

# Joint Estimation of DNA Copy Number from Multiple Platforms

Nancy R. Zhang <sup>1</sup>, Yasin Senbabaoglu <sup>2</sup>, and Jun Z. Li <sup>3</sup>

## Abstract

DNA copy number variants (CNV) are gains and losses of segments of chromosomes, and comprise an important class of genetic variation. Recently, various microarray hybridization based techniques have been developed for high throughput measurement of DNA copy number. In many studies, multiple technical platforms or different versions of the same platform were used to interrogate the same samples; and it became necessary to pool information across these multiple sources to derive a consensus molecular profile for each sample. An integrated analysis is expected to maximize resolution and accuracy, yet currently there is no well formulated statistical method to address the between-platform differences in probe design, assay methods, sensitivity, and analytical complexity.

The conventional approach is to apply one of the CNV detection (a.k.a. “segmentation”) algorithms to search for DNA segments of altered signal intensity. The results from three platforms are combined after segmentation. Here we propose a new method, Multi-Platform Circular Binary Segmentation (MPCBS), which pools statistical evidence across platforms during segmentation, and does not require pre-standardization of different data sources. It involves a weighted sum of  $t$ -statistics, which arises naturally from the generalized log-likelihood ratio of a multi-platform model. We show by comparing the integrated analysis of Affymetrix and Illumina SNP array data with fosmid clone end-sequencing results on 8 HapMap samples that MPCBS achieves improved spatial resolution, detection power, and provide a natural consensus across platforms. We also apply the new method to analyze the multi-platform data from TCGA.

The R package for MPCBS is registered on R-Forge under project name MPCBS.

## 1 Introduction

In recent years, more and more genetic studies have relied on collecting genome-scale data on DNA variants. With the rapid influx of large datasets came the increasingly common problem of data integration when multiple technical platforms (or different versions of the same platform) were used to interrogate the same biological samples. For example, the Cancer Genome Atlas (TCGA) project, an NIH-funded initiative to characterize DNA, RNA, and epigenetic abnormalities in tumors, have adopted three independent platforms for studying

---

<sup>1</sup>To whom correspondence should be addressed. Department of Statistics, Stanford University. Email: nzhang@stanford.edu

<sup>2</sup>Program in Bioinformatics, University of Michigan.

<sup>3</sup>Department of Human Genetics, University of Michigan.

DNA copy number variants (CNVs) in its pilot phase: Affymetrix SNP 6.0 arrays, Illumina HumanHap 550K SNP arrays, and Agilent CGH 244K arrays. The conventional approach for analyzing these data is to apply one of the CNV detection (a.k.a. “segmentation”) algorithms to search for genomic intervals of altered signal intensity within the data from each platform separately. The segmentation results from three platforms are then combined. However, when the platforms disagree on the calling of a CNV, it is difficult to decide what the consensus should be. Furthermore, the reported DNA copy numbers (i.e. the location and magnitude of the changes) are often different in different platforms. At the fundamental levels, the three platforms represent three distinct marker panels and vastly different molecular assay methods:

- Illumina produces allele-specific data, Agilent produces only the total intensity, whereas Affymetrix has both allele-resolved SNP probes and invariant CNV probes, thus effectively containing two sub-platforms.
- Agilent produces two-color ratio data in a test/reference format, while the other two measure each sample independently.
- In regions of high-fold amplification, Illumina and Affymetrix tend to have more pronounced signal saturation. In fact, all three platforms estimate the true levels of copy number change with different scaling factors, which may be non-linear and may vary across chromosomes or samples.
- The three methods produce data values with distinct noise characteristics, with different proportions of low-quality SNPs and distinct local signal trends that are partly due to the sample amplification procedures used.
- For some, such as the Illumina data, the default normalization procedure is not tailored to copy number analysis.

In short, each of the three platforms has its advantages and disadvantages, but together they produce a balanced genomewide survey for each sample, and represent a much denser coverage than each platform does alone. If the data from the three platforms are separately segmented, it is difficult to combine their respective segment summaries because, for the same underlying event, they will report different magnitudes, with different boundaries and different degrees of uncertainty. An integrated analysis, where information from all platforms are used at the same time to detect CNVs and to estimate the levels of change is expected to maximize resolution and accuracy. Currently, however, there is no well formulated statistical method to address the between-platform differences in probe design, assay methods, sensitivity, and analytical complexity. Simply combining the three data series without proper normalization will not yield better segmentation results. Yet when the underlying true copy number is not known, it is difficult to determine how to normalize the data given the uneven coverage between the platforms at any genomic region.

In order to tackle the increasingly common problem of data integration across multiple sources we propose a new method, multi-platform Circular Binary Segmentation (MPCBS). This method relies on a weighted  $t$  statistic to scan for copy number changes. MPCBS sums statistical evidence across platforms with proper scaling, and does not require a pre-standardization of different data sources. The statistics are based on maximizing the likelihood of a simple multi-platform model, with the dimension of the model (i.e. the number of segments) chosen by maximizing a generalized form of the modified BIC criterion proposed in Zhang and Siegmund (2007). Platform specific quantities such as noise variances and response ratios are also estimated by our method. Importantly, the method provides a single, platform-free consensus profile for each sample for downstream analyses.

