chromosome into regions of amplification, deletion or no change [13]. Many studies conducted for this purpose have concentrated on smoothing the copy number variations (CNV) throughout the whole genome [2, 5, 9, 10, 15, 17, 19, 24, 27], and CNV was defined as a duplication or deletion event involving >1 kb of DNA [6]. aCGH methods have been applied to identify chromosomal aberration in OSCC [7, 18, 21, 22, 25, 26].

Previous studies in this field have usually focused on each experimental group and detected significant regions by comparing CNV patterns among different experimental groups or the relative CNV. Therefore, they did not

consider whole samples, which were included in different experimental groups.

In this study, we used two small aCGH data sets from OSCC patients, which were collected at different time periods. The two data sets were combined after discretization, because the previous study showed that classification was improved when the data set was combined after discretization [12]. The Chi-square test can be commonly used to detect differentially expressed genes after discretization of expression intensities in microarray experiments. For the matched dataset, however, the McNemar test should be used instead of the Chi-square test.

Based on these observations, we proposed a method, shifted McNemar test, which detects significant regions by considering different experimental groups and the region size at the same time from aCGH data. The proposed method can identify significant regions, and classification accuracies can be improved by these selected regions. In addition, the relationship between the detected genomic regions and the progress of OSCC can be investigated using this novel method, which will be the topic of a subsequent study.

2 Methods

2.1 Data set

Surgical OSCC tissues and their surgical margin tissues were obtained from 11 OSCC patients. The oral cell carcinoma tissue and its marginal tissue were called "tumor" and "dysplasia", respectively, in this study. Experiments on four of the 11 patients were conducted in 2007 and the experiments on the remaining seven patients were conducted in 2008. The clinical features of the 11 samples used in this study are summarized in Table 1.

2.2 Microarray-CGH labeling and hybridization

In the aCGH experiments, we used 60 mer in situ synthesized oligonucleotide arrays designed and produced by Agilent Technologies (Santa Clara, CA, USA), containing 44k probes. As a reference sample, human genomic (male/female) DNA (Promega Corporation, Madison, WI, USA) was used. All array hybridization was performed according to Agilent's recommended protocols. Briefly, three ug gDNA was digested with restriction enzymes *Alu*I and *Rsa*I and fluorescently labeled using the Agilent DNA Labeling kit. Test samples and reference samples were fluorescently labeled with Cy3 or Cy5 dUTP. Labeled DNA were denatured and pre-annealed with Cot-1 DNA and Agilent blocking reagent prior to hybridization for 40 h at 20 rpm in a 65°C Agilent hybridization oven. Standard procedures were followed when washing. Hybridized arrays were scanned at a 5 μm resolution with an Agilent G2505A scanner. Scanned image analysis was performed using Feature Extraction Software 9.1.1.1 (Agilent Technologies), with the CGH-v4_91 protocol for background substraction and normalization. All array data passed Agilent recommended quality metrics.

2.3 Shifted McNemar test

For the matched-pairs data, the McNemar test [20] has been traditionally applied only to the case in which there are two possible categories for the outcome. In practice, however, the outcomes can be classified into multiple categories. Under this situation, the McNemar test was extended for data that contained more than three categories [1].

For example, in a study a test is performed before treatment and after treatment. The results of the test are coded "+" and "−". Using this approach, we can test if

Table 1 Clinical features of patients

Dataset	Gender	Age	Tumor primary site	Diagnosis	Grade	LN invasion	Patho. stage
Data 2008	M	46	Medial Pterygoid	SCC	G2	+	T3N2bM0
	M	59	Lower gingiva	SCC	G2	−	T3N0M0
	M	62	Upper gingiva	SCC	G2	+	T2N0M0
	F	79	Lower gingiva	SCC	G2	−	T4N0M0
	M	64	Oral tongue	SCC	G2	+	T3N2bM0
	F	45	Oral tongue	SCC	G1-2	+	T2N0M0
	M	61	Buccal cheek	SCC	G2	−	T3N0M0
Data 2007	M	62	Oral tongue	SCC	G2	+	T3N1M0
	M	64	Lower gingiva	SCC	G1-2	−	T2N0M0
	M	50	Lower gingiva	SCC	G2	+	T4N2bM0
	F	69	Oral tongue	SCC	G1-2	−	T2N0M0

Table 2 The summarized frequency table for McNemar test

		Before	
		+	−
After	+	*A*	*B*
	−	*C*	*D*

A and *D* represent the number of patients not changed after treatment. *B* and *C* represent the number of patients changed after treatment

there was a significant change in the result before and after treatment. When doing the test, the categorized values can be summarized as shown in Table 2.

A and *D* represent the number of patients not changed after treatment. *B* and *C* represent the number of patients changed after treatment.

