

DNA Copy Number Profiling in Normal and Tumor Genomes

Nancy R. Zhang¹

¹Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305-4065, USA

1.1 Introduction

For a biological sample, the DNA copy number of a genomic region is the number of copies of the DNA in that region within the genome of the sample, relative to either a single control sample or a pool of population reference samples. Within the last decade, significant advances in DNA array technology has enabled the genome-wide fine scale measurement of DNA copy number in a high throughput manner (Pinkel et al., 1998; Pollack et al., 1999; Snijders et al., 2001; Bignell et al., 2004; Peiffer et al., 2006). This enables systematic studies which can lead to a better understanding of the role of DNA copy number changes in human disease and in phenotypic variation in the human population. These high throughput experiments produce large amounts of data that are rich in structure, motivating the development of new statistical methods for their analysis. This chapter reviews the computational and statistical problems that arise in DNA copy number data and surveys recent advances in their treatment.

First, we review some terms and general concepts relating to DNA copy number. A copy number variant (CNV) is defined as a genomic region where the DNA copy number differs between two or more individuals from a population. CNVs that have so far been catalogued are by convention larger than 1 kilobase, although technologies based on high throughput sequencing (Shendure et al., 2004) and denser arrays (Ishkanian et al., 2004) can detect shorter CNVs. Within the last five years, many studies (Khaja et al., 2007; Redon et al., 2006; Conrad et al., 2006; McCarroll et al., 2006; ?) have shown that CNVs are a common type of genetic variation in the human population, with the fraction of the genome covered by CNVs estimated to be between 2% (Cooper et al., 2007) and 15% (Estivill & Armengol, 2007). Like single nucleotide polymorphisms (SNPs), variants in copy number segregate in a

Mendelian fashion and contribute to phenotypic variation. Considering that they cover significantly more genomic territory in terms of base pairs, and that they are more likely than SNPs to have a deleterious effect, CNVs are now routinely used alongside SNPs in genetic association studies.

Changes in DNA copy number have also been highly implicated in tumor genomes. Some of these changes are inherited, but many are due to somatic mutations that occur during the clonal development of the tumor. The copy number changes in tumor genomes are often referred to as copy number aberrations (CNAs), to differentiate them from inherited CNVs. CNAs are usually larger in size than CNVs, often involving gains and losses of entire chromosome arms. Their role in tumor development is not clear, although high fold amplification of genomic regions containing oncogenes and deletion of regions containing tumor suppressor genes have been widely documented. For example, a search using the terms “copy number” and “tumor” brings up 4421 articles in Pubmed. These evidence suggest that at least some CNAs play a role in driving tumor progression.

Given the raw DNA copy number data from a single sample, an immediate challenge lies in estimating the true underlying copy number from the noisy measurements. This problem, often referred to as segmentation of total copy number, has drawn considerable attention and is reviewed in Section 1.2. For data from some array platforms, such as the Affymetrix and Illumina genotyping arrays, it is possible to tease apart the underlying copy numbers of the two distinct sets of chromosomes inherited from the two biological parents. This problem, which we refer to as parent- or allele- specific copy number estimation, is motivated and reviewed in Section 1.3. In both total copy number and parent-specific copy number estimation, it is important to distinguish between tumor and normal samples in the formulation of the statistical model. This is a theme that will be re-iterated in this chapter.

In many studies, multiple technical platforms or different versions of the same platform are being used to interrogate the same biological samples. Pooling information across these multiple sources can give a more accurate consensus molecular profile for each sample. Section 1.4 looks at recent approaches to multi-platform integration. A more complex problem is the joint analysis of multiple copy number profiles, each coming from a different biological sample. There can be many different goals in such cross-sample analyses, which deserve different statistical approaches. Section 1.4 reviews the modeling issues and recent developments in cross-sample models for DNA copy number.

1.2 Total Copy Number Estimation for One Sample

The total DNA copy number data for any given sample comes in the form of a sequence $\{(x_i, y_i) : i = 1, \dots, n\}$, where n is the number of probes and x_i and y_i are respectively the genome location and normalized intensity for probe i . “Probe” and “normalized intensity” mean different things for

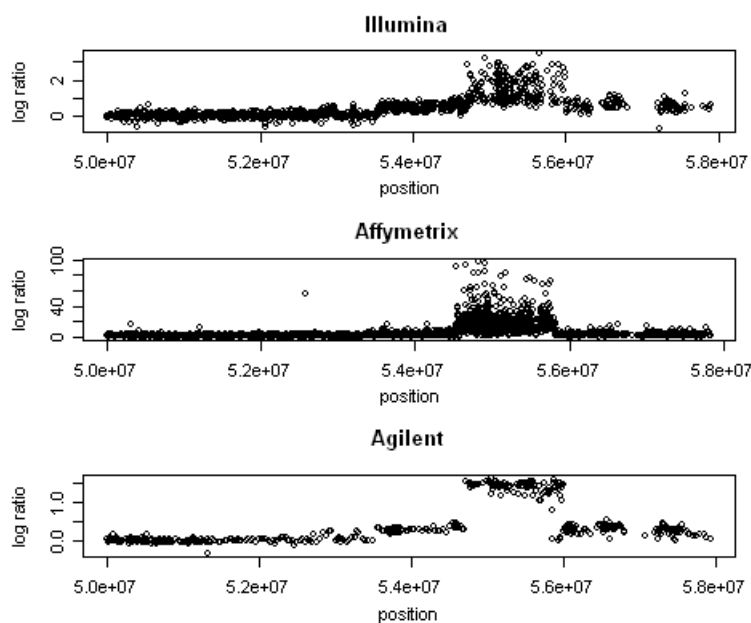


Fig. 1.1. Copy number data for a tumor sample assayed on the Agilent, Illumina, and Affymetrix platforms.

different experimental platforms, and the reader is referred to (Pinkel et al., 1998; Pollack et al., 1999; Snijders et al., 2001; Bignell et al., 2004; Peiffer et al., 2006) for more details. The term “total copy number” refers to the sum of the copy numbers for the chromosomes inherited from the two biological parents. If this number varies over the cells in the sample, then the intensity is a reflection of average copy number over all of the cells. Thus, although total copy number for each individual cell is integer valued, when the sample is genetically heterogeneous the average copy number can vary over a continuous scale.

The appropriate preprocessing procedure that is necessary to normalize the intensity measurements depends on the technical platform that generated the data, see Peiffer et al. (2006); Bengtsson et al. (2008) for some examples of non-trivial pre-processing procedures. The data from most platforms is in the form of a log ratio of the DNA quantity in the target sample versus the DNA quantity in an appropriate control. The “normal” state, where the copy number in the target agrees with that in the control, should have mean 0. A contiguous stretch of measurements that are on average higher (or lower) than 0 suggests a gain (or loss) in copy number. Figure 1.1 shows an example copy number profile for a genomic region from a tumor sample, assayed on

three different platforms. Note that different experimental platforms vary in noise variance, responsiveness to signal, and location of probes. Section 1.4 examines these differences between platforms in more detail.

