World Academy of Science, Engineering and Technology
International Journal of Biotechnology and Bioengineering
Vol:5, No:3, 2011

A Pairwise-Gaussian-Merging Approach: Towards Genome Segmentation for Copy Number Analysis

Chih-Hao Chen, Hsing-Chung Lee, Qingdong Ling, Hsiao-Jung Chen, Sun-Chong Wang, Li-Ching Wu, and H.C. Lee

***Abstract*—**Segmentation, filtering out of measurement errors and identification of breakpoints are integral parts of any analysis of microarray data for the detection of copy number variation (CNV). Existing algorithms designed for these tasks have had some successes in the past, but they tend to be $\mathcal{O}(N^2)$ in either computation time or memory requirement, or both, and the rapid advance of microarray resolution has practically rendered such algorithms useless. Here we propose an algorithm, SAD, that is much faster and much less thirsty for memory – $\mathcal{O}(N)$ in both computation time and memory requirement -- and offers higher accuracy. The two key ingredients of SAD are the fundamental assumption in statistics that measurement errors are normally distributed and the mathematical relation that the product of two Gaussians is another Gaussian (function). We have produced a computer program for analyzing CNV based on SAD. In addition to being fast and small it offers two important features: quantitative statistics for predictions and, with only two user-decided parameters, ease of use. Its speed shows little dependence on genomic profile. Running on an average modern computer, it completes CNV analyses for a 262 thousand-probe array in ~1 second and a 1.8 million-probe array in 9 seconds.

***Keywords*—**Cancer, pathogenesis, chromosomal aberration, copy number variation, segmentation analysis.

I. Introduction

LOCATING chromosomal aberrations in comparative genomic DNA samples is an important step in understanding the pathogenesis of many diseases. Amplification or deletion of chromosomal segments can lead to abnormal mRNA transcript levels and results in malfunctioning of cellular processes. This is especially true in cancer, where an enormous amount of efforts and resources has been dedicated to the detailed characterization of the chromosomal abnormalities caused by its various types.

Array comparative genomic hybridization (CGH) is a high-throughput technique developed for measuring such changes [1]-[3]. CGH arrays using BAC (Bacterial Artificial Chromosome) clones have resolutions of the order of 1Mb [2]. Those using cDNA and oligonucleotide as probes [4][5] are less robust than BACs for large segments, but offer much higher resolutions (in the order of 50-100kb). In particular, oligonucleotide arrays allow design flexibility and greater coverage and provide good sensitivity [5]. Tiling on custom arrays is also available now for even finer resolution of specific regions and allow the detection of micro-amplifications and deletions [6][7]. The drastic improvement in resolution has led to a corresponding increase in the number of probes on an array; modern high-resolution arrays now easily exceed one million probes. Arrays of such size and larger exact a severe requirement on the speed and accuracy of algorithms used to analyze the arrays, and have practically rendered useless existing algorithms that are $\mathcal{O}(N^2)$ (where N is array size) in computation time or memory requirement. Here, we propose a novel algorithm – Segmentation Analysis of DNA (SAD) – for studying copy number variation in high-resolution arrays. SAD has an extremely simple formulation, has in essence a single parameter and, compared with algorithms found in the literature, easier to understand, simpler to use, provides clearer statistical interpretation for its results, requires less memory, offers better accuracy, and is vastly faster in computation speed.

The design of SAD is based on three ingredients: (1) the assumption that measurement errors are normally distributed, (2) a clustering procedure based on Gaussian merging, and (3) t-statistic. The assumption in (1) is justified in Appendix A. SAD views every piece of raw datum as a statistical event from which the true value can be predicted via a normal, probability distribution function (PDF), or simply a Gaussian, whose variance (denoted by $\tilde{\sigma}$) is extracted from the array data. Our algorithm employs a key property of Gaussians: two Gaussians can be *algebraically* merged into a new one. By combining pair-wise merging of Gaussians with nearest-neighbor

Chih-Hao Chen is with Graduate Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan 32001 and National Center for Theoretical Science, Shinchu, Taiwan 30043 (corresponding author to provide phone: +886 3 4227151 ext 65390; e-mail: chih_hao_chen@ yahoo.com).

Hsing-Chung Lee is with Department of Surgery, Cathay General Hospital, Taipei, Taiwan.

Qingdong Ling is with Graduate Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan 32001 and Cathay Medical Research Institute, Cathay General Hospital, Taipei, Taiwan.

Hsiao-Jung Chen is with Graduate Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan 32001.

Sun-Chung Wang is with Graduate Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan 32001.

Li-Ching Wu is with Graduate Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan 32001.

H.C. Lee is with Graduate Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan 32001, National Center for Theoretical Science, Shinchu, Taiwan 30043 and Department of Physics, National Central University, Chungli, Taiwan 32001.

clustering we cluster array data into segments according to copy numbers. In an iterating process, two neighboring Gaussians are merged if their *resolvability* t (essentially an absolute t-statistic) is less than a threshold value t_{min}. In the result, each segment is assigned an *aberrance* z defined via its associated Gaussian. We call this technique Pair-wise Gaussian Merging (PGM). The operational principles of PGM are schematically illustrated in Fig. 1. In this case, the original ten pieces of data are predicted by SAD to have an underlying structure of two segments. A detailed description of PGM is given in Methods and Materials.

SAD has one essential parameter and an optional one. The essential parameter is t_{min}. In addition to its role in terminating iteration, t_{min} also provides quantitative statistics for aberrations identified by SAD. The optional parameter is sampling size N_s, designed into SAD to avoid sampling the entire array (with N probes) repeatedly during iterations. For microarray with N<100 k the option is not needed (N_s will be set automatically to N), since on a typical modern computer SAD is likely to finish the computation in less than one second. For larger microarrays, evoking the option by selecting an $1 \ll N_s \ll N$ speeds up the computation by a factor of approximately N/N_s, so that SAD becomes $\mathcal{O}(N)$ in computation time and memory requirement. Our tests show that there is little sacrifice in accuracy when $N_s \geq 100$, and we recommend setting N_s=100 when N is large.

