Cancer Informatics

METHODOLOGY

Improved Statistical Analysis for Array CGH-Based DNA Copy Number Aberrations

Hongmei Jiang[1], Zhong-Zheng Zhu[2], Yue Yu[3], Simon Lin[4] and Lifang Hou[3,5]

[1]Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, IL 60208, USA. [2]Department of Oncology, No. 113 Hospital of People's Liberation Army, Ningbo 315040, China. [3]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA. [4]The Biomedical Informatics Center, Northwestern University, Chicago, IL 60611, USA. [5]Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL 60611, USA. Corresponding author email: hongmei@northwestern.edu

Abstract: Array-based comparative genomic hybridization (aCGH) allows measuring DNA copy number at the whole genome scale. In cancer studies, one may be interested in identifying DNA copy number aberrations (CNAs) associated with certain clinicopathological characteristics such as cancer metastasis. We proposed to define test regions based on copy number pattern profiles across multiple samples, using either smoothed $\log_2$-ratio or discrete data of copy number gain/loss calls. Association test performed on the refined test regions instead of the probes has improved power due to reduced number of tests. We also compared three types of measurement of copy number levels, normalized $\log_2$-ratio, smoothed $\log_2$-ratio, and copy number gain or loss calls in statistical hypothesis testing. The relative strengths and weaknesses of the proposed method were demonstrated using both simulation studies and real data analysis of a liver cancer study.

Keywords: aCGH, DNA copy number aberration (CNA), downstream analysis, gain/loss calls, segmentation

Cancer Informatics 2011:10 249–258

doi: 10.4137/CIN.S8019

This article is available from http://www.la-press.com.

Introduction

DNA copy number is the number of copies of DNA at a genome region. Gain and/or loss of chromosomal regions have been associated with various human diseases including cancer.[1] Therefore detecting and mapping DNA copy number changes provide a systematic approach to understand the association between DNA copy number abnormality and human disease. DNA copy number changes in tumor tissues are often referred to as DNA copy number aberrations (CNAs) which vary in size from 1 Kb up to one complete chromosome arm. High resolution array-based comparative genomic hybridization (aCGH) allows measuring DNA copy number at hundreds of thousands of probes or locations throughout the genome.[2,3]

Different methods and algorithms have been developed to divide the genome into regions with the same copy number level for a single sample using diversity techniques such as Hidden Markov Models,[4] circular binary segmentation,[5,6] mixture models,[7] and Bayesian change point analysis.[8] Recently research has been devoted to identify recurrent CNAs that are shared across multiple samples. This is an important issue because chromosome regions where recurrent CNAs occur may play important roles in the development and progression of cancer and other human diseases. Rouveirol et al[9] computed the recurrent minimal genomic alterations based on discretized CGH profiles. Shah et al[10] extended single sample HMMs to multiple samples to identify recurrent CNAs by jointly inferring CNA patterns and making gain/loss calls. Zhang et al[11] proposed simple scan and segmentation algorithms based on the sum of the chi-square statistics for each individual sample to give a sparse and intuitive cross-sample summary. Multiple-sample segmentation and detection of recurrent CNAs are still challenging research areas both computationally and conceptually. For more references, see the review paper.[12]

In clinical cancer studies, clinicians and researchers may be interested in identifying CNAs associated with patients' clinical outcomes and tumor pathological characteristics. Unfortunately, little research has devoted on this type of downstream analysis, which has more clinical significance and is different from identifying recurrent CNAs as in most of the previous studies. Moreover, some studies have performed downstream analysis on whole chromosome level or chromosome arm level. Due to a lack of fine mapping of chromosome regions, this approach may not be efficient if only some regions rather than complete arm of the chromosome are associated with the interested clinical events. On the other hand, if association test is performed for each probe, a large number of tests (hundreds of thousands up to millions) are inevitable. In addition, probes adjacent on chromosome are more likely to have the same copy number status and will result in highly correlated tests. Here, we proposed to define test regions based on DNA copy number pattern profiles across multiple samples, and perform association tests on these test regions instead of individual probes. Furthermore, excluding non-variant test regions would result in smaller number of tests and improved power. We also compared the use of normalized log-ratio, smoothed log-ratio, and discrete copy number gain/loss calls in downstream analysis. The relative strengths and weaknesses of these measurements and the proposed method were evaluated using simulation studies and real data analysis of a liver cancer study.

Methods

Measurements of copy number level

Normalized $\log_2$-ratio: In an aCGH experiment, the test sample and the reference sample are labelled with different dyes and co-hybridized on a microarray, which contains hundreds of thousands to millions of probes depending on the array platform. Usually, the test sample is labelled with Cy5 and the reference sample with Cy3. Log_2-ratio of background-corrected Cy5 signal to background-corrected Cy3 signal is computed for each probe on the array. The resulting ratio at a probe provides an estimate to the ratio of copy numbers of the corresponding DNA sequences in the test and the reference samples. In general a normalization step is needed to remove systematic array effect.