## 2 Multiplatform Model and Methods Overview

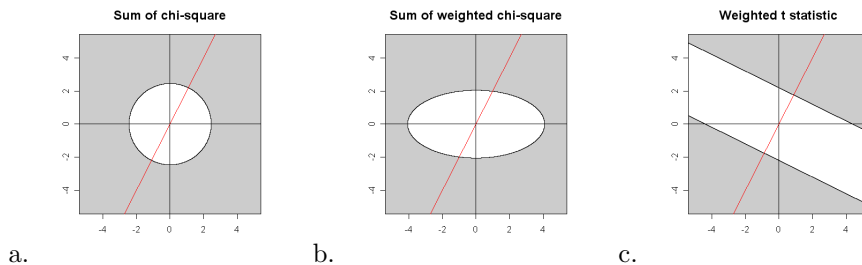
Let the platforms be indexed by  $k = 1, \dots, K$ , with  $K$  being the total number of platforms. We observe data  $\mathbf{y}_k = y_{k1}, \dots, y_{kn_k}$  for the  $n_k$  snps/clones on the  $k$ -th platform, which have ordered locations  $(t_{k1}, \dots, t_{kn_k})$  along a chromosome. We assume that for each platform, the data has been normalized to be centered at 0 for “normal” copy number and to have Gaussian (or near-Gaussian) noise. Actual data must be transformed with missing values imputed, sometimes with extreme outliers truncated in order to conform to Gaussian noise. In some studies, the “normal” diploid state of the genome is difficult to determine, such as when an entire chromosome has been amplified. When this occurs, other types of information, such as allelic ratios from SNP arrays, or intensity ratios from two-color aCGH experiments, will be needed to help assign the correct absolute copy number to each segment. Such complications are expected to affect all platforms. Here we deal with the integration of multiple platforms in detecting *changes* in CNV and only need to assume that the baseline “normal” state is shared in common across platforms.

The fact that all  $\{\mathbf{y}_k : k = 1, \dots, K\}$  are assaying the same biological sample implies that at any genomic location  $t$  there is only one true underlying copy number  $\mu_t$  for all platforms. we define the observed intensity level for the  $i$ -th probe of the  $k$ -th platform consisting of a signal  $f_k(\mu_{t_{k,i}})$  plus a noise term that has platform specific variance  $\sigma_k^2$ . Specifically, we assume the following model for the data:

$$y_{ki} = f_k(\mu_{t_{k,i}}) + \epsilon_{k,i}, \quad (2.1)$$

where the noise term  $\epsilon_{k,i}$  are independently distributed  $N(0, \sigma_k^2)$ . We call  $f_k(\cdot)$ , which quantifies the dependence of the mean intensity on the underlying copy number, the response function of platform  $k$ .

We model the true copy number as a piecewise constant function, i.e. constant within a segment, and yet may change to a different level at a “change-point”. For a chromosome of length  $T$ , we assume that there exists a series of



**Figure 1.** Comparison of the null hypothesis rejection regions between the sum of chi-square statistic (3.6), the weighted sum of chi-square statistic (3.7), and the weighted  $t$ -statistic (3.3) on  $K = 2$  platforms. In all figures, the axes are the magnitudes of the  $X$  variables (3.4) for platforms 1 and 2. A significance level of 0.05 is used to determine the decision boundaries of all three statistics. For Figures (b) and (c), weights of  $\delta_1 = 1$ ,  $\delta_2 = 2$  are used. The red line shows the direction of the weight vector  $\delta = (\delta_1, \delta_2)$ .

change-points  $0 = \tau_0 < \tau_1 < \dots, < \tau_m < T$  such that within each interval,

$$\mu_t = \theta_i, \quad t \in [\tau_i, \tau_{i+1}). \quad (2.2)$$

The magnitude parameters  $\theta = (\theta_0, \dots, \theta_m)$  and change-points  $\tau = (\tau_1, \dots, \tau_m)$  are all unknown and, like the response functions, must be estimated from the data.

For this paper, we assume that the response function is linear, i.e.  $f_k(\mu) = r_k \mu$ . The parameter  $r_k$ , which we call the response ratio, describes the ratio between the change in signal intensity and the underlying copy number change for platform  $k$ . The linearity assumption allows for simple and intuitive test statistics and fast scanning algorithms. Empirically, the platform response functions are observed to be linear for low-amplitude changes, and nonlinear for high amplitude changes. The high-amplitude changes usually have high statistical significance and are relatively less affected by this simplification in modeling, and it is the low amplitude, statistically borderline cases where we hope to boost power through multi-platform integration.

When the platform specific response ratios  $r_k$  are known, the breakpoints  $\tau$  and true copy numbers  $\theta$  can be estimated through a likelihood based recursive segmentation procedure that builds on the conceptual foundations of Olshen *et al.* (2004) and Vostrikova (1981), which we describe in Section 3.1. Conversely, when  $\tau$  and  $\theta$  is given,  $f_k$  can also be easily estimated using the procedures described in Section 3.4. Since both are usually unknown, we propose the iterative procedure described in Section 3.5.

### 3 Methods

#### 3.1 Pooling Evidence by Weighted $t$ -statistics

First consider the case where the goal is to test whether there is a CNV at a window from  $s$  to  $t$ . Under the *null* hypothesis that there is no CNV, the data within this region should have baseline mean  $f_k(0) = 0$ , i.e.

$$H_0 : y_{ki} \sim N(0, \sigma_k^2) \quad \text{for } k = 1, \dots, K; \quad \text{and } i : s \leq t_{ki} < t. \quad (3.1)$$

If there is a gain (or loss) of magnitude  $\mu$ , each platform should respond with signal  $f_k(\mu) = r_k\mu$ . The signal is a mean shift in a *common direction* for all platforms, with the observed magnitude of shift being  $r_k\mu$  for platform  $k$ , i.e.

$$H_A : y_{ki} \sim N(r_k\mu, \sigma_k^2) \quad \text{for } k = 1, \dots, K; \quad \text{and } i : s \leq t_{ki} < t. \quad (3.2)$$