II. Methods and Materials

A. Pair-wise Gaussian Merging

A Gaussian of mean μ and variance σ is defined as:

$$G\left(y;\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(y-\mu\right)^2}{2\sigma^2}\right). \quad (1)$$

Gaussians have a very useful property: a product of Gaussians yields a new Gaussian.

$$\prod_i G\left(y;\mu_i,\sigma_i^2\right)\propto G\left(y;\mu,\sigma^2\right);$$
$$\frac{1}{\sigma^2}=\sum_i\frac{1}{\sigma_i^2};\quad \frac{\mu}{\sigma^2}=\sum_i\frac{\mu_i}{\sigma_i^2}. \quad (2)$$

We apply this Gaussian-Merging technique (GM) to data analysis as follows.

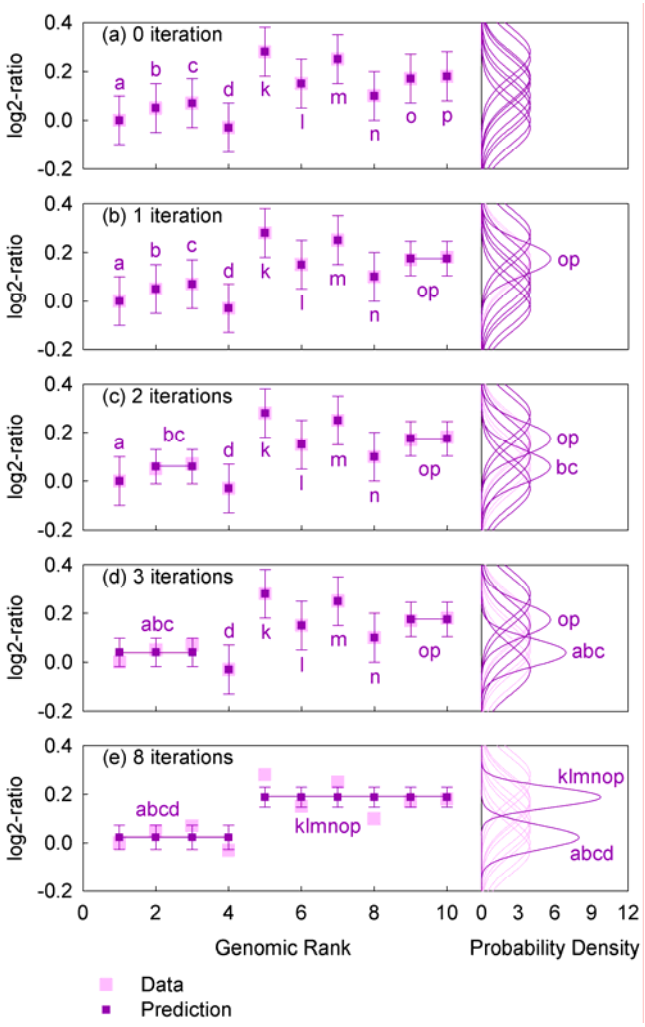

Fig. 1 Schematic illustration of PGM applied to genome segmentation. Frames on the left display the log2-ratios of probes (data) and clusters (prediction at various stages) and those on the right display associated Gaussians. In (a), each piece of raw data is treated as a Gaussian with variance $\tilde{\sigma}$. In (b), data o and p, the nearest neighboring pair, are merged in the first iteration. (c) and (d) show second and third iterations, respectively. In (e), merging stops after eight iterations because the remaining cluster pair satisfies $t \geq t_{min}$

Given a set of observations $\Omega=\{\mu_i|i=1,N\}$ of a quantity y acquired by measurement known to produce Gaussian noise of variance $\tilde{\sigma}$. Based on Ω, we want to formulate a PDF $f_\Omega(y)$ for predicting the true value. Considering a single observation μ_i first, because the variance of the noise is known, we have

$$f_{\{\mu_i\}}(y)=G\left(y;\mu_i,\tilde{\sigma}^2\right). \quad (3)$$

We can therefore associate a Gaussian with each observation. Considering Ω altogether, in terms of conditional probability and joint probability,

$$f_\Omega\left(y\right)=P\left(y\mid\Omega\right)=\frac{P\left(y\cap\Omega\right)}{P\left(\Omega\right)}$$
$$\propto P\left(y\cap\Omega\right)=\prod_i f_{\{y\}}\left(\mu_i\right)=\prod_i f_{\{\mu_i\}}\left(y\right), \quad (4)$$

where the third relation uses the fact that $P(\Omega)$ is independent of y and the last one is based on (1). (2)-(4) now yield

$$f_\Omega = G\left(y;\mu,\sigma^2\right);\ \ \mu = \sum_i \mu_i / N;\ \ \sigma^2 = \tilde{\sigma}^2 / N. \qquad (5)$$

The formulations of both μ and σ are intuitively understood: μ is the mean of the observations and σ^2 scales as required by the Central Limit Theory.

GM can be applied pair-wisely in a particular order to cluster data. Let $G_k \equiv G(y;\mu_k,\sigma_k^2)$, where k=1 and 2, be two Gaussians based on populations of n_k observations and $G(y;\mu,\sigma^2)$ be their merging product. We define resolvability t as

$$\begin{aligned} t(G_1, G_2) &\equiv \left(\left(\frac{\mu-\mu_1}{\sigma_1} \right)^2 + \left(\frac{\mu-\mu_2}{\sigma_2} \right)^2 \right)^{1/2} \\ &= \frac{|\mu_1-\mu_2|}{\tilde{\sigma}} \left(\frac{1}{n_1}+\frac{1}{n_2} \right)^{-1/2}, \end{aligned} \qquad (6)$$

where (5) is employed to obtain the third expression. Note that t is essentially an absolute t-statistic used in testing whether the means of two normally distributed populations (of equal variance) are equal. Given t_{min} as a threshold value for t, we say G_1 and G_2 are resolvable if $t(G_1,G_2) \geq t_{min}$ or unresolvable otherwise. A set of Gaussians can be clustered by carrying out GM using the following procedure: (1) Select t_{min}. (2) Identify the unresolvable pair of Gaussians with the smallest t and replace it with the merging product. (3) Iterate step (2) until all remaining pairs are resolvable. We call this technique Pair-wise Gaussian Merging (PGM) and remark that PGM is a type of nearest-neighbor clustering using t as distance.