The observed intensities are noisy surrogates of the true copy number at the measured positions. Since chromosomes are gained and lost in segments, adjacent positions in the genome are highly likely to have the same underlying copy number. This is why change-point models (Olshen et al., 2004; Venkatraman & Olshen, 2007; Zhang & Siegmund, 2007; Picard et al., 2005; Wen et al., 2006), smoothing methods (Hupé et al., 2004; Broët & Richardson, 2006; Lai et al., 2007; Tibshirani & Wang, 2008), Haar-based wavelets (Hsu et al., 2005), spatially restricted clustering (Wang et al., 2005; Xing et al., 2007), and various formulations of hidden Markov models (Fridlyand et al., 2004; Lai et al., 2007; Guha et al., 2006; Engler et al., 2006; Beroukhim et al., 2006; Colella et al., 2007) have been proposed for the estimation of DNA copy number. Lai et al. (2005) and Willenbrock & Fridlyand (2005) reviewed and compared the performance of existing approaches in 2005. In this chapter, we review the change-point formulation for this problem that underlies the Circular Binary Segmentation (CBS) algorithm (Olshen et al., 2004; Venkatraman & Olshen, 2007), which was found to be one of the most accurate methods by both Lai et al. (2005) and Willenbrock & Fridlyand (2005). We then summarize the numerous hidden Markov model based approaches, which, as we will see in Section 1.3, generalize naturally to model the more complex data from genotyping arrays.

Since the location of the probes, at a coarse global scale, is approximately uniformly distributed in the genome, their location information $\{x_i : i = 1, \dots, n\}$ is often ignored in the segmentation process. Then, a simple change-point model for the sequence of intensities is

$$y_i = \mu_i + \epsilon_i, \quad i = 1, \dots, n, \quad (1.1)$$

where $\mu = \{\mu_i : i = 1, \dots, n\}$ is a piecewise constant function of i , and $\{\epsilon_i : i = 1, \dots, n\}$ are i.i.d. errors. To describe μ , we assume that there exists a series of change-points $0 = \tau_0 < \tau_1 < \dots < \tau_m < \tau_{m+1} = n$ such that

$$\mu_t = \theta_i, \quad t \in [\tau_i, \tau_{i+1}), \quad i = 0, \dots, m. \quad (1.2)$$

For inference, the errors are usually assumed to be Gaussian, although this assumption is not crucial if the distances between successive τ_j 's are large. Under this model, the segmentation problem reduces to estimating the change-points and the means within each segment. The number of change-points m is also not known and has been observed to range from below 10 to above 100 in some tumor samples.

If the values of the change-points τ are known, then θ_j can be estimated by the mean of the observations that fall in the j -th segment. To estimate τ , the CBS algorithm employs a greedy top-down approach that recursively applies the generalized likelihood ratio statistic for testing a square wave change. In

more detail, for any interval $1 \leq a < b \leq n$, let the null hypothesis be that the observations are i.i.d. Gaussian and let the alternative be that there is a sub-interval with a change in mean and no change in variance. The generalized likelihood ratio statistic is

$$\max_{a < s < t < b} Z_{s,t}, \quad \text{where } Z_{s,t} = \frac{S_t - S_s - \frac{t-s}{b-a}(S_b - S_a)}{\hat{\sigma} \sqrt{(t-s)[1 - (t-s)/(b-a)]}}, \quad (1.3)$$

and $S_j = y_1 + \dots + y_j$. CBS starts by setting $a = 1$, $b = n$. Let z^{obs} be the observed maximum of $Z_{s,t}$, and (s^*, t^*) be the maximizing interval. If the p-value of the scan, $P(\max_{a < s < t < b} Z_{s,t} > z^{obs})$, is smaller than some pre-chosen threshold α , then the maximizing interval is reported and the intervals $[a, s^*)$, $[s^*, t^*)$, $[t^*, b]$ are recursively scanned using the same procedure. The recursion stops when none of the subregions contain a square wave change that is significant at the level α .

The p-value for the scan statistic in (1.3) can be computed using asymptotic approximations given by James et al. (1987), which is quite accurate when the interval is large. Venkatraman & Olshen (2007) give computationally efficient permutation based procedures for computing the p-values for small intervals. Alternatively, Zhang & Siegmund (2007) proposed a modified BIC criterion for estimating m , and showed that, when used in conjunction with CBS, has more accurate off-the-shelf performance than p-value based thresholds.

In contrast to the change-point formulation, hidden Markov model based methods assume that y_i are emitted by an underlying Markov chain with state sequence $S = \{S_i : i = 1, \dots, n\}$. Different published methods assume different dynamics for the underlying Markov chain. The earliest method (Fridlyand et al., 2004) assumes that S is a discrete state Markov chain, and obtains a segmentation using the Viterbi algorithm. The discrete state model works well for detecting inherited CNVs in normal samples, but is not flexible enough for tumor samples, where due to sample heterogeneity it is hard to predict how many states there should be in the underlying chain. To better accommodate fractional copy number changes, Lai et al. (2007) assumes that S is a continuous valued Markov jump process with a baseline state and a changed state, where every time a jump is made to the changed state S takes on a new Gaussian value. Exact recursive equations for the posterior expectation of S given the entire observed y sequence are given in Lai et al. (2007), along with a fast linear time approximation. The Bayesian procedure also allows computation of confidence intervals such as the expected copy number at each position and the total number of CNVs. The hidden Markov models in Guha et al. (2006); Engler et al. (2006) also assume continuous valued jumps, but uses pseudo-likelihood based approaches or estimates the underlying states using Markov chain Monte Carlo.

The fundamental difference between the change-point approach and hidden Markov models is the necessary assumptions about the length and magnitude of jumps. For a hidden Markov model, one must explicitly specify the

waiting time distribution between jumps and the distribution of the underlying state sequence. If reliable prior information in this regard is available, then hidden Markov models can more flexibly incorporate them. However, when prior information is not available, they must either be specified arbitrarily by the user or estimated from the data. The change-point formulation underlying CBS relies on the least assumptions, which is why it is a more robust off-the-shelf method.

The methods mentioned so far use only the total intensity data. Data from some platforms, such as Illumina and Affymetrix genotyping arrays and Molecular Inversion Probes, can reveal more information. These platforms measure, for targeted biallelic single nucleotide polymorphisms, the quantities of both alleles. The raw total copy number sequence $\{y_i : i = 1, \dots, n\}$ is obtained from these platforms by summing the allele intensities and then normalizing to population controls. This essentially reduces a two dimensional data sequence to one dimension, resulting in a loss of information. For Illumina data, for example, the *B*-allele frequency, defined as the normalized ratio of the quantity of the *B*-allele to the total quantity of both alleles, seems to be more informative for detecting low amplitude jumps (Peiffer et al., 2006). Methods for detecting inherited CNVs can gain power by incorporating the the *B*-allele frequency, as done in the softwares QuantiSNP (Colella et al., 2007) and PennCNV (Wang et al., 2007). More details of these methods, in the context of parent specific copy number estimation, are given in the next section.

1.3 Parent Specific Copy Number Estimation

The genome of each somatic human cell normally contains two copies of each of the 22 autosomes, one inherited from each biological parent. At any genome location, one or both of these two chromosomes may gain or lose copies. The methods described in the last section use only the total intensity data $\{y_i\}$, which measure the sum of the copy numbers of the two inherited chromosomes. These methods do not reveal whether both chromosomes have changed copies, and at polymorphic loci, which of the alleles have been affected. The extraction of this information is important, because the allele-specific nature of amplification and deletion is highly relevant for the biological understanding of cancer, and represents a key advantage of SNP array-based assays in comparison to the conventional arrayCGH experiments. Furthermore, copy neutral loss-of-heterozygosity events, defined as the simultaneous gain of one copy and loss of the other copy of the inherited chromosomes, have long been implicated in the actions of oncogenes. These events can only be detected using allelic-specific measurements.

For simplicity, we consider only biallelic SNPs, and let the alleles be arbitrarily labeled A and B. Allele “A” may refer to different bases at different SNPs, and may also reside on different chromosomes across adjacent SNPs.