Smoothed $\log_2$-ratio: A smoothing or segment detection method is applied to normalized $\log_2$-ratios to divide the genome into regions with the same copy number levels. Adjacent regions have different copy number levels. The segmented or smoothed $\log_2$-ratio is assigned to be the median or average of the normalized $\log_2$-ratios of the probes contained in that region.

Gain or loss calls: DNA copy number gain or loss calls are usually made based on the smoothed $\log_2$-ratios.

For the regions where no copy number changes occur the smoothed $\log_2$-ratios are close to 0. When there are copy number gains (or losses), the smoothed $\log_2$-ratios are significantly greater (or less) than 0. However there are no hard thresholds on choosing the significance level to make gain or loss calls. Users decide whether there is sufficient evidence to call a probe having copy number gained or lost. For example, Aguirre et al[13] considered a segment having gain or loss if the corresponding smoothed $\log_2$-ratio is more than four standard deviations away from the middle 50% quantile of data; Willenbrock and Fridlyand[14] defined the threshold using three times median absolute deviation (MAD) of difference between observed and smoothed $\log_2$-ratios.

Defining test regions

When smoothed $\log_2$-ratios or gain/loss calls are used, we define "test regions" on which downstream analysis will be performed. Figure 1 shows DNA copy number profiles for three subjects on 90 contiguous probes. Each dot represents the normalized $\log_2$-ratio at a probe. The solid (red) lines represent the smoothed mean $\log_2$-ratios. There are three, two and two segments for subject 1, 2 and 3, respectively, in this genomic region. When the test regions are defined based on the smoothed $\log_2$-ratios, there are three test regions consisting of probes 1–30, 31–60, and 61–90, respectively. When gain/loss calls are used, test regions depend on what criterion is used. Suppose we consider copy number changed when the absolute value of smoothed $\log_2$-ratio is greater than 0.50, then there is no copy number change for probes 1 to 60, and there is copy number gain for probes 61–90, for subject 1. Thus there are two test regions consisting of probes 1–60, and 61–90, respectively. The set of probes in a test region remains constant within a sample and shares the same profile across samples, therefore the association tests for these probes will be exactly the same.

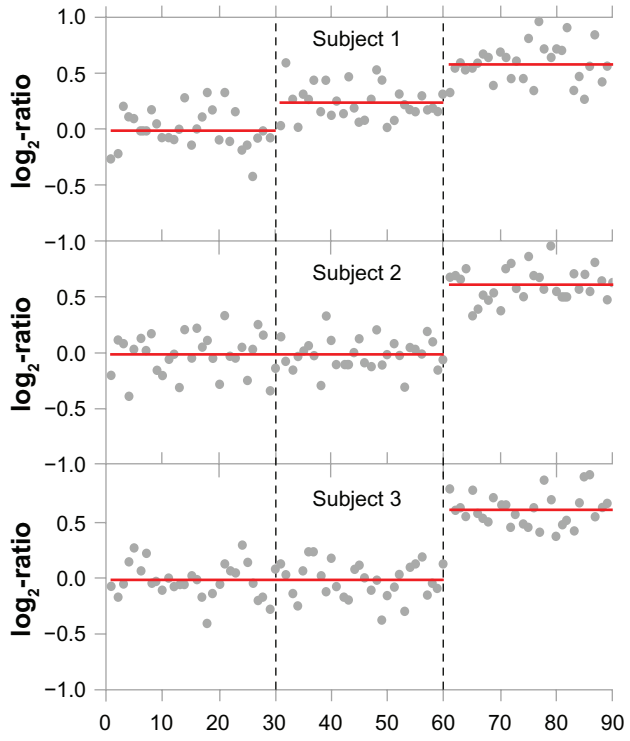

Figure 1. An example of DNA copy number profiles for three subjects on 90 consecutive probes. The gray dots are the raw $\log_2$-ratios of copy number measurements, the red straight lines represent the smoothed $\log_2$-ratios for the identified segments, and the vertical dashed black lines represent the test regions based on the smoothed $\log_2$-ratios.

We formally define a test region as a set of contiguous probes on the same chromosome in the following. Let s_{ij} represents the smoothed $\log_2$-ratio of DNA copy number for sample i ($i = 1, \ldots, n$) at probe j ($j = 1, \ldots, G$). Then a test region t of size $|t_j|$ is defined by

$$t = \{\text{Contiguous } t_1, t_2, \ldots, t_j : S_{i,t_1} = S_{i,t_2} = \ldots S_{i,tj}, \quad \text{for all} \quad i = 1, \ldots, n\} \tag{1}$$

where $|t_j|$ is the number of probes covered by this region. In addition we also require that DNA copy number profiles in any two neighboring test regions are different from each other. In this sense we are defining the maximum common region of the probes. It is possible that two or more non-adjacent test regions have the same copy number profile. When the gain/loss calls, z_{ij}, are used, the test regions can be defined in the same way as in (1) by replacing s_{ij} with z_{ij}. In general, the number of test regions using smoothed $\log_2$-ratios is bigger than that using gain/loss calls, because different values of smoothed $\log_2$-ratios above a threshold could be called as gain simultaneously.