B. The SAD Algorithm

SAD has two modes: the linear mode (LM) for low-resolution arrays or when computation time is not a concern, and the parallel mode (PM) when computation time is a concern. LM has a single parameter t_{min} while PM has an additional parameter N_s whose default value of 100 is highly recommended.

The steps in LM are: (1) Computation of $\tilde{\sigma}$. Let $\{\mu_i \mid i=1,N\}$ be the initial data of *log*2-ratios, $q_i=\mu_{i+1}-\mu_i$ and q_{IQR} be the interquartile range of q_i (or the difference between the 25th and 75th percentiles of the ranked q_i's). $\tilde{\sigma}$ is computed using

$$\tilde{\sigma} = \frac{q_{IQR}}{1.349\sqrt{2}}, \qquad (7)$$

where 1.349 is the interquartile range of $G(y;0,1^2)$ [17]. Consider each probe a cluster and associate $G(y;\mu_i,\tilde{\sigma}^2)$ with the i^{th} cluster. (2) Selection of t_{min}. This stipulates when PGM iteration stops and requires the statistical insights discussed in the following subsections. (3) PGM Phase I. Perform chromosome-wide PGM to all adjacent cluster pairs. At the end of this phase the remaining clusters are either multi-probe or single-probe; a single-probe cluster is called a loner and may hamper the merging of its two unresolvable neighbors. (4) PGM Phase II. Perform a second round of chromosome-wide PGM, now merging unresolvable cluster pairs that are either loner-divided or adjacent. When a loner-divided pair is merged, the loner is called an outlier. Outliers are excluded from the subsequent calculation of PGM. At the end of this phase each resultant cluster is called a segment and has an associated Gaussian for predicting its true *log*2-ratio.

As PGM involves very little computation, LM is inherently a very fast algorithm. However, the problem size of either $\tilde{\sigma}$ computation or LM are $\mathcal{O}(N^2)$, which implies long computation time when N is very large. PM is designed to reduce the problem size to $\mathcal{O}(N)$ with little sacrifice in accuracy. In PM, a sampling size N_s is selected and the algorithm is adjusted accordingly. In step (1) $\tilde{\sigma}$ is computed using only the first N_s probes. This reduces its problem size to $\mathcal{O}(N_s^2)$. In (3) and (4), before each iteration the current cluster set is partitioned to subsets of N_s contiguous clusters, plus a remainder. All subsets are processed in parallel and the most unresolvable pair of each subset is merged at each iteration. After each iteration, the current cluster set is circularly re-partitioned with the beginning of the remainder in the previous iteration taken as the starting point. Each iteration reduces the number of clusters by a factor of $\approx(1-N_s^{-1})$, making the problem size $\mathcal{O}(NN_s)$. That is, PM is N/N_s times faster than LM.

C. Reliability of a breakpoint

A breakpoint is the site where two adjacent resolvable segments meet. Given a null hypothesis (henceforth referred to as NHB) that the breakpoint does not exist, or in other words, the two segments have the same mean, the resolvability t as defined in (6) quantifies the statistical significance.

D. Aberrance of a segment

After clustering, each resultant segment is associated with a Gaussian $G(y;\mu,\sigma^2)$. We remark that $G(y;\mu,\sigma^2)$ is the PDF for predicting the true *log*2-ratio rather than the *log*2-ratio distribution of the probes within. We define aberrance z as

$$z = |\mu| / \sigma \qquad (8)$$

for measuring the aberrance of a segment. Given a null hypothesis (henceforth referred to as NHS) that the segment is normal, or μ=0, z is same as the z-value that quantifies the statistical significance. If need be, users are advised to prioritize further examinations of detected segments according to z.

E. Selection of t_{min}

Setting a value for t_{min} is essentially equivalent to setting a statistical level for rejecting NHB and NHS. For NHB this is because all remaining adjacent segment pairs are resolvable,

$$t_{\min} \leq t. \qquad (9)$$

For NHS this can be seen by considering an aberrant segment $G(y;\mu,\sigma^2)$ of width n wedged between two much wider disomic neighbors $G(y;0,\sigma_k^2)$ of width n_k, k=1 or 2. Let t_k be its resolvability with its k^{th} neighbor. Combining $t_{min} \leq t_k$, $n_k \gg n$, (5), (6) and (8), we get

$$t_{\min} \le t_k \equiv \frac{|\mu|}{\tilde{\sigma}}\left(\frac{1}{n}+\frac{1}{n_k}\right)^{-1/2} \approx \frac{|\mu|}{\tilde{\sigma}}\sqrt{n} = \frac{|\mu|}{\sigma} = z. \quad (10)$$

By noting $\mathrm{SNR} \equiv |\mu|/\tilde{\sigma}$, (10) can be rewritten as

$$t_{\min} \le SNR \cdot \sqrt{n}. \quad (11)$$

With an estimated SNR, t_{min} thereby determines the lower bound of aberration width. A smaller t_{min} on one hand facilitates detection of narrower segments. On the other hand it is more likely to yield false positives.

These insights make parameter t_{min} intuitively comprehensible and thus facilitate user parameter tuning.

III. Results

SAD can be run in two modes: the linear mode (LM) and the parallel mode (PM). N_s needs to be specified only in PM. Hereafter, we designate each set of SAD parameters by SAD(t_{min},N_s) and refer to LM as SAD(t_{min},−). All calculations reported here are carried out with an executable computer program written in Visual C++ that run on a computer with Intel Core 2 CPU 6420 (2.13GHz), 2GBs of DDRII 667 memory, and Windows XP as operating system. The program uses a single core on a multiple-core CPU as it runs as a single thread.

In a comparative analysis [8] (hereafter referred to as LJKP), eleven algorithms are tested for accuracy using simulated data in terms of receiver operating characteristic (ROC) and otherwise compared using real Glioblastoma Multiforme (GBM) data. For the ROC tests LJKP finds that the three smoothing-only algorithms (SO, which does not do clustering) – lowess, wavelet [9], and quantreg [10] – give better overall results while among the other eight estimation-performing algorithms (EP, which do clustering) CGHseg [11] and CBS [12] have the best overall ROC performance. At moderate to low signal-to-noise ratios (SNRs) ChARM [13], ACE [15] and HMM [16] are the low performers. For the GBM comparison LJKP finds that ChARM and HMM disagree with others in identified global aberrant regions, and that CGHseg, GLAD [17] wavelet, GA [18] and quantreg are best in locating amplifications and CLAC [19], ChARM and HMM are the poorest. Among the algorithm tested by LJKP, only CLAC and ACE provide quantitative statistics for the identified aberrations. This important feature cannot be added to algorithms based only on smoothing, such as the three SOs. In the rest of this section, we first duplicate the ROC tests in LJKP and test SAD against the eight EP algorithms and then make detailed comparisons of SAD with the top LJKP performers − CGHseg, CBS and GLAD − in terms of accuracy, speed and memory.