A, *B*, *C*, and *D* represent the frequencies of satisfying the two conditions in the "before" and "after", for each patient. McNemar Chi-square statistic was calculated using the values of *B* and *C*, because these two values represent the change between "before" and "after". Therefore, the test statistic for the McNemar test was calculated as Eq. 1.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{\left(B-\frac{B+C}{2}\right)^2}{\frac{B+C}{2}} + \frac{\left(C-\frac{B+C}{2}\right)^2}{\frac{B+C}{2}} = \frac{(B-C)^2}{(B+C)} \quad (1)$$

The degree of freedom is $n \times (n-1)/2$, where n is the number of pairs. In this equation, "*O*" and "*E*" represent "observed value" and "expected value", respectively. Here, the expected value of *B* and *C* is $(B + C)/2$, if there is no change between "before" and "after". Therefore, we can detect the probes which change significantly in the process from dysplasia to tumor, using McNemar test. McNemar test is applied to whole chromosome shifting probe by probe. Therefore, we named the method "shifted McNemar". Since the shifted McNemar test is executed shifting probe by probe (Fig. 1a), the selected significant regions can be partially duplicated as shown in the left-hand side of Fig. 1b. In this case, we integrated such regions and extended the region size (right-hand side of Fig. 1b).

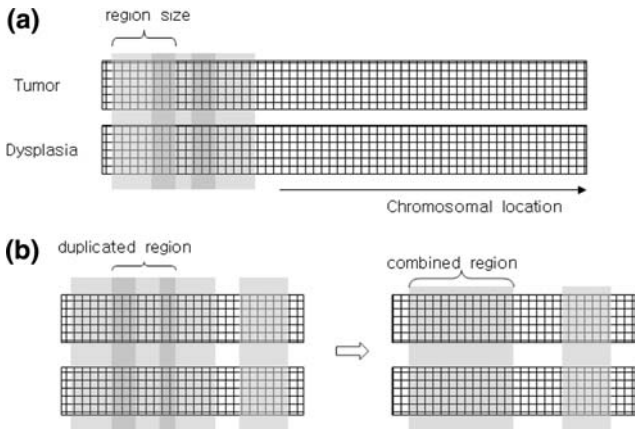

Fig. 1 Data structure of aCGH data. The *horizontal* and *vertical* axes represent chromosomal location and different patient groups, respectively. **a** The analysis is executed by area. **b** The duplicated parts of the selected regions are combined

2.4 Discretization of copy number variations (CNV) for McNemar test

Let the expression intensities of patient i be X_{i1}, X_{i2}, …, X_{in}, when there are n probes for a patient. Then, the order statistics of n expression intensities were represented as $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$. The ordered intensities can be categorized into three levels by the lower quartile (Q1, 25% upper value) and upper quartile (Q3, 75% upper value) for each patient (or experiment). By categorizing the expression intensities for each experiment, some bias, which can exist between different data sets, may be adjusted.

Table 3 The process for categorization of the raw expression intensities of a probe for 7 paired experiments

T1	*T2*	*T3*	*T4*	*T5*	*T6*	*T7*	*D1*	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*
(a) *Raw data*													
−0.05	0.031	0.242	0.189	−0.051	−0.08	0.171	−0.339	0.342	0.425	0.049	0.094	0.548	0.271
(b) *The categorized values (for each experiment) of the raw expression intensities*													
2	2	3	2	2	2	3	2	3	3	2	2	3	3
(c) *The categorized values of five probes*													
2	2	3	2	2	2	3	2	3	3	2	2	3	3
1	1	1	2	1	1	1	1	1	1	2	2	1	1
2	2	3	3	2	2	3	2	2	3	2	3	2	3
3	2	3	2	2	1	3	1	2	3	1	3	3	3
3	2	1	1	2	1	2	1	3	2	1	3	3	2

(a) represents the raw intensities and (b) represents the categorized values and the raw intensities were categorized for each experiment. (c) Shows the categorized intensities of the consecutive five probes

Table 4 The summarized frequencies of the consecutive five probes for McNemar test

		Dysplasia (before)		
		1	2	3
Tumor (after)	1	20	2	2
	2	3	12	2
	3	2	6	6

The sum of frequencies is 55

In real data, the expression intensities were categorized and summarized by the following steps.

(i) The raw expression intensities were categorized as shown in Table 3. We used raw intensities of Data2008 as example. Hence, example data shows seven tumors (T) and seven dysplasia (D).
(ii) The categorized values for 11 paired experiments can be summarized in the form of table as shown in Table 4. We used combined dataset, Data2007 and Data2008.

The sample size of the dataset used in this study was 11, which may be too small for the statistical test. However, this problem can be resolved by considering the region size. For example, if we consider five as the region size, this is the same as having a sample size of 55.

2.5 Evaluation of the proposed method

2.5.1 *Inter-correlation within the selected region*

The inter-correlation coefficient can be used for exploring the homogeneity within the selected region by the proposed method. To compare the homogeneity between the regions selected using the proposed method and a random method, we used 100 randomly selected regions that contain consecutive probes.