Since the generalized likelihood ratio statistic maximizes the power over all statistical tests for this model, we will use the likelihood based framework to test this hypothesis. Let  $n_k(s, t) = |\{i : t_{k,i} \in (s, t]\}|$  be the number of probes from the  $k$ -th platform that falls within  $(s, t]$ . Let  $\bar{y}_{k,(s,t]}$  denote the mean intensity of probes that map within  $(s, t]$ . It can be shown (see the appendix) that under this formulation, the log generalized likelihood ratio statistic is a weighted sum of platform specific terms:

$$Z(s, t) = \frac{\left[ \sum_{k=1}^K \delta_{k,s,t} X_{k,s,t} \right]^2}{\sum_{k=1}^K \delta_{k,s,t}^2}, \quad (3.3)$$

where

$$X_{k,s,t} = \frac{\bar{y}_{k,[s,t]} - \bar{y}_{k,[s,t]^c}}{\sigma_k \sqrt{n_k(s, t)^{-1} + [n_k - n_k(s, t)]^{-1}}}, \quad (3.4)$$

if  $\sigma_k$  is estimated from the data, is the  $t$ -statistic for testing for a change using only the data from platform  $k$ . The weights

$$\delta_{k,s,t} = r_k \sqrt{n_k(s, t)} / \sigma_k \quad (3.5)$$

is proportional to the response ratio  $r_k$ , the square root of the number of probes from that platform that falls into  $[s, t)$ , and the inverse of the error standard deviation  $\sigma_k$ . When there is only one platform, the statistic (3.3) is equivalent to the chi-square statistic used in the Circular Binary Segmentation algorithm of Olshen *et al.* (2004). Usually  $\sigma_k$  is unknown and must be estimated from the data as well, we replace it with an estimate  $\hat{\sigma}_k$  in (3.4) and (3.5). In the simplest case we assume a common variance for all probes of a given platform, the number of data points used to estimate  $\sigma$  is very large and thus  $\hat{\sigma}_k$  is very precise and for all practical purposes can be treated as a known quantity. In situations where  $\sigma^k$  is dependent on the underlying copy number or differs between genomic regions, a generalized likelihood ratio statistic similar to (3.3) can also be computed.

Note that the statistic (3.3), which we call the *weighted t-statistic*, is different from the sum-of-chisquares statistic proposed in Zhang *et al.* (2008) for multi-sample segmentation, where each sample comes from a different biological source assayed on the same experimental platform. The statistic used in Zhang *et al.* (2008) is the sum of chi-square from  $N$  samples,

$$Z^{SC}(s, t) = \frac{1}{N} \sum_{n=1}^N X_{n,s,t}^2. \quad (3.6)$$

Intuitively, one may be tempted to extend the above formula to the multi-platform case by proposing a weighted form

$$Z^{SWC}(s, t) = \frac{\sum_{k=1}^K \delta_{k,s,t}^2 X_{k,s,t}^2}{\sum_{k=1}^K \delta_{k,s,t}^2} \quad (3.7)$$

that does not treat all platforms equally. When pooling data across independent biological samples, we do not expect all samples to carry the same CNV, and often both deletions and amplifications can be observed between the samples at the same genome location. Thus, the statistic (3.6) should not “reward” agreement in direction of change between samples. But the drawback of the weighted version (3.7) is that it also does not reward agreement. In contrast, the statistic in (3.3) rewards agreement and penalizes disagreement. For example, consider the case of  $K = 2$ , where (3.3) simplifies to  $(\delta_1^2 X_1^2 + \delta_2^2 X_2^2 + 2\delta_1\delta_2 X_1 X_2)/2$ . If the signs of  $X_1$  and  $X_2$  agree, this statistic is always larger than (3.6), while if the signs disagree, it is smaller. This makes the weighted  $t$ -statistic more suitable for pooling evidence across multiple samples that come from the same biological source.

The difference between the three statistics is shown graphically in Figure 1, where we illustrate the simple case of two platforms with the response ratio of the second platform being twice that of the first platform. Note that all three statistics are functions of  $X = (X_1, X_2)$ , which, assuming that  $\sigma_k$  is known, is bivariate Gaussian with mean 0 and identity covariance matrix under the null hypothesis. Figures 1(a-c) show in gray the region in the  $(X_1, X_2)$  plane where the null hypothesis will be rejected. That is,  $X$  needs to fall in to the gray region to make a CNV call. For example, in Figure 1a, which depicts the situation in (3.6), the gray region is  $\{X : Z^{SC}(X) > t_\alpha^{SC}\}$ , where  $t_\alpha^{SC}$  is a threshold chosen for the test to have significance level  $\alpha$ . In Figure 1b, which depicts the situation in (3.7) the weights  $\delta_2/\delta_1 = 2$  favor evidence from  $X_2$  over evidence from  $X_1$ , giving an elliptical boundary. In Figure 1c, which depicts the situation in (3.3), the boundary of the rejection boundary is  $\{X : \delta'X > t_\alpha\}$ , which is perpendicular to the vector  $\delta_2/\delta_1$ . Importantly, note that (c) awards agreement between the two platforms, while (a,b) treat all quadrants of the plane equally. The statistic (3.3, Figure 1c) also allows one platform to dominate the others: In the case where the directions disagree, e.g. in the upper left or lower right quadrants, the consensus can still be made according to the dominant platform.

### 3.2 Recursive Segmentation Procedure

In the previous section, we described the statistic used to test whether a specific interval  $[s, t)$  constitutes a CNV. In reality, there can be multiple change-points in the chromosome copy number. To detect all change-points, we adopted a framework that is similar to Vostrikova (1981), Olshen *et al.* (2004), and Zhang and Siegmund (2007). Vostrikova (1981) proved the consistency of binary segmentation algorithms. Olshen *et al.* (2004) proposed an improvement, called circular binary segmentation, that works better in detecting small intervals of change in the middle of long regions. Zhang and Siegmund (2007) proposed a BIC criterion for deciding the number of segments. Both Olshen *et al.* (2004) and Zhang and Siegmund (2007) showed that these types of procedures work well on DNA copy number data. Two independent comparative reviews by Willenbrock and Fridlyand (2005) and Lai *et al.* (2005) concluded that the CBS algorithm of Olshen *et al.* (2004) is one of the best performing single platform segmentation methods. This motivated us to extend this approach to the case of multiple platforms.

