[Frontiers in Bioscience 11, 1311-1322, May 1, 2006]

LS-CAP: an algorithm for identifying cytogenetic aberrations in hepatocellular carcinoma using microarray data

Xianmin He [1], Qing Wei [2], Meiqian Sun [2], Xuping Fu [2], Sichang Fan [1], and Yao Li [2]

[1] Department of Health Statistics, Second Military Medical University, Shanghai, 200433, P.R China, [2] State Key Laboratory of Genetic Engineering, Institute of Genetics, School of Life Sciences, Fudan University, Shanghai,200433, P.R China

TABLE OF CONTENTS

1. ABSTRACT

Biological techniques such as Array-Comparative genomic hybridization (CGH), fluorescent *in situ* hybridization (FISH) and affymetrix single nucleotide pleomorphism (SNP) array have been used to detect cytogenetic aberrations. However, on genomic scale, these techniques are labor intensive and time consuming. Comparative genomic microarray analysis (CGMA) has been used to identify cytogenetic changes in hepatocellular carcinoma (HCC) using gene expression microarray data. However, CGMA algorithm can not give precise localization of aberrations, fails to identify small cytogenetic changes, and exhibits false negatives and positives. Locally un-weighted smoothing cytogenetic aberrations prediction (LS-CAP) based on local smoothing and binomial distribution can be expected to address these problems. LS-CAP algorithm was built and used on HCC microarray profiles. Eighteen cytogenetic abnormalities were identified, among them 5 were reported previously, and 12 were proven by CGH studies. LS-CAP effectively reduced the false negatives and positives, and precisely located small fragments with cytogenetic aberrations.

2. INTRODUCTION

Amplification and deletion of genetic regions frequently contribute to tumorigenesis. Characterization of DNA copy-number changes is important for understanding pathogenesis mechanism and diagnosis of cancer (1). Cytogenetic profiling techniques such as CGH have been used to detect cytogenetic abnormalities on a whole genome. However, the application of CGH is restricted by its mapping resolution, while higher resolution techniques, such as Array-CGH, FISH and affymetrix SNP, are labor-intensive and time-consumptive on a genome scale (2-10).

Besides directly biological detecting techniques mentioned, cytogenetic abnormalities can also be identified indirectly through predicting regional gene expression biases using microarray profiles, because these regional biases are mainly caused by chromosomal gain or loss. In genetic regions with amplification or deletion, the copy-numbers of genes would increase or decrease. Correspondingly, quantity of genes hybridized on the slide also would increase or decrease when competing with the reference, and the genes would be up or down-regulated.

Compared with the average level of differentially expression genes (DEs) on genomic scale, which is recognized as being caused by random factors, cytogenetically aberrational regions have higher proportions of DEs. The algorithms that can identify the regional gene expression biases are appropriate for identifying cytogenetic aberrations using microarray data, and CGMA is such a technique. Through analyzing gene expression profiles, CGMA had been used to predict cytogenetic abnormalities in HCC and proven to be similar to CGH in identifying cytogenetic aberrations (2, 11). CGMA predicted chromosomal abnormalities via organizing gene expression data by genetic mapping location and scanning for regional gene expression biases, in which a disproportionate number of genes change in the same relative direction. Despite its advantage in identifying genetic abnormalities indirectly, CGMA algorithm had several disadvantages. First, it could not precisely locate the regional gene expression biases; second, it might fail to identify some relatively small pieces of loss or gain, and would produce false negatives if the numbers of down-regulated and up-regulated genes happen to be the same or similar in different cytogenetic abnormalities regions within the same chromosomal arm; and third, it might produce false positives as the statistic was just for testing the difference of proportions.

Local smoothing combined with binomial distribution theory expects to solve the problems mentioned above. Local smoothing is a useful method for curve fitting: the original data points are always noisy and inaccurate, so, during smoothing process, each point is estimated by neighboring data points defined within the span (the size of slide-window), and so the smoothed values are more accurate and resistant to outliers. Methods for estimating the smoothing values include averaging, weighted-averaging, linear polynomial regression, and quadratic polynomial regression. The main advantages of smoothing methods are its robust estimation for data points and flexibility of model, so it has wide applications in curve fitting (12, 13). Multiple span moving binomial test based on smoothing theory has recently been used to identify genetic abnormalities, such as IR-CGMA algorithm. (14).

We developed a new approach, which was called locally un-weighted smoothing cytogenetic aberrations prediction (LS-CAP), for predicting cytogenetic abnormalities by moving the slide-window smoothly along the chromosome and testing the difference between the DEs rates of locally chromosomal region and corresponding population rates on genomic scale. LS-CAP approach was performed on 104 HCC microarray profiles, and the results showed that, compared with CGMA, LS-CAP approach had advantages in flexibility in choosing slide-window size and standard for identifying DEs, sensitivity in predicting and accuracy in locating of relatively small pieces of cytogennetic changes with independent statistic for down and up-regulated DEs, and effectiveness in reducing false negatives and positives of prediction. The comparison leaded us to conclude that LS-CAP could be a powerful tool for identifying cytogenetic aberrations using microarray data, and might be a useful alternative to Array-CGH and FISH techniques.

3. ALGORITHMS

CGMA algorithm predicts frequent cytogenetic aberrations using microarray data through setting up the following statistic for each chromosomal arm.

$$Z = \frac{2x - n}{\sqrt{n}} \quad (1)$$

in which, x denotes the number of up (or down) regulated genes on the chromosomal arm; n denotes the sum of up and down-regulated genes on the arm. The cytogenentic aberrations in one chromosomal arm have statistical meaning with the Z score being greater than 1.96 or less than -1.96. Statistically, if we take x as the number of up-regulated genes, the arm is recognized having amplifications in the significant level of 0.05 when $Z \geq 1.96$; on the contrary, if x is the number of down-regulated genes, the arm is considered having deletions in the significant level of 0.05 when $Z \leq -1.96$.

CGMA algorithm is simple and useful except two aspects of disadvantages: (1) It sets up Z statistic for testing the difference of proportions between up and down-regulated genes in all DEs on the chromosomal arm, not the difference of DEs rates between sample and population, then two problems may be rendered. First, if amplification and deletion both occurred on one same chromosomal arm, however, in different genetic regions, it is probable that, despite the biological meaning of the amplification and deletion, Z is not statistically significant, and CGMA can't identify these cytogentical abnormalities. Second, if rates of up and down-regulated genes on one chromosomal arm are all low and there are no essential amplifications and deletions, but the difference between the proportion of up and down-regulated genes in all DEs on the chromosomal arm might be statistically significant ($|Z| \geq 1.96$). Consequently, a conclusion will be drawn in light of CGMA that the corresponding chromosomal arm having amplifications or deletions. (2) As CGMA algorithm is based on one whole chromosomal arm, it is difficult to identify the small pieces of gains or losses and locate cytogenetic aberrations precisely.