Identifying non-variable test regions

The purpose of hypothesis testing discussed in this paper is to identify genome regions which are associated with certain clinicopathological characteristics; therefore non-variable test regions will be excluded from further analysis. For example, the test region containing probes 61–90 (Fig. 1) are recurrent CNAs

which are an important characteristic for identifying disease-related biomarkers. With all subjects having the same copy number status, it indicates that this region is not associated with the interested clinical event and it will not be tested. Similarly the first test region (Fig. 1) where probes 1–30 have no copy number changes, is not of interest either and excluded from further analysis. When gain/loss calls are used to identify CNAs with clinical significance, usually only probes with CNA frequency greater than a threshold (such as >10% as used in)[15] are included in the analysis for clinical application value.

When smoothed $\log_2$-ratios are used in the analysis, it is not obvious to identify non-variable test regions as using discrete gain/loss calls. When each biological sample has technical replicates, one can separate the variation between the samples from the measurement errors which will help us evaluate the variation of the probes across samples. When there is no technical replicate, one can compute the variation or standard deviation of the $\log_2$-ratios across samples by employing similar strategy as in microarray studies. If the variation is small, it is more likely that the corresponding region is a non-variable test region.

Performing association tests

After defining test regions and excluding non-variable ones, association tests are performed. In this paper, as an example, survival analysis using Cox proportional hazards model[16] for smoothed $\log_2$-ratios or log-rank test[17] for gain/loss calls was applied to identify CNAs associated with cancer metastasis. Let $y_i = \min(T_i, C_i)$ be the observed time, where T_i is the event time and C_i is the censoring time for the ith subject; let S_{ir} be the copy number measurement for the ith subject at region r. The hazard function $h(t)$ is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \le t + \Delta t \mid T \ge t)}{\Delta t},$$

and it has the following proportional hazards structure

$$h_i(t) = h_0(t)\exp(\beta_r \cdot S_{ir}), \qquad (2)$$

where $h_0(t)$ is the baseline hazard function and β_r is the log relative hazard for the rth test region, and i is the subscript for the ith subject. As some clinical variables such as tumor size and clinical stage might be associated with the interested event, the association between CNA and event time can be evaluated after adjusting these covariates by extending Equation (2) to

$$h_i(t) = h_0(t)\exp(\beta_r \cdot S_{ir} + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \cdots + \beta_k \cdot x_{ik}), \qquad (3)$$

where x_{ij}'s are the covariates. When the number of covariates k is bigger comparing with the number of subjects, high-dimensional variable selection and regularization techniques such as Lasso[18] and SCAD,[19] have to be employed. Alternative approach would be performing variable selection on the clinical variables first and identifying a few important variables to be included in Equation (3) such that k will be small and the conventional estimation of the parameters for Cox model can be used.

For each test region, the P-value for testing whether β_r is 0 in Equation (2) or (3) is computed. The adaptive false discovery rate (FDR) controlling procedure[20,21] is applied to the P-values to identify clinical event-associated genome regions. Here the proportion of test regions which are not associated with clinical event is estimated by the bootstrap method[22,23] and used in the adaptive FDR controlling procedure.

Liver cancer data

The proposed methods will be applied to a liver cancer data set. Sixty-three newly diagnosed HCC patients who underwent radically surgical therapy at Eastern Hepatobiliary Surgery Hospital, Shanghai, China, from December 2007 to March 2008 were recruited. All patients were ethnic Han Chinese, and none had received radiation therapy or chemotherapy before surgery. A total of 63 tumor tissue samples and 11 matched surrounding non-tumor liver tissue samples were frozen in liquid nitrogen within 1 h after surgical removal and kept at −80 °C until DNA extraction. Final diagnoses were pathologically confirmed by pathologists from H&E-stained slides. Demographic data were obtained through in-person interviews at the first hospital admission. Information on tumor characteristics, such as tumor differentiation, envelope status, thrombus status and tumor stage, was made on the basis of pathological and medical reports. The 63 patients have been followed for 2.5 years on average after the surgery and the clinical events including distant metastasis were recorded. In total, 8 (12.7%) subjects

developed distant metastasis during the follow-up, among which 7 subjects developed lung metastasis and 1 developed bone metastasis. Diagnosis and presence of distant metastases were based on: (1) positive findings on cytological or pathological examination, and/or (2) positive images on ultrasonography, CT, PET-CT, MRI and/or ECT bone scan. The study protocol was approved by the Institutional Review Board of the participant hospital, and written informed consent for this study was obtained from all patients.