The Multi-platform CBS (MPCBS) algorithm will be described in detail in the appendix. Here, we give an intuitive overview using the following notation: Let  $\mathcal{Z}$  be an ordered vector of likelihood ratio statistics, and let  $\mathcal{R}$  be the corresponding ordered set of segments  $\{(i, j) : 0 < i < j < T\}$ . For an ordered set  $\mathcal{Z}$ , we mean by  $\mathcal{Z}[i]$  the  $i$ -th element of  $\mathcal{Z}$ . We define by  $\mathcal{Z}[i : j]$  the ordered subset  $\{\mathcal{Z}[i], \mathcal{Z}[i + 1], \dots, \mathcal{Z}[j]\}$  if  $i \leq j$ , or the empty set if  $i > j$ . For any set  $S$ , we denote by  $|S|$  the number of elements in  $S$ . Let  $M$  be the maximum number of change-points tolerated, which is usually determined by computational resources.

The algorithm proceeds as follows:  $S_k$  is the list of estimated change-points in the  $k$ -th iteration, which is initialized to contain only  $\{0, T\}$ . The entire dataset is scanned for the window  $[s^*, t^*)$  that maximizes  $Z(s, t)$ , that is, where the evidence for a change is the strongest. This window is added to  $S_k$ . Then, the region to the left of  $s^*$ , between  $s^*$  and  $t^*$  and to the right of  $t^*$  are each scanned for a sub-segment that maximizes  $Z(s, t)$ , these maximum values are called  $Z_L$ ,  $Z_C$ , and  $Z_R$  respectively. The corresponding locations of the maximum are  $R_L$ ,  $R_C$ , and  $R_R$ . These are kept in the ordered lists  $\mathcal{Z}$  and  $\mathcal{R}$ . At each iteration  $k$  of the algorithm, the region whose maximum weighted  $t$  statistic is the largest, i.e.  $i^* = \arg \max_i \mathcal{Z}[i]$ , is determined. The change-points from that region that achieve this maximum, i.e.  $(s^*, t^*) = \mathcal{R}[i^*]$ , are added to  $S_k$ . Since  $s^*, t^*$  splits a previously contiguous region into three regions,  $\mathcal{Z}$  and  $\mathcal{R}$  must be updated to include the maximal  $Z$  values and maximizing change-points for the new regions to the left, center, and right of the new change points. This process is repeated until  $S_k$  has at least  $M$  change-points in addition to  $\{0, T\}$ . Finally, the BIC criterion is used to determine a best estimate of the number of change-points and the final segmentation.

### 3.3 Estimating the Number of Segments

To estimate the number of change-points, we use a modified form of the classic BIC criterion that extends Zhang and Siegmund (2007). In Zhang and Siegmund (2007), it was shown that the modified BIC, when used on top of the CBS procedure of Olshen *et al.* (2004), improves its performance for DNA copy number data.

To describe the extension of Zhang and Siegmund (2007) to the case of multiple platforms, we first define several quantities. For a given genome position  $t$ , let  $n_k(t) = |\{i : t_{k,i} < t\}|$  be the number of probes from the  $k$ -th platform whose mapping position is smaller than  $t$ . Let

$$S_{k,t} = \sum_{i=1}^{n_k(t)} y_{k,i}$$

be the sum of the intensities of all probes in the region  $[0, t)$  for platform  $k$ . For a given set of estimated change-points  $\hat{\tau} = (\hat{\tau}_0 = 0 < \hat{\tau}_1 < \dots < \hat{\tau}_k = T)$ , let  $\delta_{k,i} = r_k \sqrt{n_k(\hat{\tau}_i)}/\sigma_k$ ,

$$X_{k,i} = \frac{S_{k,\hat{\tau}_i} - n_k(\hat{\tau}_i)S_{k,\hat{\tau}_{i+1}}/n_k(\hat{\tau}_{i+1})}{\hat{\sigma}_k \sqrt{n_k(\hat{\tau}_i)[1 - n_k(\hat{\tau}_i)/n_k(\hat{\tau}_{i+1})]}},$$

and

$$U_i(\hat{\tau}) = \frac{\sum_{k=1}^K \delta_{k,i} X_{k,i}}{\left(\sum_{k=1}^K \delta_{k,i}^2\right)^{1/2}}.$$

$X_{k,i}$  is simply the  $t$  statistic for testing that the change in mean at  $\hat{\tau}_i$  is not zero.  $U_i(\hat{\tau})$  is a weighted sum of  $X_{k,i}$ , just as (3.3) is a weighted sum of (3.4). Let  $N$  be the total number of distinct values in  $\{t_{k,i} : 1 \leq k \leq K, 1 \leq i \leq n_k\}$ , that is, the number of different probe locations from all  $K$  platforms. For any natural number  $n$ ,  $n!$  denotes the factorial of  $n$ . It is possible to show using arguments similar to Zhang and Siegmund (2007) that

$$\frac{1}{2} \sum_{i=1}^m U_i(\tau)^2 - \frac{1}{2} \sum_{i=0}^m \log \left[ \sum_{k=1}^K n_k(\hat{\tau}_i, \hat{\tau}_{i+1}) \right] - \log \frac{N!}{m!(N-m)!}. \quad (3.8)$$

is asymptotically within an  $O_p(1)$  error term of the Bayes factor for comparing the model with  $k$  change-points versus the null model. The number of change-points should be selected to maximize the BIC.