If we have a series of n measurements of probe X and probe Y written as x_i and y_i where $i = 1, 2, \ldots, n$, then the Spearman correlation coefficient can be used to estimate the correlation of X and Y. When x_i, y_i are converted into ranks $x_{(i)}$, $y_{(i)}$ and the differences $d_i = x_{(i)} - y_{(i)}$, the Spearman correlation coefficient was calculated as Eq. 2.

$$r_{xy} = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (2)$$

The mean value of the calculated pair-wise correlation coefficients was used as the inter-correlation among probes within the selected region.

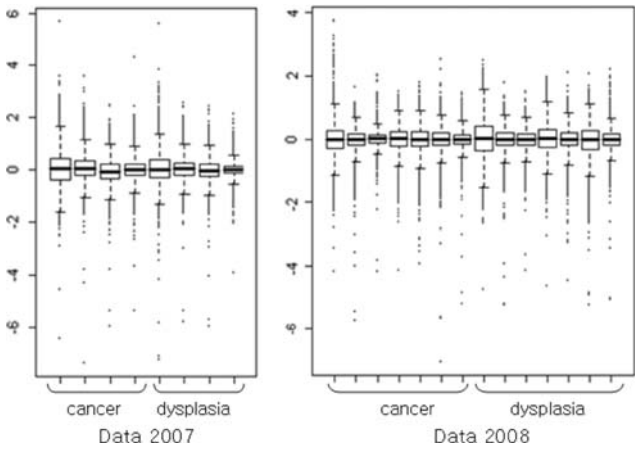

Fig. 2 Distributions of CNVs for each sample of data 2007 and data 2008. Data 2007 and Data 2008 contain four paired and seven paired tissues, respectively

2.5.2 *Classification accuracy*

For evaluating the classification accuracy of the selected regions, we calculated the OOB error rate using the randomForest (RF) test, which was included in the R package (http://www.r-project.org). For comparison, we randomly selected 100 regions that contain consecutive probes.

3 Results

The distributions of CNVs for each sample of Data 2007 and Data 2008 were shown in Fig. 2.

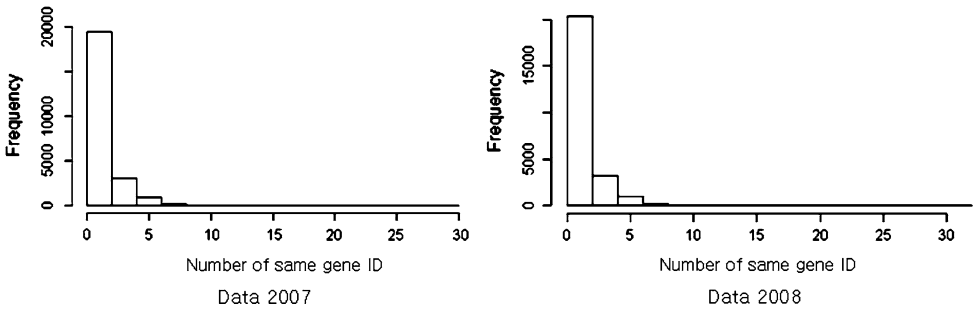

Fig. 3 The distribution of the numbers of replications in data 2007 and data 2008. The *horizontal* and *vertical* axes represent the number of replications and the frequency of the same number of replications, respectively

The CNVs of the dysplasia group were similarly distributed with those of the tumor group, and the individual variations of CNVs were shown. We discretized the raw CNVs without normalization, so that the ranks of CNVs would be retained.

3.1 Decision of appropriate region size

To determine the appropriate the region size, we explored the distribution of frequencies of each probe. As shown in Fig. 3, most frequencies were less than five; therefore, we decided that the appropriate region size was five.

The average distance from probe to probe was 28530 bp in the 44k chip used for this study, and the maximum distance between the five probes was 143 kb. Based on previous information obtained from Database of Genomic Variants (DGV, http://projects.tcag.ca/variation/), 143 kb is an appropriate size for considering the probe set as CNV.