Each of these 74 samples was labelled with Cy5 and was co-hybridized with pooled normal controls onto an Agilent Human Genome CGH Microarray Kit 244A, and the copy number levels for each of the 244K probes on the array were measured. Using the raw aCGH data, we first computed the $\log_2$-ratio of background-corrected Cy5 signal to background-corrected Cy3 signal for each probe on each of the 74 arrays. A simple median normalization approach was used to adjust the $\log_2$-ratios such that their median was 0 for each array. The CBS algorithm was applied to each sample separately to identify segments having different copy number levels. The default values of the parameters in the DNAcopy package were used. The outputs are the starting and ending positions, and the smoothed $\log_2$-ratios for each identified segment. The copy number levels are expected to be normal for non-tumor samples. Therefore the gain/loss calls (z_{ij}) for the tumor samples were determined using three standard deviations of the normalized $\log_2$-ratios of non-tumor samples as in the following:

$$z_{ij} = \begin{cases} 1 & s_{ij} \geq 3\sigma \\ 0 & |s_{ij}| \leq 3\sigma, \\ -1 & s_{ij} < -3\sigma \end{cases}$$

where s_{ij} is the smoothed $\log_2$-ratio for sample i at probe j, and $\sigma = 0.167$ is the average standard deviation of the 11 non-tumor samples. That is, copy number gain is called when the smoothed $\log_2$-ratio is greater than 0.5, and copy number loss is called when the smoothed $\log_2$-ratio is less than –0.5.

Results

We applied the proposed method on both simulated data and a liver cancer data set. We realize that the accuracy of defined test regions depend on which segmentation algorithm is used to locate DNA copy number changes. However the sensitivity and the specificity of different algorithms are not the focus of this paper. Here, the R package DNAcopy which implements the non-parametric CBS algorithm[5,6] was used for segmentation as it has been shown to have good operational characteristics.[14,24]

Simulation study

We performed simulation study on 90 consecutive probes. The copy number status for probes 1 to 30 and 61 to 90 are normal for all 63 subjects. For probes 31 to 60, on average 50% of the 63 samples have copy number aberrations. Suppose the clinical event of interest is cancer metastasis. When a probe or region is associated with the clinical event, subjects have different chance to get metastasis depending on their copy number status. To be specific, the probability of having cancer metastasis is 2/3 for subjects who have CNAs, and 1/3 for subjects who have normal copy number status. The raw $\log_2$-ratios were generated from $N(\mu,\sigma^2)$, where $\mu = 0$ for normal copy number status, $\mu = \log_2(3/2) = 0.585$ for CNAs, and $\sigma = 0.167$ was estimated from the liver cancer study discussed below. Segmentation was performed on these 63 profiles and a copy number gain was called when the smoothed $\log_2$-raio is greater than 0.5. Logistic regression was applied to each probe when the raw $\log_2$-ratios were used and to each region when the smoothed $\log_2$-ratios were used. For the gain/loss calls, Fisher's chi-square test was applied at each smoothed region. The false discovery rate (FDR) controlling procedure developed by[20] was used to control the proportion of falsely rejected probes or regions.

Figure 2 shows the mean sensitivity over 1000 simulated datasets for different thresholds of FDR, where sensitivity was computed as the proportion of correctly rejected probes among the total number of truly metastasis-associated probes. It is evident that both analysis, smoothed $\log_2$-ratio and discretized gain/loss calls, performed on segmented regions resulted in greatly improved power and sensitivity comparing with the analysis using $\log_2$-ratios on the probe level, while smoothed $\log_2$-ratio yielded the highest sensitivity. Figure 3 shows the average power over 1000 simulated datasets for each of the 90 probes at FDR significance level 0.05. It can be seen that region-based testing methods yielded improved power than

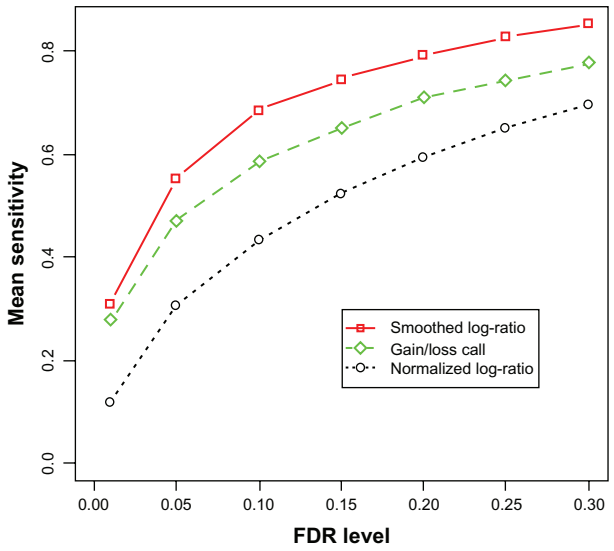

Figure 2. Average sensitivity versus different FDR thresholds based on 1000 simulations. Effects of the three types of DNA copy number measurements on the mean sensitiviy: smoothed $\log_2$-ratio (square), gain/loss call (diamond), and raw $\log_2$-ratio (circle).