The first term of the modified BIC is the maximized likelihood, and is thus the same as the first term of the classic BIC criterion. The second and third terms are penalties that increase with the number of change-points. The second term penalizes the  $\theta$  parameters by summing up the logarithm of the effective sample size for estimating each  $\theta_i$ . The third term is the logarithm of the total number of ways to select  $m$  change-points from  $N$  possible values, which penalizes the change-points parameters  $\tau$ .

With the modified BIC, there is no need for a user specified p-value threshold. The trade-off between false-positive and false-negatives is automatically decided by the modified BIC.

### 3.4 Estimating the Platform-Specific Response Ratio

In this section we discuss the situation where the segmentation is known, and we would like to estimate the platform specific response ratios  $r = (r_1, \dots, r_K)$ , the baseline levels  $\alpha = (\alpha_1, \dots, \alpha_K)$ , and the underlying copy numbers  $\theta = (\theta_1, \dots, \theta_m)$ . For each  $(\hat{\tau}_i, \hat{\tau}_{i+1})$ , the data from platform  $k$  that fall within the segment can be used to obtain an estimate of  $f_k(\theta_i)$ :

$$\hat{f}_{k,i} = n_k(\hat{\tau}_i, \hat{\tau}_{i+1})^{-1} \sum_{j:t_{k,j} \in [\tau_i, \tau_{i+1})} y_{k,i}, \quad (3.9)$$

For each  $i$  and  $k$ ,  $\hat{f}_{k,i} \sim N(f_k(\theta_i), v_{k,i})$ , where  $v_{k,i} = \sigma_k^2/n_k(\hat{\tau}_i, \hat{\tau}_{i+1})$  is proportional to the noise variance of the  $k$ -th platform and inversely proportional to the number of probes in that platform that lies in the  $i$ -th segment. Thus, the negative log-likelihood of the data is

$$\frac{1}{2} \sum_{i=0}^m \sum_{k=1}^K v_{k,i}^{-1} (\hat{f}_{i,k} - \alpha_k - r_k \theta_i)^2. \quad (3.10)$$

The unknown parameter vectors  $r$  and  $\theta$  should be chosen to minimize the above weighted sum of squares.

If the variances  $v_{k,i}$  were identical across  $i$  and  $k$ ,  $r$  and  $\theta$  can be estimated through the singular value decomposition of the matrix  $F = (f_{i,k})$  or through a robust approach such as median polish. This model would then be similar to those proposed in Irizarry *et al.* (2003) and Li and Wong (2001) for model-based probe set summary of Affymetrix Genechip data. However, the differences in variances should not be ignored, because segments with less data, for which we are less sure of the mean estimate, should be down-weighted. Similarly, platforms with higher noise variance should also be down-weighted compared to platforms with smaller noise.

There are many ways to modify existing approaches to minimize (3.10). We take the following simple iterative approach: Note that for any fixed value of  $r$ , the corresponding minimizer  $\hat{\theta}(r)$  can be found through a weighted least squares regression. The same is true if we minimize with respect to  $r$  when the value of  $\theta$  is held fixed. Thus, joint optimization of  $r$  and  $\theta$  is achieved through a simple block update procedure which we detail in the appendix.

Sample ID	Affymetrix CBS		Illumina CBS		Multi-platform CBS	
	Total	% Conc.	Total	% Conc.	Total	% Conc.
NA18517	156	24	92	12	81	20
NA18507	128	27	138	7	52	20
NA18956	182	37	98	20	159	38
NA19240	181	40	92	10	82	30
NA18555	93	37	27	5	109	37
NA12878	157	52	50	12	139	57
NA19129	193	44	66	2	128	35
NA12156	122	42	39	7	106	41

**Table 1.** Concordance of calls made by running CBS separately on Affymetrix and Illumina platforms, compared with integrated call made by multi-platform CBS. For each method, the first column (Total) is the total number of CNVs called. The second column (Conc.) is the number of calls among the total that overlaps with a fosmid call reported in Kidd *et al.* (2008). The third column (% Conc.) is the percent of concordant calls, i.e. concordant divided by total. Each row is a separate Hapmap sample.

### 3.5 Iterative Joint Estimation

Sections 3.1-3.3 detail a method for segmenting the data when the platform-specific signal response functions are known. Then, Section 3.4 describe a method for estimating the response functions with the segmentation given. In most cases both the segmentation and the response functions are unknown. The algorithm below is an iterative procedure that jointly estimates both quantities from the data.

#### Multi-platform Joint Segmentation.

Fix stopping threshold  $\varepsilon$ . Initialize  $f_k^{(0)}(\mu) = \mu$  for  $k = 1, \dots, K$ . Set  $i \leftarrow 0$ .

Iterate:

1. Estimate the segmentation  $\tau^{(i)}$  using MPCBS assuming response functions  $f^{(i)}$ .
2. Estimate  $f^{(i+1)}$  as described in Section 3.4 assuming the segmentation  $\tau^{(i)}$ .
3. If  $\|f^{(i+1)} - f^{(i)}\| < \varepsilon$ , exit loop and report:

$$\hat{\tau} = \tau^{(i)}, \quad \hat{f}_k = f_k^{(i)}, \quad k = 1, \dots, K.$$

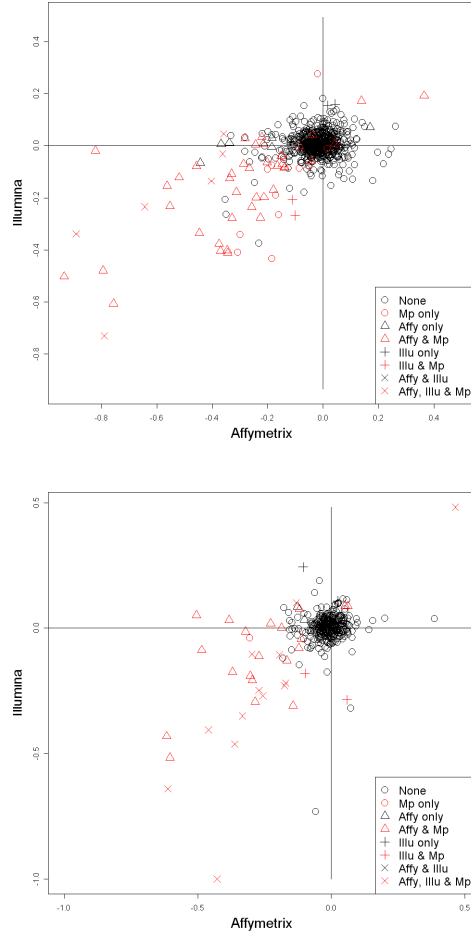
otherwise, set  $i \leftarrow i + 1$ .