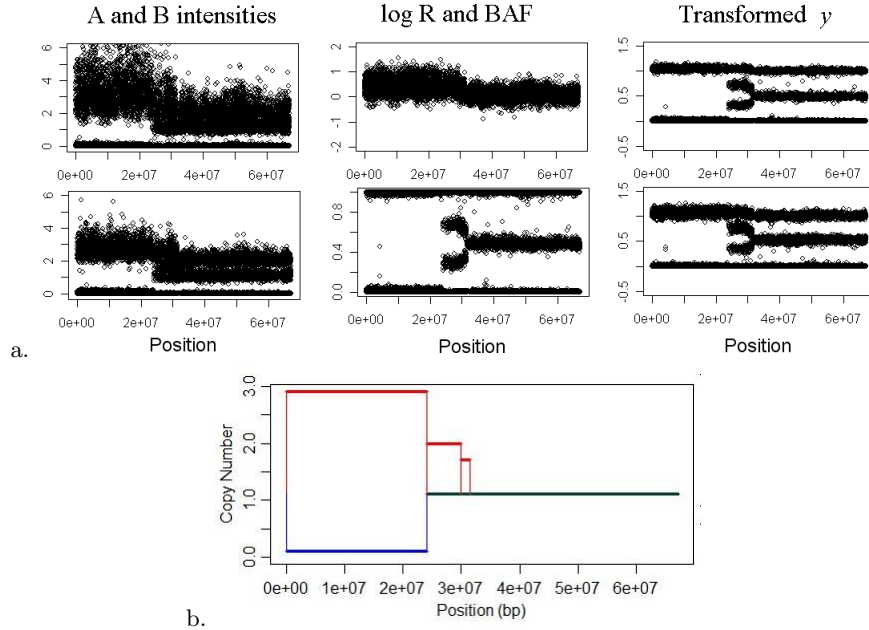


Fig. 1.2. (a) An example data sequence taken from a TCGA glioblastoma sample. The left panel shows the A and B allele intensities. The middle panel shows the log R and B-allele frequencies. The right panel shows the transformed bivariate y sequence that is used by the new model. The estimated major and minor copy numbers for the example region from TCGA sample 0258 chromosome 2 shown in (b).

Genotyping platforms give, at selected SNPs, a bivariate measurement quantifying each of the alleles A and B. A sequence of normalized A and B intensities for an example region of a tumor sample assayed using the Illumina platform is shown in the left panel of Figure 1.2. The log ratio, which is the normalized sum of the A and B intensities, and the B-allele frequency are shown in the middle panel of the figure.

The allele-specific measurement at each SNP follows a mixture distribution that depend on the genotype of the sample at that SNP. The genotype is usually unknown, and must also be inferred from the data. Without the genotype information, adjacent allele-specific measurements can not be smoothed. Thus, conventional change-point models can not be applied directly to this problem. However, the parent-specific copy number, which we define as a bivariate quantity that distinguishes between the chromosomes inherited from the two parents, is smooth across adjacent positions on a chromosome. Without family data, it is impossible to distinguish which chromosome is maternal and which is paternal. Thus the parent specific copy numbers are exchangeable. When there is an imbalance in copy number, the chromosome with the

Hidden state	Copy number	Number of genotypes	Interpretation
1	0	0	Full deletion
2	1	1	Single copy loss
3	2	3	Normal
4	2	2	Normal LOH
5	3	4	Single copy gain
6	4	5	Double copy gain

Table 1.1. Hidden states, associated copy numbers, genotype states, and biological interpretation in Colella et al. (2007),

higher copy number is called the “major” chromosome and the other is called the “minor” chromosome.

LaFramboise et al. (2005) is one of the earlier methods that make use of allele-specific data. They applied existing segmentation algorithms to the total copy number. Then, the B allele frequency is used to estimate the allele-specific copy number and loss of heterozygosity status for each segment. Since only the total copy number is used in the initial segmentation, this approach misses copy neutral loss-of-heterozygosity events.

Discrete-state hidden Markov models (Lin et al., 2004; Wang et al., 2007; Colella et al., 2007) have also been applied with success to this problem. In these approaches, the hidden states representing changes in whole copy number or generalized genotypes such as AA, AB, BB, A-, B-, AAB, ABB, *etc.* Colella et al. (2007) is one of the earliest methods in this category. Designed for Illumina data, it is based on a hidden Markov model with six underlying states described in Table 1.1. Within each state, the log ratio and B-allele frequency are assumed to be independent. The log ratio is assumed to be a mixture of a uniform distribution and a Gaussian distribution with state dependent mean and variance. The uniform distribution acts as a non-informative state for capturing outliers in the data. The B-allele frequency follows a mixture distribution that depends on the unknown genotype, with also a uniform component for robustness against outliers. The parameters of this model can be estimated by maximizing the marginal likelihood, thus giving it some desirable frequentist properties while also allowing for flexible Bayesian type inference. The PennCNV software (Wang et al., 2007) uses a similar model .

Methods based on discrete state hidden Markov models are designed for detecting copy number variants in normal tissue, where the assumption of idealized unit-copy changes holds because the cells within the samples are usually homogeneous. By requiring a fixed set of pre-defined discrete states, these methods do not perform well on heterogeneous tumor samples, which produce data with apparently fractional copy number changes. Through simulated titration studies, Staaf et al. (2008) show that methods relying on idealized genotype states lose sensitivity when tumors are diluted with normal cells.

s_t	x_t^A	x_t^B
AA	$\theta_t^1 + \theta_t^2$	0
AB	θ_t^1	θ_t^2
BA	θ_t^2	θ_t^1
BB	0	$\theta_t^1 + \theta_t^2$

Table 1.2. Relationship between the inherited allele configuration s_t and the true allele specific copy numbers x_t in Chen et al. (2009).

To treat the heterogeneity in tumor samples, Chen et al. (2009) propose a continuous-state hidden Markov model to simultaneously estimate the parent specific DNA copy numbers and the unknown genotypes. To describe their model, let $y = \{y_t = (y_t^A, y_t^B) : t = 1, \dots, n\}$ be the normalized intensity values for alleles A and B at n SNPs ordered by their location in a reference genome. Let $\theta = \{\theta_t = (\theta_t^1, \theta_t^2) : t = 1, \dots, n\}$ be the underlying parent specific copy numbers, and $s_t \in S = \{AA, AB, BA, BB\}$ be the configuration at SNP t specifying the alleles carried by the inherited chromosomes. Let x_t be the true copy numbers of alleles A and B at SNP t . The value of x_t is determined by θ_t and s_t , with the mapping shown in Table 1.2. Note that when a somatic event causes a change in copy number in one or both parental chromosomes at SNP t , the allele-specific copy numbers x_t change, but s_t remains fixed. For example, if $s_t = AB$, and if θ_t^1 were amplified two-fold, then the true copy number of allele A would be 2, but s_t would still be AB. The *observed* allele specific intensities y_t are assumed to be equal to the true allele specific quantities plus an independent measurement error,

$$y_t = x_t + \epsilon_t, \quad (1.4)$$

where $\epsilon_t \sim N(0, \Sigma_{s_t})$ and Σ_{s_t} are state specific error covariance matrices. This model is summarized in Figure 1.3. The inherited allele configurations s_t are assumed to be i.i.d. multinomial with parameters $(p_t^{AA}, p_t^{BA}, p_t^{AB}, p_t^{BB})$, which can be obtained from the genotyping data of a set of normal control samples. The dynamics of θ is modeled as a Markov jump process that generalizes the 1-dimensional model in Lai et al. (2007) to two dimensions. Conceptually, each time a jump occurs, θ_t can either go to a baseline state which consists of a point mass at a pre-defined value, or chooses a new value in \mathfrak{R}^2 according to a bivariate Gaussian distribution. Chen et al. (2009) generalizes the estimation procedure in Lai et al. (2007) to simultaneously estimate s and θ from y . The parameters of the hidden markov model can be estimated from the data via expectation maximization. A plot of the major and minor copy numbers estimated from their procedure for the example region given in Figure 1.2a is given in Figure 1.2b. From the plot, we see that this region contains an unbalanced gain of two copies in one chromosome coupled with an almost complete deletion of the other chromosome. This is immediately followed by gain of one chromosome with the other at normal level. Without analyzing allele-specific data, we would not have known that the gain in total copy