probe-based tests for every probe, although the power reduced for probes close to the breakpoints of the regions due to the variation of the segmentation methods. Table 1 shows the average power comparisons using smoothed log-ratio, gain/loss call, and raw log-ratio for probes 31 to 60 which are truly associated with the event, at FDR level 0.05 for sample sizes 63 and 100 respectively. It is clear that when the sample size increases, the statistical power increases no matter what measurement is used (smoothed log-ratio, gain/loss call, or raw log-ratio), and the performance of the gain/loss calls gets closer to that for smoothed log-ratios. However the underlying assumption for gain/loss calls being more powerful than the raw log-ratios is that the gain/loss calls are made correctly. Otherwise, using gain/loss calls may lose power or even lead to the wrong conclusions.

Results for liver cancer data

Using normalized $\log_2$-ratios

In this paper, we focus on identifying CNAs which are associated with distant metastasis using three different types of measurements of copy number levels. We first performed the analysis using normalized $\log_2$-ratios. *P*-values were computed using Cox proportional hazard model for each of the 226,000 probes (probes mapped to the sex chromosomes were excluded) for the purpose of identifying CNAs associated with cancer metastasis. For the multiple corrections, the R package q value[22] was used to compute the q-values, and the bootstrap method was used to estimate the proportion of true null hypotheses. The smallest q-value is about 0.269, indicating none of

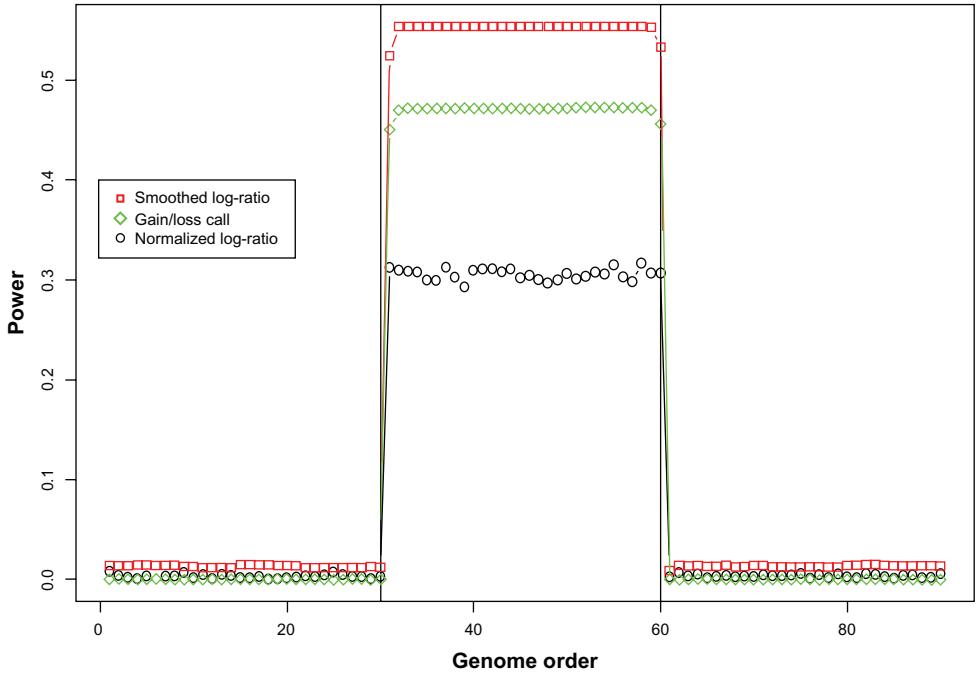

Figure 3. Average power at each probe using FDR significance level 0.05 based on 1000 simulations using 63 subjects. The three types of DNA copy number measurements: smoothed $\log_2$-ratio (square), gain/loss call (diamond), and raw $\log_2$-ratio (circle), are compared with respect to average power at probe.

Table 1. Average power at each probe for the three measurements (smoothed log-ratio, gain/loss call, and raw log-ratio) at FDR significance level 0.05 based on 1000 simulations using 63 subjects and 100 subjects, respectively.

	Position	31	32	45	58	59	60
n = 100	Smoothed log-ratio	0.779	0.795	0.796	0.796	0.796	0.778
	Gain/Loss call	0.730	0.746	0.745	0.745	0.743	0.720
n = 63	Raw log-ratio	0.581	0.566	0.594	0.589	0.57	0.566
	Smoothed log-ratio	0.521	0.542	0.543	0.542	0.54	0.515
	Gain/Loss call	0.438	0.457	0.460	0.460	0.459	0.436
	Raw log-ratio	0.286	0.29	0.299	0.292	0.294	0.295

the 226,000 probes is statistically significant if the false discovery rate (FDR) level chosen is less than 0.26. As we mentioned earlier, adjacent probes result in correlated tests. The FDR controlling procedure employed here did not take into account the correlation, which may lead to conservative results.