In the algorithm above,  $f_k^{(i)}$  and  $\tau^{(i)}$  are respectively the response function and the segmentation estimated in the  $i$ -th iteration. The response functions are initialized to be identity. Thus, we start by treating all platforms equally, which in most cases already gives a decent segmentation. After the first iteration, the estimated segmentation can be used to obtain a more accurate estimate of the response functions, which can then be used to improve the segmentation. The estimates of  $f_k$  usually stabilize within a few iterations.

## 4 Results

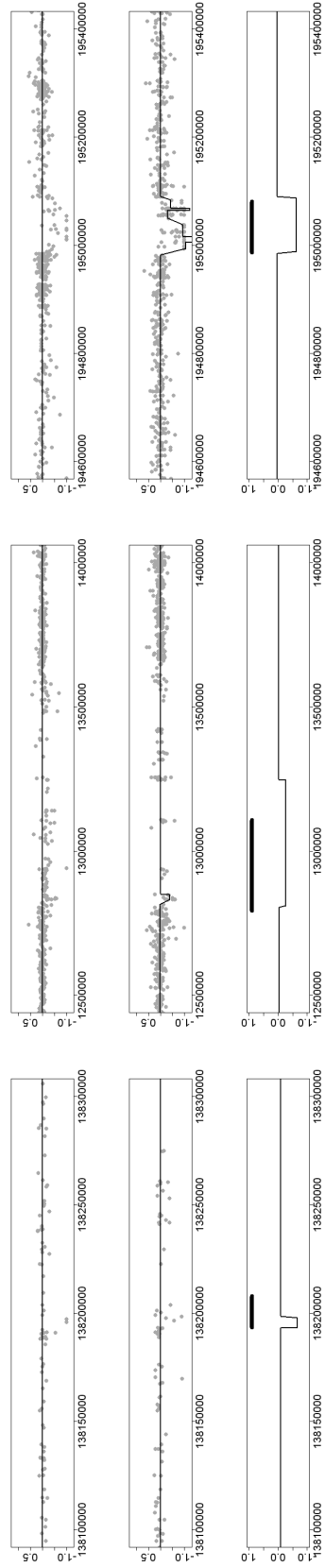
### 4.1 Comparison with Kidd *et al.* (2008) Fosmid Sequencing

We applied our approach to the eight Hapmap samples analyzed in Kidd *et al.* (2008) using fosmid clone end-sequencing. The same Hapmap samples have both been analyzed by Illumina 1M Duo and Affymetrix 6.0 genotyping chips. We used MPCBS to combine the two platforms in making joint CNV calls, and compared these calls with those made by running CBS on each individual platform separately. Table 1 shows, for both individual CBS analysis and MPCBS analysis, the total number of calls, the number of calls that overlap with a fosmid call from Kidd *et al.* (2008), and the percentage of total calls that overlap



**Figure 2.** Mean probe intensities within fosmid CNV calls for Affymetrix versus for Illumina in samples NA18956 and NA12878. The points are colored and shaped based on the combination of Affymetrix, Illumina, the integrated method that detected it.

with a fosmid call. We see from these results that concordance with fosmid is very low across all methods. The low concordance with fosmid detected CNVs has also been reported previously, see, for example Cooper *et al.* (2008) and McCarroll *et al.* (2008). Importantly, in seven out of the eight samples, multiplatform CBS gives a higher concordance rate with fosmid results than either Affymetrix and Illumina does alone. In general, Affymetrix discovers many more segments than Illumina, with many more concordant calls, likely due to having more probes than the Illumina chip.



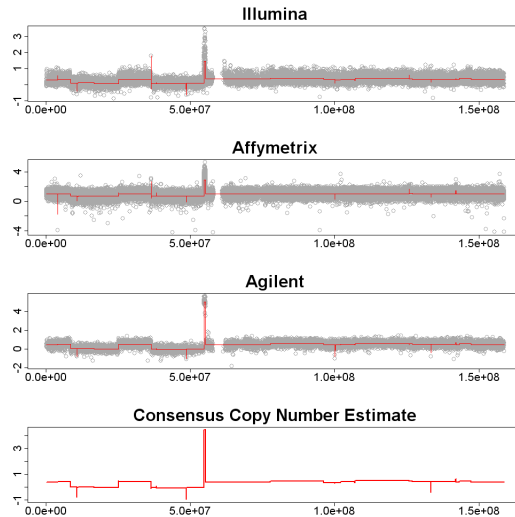
**Figure 3.** Examples of regions detected by MP CBS. For each panel, the top plot shows the Illumina data with CBS fit, the middle plot shows the Affymetrix data with CBS fit, and the bottom plot shows the MPCBS consensus estimate along with thick horizontal lines depicting the fosmid CNV call.