In our research, an elaborate algorithm named LS-CAP to identify the cytogenetic abnormalities was developed, which kept the advantages of CGMA: it set up independent statistic for up and down regulated genes, and the statistic was used to test the difference of rates between sample and population, not the proportions of up and down-regulated genes in all DEs; Additionally, LS-CAP addressed the issues over CGMA algorithm: local smoothing method was used to identify relatively small pieces of regional gene expression biases and locate these aberrations precisely.

Detailed procedures of LS-CAP algorithm are as following:

1. All genes are mapped and sorted according to their locations on chromosomes.

2. A standard for identifying DEs is built. Many methods have been used to identify DEs, from the simple forms of fold change, t-test, regularized t-test to more complex ones, such as SAM and random variance model (15-23). Fold change method was used in this paper for its simplicity, and different standards for fold change (1.8, 1.9, 2.0, 2.1, 2.2) were compared.

3. Compute population rates of down and up-regulated genes on genomic scale π_d and π_u.

$$\pi_d = {}^{n_d}\!/_{N}, \quad \pi_u = {}^{n_u}\!/_{N} \qquad (2)$$

In which, n_d and n_u denote the total numbers of down-regulated genes and up-regulated genes of all chromosomes in the study, and the N denotes the total number of genes measured in the study. Population rates, π_d and π_u, are defined as the rates of differentially expressed genes on the genome scale, which are recognized as differential expression caused just by random factors, and act as standard for comparing with sample rates.

4. For one chromosomal arm j,

I. Compute sample down and up-regulated rates p_{di} and p_{ui} for each position in the chromosomal arm centered by gene i within the slide-window. Here, the local region used to calculate p_{di} and p_{ui} on chromosome is called slide-window, and the number of genes included in the slide-window is called the size of slide-window or span, denoted as L, and $N_{(j)}$ refers the number of genes on the chromosomal arm j.

$$p_{ui} = \begin{cases} U(i-0.5L, i+0.5L)/_{L+1}, & 0.5L \le i \le N_{(j)} - 0.5L \\ U(0, i+0.5L)/_{0.5L+i}, & 0 < i < 0.5L \\ U(i-0.5L, N_{(j)})/_{N_{(j)}+0.5L-i}, & N_{(j)} - 0.5L < i < N_{(j)} \end{cases} \qquad (3)$$

In which, $U(\ )$ refer the number of up-regulated genes in the slide-window centered by gene i on chromosomal arm j. Formula for p_{di} is similar, with $_{U(\ )}$ replaced by $_{D(\ )}$.

II. Compute the statistic Z_{di} and Z_{ui}. Binomial theory was used to test the rate difference between sample and population with null and alternative hypothesis being $p_i = \pi$ and $p_i \neq \pi$ respectively. The restrain, np and n(1-p) are both greater than 5, were satisfied in the paper.

$$Z_{ui} = \frac{p_{ui} - \pi_u}{\sqrt{\pi_u(1-\pi_u)/n_i}} \qquad (4)$$

$$Z_{di} = \frac{p_{di} - \pi_d}{\sqrt{\pi_d(1-\pi_d)/n_i}} \qquad (5)$$

We denote Z_{ui} as the Z score statistic for up-regulated genes, where, p_{ui} is the proportion of up-regulated genes for the chromosomal region centered by gene i, π_u is the corresponding population rate caused just by chance, and n_i denotes the number of genes in the slide-window. Similarly, Equation (5) generates the LS-CAP statistic Z_{di} for down-regulated genes.

Moving gene i one by one, a pair of Z values, Z_{ui} and Z_{di}, for each position on chromosomes are acquired. When span is equal or larger than the number of genes on a whole chromosomal arm, the paired Z values on that arm would be all the same, then we defined the algorithm as LSW-CAP (locally un-weighted smoothing cytogenetic aberrations prediction based on whole chromosomal arm). Otherwise, when span is smaller than the number of genes on a whole chromosomal arm, we define the algorithm as LSS-CAP (locally un-weighted smoothing cytogenetic aberrations prediction based on slide-window).

5. Set up statistically significant standard as $Z = 1.96$ (P=0.05), and regions with $Z \ge 1.96$ are identified as regional gene expression biases, for example, the region will be considered having amplifications if $Z_{ui} \ge 1.96$, because the difference between rates of up-regulated genes in the region and corresponding population rate is beyond the range caused just by chance. Similarly, the region will be considered having deletions if $Z_{di} \ge 1.96$.

4. IMPLEMENTION

4.1. Data

Primary HCC is a common cancer and the fourth leading cause of death from caner wide (24). Here we applied LS-CGMA approach on HCC profiles dataset. Normalized, log-transformed gene-expression data for 104 HCC samples and 76 corresponding non-cancerous liver gene expression profiles were obtained from the Stanford Microarray Database (http://genome-www5.stanford.edu) (25, 26).

Our aim was identifying cytogenetic aberrations of HCC tissue relatively to correspondingly normal liver tissue, however, indirect design was used in the experiment in which the common reference (pooled cell-line) was implemented. To compare gene expression levels of HCC with those of surrounding non-cancerous tissue, the HCC gene expression data were mathematically transformed into levels relative to the corresponding normal tissue, in stead of the original reference of pooled cell-line. Since the reference sample in the two experiments were same, the

resulting new logarithm ratio for each gene, tumor verse normal (T/N), was estimated.

$$\log_2(R_i) = \log_2(T_i/N_i) = \log_2(T_i/U_i) - \log_2(N_i/U_i) \quad (6)$$

in which, $\log_2(T_i/U_i)$ and $\log_2(N_i/U_i)$ were log-transformed HCC and non-cancerous tissue ratios respectively, with pooled cell-line as common reference. If an HCC sample did not have a corresponding non-cancerous sample, the global mean of the non-cancerous tissue gene expression ratios were used.

The results of mapping genes to get their genetic locations were achieved through SOURCE tool (http://genome-www5.stanford.edu/cgi-bin/source/sourceBatchSearch) with identifiers as GeneBank Accession, CloneID and UniGene Name. Totally, 17238 genes had been located (with a few genes locating at multiple chromosomal regions). For the number of genes in Yq (only 14) and Yp (only 6) were insufficient for LS-CAP approach, these two chromosomal arms were excluded in subsequent analysis. The actual number of genes in analysis was 17218.

4.2. LSW-CAP analysis and results

We started from LSW-CAP algorithm, the special and simplest form of LS-CAP. Like CGMA, the LSW-CAP algorithm was also based upon each whole chromosomal arm, and a pair of statistics for down and up-regulation, $Z_{d(j)}$ and $Z_{u(j)}$, were obtained for each arm j. Under LSW-CAP analysis on 104 HCC samples**,** Gene expression biases for each chromosomal arm were shown in Figure 1 and 2. To estimate the expression biases comprehensively from multiple samples, the following two procedures were processed: (1) Calculated average log-expression ratio for each gene on 104 samples, and performed LSW-CAP analysis using the average, the Z-values based on the average were shown as the most right columns in Figure 1 and 2, labeled as 'AVERAGE'. (2) Calculated the proportion of $Z \geq 1.96$ in 104 samples for each arm, the results were shown as 'proportion ($Z \geq 1.96$)' columns, and the arms were labeled with red if significant gene expression biases occurred for at least 1/3 of all the samples.