Table 5 Summary of 73 probes in the selected regions

ChrName	Cytoband	FeatureNum	Name.of.Gene	ChrName	Cytoband	FeatureNum	Name.of.Gene
1	1q21.3	43801	SHE	8	8q22.2	2460	STK3
1	1q21.3	7699	TDRD10	8		14506	chr8:100025224-100025283
1	1q21.3	21721	UBE2Q1	8	8q22.2	1001	OSR2
1	1q21.3	1751	CHRNB2	8		10023	chr8:108261475-108261534
1	1q21.1-q21.2	38385	ADAR	8	8q22.3-q23	36353	ANGPT1
1		8713	X52229	8	8q22.3-q23	42716	ANGPT1
2	2p24.2	26570	RDH14	8	8q22.3-q23	20	ANGPT1
2	2p24.2	15383	NT5C1B	8	8q22.3-q23	36778	ANGPT1
2	2p24.2	25802	NT5C1B	8	8q24.21	23708	CCDC26
2		31122	chr2:019074472-019074531	8		15122	chr8:130792478-130792537
2	2p24.1	21013	OSR1	8	8q24.1-q24.2	39908	MLZE
2	2p24.1	8723	OSR1	8	8q24.1-q24.2	7843	MLZE
2		40239	chr2:019600677-019600736	8	8q24.21	19488	FAM49B
2	2q31.1	2831	SSB	11	11p11.2	28303	CREB3L1
2	2q31.1	38025	SSB	11	11p11.2	12588	CREB3L1
2	2q31.1	31677	METTL5	11	11p11.2	21355	DGKZ
2		30574	chr2:170536301-170536360	11	11p11.2	15863	DGKZ
2		38866	ZNF650	11	11p11.2	25714	DGKZ
3		17948	chr3:177181690-177181749	14	14q21	10157	RPL36AL
3		34719	chr3:177292490-177292549	14	14q21	34583	MGAT2
3		32150	chr3:177640401-177640460	14	14q22.1	11685	C14orf104
3		40691	chr3:177754989-177755048	14	14q22.1	25735	C14orf104
3		8731	chr3:177913624-177913683	14		10674	chr14:049179899-049179958
4		38239	chr4:006875302-006875361	14	14q21-q22	14935	POLE2
4	4p16.1	2538	KIAA0232	14	14q23.3	25956	MPP5
4	4p16.1	6704	TBC1D14	14	14q23.3	15736	MPP5
4	4p16.1	18434	TBC1D14	14		8615	CR619369
4	4p16.1	14604	TBC1D14	14	14q23-q24.2	18828	ATP6V1D
4	4p16.1	27674	CCDC96	14	14q23.3	25470	EIF2S1
4	4p16	30511	GRPEL1	22	22q13.33	27637	SAPS2
8		10992	chr8:062949506-062949565	22	22q13.33	25303	SBF1
8		38151	chr8:063235271-063235330	22	22q13.33	16249	ADM2
8		1409	FAM77D	22	22q13.3	40930	MIOX
8		6184	FAM77D	22		13684	BC002942
8		21587	FAM77D	22		6743	hCAP-H2
8	8q22.2	7832	STK3	22		9546	ECGF1
8	8q22.2	42288	STK3				

3.2 Description of probes in the selected region

Using the shifted McNemar test, 21 regions were detected, which contained 73 probes. Here, we used region size and p-value by 5 and 0.01, respectively. The number of regions, therefore, can be increased or decreased according to p-value. These selected probes were described in Table 5. The first column represented the probe number in the chip used for this study.

From the "Genomic Variant" Database (http://projects.tcag.ca/variation/), we confirmed that the selected probes, including ADAR, RDH14, NT5C1B, SSB, METTL5, KIAA0232, TBC1D14, CCDC96, GRPEL1, ANGPT1, FAM49B, POLE2, MPP5, ATP6V1D, EIF2S1, ADM2, and MIOX, had copy number variations in the previous studies. In addition, MLZE on 8q24.21 is known to be expressed in metastatic melanoma cell [28].

3.3 Comparison of inter-correlation among probes within the selected regions

To compare inter-correlations, we used the mean value of the pairwise correlation coefficients among probes within the selected region.

The inter-correlation within the regions selected by the proposed method was significantly higher than those determined by a random method (Fig. 4, p-value = 0.002870). This result indicates that the proposed method selected significantly meaningful probes, which were homogeneous as well as consecutive within a region.

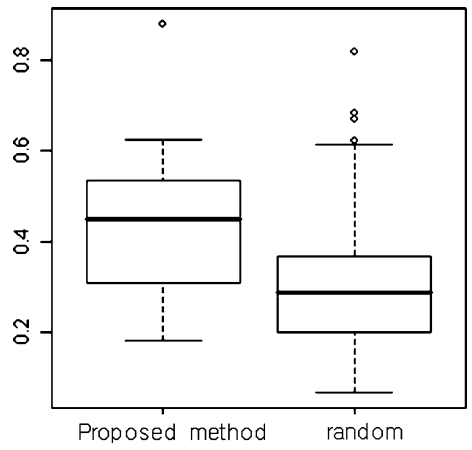

Fig. 4 Comparison of the correlation coefficients between the selected regions by the proposed method and a random method. The correlation coefficients determined using these two methods were significantly different

3.4 Exploration of CNV in the selected regions between different experimental groups

The CNV patterns of the selected regions in chromosome 2 and chromosome 8 were investigated.