A. Accuracy

We calculate the ROC curves of SAD in the same way as in LJKP except that, for better statistics, we generate 10,000 rather than 100 simulated chromosomes (of 100 probes) for each parameter set. The results, shown in Fig. A2, indicate that a higher t_{min} is more suitable for easy situations (wide aberration and large SNR) while a lower t_{min} better facilitates aberration detection in difficult situations (narrow aberration and small SNR). Referring to Fig. 2 in LJKP, we see that in easy situations SAD(2.5-4,100) match the performance of CGHseg, the best among the eight EPs, and in difficult situations SAD(1.5-2.5,100) outperforms CGHseg. Compared to the other seven EPs, SAD performs significantly better. How to select t_{min} is discussed in details in Methods.

Fig. 2 focuses on the important part of the comparison and shows SAD(t_{min},100), t_{min}=1.5, 2, and 2.5, side by side with the eight EPs (as given in LJKP) for two difficult situations, (SNR,Width)=(2,5) and (1,10). We understand that the LJKP results are obtained from the relevant software packages using default parameters, which may not be optimal for the cases considered.

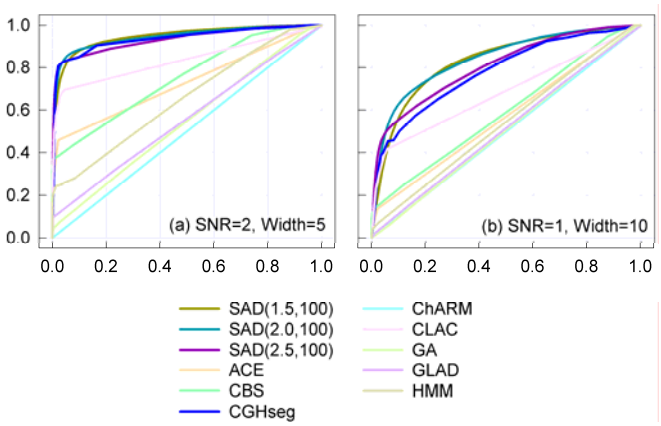

Fig. 2 ROC curves of SAD(1.5,100), SAD(2.0,100) and SAD(2.5,100) compared with the eight EPs in LJKP in two difficult situations, (SNR,Width) = (2,5) in (a) and (1,10) in (b)

PM differs from LM in clustering order. Although a smaller N_s facilitates higher computation speed, it incurs more error. This error is estimated in Appendix B and, as suggested by Fig. A3, is negligible for $N_s \ge 100$.

B. Speed and Memory

In Fig. 3 we compare SAD(10,100) to CGHseg, CBS and GLAD (with their default parameters) in speed and memory. Simulated chromosomes for this calculation are generated as in the following. SNR=2. Each simulated chromosome has either one or two amplifications. To plant the amplifications, each chromosome is divided into five same-width sections first. For the one-amplification cases, the second section is amplified. For the two-amplification cases, the second and the forth are amplified. For speed, we measure computation time τ (Fig. 3a). The difference in τ between one and two amplifications reflects the dependence of speed on genomic profiles. Memory test is read from the Processes tab of Windows Task Manager and involves two steps: data loading and data processing. The reading between the two steps, denoted by κ_d, is memory used

for program and data. The maximum reading amid data processing is $\kappa_o=\kappa_d+\kappa_p$ (Fig. 3c), where κ_p (Fig. 3d) is the maximum addition for data processing. We derive the scaling exponent γ_τ (Fig. 3b) in the power-law in τ *vs.* N from data in Fig. 3a, and the scaling exponent γ_p (Fig. 3e) in κ_p *vs.* N from data Fig. 3d.

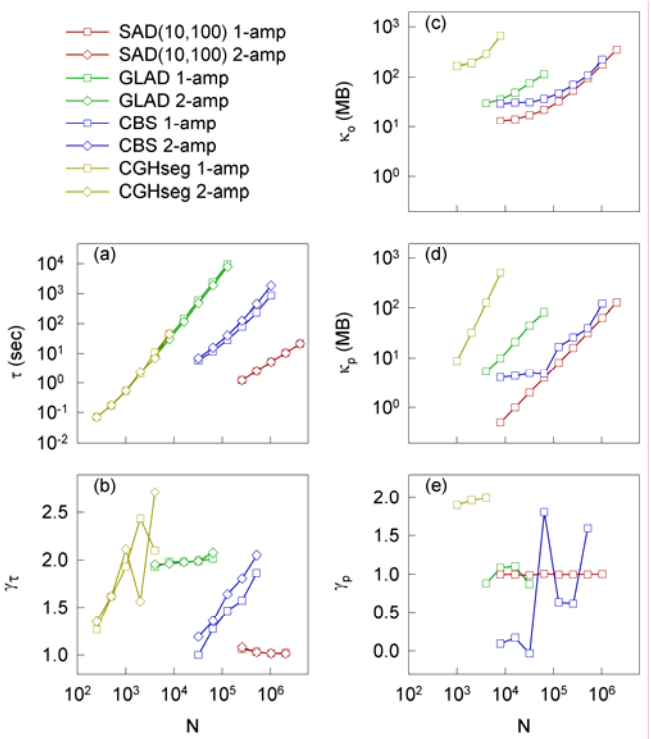

Fig. 3 Comparisons of SAD to CGHseg, CBS and GLAD in speed and memory. For speed, (a) shows computation time τ versus N and (b) shows the power-law exponent γ_τ of (a). For memory, (c) shows maximum overall memory κ_o versus N, (d) shows maximum data-processing memory κ_p versus N and (e) shows the power-law exponent γ_p of (d)

Fig. 3a shows SAD to be vastly faster than the others. Fig. 3b shows that τ for GLAD scales as $\mathcal{O}(N^2)$, and for CBS, though claimed to be $\mathcal{O}(N)$ at low resolution [20], to trend up as N increases and become $\mathcal{O}(N^2)$ at $N\approx 5\times10^5$. Speed-wise CGHseg is similar to GLAD and is generally $\mathcal{O}(N^2)$ above $N\approx 10^3$. SAD is constantly $\mathcal{O}(N)$. Speed dependence on genomic profile is significant for CBS, minor for GLAD and CGHseg, and negligible for SAD.