The larger the sample rate of DEs for each chromosomal arm was, the higher the corresponding Z statistic was. Difference between the sample and population rates was considered statistically significant when it was beyond the range of variation caused by random factors, and the corresponding chromosomal arm was concluded to be genetic changed, gain or loss of chromosomal fragment occurred in light of which statistic being lager than 1.96, $Z_{d(j)}$ or $Z_{u(j)}$. Arms in which at least 1/3 of 104 HCC samples were statistically significant included -4q (lost in 68.27% of tumor samples), -8p (50.96%), -13q (43.27%), -9p (34.62%), -6q (33.65%), -12p (33.65%), -17p (33.65%), +1q (gained in 80.77% of samples), +6p (49.04%), +8q (49.04%), +17q (36.54%) and +20q (34.62%). 4q and 1q were typical down-regulated and up-regulated biases, respectively. The bar graphs at the bottom of Figure 1 and 2 gave the detailed expression profile for all genes on the arm 4q and 1q, respectively. With 2.0 for fold change, the proportions of down regulated genes ($\log_2(ratio) \leq -1$) on arm 4q (p=0.0742, Z=10.2965) and up regulated genes ($\log_2(ratio) \geq 1$) on arm 1q (p=0.0855, Z=5.3004) were obviously higher than those of the average level on genomic scale, which were 0.0395 and 0.0203, respectively. Therefore, LSW-CAP was specially suited for exploring genetic abnormalities from the standpoint of the whole chromosomal arm, and had distinctive statistic compared with CGMA.

Included in the set of HCC samples were several cases in which multiple tumor nodules were removed from the same patient. In some of the cases such as HK63, HK64 and HK66, different nodules from the same patient had related gene expression profiles, whereas in other cases such as HK65, HK85, tumors from the same patient had distinctive profiles. Clustering and correlation analysis based on Z statistic have been performed and the resulted dendrograms are shown in Figure 2 and Figure 3.

The dendrogram (Figure 2) of hierarchical clustering confirmed the similarity relationship between gene expression profiles, in which, samples from patients HK65 (HK65.1, HK65.2 and HK65.4) and HK85 (HK85.1 and HK85.2) were each separated by other samples, whereas the samples from other patients were adjacently distributed. The gene expression patterns observed in tumor nodules HK65.2 and HK65.4 were more similar to each other than either was to the pattern observed in HK65.1, the same relationship were also found in LSW-CAP-predicted cytogenetic profiles. Results of p53 immunohistochemical staining and Southern analysis suggested that HK65.2 and HK65.4 arose from the same clone, whereas HK65.1 was distinctly different from that of HK65.2 and HK65.4. LSW-CAP identified 8 common genetic aberrations in all the three samples from HK67. In addition to those (+19q, -15q, -16p, -19p) identified by CGH and CGMA, LSW-CAP also identified aberrations in +1q, +6p, -8p and -16q. Besides, five aberrations (+5q, +2q, -4q, -12q, -21q) were found in HK67.2 and HK67.3 though not in HK67.1, among them only +5q and +2q were identified by CGMA. Although the correlation coefficients of the tumors from patient HK67 were all high, the coefficient between HK67.2 and HK67.3 was higher than that with HK67.1. Taken together, the LSW-CAP results confirmed the hypothesis that HK67.1 was the primary tumor nodule and HK67.2, HK67.3 tumor nodules probably were divergent HK67.1 subclones, and additional distinct cytogenetic changes had occurred for HK67.2, HK67.3 nodules during tumor progression. Tumor nodules from patient HK85 showed different expression profiles and distinct HBV integration sites. Similarly, the tumors from patient HK85 also showed distinct LSW-CAP-predicted cytogenetic profiles, reflecting the independent transforming mechanism (11, 25). All these results also indicated that Z statistic in the LSW-CAP algorithm was capable of extracting the information of microarray profiles correctly, the analysis results based on Z statistic of LSW-CAP algorithm were accordant with results of the direct analysis on expression profiles analysis and related biological detections.