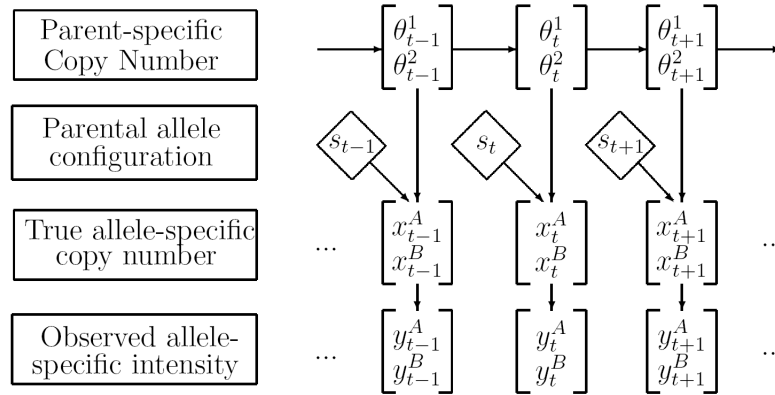


Fig. 1.3. Overview of stochastic segmentation model in Chen et al. (2009).

number at the left end of this region actually involves an almost complete loss of heterozygosity.

1.4 Integration of Multiple Array Platforms

With the rapid development of new genome-wide profiling platforms, there is now an increasing need of data integration when more than one technical platform, or different replicates on the same platform, are used to assay 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, has adopted three independent platforms for studying 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 types of data is to apply existing segmentation algorithms to search for copy number changes within the data from each platform separately. The segmentation results from all platforms then needs to be combined. However, the platforms often disagree on the calling of a change, either on its significant or on its location and magnitude. In such situations, it is difficult to decide what the consensus should be. Furthermore, by segmenting each platform separately, information is not pooled across platforms for boosting the power of hard-to-detect signals. For these reasons, Bengtsson et al. (2009) and Zhang et al. (2009a) proposed methods to integrate the data across platforms during the segmentation process.

To integrate the data during segmentation, the differences between platforms need to be resolved. Some platforms, such as Illumina and Agilent, produce allele specific data, while others, such as Agilent and cDNA arrays, produce two-color ratio data that measure only total intensity. The probes

from each platform map to a different set of locations in the genome, at different densities. For example, the Affymetrix 6.0 array has 1.8 million probes while the Agilent CGH 244K array has about one-sixth as many probes. The measurements from different platforms also have different signal to noise ratios, as can be seen from Figure 1.1. The different platforms also respond with different saturation curves (Bengtsson et al., 2009). In regions of high-fold amplification, Illumina and Affymetrix tend to have more pronounced signal saturation than Agilent. 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.

Bengtsson et al. (2009) studied the problem of differing degrees of attenuation of the true copy number changes across platforms, and proposed a method based on principal curves to correct for between platform differences. Their method brings the unsegmented intensity levels to the same scale across sources. For the same underlying true copy number, each platform should have the same mean value after this normalization procedure. Then, existing segmentation methods can be applied to a combined set of intensity levels over all sources to identify the copy number changes. However, the normalization procedure in Bengtsson et al. (2009) does not resolve the issue of differing error variances between sources, and the homogeneous variance model in (1.3) would not be optimal when applied to this combined data set.

Recently, Zhang et al. (2009) proposed a multi-platform Circular Binary Segmentation (MPCBS) procedure, which has been adopted in the processing of the TCGA glioblastoma samples. 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, which can be formulated as follows. Let the platforms be indexed by $k = 1, \dots, K$, with K being the total number of platforms. The observed data is $\mathbf{y}_k = y_{k1}, \dots, y_{kn_k}$ for the n_k snps/clones on the k -th platform, which have ordered locations $(t_{k1}, \dots, t_{kn_k})$ along a chromosome. It is assumed that for each platform, the data has been normalized to be centered at 0 for “normal” copy number and to have Gaussian noise. 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 μ_t for all platforms. Let $f_k(\cdot)$ be the response function for platform k , which quantifies the dependence of the mean intensity on the underlying copy number. The observed intensity level for the i -th probe of the k -th platform is modeled as

$$y_{ki} = f_k(\mu_{t_{k,i}}) + \epsilon_{k,i}, \quad (1.5)$$

where the noise term $\epsilon_{k,i}$ are independently distributed $N(0, \sigma_k^2)$, and σ_k^2 is the platform specific noise variance. Zhang et al. (2009) consider only linear response functions $f_k(\mu) = r_k \mu$, where the parameter r_k , called the response ratio, describes the ratio between the change in signal intensity and the un-

derlying copy number change for platform k . This linearity assumption allows for simple and intuitive test statistics and fast scanning algorithms. The true copy number μ_t is modeled as a piecewise constant function as in (1.2), where the endpoint n is replaced by the length of the chromosomal region in base pairs. 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 ratios, must be estimated from the data.

It is insightful to look at the generalized likelihood ratio statistic that arises from this cross-platform model and compare it to (1.3) for the one sample case and (1.12) for the multi-sample case given in the next section. Consider the case where $r = (r_1, \dots, r_K)$ is known, and 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 change, the data within this region should have baseline mean $f_k(0) = 0$, i.e.

$$H_0 : y_{ki} \sim N(0, \sigma_k^2) \quad \text{for } k = 1, \dots, K; \quad \text{and } i : s \leq t_{ki} < t. \quad (1.6)$$

If there is a gain (or loss) of magnitude μ , each platform should respond with signal $f_k(\mu) = r_k \mu$. The signal is thus 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.

$$H_A : y_{ki} \sim N(r_k \mu, \sigma_k^2) \quad \text{for } k = 1, \dots, K; \quad \text{and } i : s \leq t_{ki} < t. \quad (1.7)$$

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 $(s, t]$, and $\bar{y}_{k,(s,t]}$ be the mean intensity of these probes. The generalized log likelihood ratio statistic for testing H_A versus H_0 is

$$Z(s, t) = \frac{\left[\sum_{k=1}^K \delta_{k,s,t} X_{k,s,t} \right]^2}{\sum_{k=1}^K \delta_{k,s,t}^2}, \quad (1.8)$$

where

$$X_{k,s,t} = \frac{\bar{y}_{k,[s,t]} - \bar{y}_{k,[s,t]}^c}{\sigma_k \sqrt{n_k(s, t)^{-1} + [n_k - n_k(s, t)]^{-1}}}, \quad (1.9)$$

and

$$\delta_{k,s,t} = r_k \sqrt{n_k(s, t)} / \sigma_k. \quad (1.10)$$

Some easy algebra shows that $X_{k,s,t}$ is equivalent to the statistic in (1.3) computed for the k -th platform. The cross-platform statistic is then the projection of $X_{s,t} = (X_{1,s,t}, \dots, X_{K,s,t})$ on to the vector $\delta = (\delta_{1,s,t}, \dots, \delta_{K,s,t})$. We thus call (1.8) the projected χ^2 statistic. It can also be viewed as the squared norm of a weighted sum of t -test statistics, where the weight $\delta_{k,s,t}$ for platform k 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 σ_k . When there is only one platform, the statistic (1.8) is equivalent to the chi-square statistic used in the Circular Binary Segmentation algorithm of Olshen et al. (2004). As for CBS, σ_k is usually unknown and must be estimated from the data. For more details on the estimation of the platform response ratios, see Zhang et al. (2009a).