When the FDR level was increased to 0.30, there were 2441 statistically significant probes. CNAs usually occur on a relatively large genome region, one or two probes alone being statistically significant do not give reliable results for CNAs. Number of statistically significant probes was reduced from 2441 to 695 by restricting our examination to genome regions with at least three consecutive statistically significant probes.

Using smoothed $\log_2$-ratios

The total 226,000 probes were divided into 13,566 test regions after dropping the segments on sex chromosomes. Excluding the test regions which contain one or two probes only, we performed analysis on the remaining 10,258 test regions. At FDR level 0.10, we identified 119 statistically significant test regions covering a total of 2593 probes. The number of probes covered by each of the 119 test regions varies from 3 to 114.

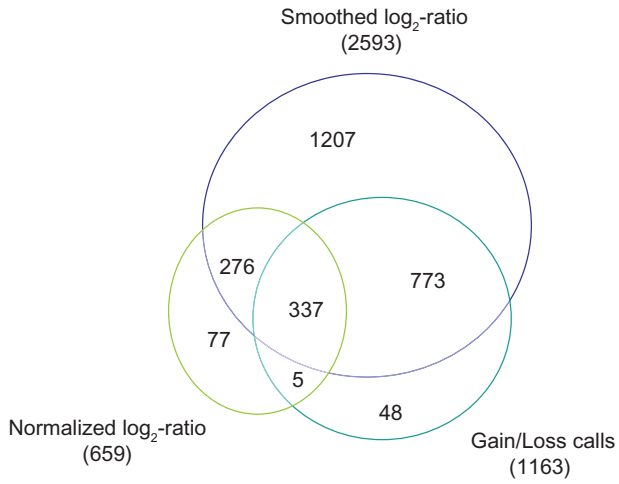

Figure 4. Numbers of statistically significant probes for the liver cancer study. The three types of DNA copy number measurements: smoothed $\log_2$-ratio, gain/loss call, and raw $\log_2$-ratio, are compared in terms of identification of statistically significant probes.

Using gain and loss calls

We performed log-rank test using gain/loss calls to identify chromosome regions on which patients with different copy number status have different survival curves of metastasis. Non-variable test regions where the copy number levels are normal for all 63 patients were excluded from further analysis. We also excluded the test regions containing one or two probes only.

Table 2. Clinicopathological characteristics and their association with metastasis.

Clinical variables	Number of patients	*P*-value[2]
Age		0.096
≤50	31	
>50	32	
Sex		0.045
Female	17	
Male	46	
Liver cirrhosis		0.207
No	25	
Yes	38	
Tumor differentiation[1]		0.157
II	13	
III	50	
Complete envelope		0.866
No	41	
Yes	22	
Cancer thrombus		0.474
Negative	32	
Positive	31	
Tumor stage		0.008
I	26	
II	13	
III or IV	24	

Notes: [1]According to the Edmondson-Steiner grading system; [2]Cox proportional hazard regression model, adjusted for each other.

Table 3. The copy number aberrations (CNAs) in association with metastasis.

Cytoband[1]	CNA type	Map position (start–end)[1]	Size (bp)	HR (95% CI)[2]	*P*-value[2]
3p14.2	Loss	60996566–61019967	23 402	8.94 (1.73–46.15)	0.009
5q13.2	Loss	70622715–50716337	93 623	5.24 (1.18–23.24)	0.029
7p15.2	Loss	26136974–46162024	25 051	14.64 (1.52–141.21)	0.020
12p13.31-11.22	Loss	9626736–69596274	19 969 539	9.19 (1.41–60.08)	0.021

Notes: [1]Cytoband and map position are based on the public UCSC database [Human Genome Browser, May 2004 Assembly (hg 17)]; [2]Cox proportional hazard model with the adjustment for sex and tumor stage.

Finally, from the clinical standpoint, a log-rank test is performed on a test region if and only if there are at least 6 subjects ($6/63 \approx 10\%$) in at least two of the three copy number status (gain, normal, and loss). If the frequency of CNA is less than 10% for a copy number status at a test region, patients having the corresponding status will be excluded from the log-rank test on that region. This is very important as extreme values may have a huge effect on the log-rank test. It is important to note that different patients may be excluded at different test regions. To summarize, there were 4190 test regions and the log-rank tests were performed on 1409 of them, total of 10 statistically significant test regions covering 1163 probes were yielded at FDR level 0.10.