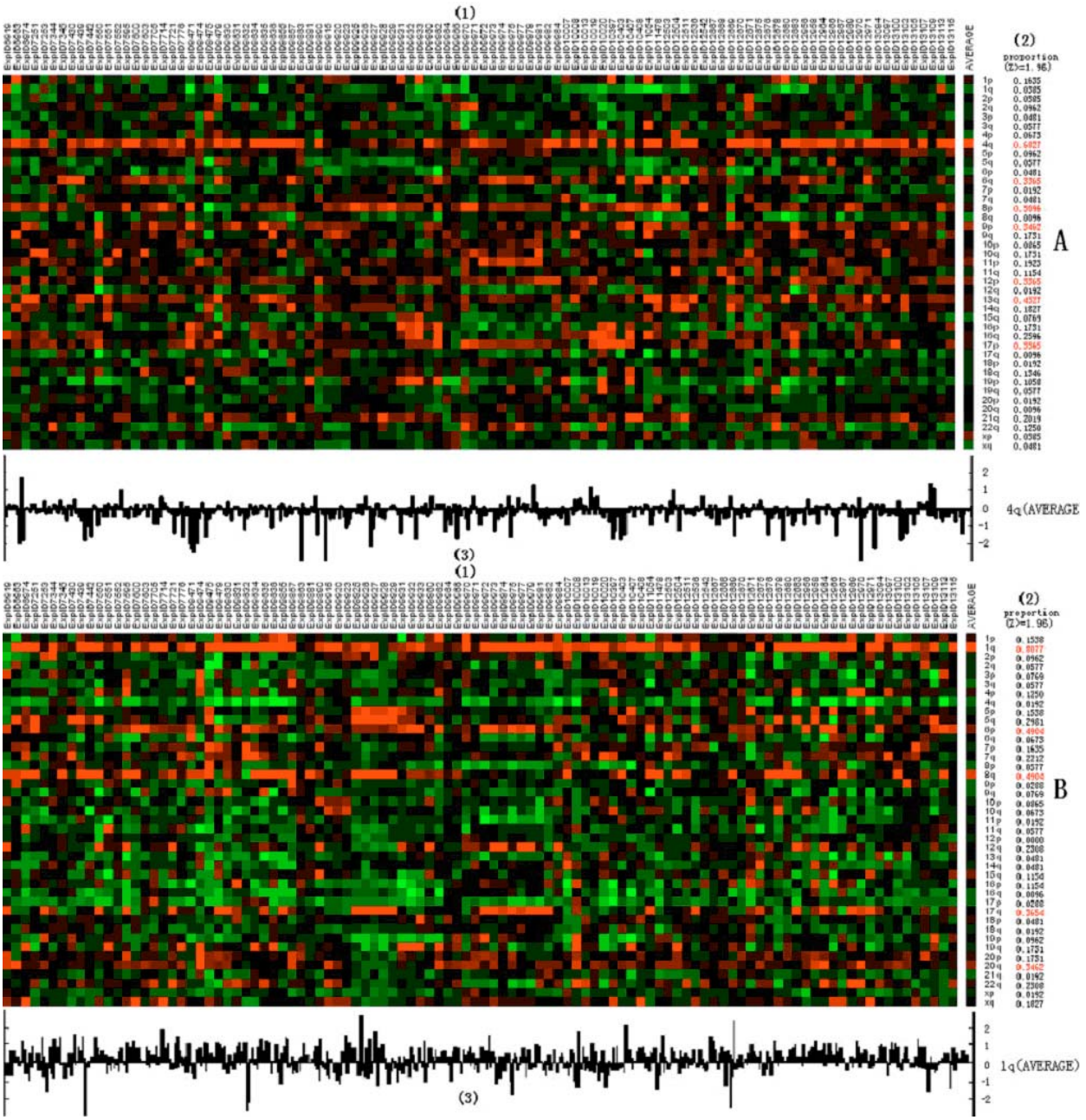

Figure 1. Z statistics of down and up regulated genes for each chromosomal arm in 104 HCC samples (subplot A and B). For subplot A, (1) Each block of the Z-statistic profile (left-top) gave the graphic information of down regulated genes on corresponding sample and chromosomal arm. Red blocks indicated higher proportion of down regulated genes compared with the average level, black ones indicated non-significantly different proportion, and green ones indicated lower proportion of down regulated genes; the column labeled as 'AVERAGE' indicated the average level of down regulated genes proportion among 104 HCC samples. (2) Each value in column 'proportion ($Z \geq 1.96$)' (right-top) indicated the proportion of $Z \geq 1.96$ in 104 samples for the arm, the values were labeled with red if the proportions were greater than 1/3. (3) Bar graph (bottom) indicated detailed gene expression level ($\log_2(ratio)$) for all genes on chromosomal arm 4q according to their genetic mapping locations. The notes for subplot B are same.

4.3. LSS-CAP analysis and results

Two parameters needed to be set up in LSS-CAP algorithm, the slide-window and the standard for discriminating differentially expressed genes. The size of slide-window was a crucial parameter for LSS-CAP, which affected the outcome of predicting. Smaller the span was, more sensitive the variation of predicted Z-curve was, and higher the false positive rate caused by a random factor

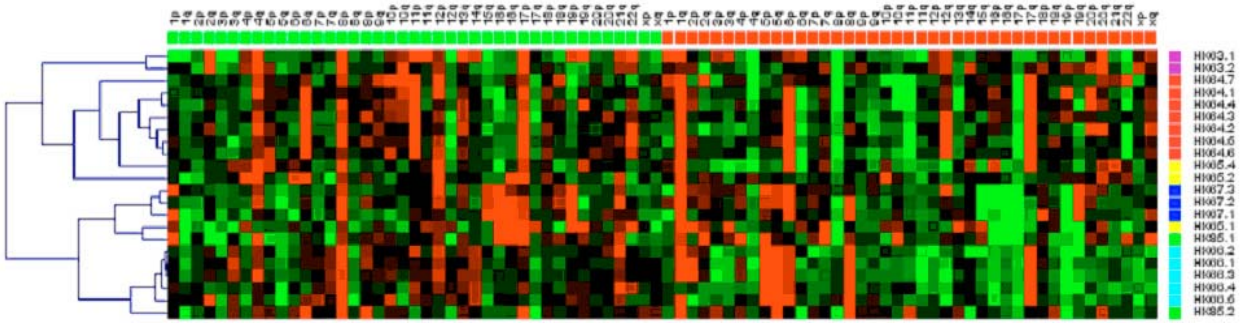

Figure 2. Dendrogram of hierarchical clustering for partial correlated samples using Z statistic. Dendrogram of hierarchical clustering was constructed with Pearson correlation and average linkage, and variables used here were Z statistic of down regulated genes proportions (left half with blue header) and up regulated genes proportions (right half with red header) for all arms. Additionally, samples from different patients were labeled with headers of different colors (second column from the most right). Samples from patients HK65 and HK85 were each separated by other samples, whereas the samples from other patients were adjacently distributed.

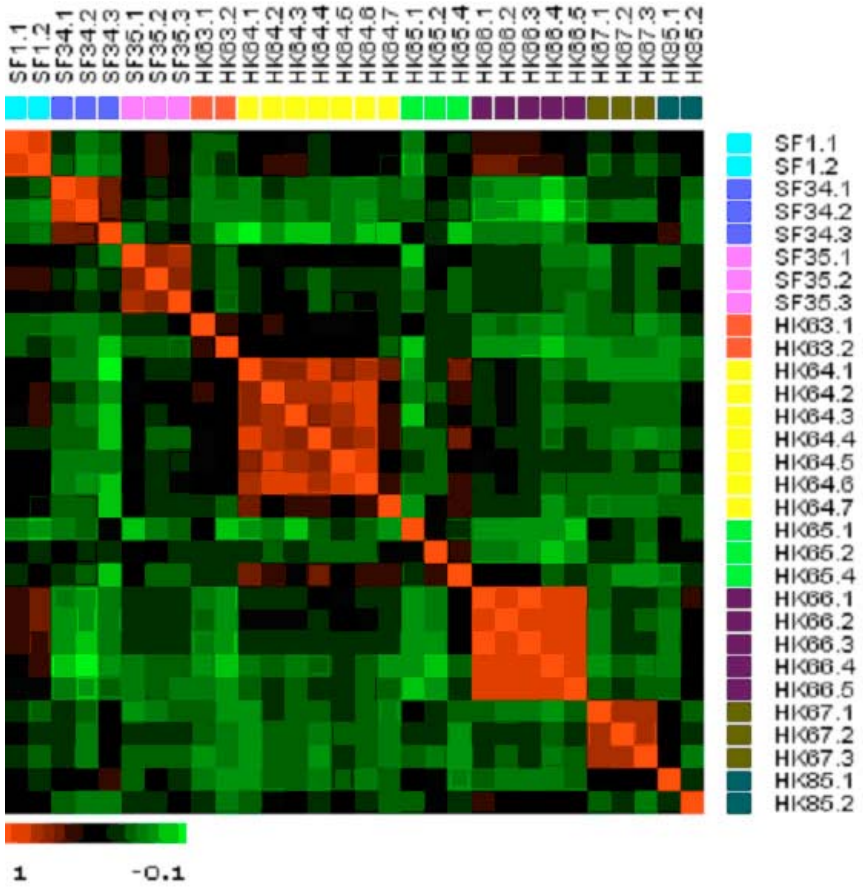

Figure 3. Pairwise Pearson correlation coefficients for partial correlated samples using Z statistic. The pairwise Pearson correlation coefficients were in the range [-0.1, 1], and the legend could be seen at the left-bottom. Samples from same patients were labeled with same color for the headers, and the color of cross point indicated the strength of correlation and its direction.

was. On the contrary, bigger the span was more stable the variation of predicted Z-curve was, and more robust against the influence of random factors the algorithm was. Different sizes of span between 100 and 300 had been tested in this study according to biological knowledge and the results suggested that the span between 150 and 250 was appropriate for our data. Additionally, the calculating method for the chromosomal arm ends (within 0.5span from the end) was detailed in ALGORITHMS. As to the standard for discriminating DEs, different fold changes