Is the low concordance between Affymetrix, Illumina, and Fosmid CNV calls due to inherent disagreement in the raw data, or low sensitivity or specificity of the statistical method? To investigate this issue, for each fosmid CNV call, we computed the mean intensity of the Affymetrix or Illumina probes mapping within each fosmid CNV call. We would expect that if the absolute change in mean probe intensity is high for a given platform, and if the segment spans a sufficient number of probes, the CNV is more likely to be also called by that platform. Alternatively, if the mean probe intensity within the fosmid CNV call is indistinguishable from baseline, it would be missed by that platform. Figure 2 shows the Affymetrix versus Illumina mean intensity plot for two of the eight samples. Each point corresponds to a fosmid CNV. The points colored in red are fosmid CNVs also detected by MPCBS, i.e. overlapping one of the CNVs called by MPCBS. The shapes of the points reflect whether they are detected by none of the individual platforms alone, by only Affymetrix, or only Illumina, or by both Affymetrix and Illumina. Most of the fosmid CNV calls do not have a shift in intensity in any platform, suggesting that the microarray based assays are noisy and prone to cross hybridization, especially in repetitive regions or regions with complex rearrangements (Cooper *et al.*, 2008). By combining information from the Affymetrix and Illumina platforms, MPCBS is also able to make calls that were not identified in either platform alone.

Figure 3 shows four examples of CNV calls made by multi-platform CBS that is missed by one or both of the individual platforms. In the first two examples shown in the top left and top right panels, the number of probes in each platform is too few to make a call. However, combining the two platforms, multi-platform CBS makes a call that partially overlaps with a fosmid call. In the examples on the bottom left and bottom right panels, multi-platform CBS improves on the boundaries of the Affymetrix call.

## 4.2 TCGA Cancer Data

To provide an example of application to somatic CNVs, we analyze a data set from TCGA samples. Intensity data from three platforms, Illumina 550 K, Affymetrix 6.0 and Agilent 244K were downloaded from TCGA data portal. The segmentation result for CBS and MPCBS on Chromosome 7 of the data is shown in Figure 4. The top three panels show the results for the standard approach, which is to call CNVs for each platform separately. But to integrate the three CBS datasets and generate a consensus CNV result for each sample one is faced with the difficulty that for a true underlying CNV, the three segmentation summaries may not have all detected the CNV, and even when they do, they will report different magnitudes, different boundaries and different degrees of uncertainty. The MPCBS result in the bottom panel provides a natural consensus estimate without the problem of having to decide how to integrate the three CBS segmentation results.



**Figure 4.** Result of MPCBS on a TCGA sample. The top three plots show Illumina, Affymetrix, and Agilent data with CBS fit. Bottom panel shows multi-platform consensus.

## 5 Discussion

We have proposed a model for the joint analysis of DNA copy number data coming from multiple experimental platforms. Under simplifying assumptions, the maximum likelihood framework under this model can lead to an easily interpretable statistic and a computationally tractable algorithm for combining evidence across platforms during segmentation. By comparing to fosmid clone end-sequencing data on eight Hapmap samples, we showed that MPCBS gives more accurate copy number calls. This method has also been applied to TCGA data, where it provides consensus copy number estimates that provide a natural summary of data from Affymetrix, Illumina, and Agilent platforms.

A main feature of MPCBS is that it combines scan statistics from multiple platforms in a weighted fashion, thus without requiring pre-standardization across different data sources. For a given underlying copy number change, platform A may report a higher level of absolute change in signal intensity than platform B, but if A also shows a higher level of noise, or fewer probes in the genomic region in question, the scan statistics of A may not be larger than those of B because such statistics are scaled appropriately within each platform before being combined in MPCBS. However, careful normalization and standardization across platforms is still desirable when running MPCBS. This is because while segmentation per se is not sensitive to absolute signals of different platforms, the mean level of change reported by MPCBS can still be sensitive to the scale of different platforms. Recently, Bengtsson *et al.* (2009) proposed a joint normalization method for bringing different platforms to the same scale

and for addressing the issue of non-linear scaling between platforms. While the method of Bengtsson et al. is not concerned with joint segmentation, it can be coupled to MPCBS so that the mean level of copy number change reported by MPCBS is an even better approximation of the consensus level of change. We expect that the segmentation result will alter only slightly when using data pre-processed by the method of Bengtsson et al. mainly because the current version of MPCBS has not considered non-linear response functions. In short, we recommend pre-standardization of the scale of copy number changes across platforms before running MPCBS. This would have little impact on segmentation but may improve the mean copy number change reported.

MPCBS can be applied also to the situation when a biological sample is assayed multiple times on the same experimental platform. In this case, the platform response ratios are identity and need not be estimated from the data.

## Acknowledgement

The authors thank Terry Speed, Henrik Bengtsson for helpful discussions related to this work.

## 6 Appendix

### 6.1 Derivation of the Likelihood Ratio Statistic (3.3)

To show that the likelihood ratio statistic gives (3.3): For simplicity of notation we suppress the location indices  $[s, t]$ . Since this is a Gaussian mean shift model, the log likelihood ratio between  $H_A$  and  $H_0$  is

$$l_A(\mu) - l_0 = \sum_{k=1}^K [\mu \delta_k X_k / \sigma_k - \mu^2 \delta_k^2 / (2\sigma_k^2)]. \quad (6.1)$$

Differentiating the above with respect to  $\mu$  and setting the derivative to 0, we get  $\hat{\mu} = \tilde{\delta}' X / \tilde{\delta}' \tilde{\delta}$ , where  $\tilde{\delta} = (\delta_1 / \sigma_1, \dots, \delta_K / \sigma_K)$ . Plugging this value back into (6.1), we have  $l_A(\hat{\mu}) - l_0$  equals  $(\tilde{\delta}' X)^2 / (2\tilde{\delta}' \tilde{\delta})$ , which is one-half of (3.3).