Figs. 3c and 3d show that SAD requires the least memory, overall (κ_o) or data-processing (κ_p). Fig. 3e shows memory requirement in SAD and GLAD scales as $\mathcal{O}(N)$, in CBS more or less so while displaying considerable irregularity, and in CGHseg scales as $\mathcal{O}(N^2)$. Moreover, in a computer with 2 GBs of memory, CGHseg ceases to function when N approximately exceeds 16,000.

In a test using real data, we run SAD(10,100) on a 1.8 million-probe Affymetrix Genome-Wide Human SNP Array 6.0 hybridized with a colorectal cancer sample [21], and measure τ=9 seconds and κ_o =323 MBs.

C. Validation on a Low-Resolution Dataset

As a low-resolution validation test and demonstration of t_{min} selection and utility, we apply SAD to a public dataset [22] which corresponds to 15 human cell strains from the NIGMS Human Genetics Cell Repository. Each cell strain has either one or two alterations, as identified by spectral karyotyping, and has been hybridized with an array CGH of 2276 BACs, spotted in triplicate.

We use (11) to select a value of t_{min}. For trisomic segments of the dataset, SNR≈0.6/0.08, where 0.6 is approximately the *log2*-ratio of a trisomic segment and 0.08 is $\tilde{\sigma}$ estimated using (7). To identify a trisomic aberration which is 2 probes or wider (because in its design, SAD identifies single-probe aberrations as outliers), t_{min}<10.5 is required. We therefore use SAD(10,100) for this calculation.

This dataset has previously been studied with GLAD [17] and CBS [12]. These are compared side-by-side with SAD(10,100) in Table 1. The performance of SAD(10,100) stand out in two features. (1) SAD(10,100) gives far fewer false positives: the average numbers of false positive breakpoints per cell strain are 4/15, 46/15, 26/15, 37/9 and 16/9 for, SAD(10,100), GLAD(λ'=8), GLAD(λ'=10), CBS(α=0.01) and CBS(α=0.01), respectively. (2) Among the three, SAD is the only one to give an aberrance z to each of the aberrant segments, including whole-chromosome ones, for showing extent of aberrance. Eight aberrations on six cell strains are whole-chromosome: GM00143/18, GM02948/13, GM03576/2, GM03576/21, GM04435/16 GM04435/21, GM07408/20 and GM10315/22. For these GLAD and CBS are silent because they are based on breakpoint detection within a chromosome.

The single-probe aberration on GM01535/12 is not detected by SAD(10,100) because t_{min}=10 is optimal for aberrations of two probes or wider. A smaller t_{min} value, say 8, identifies this aberration as an outlier but gives a total of 8 false-positive breakpoints rather than 4. GLAD identifies this aberration as an AWS outlier. CBS fails to detect it.

Two-probe or wider aberrations can be detected by SAD(10,100). As an example, the aberration on GM03563/9 which comprises the first two probes is successfully detected. So can it be detected with GLAD. CBS once again fails this detection.

The 4 false-positive breakpoints detected by SAD(10,100) come from 2 false-positive segments which happen to be at the same location on 2 different cell strains. Although karyotyping is not positive, these segments look monosomic to visual expertise.

Like the other algorithms, SAD(10,100) fails to detect the monosomic region on GM07081/15. This is because the aberration is not detected by array CGH technology [22].

For this particular dataset, the following limitation is shared by all algorithms: Breakpoints can be mistakenly located when probes at the joins of two segments have ambiguous *log*2-ratios. For instance, with SAD(10,100), RP11-88j19 of GM01535/5

TABLE I
RESULTS OF SAD(10,100) ON SNIJDERS' DATASET [22] COMPARED WITH GLAD(λ'=8), GLAD(λ'=10) [17], CBS(α=0.01) AND CBS(α=0.01) [12]. THE CONVENTION FOR SPECIFYING CHROMOSOME IN THE FIRST COLUMN IS AS FOLLOWS: A NUMBER INDICATES THE CHROMOSOME NUMBER ON WHICH THE ABERRATION IS PRESENT; A NUMBER WITH A '*' INDICATES THE ABERRATION IS WHOLE-CHROMOSOME; THE TERM 'FALSE' INDICATES THE NUMBER OF FALSE-POSITIVE BREAKPOINTS DETECTED. IN COLUMNS 2-6, 'YES' MEANS THE ABERRATION ON THE CHROMOSOME IS IDENTIFIED, 'NO' MEANS IT IS NOT, '–' MEANS NO PREDICTION WAS GIVEN AND 'NA' MEANS THE CELL STRAIN IS NOT DISCUSSED. IN COLUMN 2, THE NUMBERS IN SQUARE PARENTHESES ARE *Z* VALUES GIVEN BY SAD FOR SHOWING EXTENT OF ABERRANCE