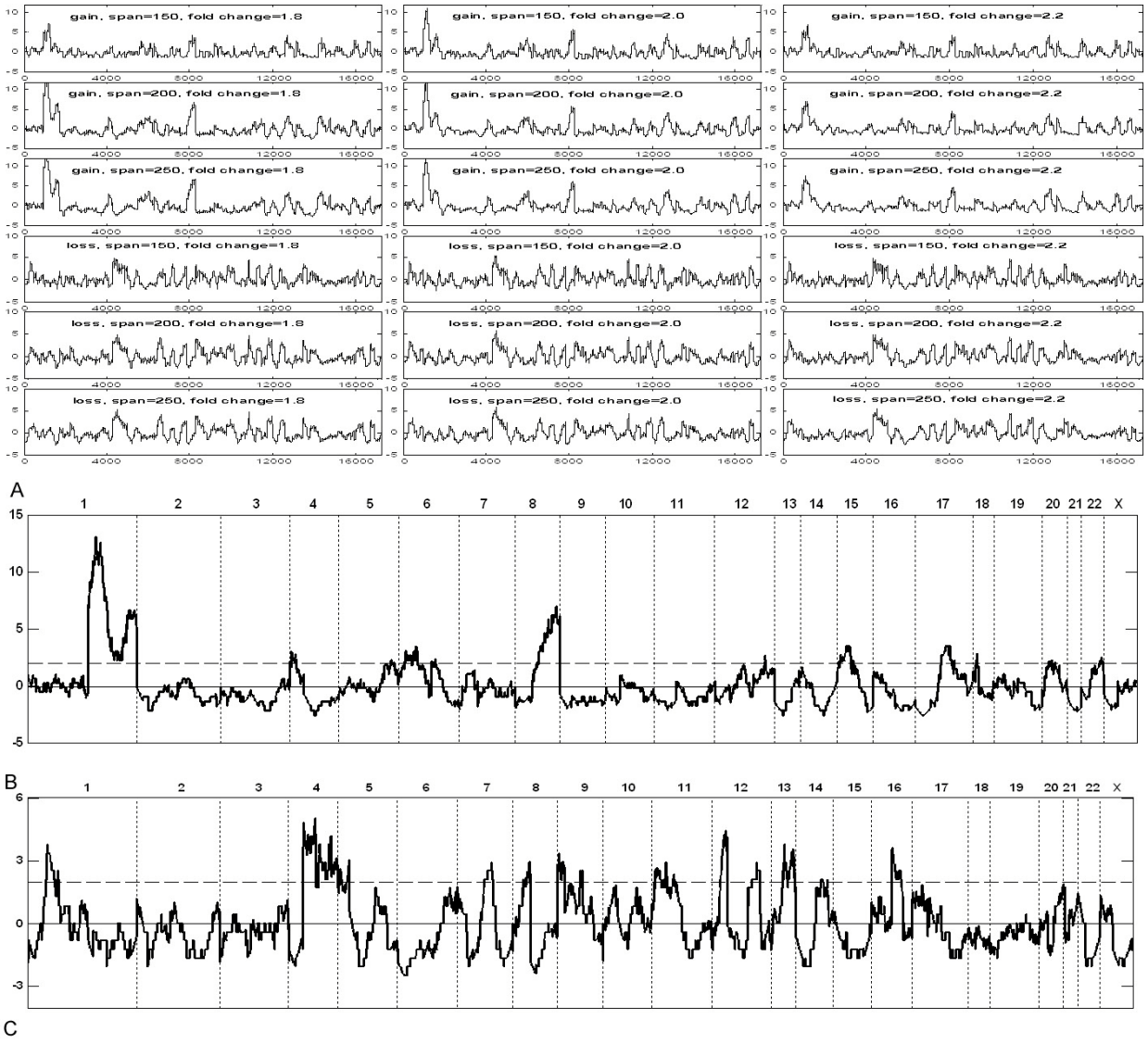

Figure 4. (a) Z statistic distributions along the whole genome with different parameters combinations. For rows, span were 150, 200 and 250, respectively; for columns, fold change were 1.9, 2.0, 2.1, respectively. The pieces with $Z \geq 1.96$ were recognized as statistically significant. (b) Zoom in of z statistic for up-regulation with span=200, fold=1.8. The horizontal axis was different chromosomes (separated with vertical dashed lines); the vertical axis was z statistic; horizontal dashed line correspond significance level (z=1.96). (c) Zoom in of z statistic for down-regulation with span=200, fold=2.2.

(1.8, 1.9, 2.0, 2.1, and 2.2) were tried, and the results were generated with a little difference. For example, biases 17p were identified upon the fold change of 1.8, but not upon 2.0. To try to make a proper and comprehensive interpretation of the Z statistic, we plotted Z statistic for different setups of span and standard for discriminating DEs as shown in Figure 4. Usually, common biases identified by different combinations of parameters were convincing.

Cytogenetic aberrations, such as amplification and deletion, could be located precisely using LSS-CAP (Figure 5). Compared with CGMA and LSW-CAP, LSS-CAP not only gave the arms in which the gene expression biases occurred, but also the extra details of each chromosomal arm. Except the biases occurred on arms 4q and 1q, on which biases occurred on the whole range of arms, all the other regional gene expression biases occurred on only part of the chromosomal arms, such as the newly found regional biases on chromosomes 15q(+), 12q(+), 22q(+), 9p(-), 12p(-) and 14q(-). 17 cytogenetically abnormal regions had been identified significantly with span being 200 and fold change being 2.0, including 7 gains and 10 losses; the precise localizations of cytogenetic aberrations which were showed to be similar to that of IR-CGMA algorithm were listed in table 1 (14).

Some of the regions have been proven to be associated with HCC in the previous studies. For example, gains on chromosomes 1q and 8q were showed to be involved in the genesis of HCC, while loss on chromosome 4q was linked to increased aggressiveness of established tumors (11, 25).