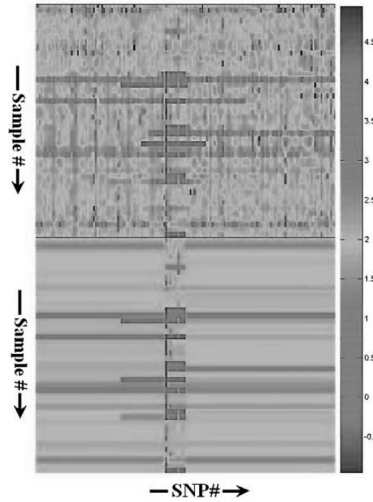


Fig. 1.4. An example of a joint segmentation of a set of tumor samples. The segmentation outputs a set of common change-points to give the best sparse summary of the set of tumors.

1.5 Modeling recurrence across samples

When the same biological sample is assayed using multiple platforms, the underlying signal for the copy number profiles from each platform should be the same. This is the concept that underlies the projected χ^2 statistic (1.8) in the MPCBS algorithm. However, when each copy number profile represents a different biological sample, the underlying signal is no longer shared. Usually, only a fraction of the samples are carriers of any given CNV. In an integrated analysis of copy number data across multiple biological samples, one is often interested in the differences across samples as well as the similarities.

Before introducing the statistical models for cross-sample analysis, we need to examine more carefully the purpose of cross-sample integration. What do we hope to achieve in such an analysis? What types of signals are we aiming to capture across samples? The answer to this question is simple for multiplatform integration, where the goal is simply to combine data across platforms to obtain a better estimate of the shared underlying change-points. In cross sample analyses, how should the concept of a shared signal be defined? When the signals are not shared, how should the variation be characterized?

One frequently encountered problem in copy number studies over a cohort of tumor samples is finding regions of recurrent aberration. Such regions, where a large number of samples of the same type of tumor have gained or lost copies, may contain genes that are key players in the development of the tumor. For example, Figure 1.4 shows a set of tumor samples, with many sam-

ples carrying overlapping deletions covering chromosome 9 (Schiffman et al., 2009). Such commonly deleted regions may carry genes that play a role in cell proliferation or delay apoptosis. Similarly, commonly amplified regions may harbor tumor suppressor genes. In Section 1.5.1 We review methods that are geared towards identifying these regions.

The methods reviewed in Section 1.5.1 combine information across samples post-segmentation. That is, each sample is segmented on its own, and the cross sample analysis sees only the segmented data. However, in some cases, such as inherited CNVs, the change-points are often shared across samples for instances of the same CNV. In such cases, aggregating information across samples can improve the power of detecting shared weak signals. Consideration of power is especially relevant to the detection of inherited CNVs, most of which are very short and may only span a few probe sets or clones, and thus are easily missed in single sample detection methods. Inherited variants are also hard to detect in the sense that they usually involve single-copy changes, as compared to aberrations in tumors which often consist of high fold amplifications and homozygous deletions. In Section 1.5.2, we discuss the aggregation of data across samples prior to or during segmentation.

In the analysis of both inherited and somatic copy number variants, it is often useful to obtain a sparse cross-sample summary of a complex region for use in downstream analyses. For example, in clinical studies we may have, along with the copy number data, variables such as survival outcome or status of other biomarkers. We may want to find chromosomal regions whose copy number status is correlated with these variables. These types of analyses are often done with gene or protein expression data, but for copy number data it is unclear what to use as the explanatory variables. If each probe were considered as a variable, then the smoothness of the underlying signal is ignored. Since copy number studies now routinely use platforms containing hundreds of thousands to over a million probes, if each probe were considered a variable we would be faced with a very large number of highly correlated variables, which would reduce the sensitivity of downstream analyses. Some studies take the average copy number over each chromosome, chromosome arm, or cytoband as the variables for downstream analysis. This clearly is a coarse method that sacrifices sensitivity. In Section 1.5.3, we describe a recently developed method for reducing a set of copy number profiles into a set of representative regions, so that the average copy number of each sample in each region gives a good summary of the cohort.

1.5.1 Post-Segmentation Procedures

In post-segmentation procedures (Newton et al., 1998; Newton & Lee, 2000; Diskin et al., 2006; Beroukhi et al., 2007; Guttman et al., 2007; Taylor et al., 2008; Rouveiol et al., 2006), each sample is first segmented separately, which reduces them to piece-wise constant sequences indicating regions of amplification, deletion, or normal copy number. Then, the samples are aligned, and

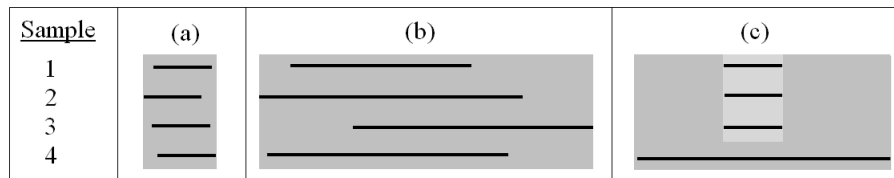


Fig. 1.5. Illustration of the concept of a footprint of in the STAC program. In each of the scenarios (a-c), the gray box indicates the footprint for the entire stack of four samples. In (c), the light gray box indicates the footprint for the stack containing only samples 1-3.

a statistical model (Newton et al. (1998); Newton & Lee (2000)) or permutation based approach (Diskin et al. (2006)) is used to identify regions of highly recurrent aberration.

One of the earliest methods is STAC (Diskin et al., 2006), which takes in a binary sequence for each sample that indicates whether the sample contains an aberration at each probe position. Consider first the simple method which considers only the prevalence of aberration at each location m , denoted by $F(m)$. The p-value of $F(m)$ can be computed by permutation, where the intervals within each profile are randomly rearranged. Let $F_i(m)$ be the prevalence over samples at location m for permutation i . Then, the p -value of a location m_0 is computed by

$$P_F(m_0) = \frac{|\{i : \max_m F_i(m) > F(m_0)\}|}{\text{total number of permutations}}.$$

Locations that are aberrant in a large number of samples have a low p-value. However, the statistic $F(m)$ does not capture the fact that it is more surprising when several samples have tight overlap for a short aberrant interval (Figure 1.5a), as opposed to when an overlap is a result of long aberrations that do not align tightly (1.5b). To differentiate between these two situations, Diskin et al. (2006) defines a “stack” to be a set of intervals that share at least one common position, and the “footprint” of a stack to be the union of the sets of positions covered by the intervals. Each of the set of intervals in Figure 1.5a,b is a stack, but Figure 1.5b has a larger footprint than Figure 1.5a. A smaller footprint provides greater evidence for localization of an important gene in the region. In Figure 1.5c, the set of all 4 intervals is a stack, but so is any subset, e.g. the intervals from only samples 1-3. The stack for samples 1-3 has a much smaller footprint than the stack containing all four samples, and may be more biologically interesting. For each position m , let S_m to be the set of all stacks with m as a common position, i.e. the set of all subsets of samples that are aberrant at position m . For each $s \in S_m$, let $P(s)$ be the p-value for the footprint of the stack s computed by permutation. Then, glossing over details, the score for the most significant footprint is computed

as $R(m) = \min_{s \in S_m} P(s)$. A related method is MSA, which builds upon the notions of frequency and footprint but relies on the original intensity data and searches over a set of possible cut-off values in the segmentation procedure. As the permutation based p-values are computationally intensive, MSA also contains algorithmic improvements which reduces the execution time.