Cell strain/chromosome	SAD(10,100)	GLAD(8)	GLAD(10)	CBS(.01)	CBS(.001)
GM00143/18*	[54.1]	–	–	NA	NA
GM00143/False	0	8	0	NA	NA
GM01524/6	Yes[35.3]	Yes	Yes	Yes	Yes
GM01524/False	0	0	0	6	2
GM01535/5	Yes[20.4]	Yes	Yes	Yes	Yes
GM01535/12	No	Yes	Yes	No	No
GM01535/False	0	0	0	2	0
GM01750/9	Yes[25.3]	Yes	Yes	Yes	Yes
GM01750/14	Yes[21.1]	Yes	Yes	Yes	Yes
GM01750/False	0	0	0	1	0
GM02948/13*	[39.5]	–	–	NA	NA
GM02948/False	0	1	0	NA	NA
GM03134/8	Yes[45.6]	Yes	Yes	Yes	Yes
GM03134/False	2	4	4	3	1
GM03563/3	Yes[48.1]	Yes	Yes	Yes	Yes
GM03563/9	Yes[18.6]	Yes	Yes	No	No
GM03563/False	0	8	4	8	5
GM03576/2*	[84.5]	–	–	NA	NA
GM03576/21*	[48.7]	–	–	NA	NA
GM03576/False	0	0	0	NA	NA
GM04435/16*	[59.5]	–	–	NA	NA
GM04435/21*	[36.6]	–	–	NA	NA
GM04435/False	2	2	2	NA	NA
GM05296/10	Yes[47.5]	Yes	Yes	Yes	Yes
GM05296/11	Yes[37.4]	Yes	Yes	Yes	Yes
GM05296/False	0	8	6	3	0
GM07081/7	Yes[58.1]	Yes	Yes	Yes	Yes
GM07081/15	No	No	No	No	No
GM07081/False	0	6	6	1	0
GM07408/20*	[113.6]	–	–	NA	NA
GM07408/False	2	2	2	NA	NA
GM10315/22*	[31.5]	–	–	NA	NA
GM10315/False	0	3	0	NA	NA
GM13031/17	Yes[30.4]	Yes	Yes	Yes	Yes
GM13031/False	0	4	4	5	3
GM13330/1	Yes[44.8]	Yes	Yes	Yes	Yes
GM13330/4	Yes[43.7]	Yes	Yes	Yes	Yes
GM13330/False	0	0	0	8	5

and RP11-237j07 of GM05296/10 are mistakenly identified as disomic. The latter is also reported with GLAD.

D. Validation on a High-Resolution Dataset

As a high-resolution validation test and demonstration of the effects of tuning t_{min}, we apply SAD on the 500K copy number sample data provided by Affymetrix (http://www.affymetrix.com). The sample data consist of 9 Tumor/Normal Pairs derived from human cancer cell lines and X Chromosome titration set (3X, 4X, and 5X). We apply SAD to the 262,217-SNP-long NSP dataset from the (CRL-5868D, CRL-5957D) pair, and show the results in Fig. 4 where we have zoomed in at the 8th chromosome. Parameter sets we use are SAD(50,100), SAD(30,100), and SAD(10,100). With SAD(50,100), which is suitable for broad aberrations, two segments are identified with *z* values of 77.9 and 196.0. With SAD(30,100), four narrow segments labeled 1, 2, 3, and 4 are identified with *z* values of 52.1, 46.1, 51.0, and 42.2 respectively, and are magnified in the insets. With SAD(10,100), even narrower segments labeled 5, 6, and 7 are spotted with *z* values of 16.5, 39.7, and 31.5, respectively. Segment 5 is the narrowest of all identified structures and comprises only two probes.

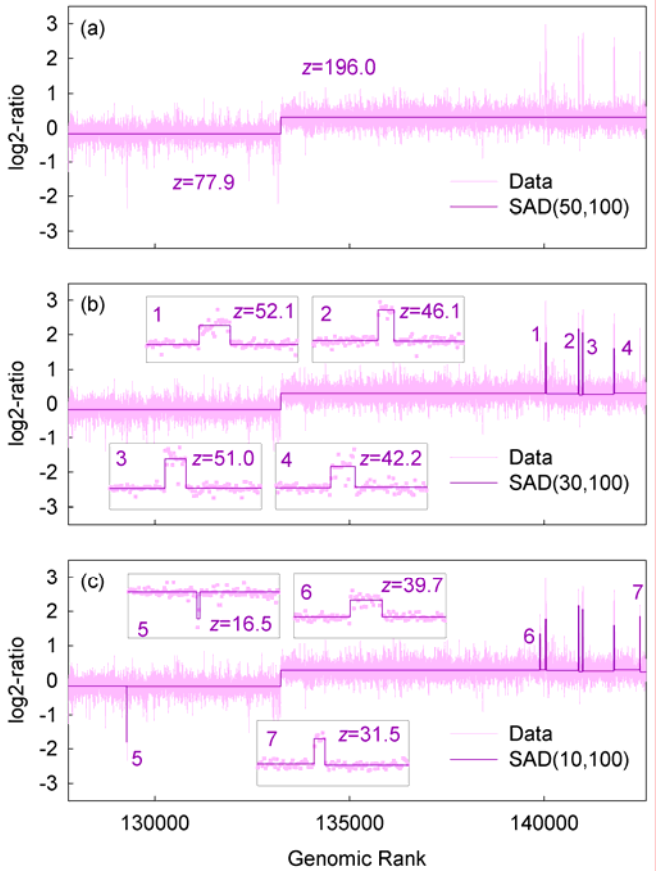

Fig. 4 A high-resolution validation test by SAD. Data are in pink and SAD predictions are in dark purple. The three frames are for the three parameter sets: SAD(50,100), SAD(30,100) and SAD(10,100). Details, with z values, of structures at the numbered sites are magnified in the insets. The two z values in the first frame, refer to the segments in the two-step structure

IV. Discussions

We have developed SAD for genome segmentation for copy number analysis and have demonstrated that, compared with existing algorithms, SAD is more accurate, far faster and parsimonious in memory use. SAD owes its computational efficiency to the way each piece of a raw datum is viewed: Not as a number that contributes to the collective nature of a segment, but as an independent statistical event that gives a PDF for predicting the segment. The rest of the algorithm is universal: the assumption that measurement errors are normally distributed, a clustering procedure based on Gaussian merging and t-statistic.

The statistical aspect of the view leads to clear statistical interpretation of its predictions and easy parameter tuning. The statistical significances of NHB and NHS are t and z, respectively. The former reflects the reliability of a breakpoint while the latter, the extent of aberrance of a segment. Parameter tuning is easy because t_{min} and N_s are intuitively comprehensible. t_{min} is the statistical level for rejecting NHB and NHS. It also defines the lower bound of aberration width if SNR is known. N_s is a balance between speed and accuracy and facilitates PM in which SAD is $\mathcal{O}(N)$ in computation time and memory requirement. We show in our test that 100 is a good default value which incurs little error.