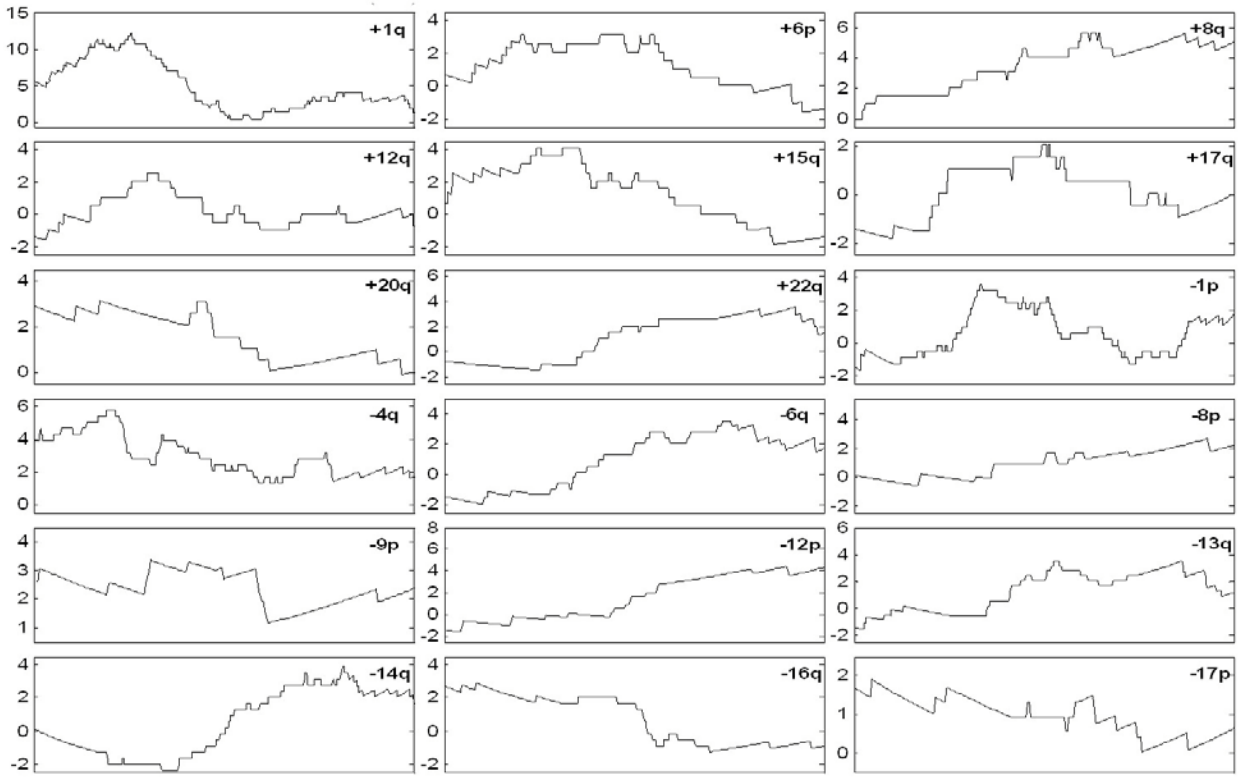

Figure 5. Details for presumably gene expression biases regions predicted by LSS-CAP with 200 genes for span and 2.0 for fold change. Every subplot gave the detailed information for each presumably loss or gain. The regions with Z being above 1.96 were significant. Except 1q and 4q, on which biases occurred on whole chromosomal scale, all the others occurred on only part of the chromosomal arm. The arm 17p was also displayed here because it could be found meaningful with other standards for fold change.

Table 1. Precise localization of cytogenetic aberrations

Loss		**Gain**	
1p	1p31-1p34	1q	1q12-1q44
4q	4q11-4q35	6p	6p11-6p22
6q	6q21-6q27	8q	8q13-8q24
8p	8p21-8p23	17q	17q12-17q23
13q	13q14-13q34	20q	20q11-20q13
16q	16q11-16q22	12q	12q13-12q21
17p	17p11-17p13	22q	22q12-22q13
9p	9p11-9p24	15q	15q11-15q24
12p	12p12-12p13		
14q	14q22-14q32		

Here, the simplest model of data was detailed. For more complicated model, data of two and above groups, we should discriminate whether each gene being differentially expressed or not with multiple testing approaches such as t test, F test, SAM and random variance model, and then perform LSS-CAP analysis.

5. DISCUSSIONS

The HCC cytogenetic aberrations resulting from different approaches including CGH, CGMA, LSW-CAP and LSS-CAP were compared (table 2). CGMA identified 13 cytogenetic aberrations totally, among them 10 regions had been proven by CGH. Among the 3 additional regions found by CGMA, one was proven by LSS-CAP, but the other two regions had been tested having no significant difference between sample and population rates with LSW-CAP and LSS-CAP approaches, and we concluded that they might be false positives of CGMA prediction. 1p and 6q, which had been proven to be losses by CGH, had not been identified by CGMA, whereas they have been successfully identified as losses by LSS-CAP, therefore CGMA might produce false negatives. Compared with LSS-CAP, CGMA can't identify the relatively small pieces of amplifications or deletions. LSW-CAP approach was also based on each whole chromosomal arm, but with distinctive testing method, therefore, though it couldn't identify the relatively small pieces of amplifications or deletions, just like CGMA, it successfully reduced the false positives as we expected: the 10 gains and losses detected by LSW-CAP were all proven by CGH. LSS-CAP approach has combined the local smoothing with binomial distribution theory for testing the difference of rates between sample and population, so it identified all the 12 cytogenetical changes detected by previous CGH studies, 11 regions proven by CGMA prediction, and 5 relatively small pieces of regional gene expression biases that have not been found in previous studies. LSS-CAP always gave more detailed information, such as the locations of small pieces of losses and gains occurred, though some of them

Table 2. Comparison of CGH, CGMA and LS-CAP on HCC

Methods	Gain										Loss									
	1q	6p	8q	17q	20q	12q	15q	22q	5q	19q	1p	4q	6q	8p	13p	16p	17p	9p	12p	14q
CGH	●	●	●	●	●	○	○	○	○	○	●	●	●	●	●	●	●	○	○	○
CGMA	●	●	●	●	●	●	○	○	●	●	○	●	○	●	●	●	●	○	○	○
LSW-CAP	●	●	●	●	●	○	○	○	○	○	○	●	●	●	●	○	●	○	○	○
LSS-CAP (200,2)	●	●	●	●	●	●	●	●	○	○	●	●	●	●	●	●	○	●	●	●

● identified, ○ unidentified.

Table 3. Differentially expressed genes proven to be correlated with HCC in previous studies and their expression levels in HCC

Gene Symbol	Locus	Fold change
GPC3,glypican 3	Xq26.1	18.66
AFP,alpha-fetoprotein precursor[2]	4q11-q13	3.23
FOXM1,forkhead box M1[2]	12p13	2.91
CCNA2,cyclin A[2]	4q25-q31	2.50
LC27,lysosomal-associated transmembrane protein 4 beta[2]	8q22.1	2.45
PCNA,cyclin-dependent kinase inhibitor 1A	20pter-p12	2.20
EGR1,early growth response 1[2]	5q31.1	-5.94
IGFBP3,insulin-like growth factor binding protein 3	7p13-p12	-5.01
ANG,angiopoietin 1 isoform b[2]	14q11.1	-3.43[1]
PTGS2, prostaglandin-endoperoxide synthase 2 precursor[2]	1q25.2-q25.3	-3.36
p28,proteasome 26S non-ATPase subunit 10 isoform 1[2]	1p35.1	-2.27[1]
FGL1,fibrinogen-like 1 precursor[2]	8p22-p21.3	-2.19
COPEB,core promoter element binding protein	10p15	-2.09

[1] genes whose regulation directions were inconsistent with previous studies; [2] genes in cytogenetic aberrations.

might be naturally dense regions of differentially expressed genes or caused by other factors in experiments.

IR-CGMA also identified similar abnormalities in HCC, such as -1q, -4q, -8q, +8q, -13q, -16q, -17p and +17q (12).