Figure 5a, b shows CNV patterns and discretized CNV patterns, respectively, of a region of chromosome 2p24, where the last row represents the McNemar Chi-square statistic. The intensities in the regions with large McNemar Chi-square statistics indicate significant differences between two experimental groups. The highlighted region was a region selected by the proposed method, and RDH14, NT5C1B and OSR1 were included in this region. It has already been reported that these genes contain copy number variations in human (http://projects.tcag.ca/variation/), and overexpression of OSR1 resulted in up-regulation of p53 activity [8]. OSR1 was also shown to activate p53 through repression of HDM2 transcription and its over-expression resulted in up-regulation of p53 activity [8]. The expression of OSR1 mRNA was significantly weakened in gastric cancer cell lines (OKAJIMA, MKN45), pancreatic cancer cell lines (PANC-1, BxPC-3, AsPC-1, PSN-1, PSN-1, Hs766T), and esophageal cancer cell lines (TE10) [11].

The three regions detected from chromosome 8 were 8q22.2, 8q22.3-q23, and q24.1-q24.2. These regions include STK3, OSR2, ANGPT1, CCDC26, MLZE, and FAM49B. It is known that ANGPT1 and FAM49B are deleted in 8q23.1 and 8q24.21, respectively (DGV). ANGPT1 was shown in the selected region of Fig. 6b. This gene was deleted in the process from dysplasia to tumor. It has also been reported that down-regulation of ANGPT1 was closely related to tumor angiogenesis and vessel maturation [16], and a high level of ANGPT1 has been associated with aggressive tumor behavior in OSCC [4].

3.5 Classification accuracy of the selected regions

To compare the classification accuracy, we calculated the out of bag (OOB) error rates using the mean values of the expression intensities in each region.

To compare the discriminative accuracy, we used the mean values of the regions selected by the proposed method and random method. We calculated OOB error rates using the number of CNV patterns, which ranged from 2 to 10. To explore the distribution of OOB error rates, we used 100 repeatedly extracted regions for each size. Figure 7 shows the distributions of OOB error rates, which were significantly different ($p < 10e\text{-}16$) regardless of the number of regions. The inter-quartile ranges of OOB error rates were narrower in the proposed method compared to the random method and the mean OOB error rates were significantly low. This result indicates that the region