Because a user-specific t_{min} can yield a significant gain in accuracy, as is demonstrated in Figs. 2, A2 and Table I, SAD users are advised to select t_{min} based on either their requirement of statistical significance (using (9) and (10)) or the typical SNR of data combined with their aberration width of interest (using (11)).

Appendix

A. The Assumption

SAD is based on a universal assumption that measurement errors are normally distributed. In a microarray dataset, each datum (in the form of *log*2-ratio) from a probe is viewed as an independent observation. The set of probes on a segment of a copy number are therefore assumed to exhibit normally distributed values of *log*2-ratio centered at the true value.

In Fig. A1 we test our assumption using the 500K copy number sample data provided by Affymetrix (http://www.affymetrix.com). Fig. A1a shows the genomic profile of chromosome 2 from the (CRL-5868D,CRL-5957D) STY pair. Two sections of the chromosome, which by visual expertise belong to two segments of different copy numbers, are examined. In Fig. A1b, the *log*2-ratio distributions of the two sections are shown with two Gaussians: $G(y;0.36,0.22^2)$ and $G(y;-0.12,0.22^2)$. The two distributions resemble the corresponding Gaussians and the variances are similar in spite of different copy numbers. In Fig. A1c, we show that, in terms of $\tilde{\sigma}$ estimation, the two sections are similar to a simulated one. The simulated section is 8000 probes wide and is randomly generated using $G(y;0,0.22^2)$. For each n taken, each of the three sections is divided into subsections of n probes, plus a remainder. Each subsection yields a $\tilde{\sigma}$. The entire set of subsections gives a mean and a standard deviation which are denoted by a square and an error bar, respectively. The three sections show similar means and similar variances. The means appear to be constant of n.

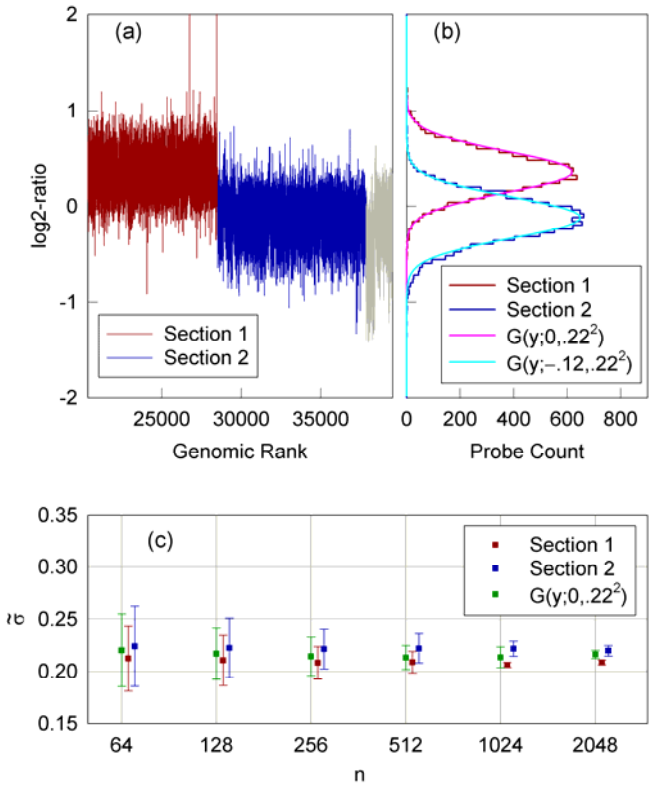

Fig. A1 A test of our assumption on the 500K copy number sample data provided by Affymetrix (http://www.affymetrix.com). (a) is the genomic profile of chromosome 2 from the (CRL-5868D,CRL-5957D) STY pair. The two colored sections are selected for analysis. (b) is *log*2-ratio distributions of the selected sections compared with two Gaussians: $G(y;0.36,0.22^2)$ and $G(y;-0.12,0.22^2)$. (c) is $\hat{\sigma}$ (estimated from subsections of n probes) of the two sections compared with a simulated section randomly generated with $G(y;0,0.22^2)$

B. Accuracy Assessment

In Fig. A2, we assess the accuracy of SAD using ROC curves. The ROC curves are calculated in the same way as in LJKP [8]. Aberration widths of 5, 10, 20 and 40 probes and SNRs of 1, 2, 3 and 4 are investigated. SNR was defined as the mean magnitude of the aberration (i.e. signal) divided by the standard deviation of the superimposed Gaussian noise. For each aberration width and SNR, we generated 10,000 simulated chromosomes (rather than 100 of LJKP), each consisting of 100 probes and with a square-wave signal profile added to the center of the chromosome. True positive rate (TPR) is defined as the number of probes inside the aberration whose fitted values are above the threshold level divided by the number of probes in the aberration. False positive rate (FPR) is defined as the number of probes outside the aberration whose fitted values are above the threshold level divided by the total number of probes outside the aberration. In order to compute the ROC curve, we vary the threshold value for aberration from the minimum *log*2-ratio value to the maximum. Each threshold value results in a TPR and a FPR, represented by a point on the ROC curve. A set of TPRs and FPRs are then plotted to reveal the algorithm's ROC profile for the particular aberration width and SNR.

For each situation of aberration width and SNR, we test parameter sets of SAD(t_{min},−), t_{min}=1.5 to 4.0 in intervals of 0.5. The panels are arranged so that data become more difficult (lower SNRs and narrower aberrations) to analyze as one goes from upper-left to lower-right. We find in Fig. A2 that a higher t_{min} is more suitable for easy situations while a lower t_{min} facilitates aberration detection in difficult situations.