In this study, 978 differentially expressed genes were identified for fold change above 2.0, including 336 up regulated genes and 642 down regulated ones. Among these genes, 498 (50.92%) were at the LS-CAP-predicted biases regions. Genes associating with HCC proven in previous studies have been found through milano program (http://milano.md.huji.ac.il. The program performs automatic searches in PubMed or the GeneRIF collection for articles containing co-occurrences of search terms with a list of genes), and the results listed in table 3. 9 of the 13 genes associating with HCC were at the gene expression biases regions, therefore there was a higher proportion of abnormally expressed genes associating with HCC locating on cytogenetic aberrations than on normal regions, and the proportion difference between cytogenetic aberrations and normal regions was statistically significant (P=0.0262). 11 of the 13 genes have the same regulation directions as in the relating studies (27-37). The other two genes, ANG and P28, were down regulated which were inconsistent with previous studies (38, 39). Variation among experiments might contribute to that, but in this paper we found that the two genes were all at down regulated biases regions, so the genetic deletions of the chromosomal arms might also partially contribute to that.

6. CONCLUSIONS

In this study, LS-CAP algorithm was developed based on locally un-weighted smoothing theory and applied to predict cytogenetic aberrations in HCC with gene expression microarray data. Two types of LS-CAP approaches had been built: LSW-CAP constructed independent statistics for rates of down and up regulated genes based on each whole chromosomal arm, whereas LSS-CAP constructed independent statistics on given size of chromosomal fragment. LSW-CAP was similar to CGMA but having distinctive statistic in the model; However, LSS-CAP had quite some outstanding features: constructed independent statistic for down and up regulated genes capable of reducing the false negatives and false positives of CGMA prediction, flexibly and thoroughly dug into gene expression data simply through an expansive setup of slide-window and fold change, precisely located

the gene expression biases regions on chromosomal arms, and sensitively identified small pieces of losses or gains that could not be acquired with CGMA and LSW-CAP. On the other side, LS-CAP had two major limitations: (1) Based on genomic scale gene expression microarray, the algorithm was not applicable to analyze the microarray data when the number of genes on a chromosomal arm was relatively small. (2) In addition to cytogenetic aberrations, the major reason for biases, there might be other unknown reasons contributing to biases such as epigenetics and random factors. These factors might cause small pieces of regional gene expression biases, but the chance was very little and the results could still provide clues for further researches.

LS-CAP can be used to analyze several types of experimental data. (1) For single profile, the experiment must be directly pairwise designed, so LS-CAP prediction can be used directly, but the only method for identifying differentially expression genes are fold change. (2) For multiple profiles, discriminate the differentially expressed genes in two or multiple groups with statistical test methods and perform LS-CAP prediction, or, perform LS-CAP prediction for each single profile first and then integrate the results.

FISH and Array-CGH are considered to be the major high-throughput and high-resolution techniques for detecting genetic aberrations. However, they are labor-intensive and time-consumptiv. In this study, we found that LS-CAP can successfully predict cytogenetic aberrations with gene expression microarray data, which means when we do studies on identifying differentially expressed genes, genes clustering or pattern recognizing with gene expression microarray data on genome scale, we can also identify genetic abnormalities using these data without additional works on FISH or Array-CGH. From all the aforementioned advantages, LS-CAP could potentially be a powerful alternative for FISH and Array-CGH.

6. ACKNOWLEDGEMENT

This research is supported by a project 30371422 from the National Natural Science Foundation of China, and also a part of grant 2002AA2Z2002 from the National High Technology Research and Development Program of China (863 Program).

7. REFERENCES

1. Kallioniemi A, O. P.Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman & D. Pinkel: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258, 818-821 (1992)

2. Haddad R, K. A. Furge, J. Miller, J. Schoumans, B. Haab, B. The, L. Barr & C. Webb: Genomic profiling and cDNA microarray analysis of human colon adenocaicinoma and associated peritoneal metastasis reveals consistent cytogenetic and transcriptional aberrations associated with progression of multiple metastases. *Appl Genomics Proteomics* 1, 123-134 (2002)

3. Huang J, W. Wei, J. Zhang, G. Liu, G. R. Bignell, M. R. Straton, P. A. Futreal, R. Wooster, K. W. Jones & M. H. Shapero: Whole genome DNA copy number changes identified by high density oligonucleatide arrays. *Hum Genomics* 1, 287-299 (2004)

4. Hughes T. R, C. J. Roberts, H. Dai, A. R. Jones, M. R. Meyer, D. Slade, J. Burchard, S. Dow, T. R. Ward, M. J Kidd, S. H. Friend & M. J. Marton: Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat Genet* 25, 333-337 (2000)

5. Marchio A, M. Meddeb, P. Pineau, G. Danglot, P. Tiollais, A. Bernheim & A. Dejean: Recurrent chromosomal abnormalities in hepatocellular carcinoma detected by comparative genomic hybridization. *Genes Chromosomes Cancer* 18, 59-65 (1997)

6. Phillips J. L, S. W. Hayward, Y. Wang, J. Vasselli, C. Pavlovich, H. Padilla-Nash, J. R. Pezullo, B. M. Ghadimi, G. D. Grossfeld, A. Rivera, W. M. Linehan, G. R. Cunha & T. Ried: The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis. *Cancer Res* 61, 8143-8149 (2001)

7. Pinkel D, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray & D. G. Albertson: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* 20, 207-211 (1998)

8. Pollack J. R, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein & P. O. Brown: Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 23, 41-46 (1999)

9. Pollack J. R, T. Sorlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A. L. Børresen-Dale & P. O. Brown: Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *PNAS* 99, 12963-12968 (2002)

10. Virtaneva K, F. A. Wright, S. M. Tanner, B. Yuan, W. J. Lemon, M. A. Caligiuri, C. D. Bloomfield, A. Chapelle & R. Krahe: Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *PNAS* 98, 1124-1129 (2001)

11. Crawley J. J & K. A. Furge: Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data. *Genome Biol* 3, research0075.1–0075.8 (2002)

12. Cleveland W. S: Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74: 829-836 (1979)

13. Cleveland W. S & S. J. Devlin: Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83, 596-610 (1988)

14. Furge K. A, K. J. Dykema, C. Ho & X. Chen: Comparison of array-based comparative genomic hybridization with gene expression-based regional expression biases to identify genetic abnormalities in hepatocellular carcinoma. *BMC Genomics* 6, 67 (2005)

15. Alizadeh A. A, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X.

Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown & L. M. Staudt: Distinctive types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511 (2000)

16. Baldi P & A. D. Long: A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17, 509-519 (2001)

17. Cui X & G. A. Churchill: Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* 4, 210 (2003)

18. DeRisi J. L, V. R. Iyer & P. O. Brown: Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680-686 (1997)

19. Lonnstedt I & T. Speed: Replicated microarray data. *Statistica Sinica* 12, 31-46 (2002)

20. Tusher V. G, R. Tibshirani & G. Chu: Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 98: 5116-5121 (2001)

21. Wong N, P. Lai, S. W. Lee, S. Fan, E. Pang, C. T. Liew, Z. Sheng, J. W. Y. Lau & P. J. Johnson: Assessment of genetic changes in hepatocellular carcinoma by comparative genomic hybridization analysis: relationship to disease stage, tumor size, and cirrhosis. *Am J Pathol* 154, 37-43 (1999)