Comparisons of methods

Since using normalized $\log_2$-ratio does not yield any statistically significant regions with FDR level below 0.269, we compared results using normalized $\log_2$-ratio at FDR level 0.30 with those for smoothed $\log_2$-ratio and gain/loss calls at FDR level 0.10 (Fig. 4). For the purpose of comparisons numbers of probes covered by statistically significant test regions were listed. The three types of measurements of copy number levels gave overlapping results, and analysis using the smoothed $\log_2$-ratio yielded the largest number of statistically significant results. The identified metastasis-associated CNAs in common by all three methods are on the short arm of chromosome 12. In addition smoothed $\log_2$-ratios identified regions on chromosome 2. Statistically significant results detected by normalized $\log_2$-ratio or gain/loss calls alone consist of a relatively small proportion of probes.

Detection of significant CNAs in association with metastasis

The association between clinicopathological variables and metastasis in 63 patients with HCC is summarized in Table 2. Among the 7 variables evaluated, only tumor stage ($P = 0.008$) and sex ($P = 0.045$) are statistically associated with metastasis. The CNAs associated with metastasis are evaluated using Cox proportional hazards model after adjusting for tumor stage and sex. As shown in Table 3, patients with loss on chromosome region 3p14.2, 5q13.2, 7p15.2, and 12p13.31-11.22 have a 8.94-fold (95% CI = 1.73–46.15), 5.24-fold (95% CI = 1.18–23.24), 14.64-fold (95% CI = 1.52–141.21), and 9.19-fold (95% CI = 1.41–60.08) increased hazard ratios of developing metastasis when compared with patients without the loss, respectively.

Discussions and Conclusions

Chromosomal aberrations such as gains, losses, structural rearrangements and other genetic mutations are hallmarks of human cancers.[3,25] Among them, genomic CNA has been regarded as an essential component in multiple types of cancers including HCC.[26] CNAs may contribute to the development and progression of various cancers by inducing gene expression alterations with or without other genetic mutations.[25] Identifying genomic CNAs in association with clinicopathological characteristics may provide some insights into the initiation and progression of cancer, and improve the diagnosis, prognosis, and treatment strategies.

In this paper we discussed the statistical framework for downstream analysis in copy number studies, using identification of CNAs associated with cancer metastasis as an example. When downstream analysis is performed at probe level and the correlation structure among adjacent probes is ignored, the power for detecting clinical event-associated CNAs is reduced. Here, we defined test regions based on DNA copy number patterns across samples, using either smoothed $\log_2$-ratios or discrete data of gain/loss calls. Segmentation could be done for each sample separately, or for all

samples simultaneously. Downstream analysis such as survival analysis is performed on these test regions instead of individual probes yielding improved power due to reduced number of tests. The advantages of using test regions instead of probes are at least twofold. One is that the number of tests performed is reduced from hundreds of thousands of probes to thousands of test regions. The other is that the adjacent test regions are not correlated with each other, and the false discover rate controlling procedures can be applied without losing power.

We further compared the effects of using different types of copy number measurements on downstream analysis. The raw log-ratios are usually very noisy even after normalization. The large number of probes makes it more difficult to detect statistically significant CNAs with limited number of subjects. New multiple testing methods which can take into account the correlations among the neighbouring probes are needed. Gain/loss calls are convenient for medical practitioners to explain the clinical significance. Survival curves for different copy number status (gain, loss or normal) can be visually presented. However the results are heavily relied on which threshold is used to call a test region having copy number gain or loss. In our analysis, we used the copy number measurements from non-tumor (control) samples as the reference. In many experiments, the control samples are not included in the experiment. In addition, summarizing $\log_2$-ratios into gain/loss calls may lose information present in the original measurements.[27] When inappropriate criterion is used, it may lead to unreliable categorization of true copy number level; hence the threshold used for making gain/loss calls has a deterministic effect on downstream analysis. For a potential improvement, the uncertainty of gain/loss calls should be incorporated into downstream analysis, which would be a topic for future research. From the simulation study and real data analysis, smoothed $\log_2$-ratio has the largest power in detecting genome regions with CNAs associated with clinical outcomes due to the reduced number of test regions from the huge number of probes, and the accurate and reliable measurement of copy number levels by borrowing strength from neighbouring probes.

Array-based CGH analyses of HCC have identified a number of recurrent chromosomal CNAs and some of these have been associated with tumor stage, differentiation, and survival.[26,28,29] However, no study has focused on the identification of the CNAs which play a key role in the determination of distant metastasis of HCC patients. We reported here for the first time that chromosome region loss on 3p14.2, 5q13.2, 7p15.2, and 12p13.31-11.22 are associated with distant metastasis of HCC. It has been reported 12p loss was frequently found in high-grade tumors and recurrent tumors in uterine leiomyosarcomas.[30] Furthermore, loss on 12p13.31 has been reported to be one of the most powerful independent markers for poor outcome in multiple myeloma.[31] Taken together, these results indicated that 12p loss is associated with the progression of malignant tumors, in support of the present observation of 12p13.31-11.22 loss in association with distant metastasis in HCC. In the 19.97 Mb window within 12p13.31-11.22, there locates many cancer-related genes including *CDKN1B* which may contribute to metastasis by copy number-induced down-regulation in gene expression level.[32] Other metastasis-associated region (3p14.2, 5q13.2, and 7p15.2.22) identified in the present study are relatively short, and contains no known gene. It should be noted that the high-resolution 244K aCGH platform used in the present study is different from low-resolution platforms, which are frequently used in clinical studies, in respect to sensitivity and specificity, and thus the extent to which our results apply to low-resolution platform detection remains to be determined.