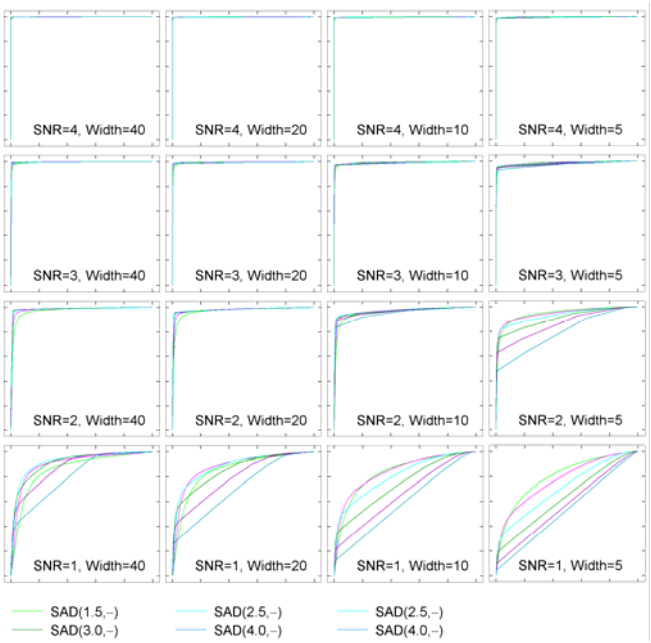

Fig. A2 ROC curves of SAD(t_{min},−) in 16 different situations: Aberration widths of 5, 10, 20 and 40 and SNRs of 1, 2, 3 and 4. t_{min} values used are 1.5, 2.0, 2.5, 3.0, 3.5 and 4.0.

Referring to Fig. 2 of LJKP, we see that in easy situations SAD(2.5-4,−) matches the performance of CGHseg, the best among the eight EPs, and in difficult situations SAD(1.5-2.5,−) outperform CGHseg.

PM of SAD differs from LM in clustering order. We compare PM to LM in accuracy as follows. Based on the aberration seen on chromosome 3 of MCF7, one of the three breast cancer cell lines evaluated using the Affymetrix 100K SNP platform [23], the specifics of this calculation are: N=8000 and SNR=3; A 250-probe-wide amplification is planted at the center of the chromosome; 10,000 simulated chromosomes are generated for each parameter set tested. SAD(t_{min},−), SAD(t_{min},400), SAD(t_{min},200), SAD(t_{min},100) and SAD(t_{min},50) are tested with t_{min}=4, 8 and 16. The results are shown in Fig. A3, where we have zoomed in on the top-left corner of ROC space. With t_{min}=4, in terms of area under curve (AUC), SAD(4,100) differs from SAD(4,−) by 6.3×10^{-6}. With t_{min}=8, the five curves are indistinguishable. With t_{min}=16, which is not shown in Fig. A3, the curves get even closer. We therefore consider the error incurred by N_s=100 negligible.

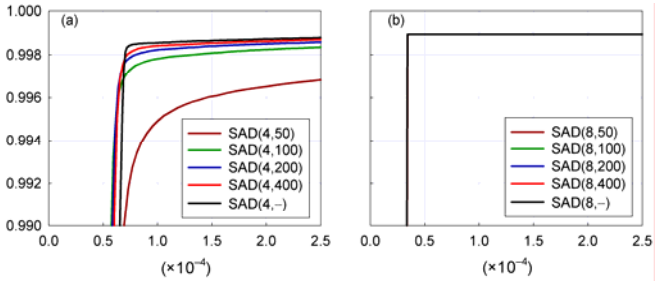

Fig. A3 Comparison of PM (N_s=50, 100, 200 and 400) with LM in error by ROC. We zoom in on the top-left corner of ROC space. (a) is for t_{min}= 4 and (b) for t_{min}= 8. In (b) the 5 curves are indistinguishable

ACKNOWLEDGMENT

This work is partly supported by grants97-2112-M-008-013 from the National Science Council (ROC) and the Cathy General Hospital-NCU Collaboration Grant 97-CGH-NCU-A1.

REFERENCES

[1] Solinas-Toldo, S. *et al*. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, 20, 399–407.
[2] Pinkel, D. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.
[3] Pinkel, D. and Albertson, D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, 37, Suppl 11–17.
[4] Pollack, J.R. *et al*. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, 23, 41–46.
[5] Brennan, C. et al. (2004) High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res.*, 64, 4744–4748.
[6] Lucito, R. *et al*. (2003) Representational oligonucleotide microarray analysis: a highresolution method to detect genome copy number variation. *Genome Res.*, 13, 2291–2305.
[7] Ishkanian, A.S. *et al*. (2004) A tiling resolution DNAmicroarray with complete coverage of the human genome. *Nat. Genet.*, 36, 299–303.
[8] Lai, W.R., Johnson, M.D., Kucherlapati, R., & Park, P.J. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, 21, 3763–3770.
[9] Hsu, L. *et al*. (2005) Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6, 211–226.
[10] Eilers, P.H.C. and de Menezes, R.X. (2005) Quantile smoothing of array CGH data. *Bioinformatics*, 21, 1146–1153.
[11] Picard, F., Robin, S., Lavielle, M., Vaisse, C. & Daudin J. (2005) A statistical approach for array CGH data analysis. *BMC Bioinforma.*, 6, 27.
[12] Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5, 557–572.
[13] Myers, C.L., Dunham, M.J., Kung, S.Y. & Troyanskaya, O.G. (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, 20, 3533–3543
[14] Wang, P., Kim, Y., Pollack, J., Narasimhan, B. & Tibshirani, R. (2005) A method for calling gains and losses in array CGH data. *Biostatistics*, 6, 45–58.
[15] Lingjærde, O.C., Baumbusch, L.O., Liestøl, K., Glad, I.K. & Børresen-Dale A. (2005) CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics*, 21, 821–822.
[16] Fridlyand,J. *et al*. (2004) Hidden Markov models approach to the analysis of array CGH data. *J. Multivariate Anal*., **90,** 132–153
[17] Hupé, P., Stransky, N., Thiery, J., Radvanyi, F. & Barillot, E. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20, 3413–3422.
[18] Jong, K. *et al*. (2003) Chromosomal breakpoint detection in human cancer. In *Lecture Notes in Computer Science*, Springer-Verlag, Berlin, Vol. 2611, pp. 54–65.
[19] Wang, P., Kim, Y., Pollack, J., Narasimhan, B. & Tibshirani, R. (2005) A method for calling gains and losses in array CGH data. *Biostatistics*, 6, 45–58.
[20] Venkatraman, E.S. and Olshen, A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23, 657-663.
[21] Lee, Hsin-Chung. Private Communication.
[22] Snijders, A.M. *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, 29, 263–264.
[23] Ting, J.C., Ye, Y., Thomas, G.H., Ruczinski, I. & Pevsner, J. (2006) Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan. BMC *Bioinformatics*, **7**, 25