Another example of a method in this category is GISTIC (Beroukhim et al., 2007). Unlike STAC and MSA, which considers only the location and length of the aberrant intervals, GISTIC also factors in the amplitude of the aberration in each sample. The rationale given for this is that the prevalence of the aberration among samples and the average amplitude of these events are both positively associated with the likelihood that a region carries driver aberrations. Beroukhim et al. (2007) define the G -score as the prevalence of the copy-number change times the average amplitude over the carriers. Permutation tests are used to compute the significance of the observed G -scores, and regions with maximal G -scores are selected. GISTIC has been applied with some success to several different cancers.

It is important to note that, since these methods use segmented data as input, the quality of their results depend on the reliability of the underlying segmentation algorithm. All segmentation methods incur errors, and these methods assume that it is more likely for biologically significant aberrations to recur across samples than experimental or statistical errors. However, many errors in segmentation are due to experimental artifacts, such as local trends, which also recur across samples at the same locations. These artifacts, if not carefully removed, may also be misconstrued as significant recurrent regions. Also, while it is likely that the recurrent regions are due to driver mutations, they may also be a result of biases in the DNA mutation or repair machinery. Thus, care must be taken in their interpretation.

1.5.2 Cross-sample Detection of Inherited Variants

The methods described in the previous section pool information across samples post segmentation. In this section, we consider methods for joint segmentation of a cohort of samples. For detecting weak signals, such joint segmentation methods have been shown to boost power Zhang et al. (2009b). Also, since different samples have different signal quality and noise characteristics, integrating them during segmentation can account for these differences.

When testing for a change in mean in a single sequence, the two quantities that affect the power of detection are the height of the jump and the width of the changed interval. The generalized likelihood ratio statistic (1.3) is a function of these two quantities. When multiple samples are simultaneously scanned, a third quantity, the number of carriers of the change, enters into the equation. When more than one sample show evidence for a change, then the change is more likely to be real. By pooling samples in the segmentation step, we are capitalizing on this fact to boost power.

There are now several models for multi-sample joint segmentation of copy number data. The HMM-based approach of Wang et al. (2008) focused on the analysis of cancer data, and does not assume the change-points to be shared across samples. The authors mention that a shared change-point model would be desirable for the detection of inherited CNVs, and they note the substantial computational task inherent in a satisfactory HMM for this problem. Such an undertaking is reported in Shah et al. (2007), where a multi-layer hierarchical hidden Markov model is used to segment all samples simultaneously. Shah et al. (2007) assumes an underlying “master” Markov chain which decides whether the samples, as a group, should enter a “changed” state. Given that the master has entered the changed state, each sample can choose, with a flip of coin, whether to jump to a shared mean level or to stay at the baseline level. This model assumes that all carriers must change in the same direction with the same magnitude. An MCMC algorithm is proposed to sample from the posterior distribution of the master state, and outputs a sequence of posterior probabilities of change.

Hidden Markov models for this problem rest on many assumptions about the relationship of the aberrations between samples, e.g. they must be in a common direction, or must be present in a majority of the samples. For a less restrictive approach, Zhang et al. (2009) proposed a simultaneous change-point model for the detection of inherited changes. To describe the model, let the observed data is a two dimensional array $\{y_{it} : i = 1, \dots, N, t = 1, \dots, T\}$, where y_{it} is the data point for the i -th sample at probe t , N is the total number of samples, and T is the total number of probes. For each sample i , the random variables $y_i = \{y_{it} : t = 1, \dots, T\}$ are mutually independent and Gaussian with mean values μ_{it} and variances σ_i^2 . The null hypothesis assumes that the means for each profile are identical across locations. The simplest alternative where there is a single changed interval assumes that there exist integer values $1 \leq \tau_1 < \tau_2 \leq T$ and at least one sample i such that

$$\mu_{it} = \mu_i + \delta_i I_{\{\tau_1 < t \leq \tau_2\}}, \quad (1.11)$$

where the δ_i are non-zero constants and μ_i is the baseline mean level for profile i . For this testing problem, a direct generalization of (1.3) is $\max_{s < t} Z(s, t)$, where

$$Z(s, t) = \sum_{i=1}^N U_i^2(s, t) \quad (1.12)$$

and $U_i(s, t)$ is the sequence specific statistic defined as in (1.3) for the i th sequence. If the variances were known, then (1.12) would be the generalized log likelihood ratio statistic for testing H_0 versus H_A . For each fixed $s < t$, the null distribution of $Z(s, t)$ is approximately χ^2 with N degrees of freedom. Large values of $\max_{s < t} Z(s, t)$ are evidence against the null hypothesis. If the null hypothesis is rejected, the maximum likelihood estimate of the location of the variant interval is $(s^*, t^*) = \operatorname{argmax}_{s, t} Z(s, t)$. Zhang et al. (2009b)

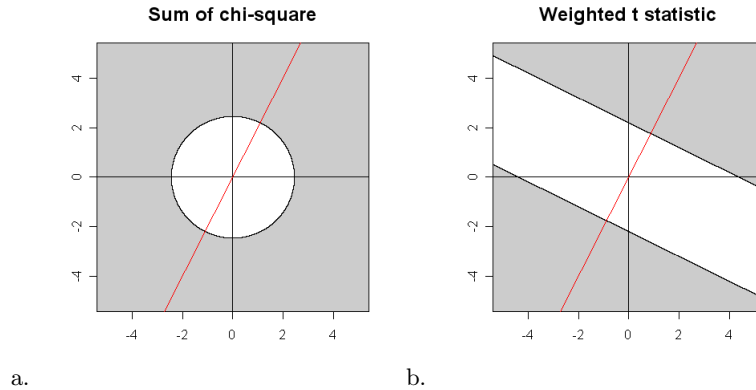


Fig. 1.6. Comparison of the null hypothesis rejection regions between the sum of chi-square statistic (1.12) and the weighted t -statistic (1.8) on $K = 2$ platforms. In all figures, the axes are the magnitudes of the X variables (1.9) for platforms 1 and 2. A significance level of 0.05 is used to determine the decision boundaries of both statistics. For Figure (b), weights of $\delta_1 = 1$, $\delta_2 = 2$ are used. The red line shows the direction of the weight vector $\delta = (\delta_1, \delta_2)$.

developed analytic p-value approximations for scans using statistics of the form (1.12)

Similar to the sum of chi-squares statistic is the “interval scores” method of Lipson et al. (2006), which uses a statistic like $Z(s, t)$ but without the squares. Thus, like Shah et al. (2007) the method focuses on common deletions or common amplifications. However, many inherited CNVs have changes of both types at any given locus. This is because the individual copy numbers are defined relative to the population average, and when two or more copy number levels exist in the population for a given locus, normalizing to the average create both “gains” and “losses” in the population.

To assess the improvement in sensitivity gained from pooling data across samples, Zhang et al. (2009b) used a set of 62 samples assayed using Illumina 550K Beadchips. The experiments were performed on DNA samples extracted from lymphoblastoid cell lines derived from 10 sets of trios consisting of a child and his/her two parents, and 16 pairs of technical replicates for 16 independent DNA samples. Zhang et al. (2009b) showed that using statistic (1.12) improves the concordance rate between replicates and between the child and parent samples, as compared to single sample scans.