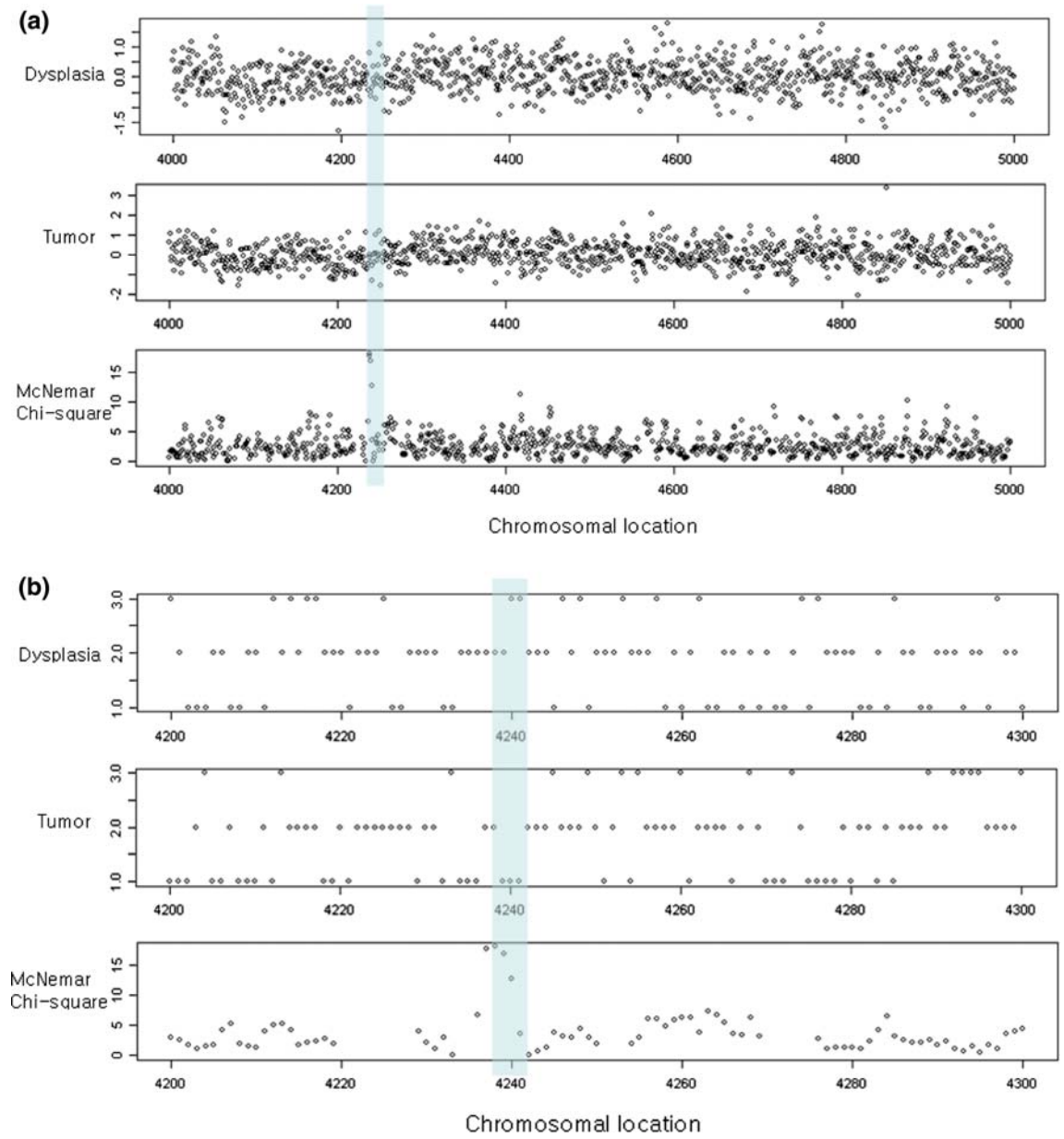

Fig. 5 CNV patterns of chromosome 2p24. The *horizontal* axis represents chromosomal locations and the *vertical* axes represent raw intensity (**a**), discretized intensity (**b**) and McNemar Chi-square statistic

selected by the proposed method could be used to accurately classify tumor and dysplasia.

However, the average OOB error rates were about 40% even in the proposed method. Based on this observation, it is probable that tumor and dysplasia were not strongly heterogeneous. Therefore, the classification accuracy can be highly improved if the proposed method is applied to the data set, which includes clearly heterogeneous experimental groups, for example, tumor and normal.

4 Discussion

The main issue in aCGH analysis is to segment the sequence of log ratios along the chromosome into regions of amplification, deletion, or no change [13]. A previous study indicated that the correlation of neighboring genomic intervals should be considered in the structural analysis of aCGH datasets [17], and the neighboring probes correlated with each other [3]. These findings indicate that the significant region would be more reliable for classification of