### 6.2 Pseudo-code for MPCBS Segmentation Algorithm

**Initialize:**

Set  $k \leftarrow 0$ ,  $S_0 \leftarrow \{0, T\}$ ,

$$Z_{\max} = \max_{0 < i < j < T} Z(i, j), \quad (s^*, t^*) = \arg \max_{0 < i < j < T} Z(i, j),$$

Set  $\mathcal{Z} \leftarrow Z_{\max}$ ,  $\mathcal{R} \leftarrow (s^*, t^*)$ ,  $BIC(0) \leftarrow 0$ .

**While**  $|S_k| - 2 < M$  **repeat:**

1. Let  $i^* \leftarrow \arg \max_i \mathcal{Z}[i]$ ,  $(s^*, t^*) \leftarrow \mathcal{R}[i^*]$ ,

$$s \leftarrow \max\{i \in S_k, i < s^*\}, \quad t \leftarrow \min\{i \in S_k, i > t^*\}.$$

For each of  $(i, j) \in \{[s, s^*], [s^*, t^*], [t^*, t]\}$ , compute

$$Z_{\max} = \max_{i < a < b < j} Z(a, b), \quad (s^*, t^*) = \arg \max_{i < a < b < j} Z(a, b).$$

Let  $Z_L$ ,  $Z_C$ , and  $Z_R$  be respectively the value of  $Z_{\max}$  computed for the left segment  $[s, s^*]$ , the center segment  $[s^*, t^*]$ , and the right segment  $[t^*, t]$ . Similarly, let  $R_L$ ,  $R_C$ ,  $R_R$  be respectively the maximizer for the left, center, and right segments.

2. Let  $L = |\mathcal{Z}|$ , Set:

$$k \leftarrow k + 1,$$

$$S_k \leftarrow S_{k-1} \cup \{s^*, t^*\},$$

$$\mathcal{Z} \leftarrow \{\mathcal{Z}[1 : i^* - 1], Z_L, Z_C, Z_R, \mathcal{Z}[i^* + 1, L]\},$$

$$\mathcal{R} \leftarrow \{\mathcal{R}[1 : i^* - 1], R_L, R_C, R_R, \mathcal{R}[i^* + 1, L]\}.$$

Set  $BIC(k)$  to be the BIC criterion (3.8) of the estimated change-points  $S_k$ .

**Finally, let  $k^* = \arg \max_{0 \leq k \leq M} BIC(k)$ . Return  $S_{k^*}$ .**

### 6.3 Block-update procedure for estimating platform response ratio

Let  $K$  be the number of platforms,  $m$  be the number of regions. We are fitting

$$\frac{1}{2} \sum_{i=0}^m \sum_{k=1}^K v_{k,i}^{-1} (\hat{f}_{i,k} - \alpha_k - r_k \theta_i)^2 \quad (6.2)$$

with the response ratio  $r_K$  for platform  $K$  constrained to be 1.

Initialize  $t \leftarrow 0$ ,

$$r^0 \leftarrow (1, \dots, 1)_{1 \times K},$$

$$\alpha^0 \leftarrow (0, \dots, 0)_{1 \times K}.$$

Repeat:

1.  $t \leftarrow t + 1$
2. Given  $r^{t-1}$ , estimate by weighted least squares

$$\theta^t \leftarrow \arg \min_{\theta} \sum_{i=0}^m \sum_{k=1}^K v_{k,i}^{-1} (\hat{f}_{i,k} - \alpha_k^t - r_k^{t-1} \theta_i)^2.$$

3. Given  $\theta^t$ , estimate by weighted least squares

$$(\alpha_{1:K-1}^t, r_{1:K-1}^t) \leftarrow \arg \min_{\alpha, r} \sum_{i=0}^m \sum_{k=1}^{K-1} v_{k,i}^{-1} (\hat{f}_{i,k} - \alpha_k - r_k \theta_i^t)^2.$$

4. For the  $K$ -th platform, keep  $r_K^t$  at 1 and set  $\alpha_K^t \leftarrow \sum_{i=1}^m (\hat{f}_{i,K} - \theta_i)$ .

5. If  $\|r^t - r^{t-1}\|/m < \epsilon$  exit loop.

Report  $r = r^t$ ,  $\theta = \theta^t$ ,  $\alpha = \alpha^t$ .

## References

- Bengtsson, H., Ray, A., Spellman, P., and Speed, T. (2009). A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. *Bioinformatics*, **25**, 861–867.
- Cooper, G. M. M., Zerr, T., Kidd, J. M. M., Eichler, E. E. E., and Nickerson, D. A. A. (2008). Systematic assessment of copy number variant detection via genome-wide snp genotyping. *Nature genetics*, **40**, 1199–1203.
- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249–264.
- Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, A. N., Tsang, P., Newman, T. L., Tüzün, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., Mckernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., Mccarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., and Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**(7191), 56–64.
- Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics*, **21**, 3763–3770.
- Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *PNAS*, **98**(1), 31–36.
- McCarroll, S. A. A., Kuruvilla, F. G. G., Korn, J. M. M., Cawley, S., Nemes, J., Wysoker, A., Shaper, M. H. H., de Bakker, P. I. W. I., Maller, J. B. B., Kirby, A., Elliott, A. L. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W. W., Rava, R., Daly, M. J. J., Gabriel, S. B. B., and Altshuler, D. (2008). Integrated detection and population-genetic analysis of snps and copy number variation. *Nature genetics*, **40**, 1166–1174.
- Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, **5**, 557–572.
- Vostrikova, L. (1981). Detecting disorder in multidimensional random process. *Soviet Mathematics Doklady*, **24**, 55–59.
- Willenbrock, H. and Fridlyand, J. (2005). A comparison study: applying segmentation to arraycgh data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.

- Zhang, N. and Siegmund, D. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, **63**, 22–32.
- Zhang, N., Siegmund, D., Ji, H., and Li, J. Z. (2008). Detecting simultaneous change- points in multiple sequences. *Technical Report, Department of Statistics, Stanford University*.