Note the differences between the projected chi-squares statistic (1.8), and the sum-of-chi-squares statistic (1.12) for multi-sample segmentation. When pooling data across independent biological samples, not all samples are expected to carry the same CNV, and often both deletions and amplifications can be observed between the samples at the same genome location. This is

why the sum-of chi-squares statistic (1.12) does not “reward” agreement in direction of change between samples. In contrast, the statistic in (1.8) rewards agreement and penalizes disagreement. The difference between the two statistics is shown graphically in Figure 1.6, where we illustrate the simple case of two samples/platforms with the response ratio of the second platform being twice that of the first platform. Note that both statistics are functions of $X = (X_1, X_2)$, which, assuming that σ_k is known, is bivariate Gaussian with mean 0 and identity covariance matrix under the null hypothesis. Figures 1(a-b) show in gray the region in the (X_1, X_2) plane where the null hypothesis will be rejected. For example, in Figure 1a, which depicts the situation in (1.12), the rejection boundary is a circle centered at the origin. In Figure 1b, which depicts the situation in (1.8), the rejection boundary is $\{X : \delta'X > t_\alpha\}$, which is perpendicular to the vector δ_2/δ_1 . Importantly, note that (b) awards agreement between the two platforms, while (a) treat all quadrants of the plane equally. The statistic (1.8, Figure 1b) 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.

1.5.3 Obtaining a Cross-sample Signature

As before, consider a set of copy number profiles $\{y_{it} : i = 1, \dots, N, t = 1, \dots, T\}$ over n samples and T probes. A sparse cross-sample signature can be defined as a set of genomic regions $R = \{(s_i, t_i) : s_i < t_i, i = 1, \dots, M\}$ and an associated $n \times M$ matrix X such that an approximation \hat{y} for y can be constructed solely using the information in R and X . To evaluate the approximation, one may use the sum of squared errors

$$\sum_{i=1}^T \sum_{j=1}^N (y_{ij} - \hat{y}_{ij})^2.$$

We seek signatures where $M \ll T$, so as to reduce the dimension of the data set. Then, the matrix X can be used in place of the matrix y in downstream clustering, classification, and regression modeling. This is still largely an open problem, with much ongoing work. Here, we describe a solution given in Zhang et al. (2009b) based on an extension of the CBS algorithm:

Algorithm 1 (Multi-sample Circular Binary Segmentation) *Fix the global significance level α , parameter p , and a maximum window $T_0 < T$. We denote by $Y_{h:k}$ the matrix $\{y_{i,t} : i = 1, \dots, N, t = h, \dots, k\}$.*

1. Initialize $T_1 = 1$ and $T_2 = T$.
2. Compute

$$Z_{max} = \max_{\substack{T_1 \leq s < t \leq T_2 \\ 1 \leq t-s \leq T_0}} \{Z(s, t)\}.$$

Let (s^*, t^*) be the maximizing interval.

3. If the p -value of Z_{max} , as computed using approximations give in in Zhang et al. (2009b), is less than α , then for each $(u, v) \in \{(T_1, s^* - 1), (s^*, t^*), (t^* + 1, T_2)\}$, do:
 - a) Determine the carriers of the variant. For all $t = u, \dots, v$, if a sample carries the variation, let $\hat{y}_{i,t} = \bar{y}_{i,u:v}$, and for the other samples let $\hat{y}_{i,t} = \bar{y}_{i,T_1:T_2}$. Let $Y'_{u:v} = Y_{u:v} - \hat{Y}_{u:v}$, where $\hat{Y}_{u:v}$ is the matrix $\{\hat{y}_{i,t} : i = 1, \dots, N, t = u, \dots, v\}$.
 - b) Repeat steps 2-3 for $T_1 = u, T_2 = v$ and the newly normalized $Y'_{u:v}$.

This algorithm, like CBS, recursively scans the genome for intervals that maximize the sum of chi-square statistic. If the p -value of the maximum is smaller than a pre-defined threshold, a cut is made at the maximizing interval. Then, the carriers of the variant, i.e. samples whose mean level actually changes in the interval, are determined. There are myriad ways of determining the carriers, some simple ad hoc solutions are given in Zhang et al. (2009b). A box-shaped signal is estimated for the carriers, while a flat line is fitted for the non-carriers. Residuals are taken, and the regions to the left, center, and right of the cut are recursively segmented using the same procedure. The algorithm outputs R in the form of a set of change-points and X in the form of the fitted value of each sample between each change-point.

The multi-sample CBS algorithm was used to analyze the 9p21 deletion in childhood leukemia Schiffman et al. (2009), and showed that pattern of deletion in this region can be used to characterize precursor B-cell acute lymphoblastic leukemia, as compared with precursor T-cell acute lymphoblastic leukemia patients.

1.6 Concluding Remarks

In this chapter, we have surveyed some of the recent developments in the analysis of DNA copy number data from microarray platforms. These analyses started with a focus on single sample total copy number segmentation. Now, new statistical and computational challenges arise in the proper extraction of allele-specific information from DNA genotyping arrays, and in the analysis of DNA copy number data from multi-sample, multi-platform experiments. These applications have inspired new developments in change-point models, especially in the formulation of simultaneous change-point models over multiple sequences. The theory underlying these models are quite general, and may prove useful to other applications, especially other types of genome-wide profiling.

There is much ongoing work on the integrated analysis of DNA copy number data with gene expression, protein expression, and methylation data. Although such studies seem well motivated on the biological side, there is a shortage of statistical models in this area. For recent progress on this problem, see Peng et al. (2009b) and Peng et al. (2009a).

Since detection of copy number variants is a zero cost by-product of the genotyping arrays used in association studies, there is strong motivation for using them in existing association studies. However, as for SNPs, much further statistical work and biological evidence is needed to determine how to utilize the copy number information in studies of genetic inheritance. Advances in this area rest on the understanding of how CNVs segregate in a population (Redon et al., 2006), and the selective pressures acting on CNVs (Cooper et al., 2008). For a nice survey on this problem, see Mccarroll (2008).

In interpreting the results from microarray-based copy number studies, we must carefully note that these assays are noisy and prone to cross hybridization, especially in repetitive regions or regions with complex rearrangements (Cooper et al., 2008). Concordance of CNVs detected using microarrays with those detected through sequencing experiments is incredibly low (Cooper et al., 2008; McCarroll et al., 2008; Zhang et al., 2009a). To date, biological confirmation of CNV detection methods has been limited to small-scale experiments involving, for example, male vs. female copy number on the X chromosome or confirmation of a few known CNVs. New sequencing-based technologies, especially paired end mapping (Korbel et al., 2007; Kidd et al., 2008), will be invaluable complementary tools in CNV discovery.

References

- BENGTSSON, H., IRIZARRY, R., CARVALHO, B. & SPEED, T. (2008). Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* **24**, 759–767.
- BENGTSSON, H., RAY, A., SPELLMAN, P. & SPEED, T. P. (2009). A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. *Bioinformatics (Oxford, England)* **25**, 861–867.
- BEROUKHIM, R., GETZ, G., NGHIEMPHU, L., BARRETINA, J., HSUEH, T., LINHART, D., VIVANCO, I., LEE, J. C., HUANG, J. H., ALEXANDER, S., DU, J., KAU, T., THOMAS, R. K., SHAH, K., SOTO, H., PERNER, S., PRENSNER, J., DEBIASI, R. M., DEMICHELIS, F., HATTON, C., RUBIN, M. A., GARRAWAY, L. A., NELSON, S. F., LIAU, L., MISCHER, T. F., MEYERSON, M., GOLUB, T. A., LANDER, E. S., MELLINGHOFF, I. K. & SELLERS, W. R. (2007). Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences*, 0710052104+.
- BEROUKHIM, R., LIN, M., PARK, Y., HAO, K., ZHAO, X., GARRAWAY, L. A., FOX, E. A., HOCHBERG, E. P., MELLINGHOFF, I. K., HOFER, M. D., DESCAZEAUD, A., RUBIN, M. A., MEYERSON, M., WONG, W. H., R., S. W. & C., L. (2006). Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide snp arrays. *PLoS Computational Biology* **2**, e41.