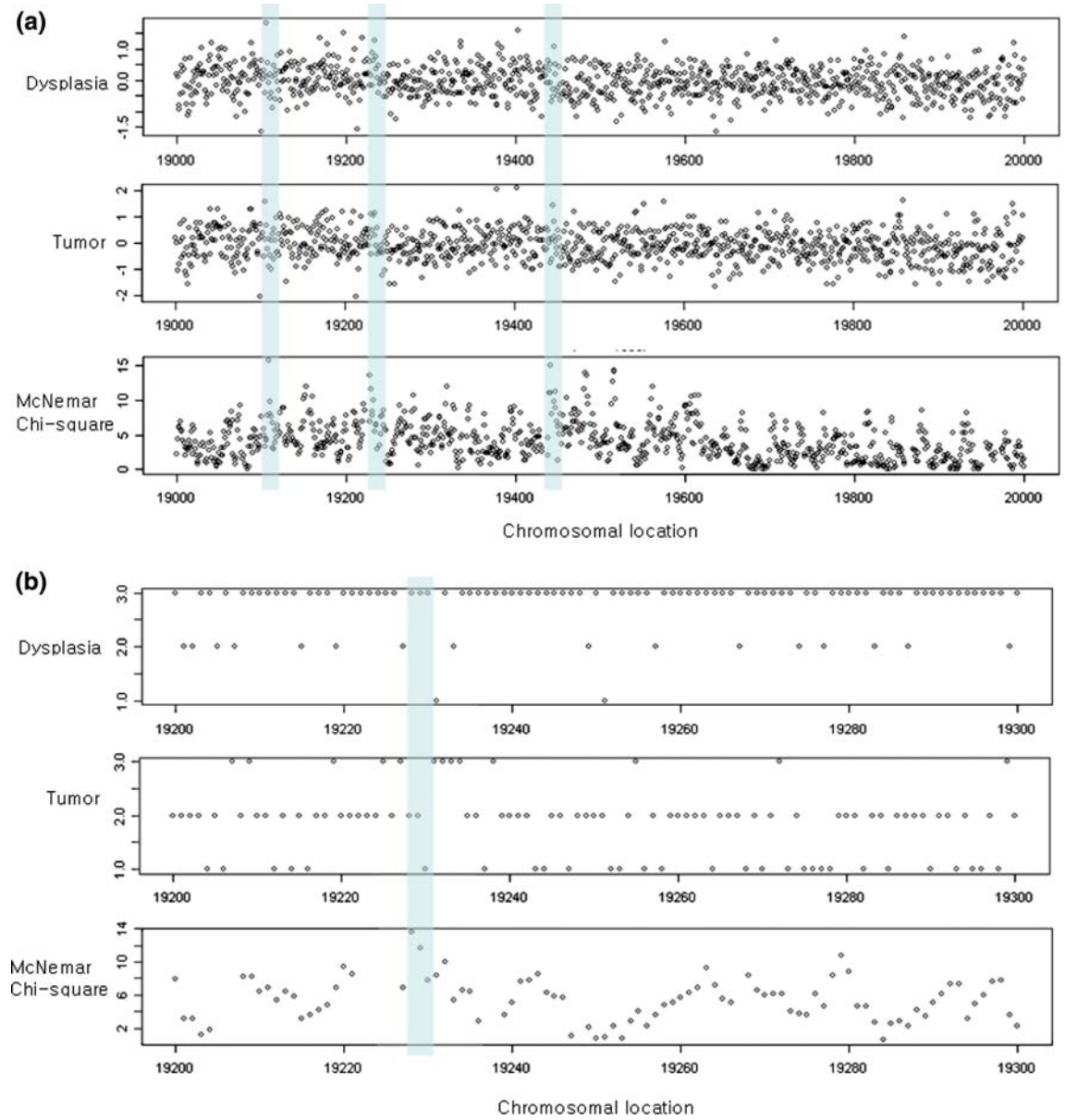

Fig. 6 CNV patterns of chromosome 8q22-q24. The *horizontal* axis represents chromosomal locations and the *vertical* axes represent raw intensity (**a**), discretized intensity (**b**) for different two experimental groups and McNemar Chi-square statistic

experimental groups, which includes the correlated neighboring probes.

In many aCGH studies, we are interested not only in the copy number variation in an experimental group but also in the comparison of groups of samples, i.e., whether there is a consistent change across the different experimental groups. Therefore, the proposed method could detect regions with significant genetic variations for comparison between different experimental groups. In addition, since we used two data sets for this study and combined these data sets before detection of significant genomic regions, we discretized the continuous CNV to minimize the bias between two data sets derived from different time periods.

The McNemar test has been commonly used to detect differentially expressed genes from the paired and discretized microarray data set. Although the general McNemar test has been applied to a gene (probe) independently in a microarray data set for significant gene selection, the proposed method, shifted McNemar test, was used for detecting significant regions, not probes, from aCGH data. In this novel extended MeNemar test, the neighboring genomic intervals, region, are taken into consideration. This method uses discretized aCGH data to identify regions that are significantly aberrant across the paired samples. Therefore, the CNV patterns of the selected regions were shown to be changed between paired samples.

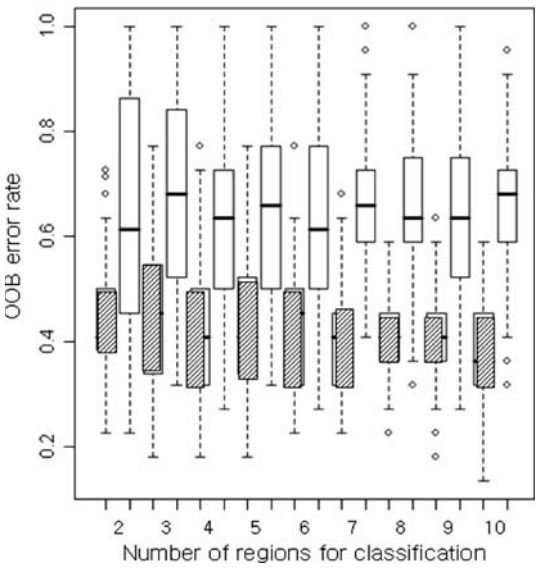

Fig. 7 Comparison of OOB error rates of the regions selected by the proposed method and random, using box plot. The *vertical* and *horizontal* axes represent OOB error rates and the number of regions used for classification, respectively. The *gray* and *white* regions represent the proposed method and random method, respectively

We illustrated the performance of the proposed method using inter-correlation within the selected regions and OOB error rates. The significant regions selected by the proposed method were strongly homogeneous, and high classification accuracies were achieved with these regions.

In conclusion, this method might be useful for identifying new candidate genes that neighbor known genes because the proposed method detects significant chromosomal regions and not independent probes. Also, the proposed method could be more useful in analyzing several data sets derived from different conditions. The candidate genes, which are selected by the proposed method, could be further analyzed based on known functionality and possible links to carcinogenesis.

Acknowledgments This work was supported by Priority Research Centers Program through the National Research Foundation of Korea(NRF), funded by the Ministry of Education, Science and Technology (2009-0094030).