22. Wright G. W & R. Simon: A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 19, 2448-2455 (2003)

23. Yang I. V, E. Chen, J. P. Hasseman, W. Liang, B. C. Frank, S. Wang, V. Sharov, A. I. Saeed, J. White, J. Li, N. H. Lee, T. J. Yeatman & J. Quackenbush: Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol* 3, research0062.1–0062.12 (2002)

24. Beasley R. P: Hepatitis B virus. The major etiology of hepatocellular carcinoma. *Cancer* 61, 1942–1956 (1988)

25. Chen X, S. T. Cheung, S. So, S. T. Fan, C. Barry, J. Higgins, K. M. Lai, J. Ji, S. Dudoit, O. L. N. Irene, R. Matt, D. Botstein & P. O. Brown: Gene expression patterns in human liver cancers. *Mol Biol Cell* 13, 1929–1939 (2002)

26. Sherlock G, T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. C. Matese, S. S. Dwight, M. Kaloper, S. Weng, H. Jin, C. A. Ball, M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein & J. M. Cherry: The Stanford Microarray Database. *Nucleic Acids Res* 29, 152-155 (2001)

27. Cheng A. S, H. L. Chan, K. F. To, W. K. Leung, K. K. Chan, C. T. Liew & J. J. Sung: Cyclooxygenase-2 pathway correlates with vascular endothelial growth factor expression and tumor angiogenesis in hepatitis B virus-associated hepatocellular carcinoma. *Int J Oncol* 24, 853–860 (2004)

28. Gramantieri L, D. Trere, P. Chieco, M. Lacchini, C. Giovannini, F. Piscaglia, A. Cavallari & L. Bolondi: In human hepatocellular carcinoma in cirrhosis proliferating cell nuclear antigen (PCNA) is involved in cell proliferation and cooperates with P21 in DNA repair. *J Hepatol* 39, 997-1003 (2003)

29. Hao M. W, Y. R. Liang, Y. F. Liu, L. Liu, M. Y. Wu & H. X. Yang: Transcription factor EGR-1 inhibits growth of hepatocellular carcinoma and esophageal carcinoma cell lines. *World J Gastroenterol* 8, 203-207 (2002)

30. Huynh H, P. K. Chow, L. L. Ooi & K. C. Soo: A possible role for insulin-like growth factor-binding protein-3 autocrine/paracrine loops in controlling hepatocellular carcinoma cell proliferation. *Cell Growth Differ* 13, 115-122 (2002)

31. Kalinina O. A, S. A. Kalinin, E. W. Polack, I. Mikaelian, S. Panda, R. H. Costa & G. R. Adami: Sustained hepatic expression of FoxM1B in transgenic mice has minimal effects on hepatocellular carcinoma development but increases cell proliferation rates in preneoplastic and early neoplastic lesions. *Oncogene* 22, 6266-6276 (2003)

32. Kremer-Tal S, H. L. Reeves, G. Narla, S. N. Thung, M. Schwartz, A. Difeo, A. Katz, J. Bruix, P. Bioulac-Sage, J. A. Martignett & S. L. Friedman: Frequent inactivation of the tumor suppressor Kruppel-like factor 6 (KLF6) in hepatocellular carcinoma. *Hepatology* 40, 1047-1052 (2004)

33. Liu X. R, R. L. Zhou, Q. Y. Zhang, Y. Zhang, Y. Y. Jin, M. Lin, J. A. Rui & D. X. Ye: Structure analysis and expressions of a novel tetratransmembrane protein, lysosoma-associated protein transmembrane 4 beta associated with hepatocellular carcinoma. *World J Gastroenterol* 10, 1555-1559 (2004)

34. Masaki T, Y. Shiratori, W. Rengifo, K. Igarashi, M. Yamagata, K. Kurokohchi, N. Uchida, Y. Miyauchi, H. Yoshiji, S. Watanabe, M. Omata & S Kuriyama: Cyclins and cyclin-dependent kinases: comparative study of hepatocellular carcinoma versus cirrhosis. *Hepatology* 37, 534-543 (2003)

35. Sung Y. K, S. Y. Hwang, M. K. Park, M. Farooq, I. S. Han, H. I. Bae, J. C. Kim & M. Kim: Glypican-3 is overexpressed in human hepatocellular carcinoma. *Cancer Sci* 94, 259-262 (2003)

36. Yan J, Y. Yu, N. Wang, Y. Chang, H. Ying, W. Liu, J. He, S. Li, W. Jiang, Y. Li, H. Liu, H. Wang & Y. Xu. LFIRE-1/HFREP-1, a liver-specific gene, is frequently downregulated and has growth suppressor activity in hepatocellular carcinoma. *Oncogene* 23, 1939-1949 (2004)

37. Yoshida S, K. Kurokohchi, K. Arima, T. Masaki, N. Hosomi, T. Funaki, M. Murota, Y. Kita, S. Watanabe & S. Kuriyama: Clinical significance of lens culinaris agglutinin-reactive fraction of serum alpha-fetoprotein in patients with hepatocellular carcinoma. *Int J Oncol* 20, 305-309 (2002)

38. Fu X. Y, H. Y. Wang, L. Tan, S. Q. Liu, H. F. Cao & M. C. Wu: Overexpression of p28/gankyrin in human hepatocellular carcinoma and its clinical significance. *World J Gastroenterol* 8: 638-643 (2002)

39. Mitsuhashi N, H. Shimizu, M. Ohtsuka, Y. Wakabayashi, H. Ito, F. Kimura, H. Yoshidome, A. Kato, Y. Nukui & M. Miyazaki: Angiopoietins and Tie-2 expression in angiogenesis and proliferation of human hepatocellular carcinoma. *Hepatology* 37, 1105-1113 (2003)

Abbreviations: CGH: comparative genomic hybridization; FISH: fluorescent in situ hybridization; SNP: single nucleotide pleomorphism; CGMA: Comparative genomic microarray analysis; HCC: hepatocellular carcinoma; DEs: differentially expression genes; LS-CAP: locally un-weighted smoothing cytogenetic aberrations prediction; LSW-CAP: locally un-weighted smoothing cytogenetic aberrations prediction based on whole chromosomal arm;

LSS-CAP: locally un-weighted smoothing cytogenetic aberrations prediction based on slide-window

Key Words: Cytogenetic Aberration, Tumor, Neoplasis, Cancer, Carcinoma, Hepatoma, Hepatocellular Carcinoma, cDNA Microarray, Comparative Genomic Microarray Analysis, Smoothing Theory

Send correspondence to: Dr He Xian Min, Department of Health Statistics, Second Military Medical University, 800 Xiangyin Road, Shang Hai, 200433, P. R. China, Fax: +86-21-2507-1486, Tel: 86-21-2507-0419, E-mail:hxmine@hotmail.com, hxmine@sina.com

http://www.bioscience.org/current/vol11.htm