In this paper, we have presented a simple but efficient dimension reduction method to identify genome regions with CNAs associated with clinical outcomes. We recommend the use of smoothed $\log_2$-ratio for downstream association test because it demonstrated the best statistical power. Even though the proposed method is demonstrated using aCGH data, the concept of test region can be applied to cancer-related CNAs studies using SNP arrays or sequencing technologies. This method has the potential to be applied for clinical screening of CNAs, thus help develop more accurate strategies in diagnosis, prognosis and therapy.

Acknowledgements

ZZZ's work was supported by the Medical Science and Technology Innovation Fund of PLA, Nanjing branch, China (No. 08MA023; No. 09MA022), and Ningbo Nature Science Foundation Program (No. 2009A610126).

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

References

1. Speleman F, Kumps C, Buysse K, et al. Copy number alterations and copy number variation in cancer: close encounters of the bad kind. *Cytogenet Genome Res*. 2008;123:176–82.
2. Pinkel D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*. 1998; 20(2):207–11.
3. Pinkel D, Albertson D. Array comparative genomic hybridization and its application in cancer. *Nat Genet*. 2005;37(Suppl S):S11–7.
4. Fridlyand J, et al. Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*. 2004;90:132–53.
5. Olshen AB, et al. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5:557–72.
6. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*. 2007;23(6):657–66.
7. Broët P, Richardson S. Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics*. 2006;22(8):911–8.
8. Erdman C, Emerson JW. A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*. 2008;24(19):2143–8.
9. Rouveirol C, et al. Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*. 2006;22:2066–73.
10. Shah S, et al. Modeling recurrrent CNA copy number alterations in array CGH data. *Bioinformatics*. 2007;23:i450–8.
11. Zhang NR, Siegmund DO, Ji H, Li J. Detecting simultaneous change-points in multiple sequences. *Biometrika*. 2010;97:631–45.
12. Rueda OM, Diaz-Uriarte R. Finding recurrent copy number alteration regions: a review of methods. *Current Bioinformatics*. 2010;5:1–17.
13. Aguirre AJ, et al. High-resolution characterization of the pancreatic adenocarcinoma genome. *PNAS*. 2004;101:9067–72.
14. Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analysis. *Bioinformatics*. 2005;21: 4084–91.
15. Korshunov A, et al. Adult and pediatric medulloblastomas are genetically distinct and require different algorithms for molecular risk stratification. *Journal of Clinical Oncology*. 2010;28(18):3054–60.
16. Cox DR. Regression models and life tables. *J R Statist Soc B*. 1972;34: 187–220.
17. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *J R Statist Soc A*. 1972;135(2):185–207.
18. Tibshirani R. The LASSO method for variable selection in the Cox model. *Statistics in Medicine*. 1997;16:385–95.
19. Fan J, Li R. Variable Selection for Cox's Proportional Hazards Model and Frailty Model. *The Annals of Statistics*. 2002;30:74–99.
20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple. *J R Statist Soc B*. 1995;57(1):289–300.
21. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*. 2000;25(1):60–83.
22. Storey JD. A direct approach to false discovery rates. *J R Statist Soc B*. 2002;64:479–98.
23. Storey JD, et al. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J R Statist Soc B*. 2004;66:187–205.
24. Lai WR, et al. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*. 2005;21:3763–70.
25. Meyerson M, et al. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*. 2010;11(10):685–96.
26. Midorikawa, et al. High-resolution mapping of copy number aberrations and identification of target genes in hepatocellular carcinoma. *Biosci Trends*. 2007;1:26–32.
27. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. *Nat Genet*. 2007;39:S37–42.
28. Woo HG, et al. Identification of potential driver genes in human liver carcinoma by genomewide screening. *Cancer Res*. 2009;9(9):4059–66.
29. Chochi, et al. A copy number gain of the 6p arm is linked with advanced hepatocellular carcinoma: an array-based comparative genomic hybridization study. *J Pathol*. 2009;217(5):677–84.
30. Hu J, et al. Genomic alterations in uterine leiomyosarcomas: potential markers for clinical diagnosis and prognosis. *Genes Chromosomes Cancer*. 2001; 31(2):117–24.
31. Avet-Loiseau, et al. Prognostic significance of copy-number alterations in multiple myeloma. *J Clin Oncol*. 2009;27(27):4585–90.
32. Hershko DD. Cyclin-dependent kinase inhibitor p27 as a prognostic biomarker and potential cancer therapeutic target. *Future Oncol*. 2010;6(12): 1837–47.