References

1. Bennett BM, Underwood RE (1970) On McNemar's test for the 2 × 2 table and its power function. Biometrics 26:339–343
2. Ben-Yaacov E, Eldar YC (2008) A fast and flexible method for the segmentation of aCGH data. Bioinformatics 24(16):i139–i145
3. Chen HI, Hsu FH, Jiang Y, Tsai MH, Yang PC, Meltzer PS, Chuang EY, Chen Y (2008) A probe-density-based analysis method for array CGH data: simulation, normalization and centralization. Bioinformatics 24(16):1749–1756
4. Chien CY, Su CY, Chuang HC, Fang FM, Huang HY, Chen CM, Chen CH, Huang CC (2008) Angiopoietin-1 and -2 expression in recurrent squamous cell carcinoma of the oral cavity. J Surg Oncol 97(3):273–277
5. Eilers PH, de Menezes RX (2005) Quantile smoothing of array CGH data. Bioinformatics 21(7):1146–1153
6. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME et al (2006) Copy number variation: new insights in genome diversity. Genome Res 16(8):949–961
7. Garnis C, Campbell J, Zhang L, Rosin MP, Lam WL (2004) OCGR array: an oral cancer genomic regional array for comparative genomic hybridization analysis. Oral Oncol 40(5):511–519
8. Huang Q, Raya A, DeJesus P, Chao SH, Quon KC, Caldwell JS, Chanda SK, Izpisua-Belmonte JC, Schultz PG (2004) Identification of p53 regulators by genome-wide functional analysis. Proc Natl Acad Sci USA 101(10):3456–3461
9. Huang J, Gusnanto A, O'Sullivan K, Staaf J, Borg A, Pawitan Y (2007) Robust smooth segmentation approach for array CGH data analysis. Bioinformatics 23(18):2463–2469
10. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 20(18):3413–3422
11. Katoh M (2002) Molecular cloning and characterization of OSR1 on human chromosome 2p24. Int J Mol Med 10(2):221–225
12. Kim KY, Ki DH, Jeung HC, Chung HC, Rha SY (2008) Improving the prediction accuracy in classification using the combined data sets by ranks of gene expressions. BMC Bioinformatics 9:283
13. Lai W, Choudhary V, Park PJ (2008) CGHweb: a tool for comparing DNA copy number segmentations from multiple algorithms. Bioinformatics 24(7):1014–1015
14. Lengauer C, Issa JP (1998) The role of epigenetics in cancer. DNA methylation, imprinting, the epigenetics of cancer—an American Association for Cancer Research Special Conference. Las Croabas, Puerto Rico, 12–16 1997 December. Mol Med Today 4(3):102–103
15. Li Y, Zhu J (2007) Analysis of array CGH data for cancer studies using fused quantile regression. Bioinformatics 23(18):2470–2476
16. Li C, Feng HC, Chen JC, Song YF (2005) Expression and significance of angiopoietin-1 and angiopoietin-2 in oral squamous cell cacinoma. Ai Zheng 24(11):1388–1393
17. Liu J, Mohammed J, Carter J, Ranka S, Kahveci T, Baudis M (2006) Distance-based clustering of CGH data. Bioinformatics 22(16):1971–1978
18. Liu CJ, Lin SC, Chen YJ, Chang KM, Chang KW (2006) Array-comparative genomic hybridization to detect genome wide changes in microdissected primary and metastatic oral squamous cell carcinomas. Mol Carcinog 45(10):721–731
19. Liu J, Ranka S, Kahveci T (2008) Classification and feature selection algorithms for multi-class CGH data. Bioinformatics 24(13):i86–i95
20. McNemar Q (1947) Note on the sampling error of the differences between correlated proportions or percentages. Psychometrika 12:53–157
21. Nakaya K, Yamagata HD, Arita N, Nakashiro KI, Nose M, Miki T, Hamakawa H (2007) Identification of homozygous deletions of tumor suppressor gene FAT in oral cancer using CGH-array. Oncogene 26(36):5300–5308
22. O'Regan EM, Toner ME, Smyth PC, Finn SP, Timon C, Cahill S, Flavin R, O'Leary JJ, Sheils O (2006) Distinct array comparative genomic hybridization profiles in oral squamous cell carcinoma occurring in young patients. Head Neck 28(4):330–338
23. Pinkel D, Albertson DG (2005) Comparative genomic hybridization. Annu Rev Genomics Hum Genet 6:331–354
24. Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, Ng R, Murphy KP (2006) Integrating copy number polymorphisms into

array CGH analysis using a robust HMM. Bioinformatics 22(14): e431–e439
25. Squire JA, Bayani J, Luk C, Unwin L, Tokunaga J, MacMillan C, Irish J, Brown D, Gullane P, Kamel-Reid S (2002) Molecular cytogenetic analysis of head and neck squamous cell carcinoma: by comparative genomic hybridization, spectral karyotyping, and expression array analysis. Head Neck 24(9):874–887
26. Suzuki E, Imoto I, Pimkhaokham A, Nakagawa T, Kamata N, Kozaki KI, Amagasa T, Inazawa J (2007) PRTFDC1, a possible tumor-suppressor gene, is frequently silenced in oral squamous-cell carcinomas by aberrant promoter hypermethylation. Oncogene 26(57):7921–7932
27. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23(6):657–663
28. Watabe K, Ito A, Asada H, Endo Y, Kobayashi T, Nakamoto K, Itami S, Takao S, Shinomura Y, Aikou T et al (2001) Structure, expression and chromosome mapping of MLZE, a novel gene which is preferentially expressed in metastatic melanoma cells. Jpn J Cancer Res 92(2):140–151