- BIGNELL, G. R., HUANG, J., GRESHOCK, J., WATT, S., BUTLER, A., WEST, S., GRIGOROVA, M., JONES, K. W., WEI, W., STRATTON, M. R., FUTREAL, P. A., WEBER, B., SHAPERO, M. H. & WOOSTER, R. (2004). High-resolution analysis of dna copy number using oligonucleotide microarrays. *Genome Research* **14**, 287–295.
- BROËT, P. & RICHARDSON, S. (2006). Detection of gene copy number changes in cgh microarrays using a spatially correlated mixture model. *Bioinformatics* **22**, 911–918.
- CHEN, H., XING, H. & ZHANG, N. R. (2009). Estimation of parent specific dna copy number in tumors using high-density genotyping arrays. *Technical Report, Department of Statistics, Stanford University*.
- COLELLA, S., YAU, C., TAYLOR, J. M., MIRZA, G., BUTLER, H., CLOUSTON, P., BASSETT, A. S., SELLER, A., HOLMES, C. C. & RAGOISSIS, J. (2007). Quantisnp: an objective bayes hidden]markov model to detect and accurately map copy number variation using snp genotyping data. *Nucleic Acids Research* **35**.
- CONRAD, D., ANDREWS, T., CARTER, N., HURLES, M., & PRITCHARD, J. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics* **38**, 75–81.
- COOPER, G. M. M., ZERR, T., KIDD, J. M. M., EICHLER, E. E. E. & NICKERSON, D. A. A. (2008). Systematic assessment of copy number variant detection via genome-wide snp genotyping. *Nature genetics* **40**, 1199–1203.
- DISKIN, S. J., ECK, T., GRESHOCK, J., MOSSE, Y. P., NAYLOR, T., STOECKERT JR., C. J., WEBER, B. L., MARIS, J. M. & GRANT, G. R. (2006). Stac: A method for testing the significance of dna copy number aberrations across multiple array-cgh experiments. *Genome Research* **16**, 1149–1158.
- ENGLER, D., MOHAPATRA, G., LOUIS, D. & BETENSKY, R. (2006). A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics* **7**, 399–421.
- ESTIVILL, X. & ARMENGOL, L. (2007). Copy number variants and common disorders: Filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genetics* **3**, e190+.
- FRIDLYAND, J., SNIJDERS, A., PINKEL, D., ALBERTSON, D. G. & JAIN, A. (2004). Application of hidden markov models to the analysis of the array-cgh data. *Journal of Multivariate Analysis* **90**, 132–153.
- GUHA, S., LI, Y. & NEUBERG, D. (2006). Bayesian hidden markov modeling of array cgh data. *Harvard University Biostatistics Working Paper Series*.
- GUTTMAN, M., MIES, C., DUDYCZ-SULICZ, K., DISKIN, S. J., BALDWIN, D. A., STOECKERT, C. J. & GRANT, G. R. (2007). Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genetics* **3**, e143+.

- HSU, L., SELF, S., GROVE, D., RANDOLPH, T., WANG, K., DELROW, J., LOO, L. & PORTER, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* **6**, 211–226.
- HUPÉ, P., STRANSKY, N., THIERY, J. P., RADVANYI, F. & BARILLOT, E. (2004). Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Bioinformatics* **20**, 3413–3422.
- ISHKANIAN, A. S., MALLOFF, C. A., WATSON, S. K., DELEEUW, R. J., CHI, B., COE, B. P., SNIJDERS, A., ALBERTSON, D. G., PINKEL, D., MARRA, M. A., LING, V., MACAULAY, C. & LAM, W. L. (2004). A tiling resolution dna microarray with complete coverage of the human genome. *Nat Genet* **36**, 299–303.
- JAMES, B., JAMES, K. & SIEGMUND, D. (1987). Tests for a change-point. *Biometrika* **74**, 71–83.
- KHAJA, R., ZHANG, J., MACDONALD, J., HE, Y., JOSEPH-GEORGE, A., WEI, J., RAFIQ, M.A. AND, Q. C., SHAGO, M., PANTANO, L., ABURATANI, H., JONES, K., REDON, R., HURLES, M., ARMENGOL, L., ESTIVILL, X., MURAL, R., LEE, C., SCHERER, S. & FEUK, L. (2007). Genome assembly comparison to identify structural variants in the human genome. *Nature Genetics* **38**, 1413–1418.
- 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. & EICHLER, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64.
- KORBEL, J. O., URBAN, A. E., AFFOURTIT, J. P., GODWIN, B., GRUBERT, F., SIMONS, J. F., KIM, P. M., PALEJEV, D., CARRIERO, N. J., DU, L., TAILLON, B. E., CHEN, Z., TANZER, A., SAUNDERS, A. C., CHI, J., YANG, F., CARTER, N. P., HURLES, M. E., WEISSMAN, S. M., HARKINS, T. T., GERSTEIN, M. B., EGHOLM, M. & SNYDER, M. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426.
- LAFRAMBOISE, T., WEIR, B. A., ZHAO, X., BEROUKHIM, R., LI, C., HARRINGTON, D., R., S. W. & MEYERSON, M. (2005). Allele-specific amplification in cancer revealed by snp array analysis. *PLoS Computational Biology* **1**, e65.
- LAI, T. L., XING, H. & ZHANG, N. R. (2007). Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics* **9**, 290–307.

- 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.
- LIN, M., WEI, L.-J., SELLERS, W. R., LIEBERFARB, M., WONG, W. H. & LI, C. (2004). dchipsnp: significance curve and clustering of snp-array-based loss-of-heterozygosity data. *Bioinformatics* **20**, 1233–1240.
- LIPSON, D., AUMANN, Y., BEN-DOR, A., LINIAL, N. & YAKHINI, Z. (2006). Efficient calculation of interval scores for dna copy number data analysis. *Journal of Computational Biology* **13**, 215–228.
- MCCARROLL, S., HADNOTT, T., PERRY, G., SABETI, P., ZODY, M., BARRETT, J., DALLAIRE, S., GABRIEL, S., LEE, C., DALY, M., ALTSHULER, D. & THE INTERNATIONAL HAPMAP CONSORTIUM (2006). Common deletion polymorphisms in the human genome. *Nature Genetics* **38**, 86–92.
- MCCARROLL, S. A. (2008). Copy-number analysis goes more than skin deep. *Nature Genetics* **40**, 5–6.
- MCCARROLL, S. A. A., KURUVILLA, F. G. G., KORN, J. M. M., CAWLEY, S., NEMESH, J., WYSOKER, A., SHAPERO, 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. & ALTSHULER, D. (2008). Integrated detection and population-genetic analysis of snps and copy number variation. *Nature genetics* **40**, 1166–1174.
- NEWTON, M., GOULD, M., REZNIKOFF, C. & HAAG, J. (1998). On the statistical analysis of allelic-loss data. *Statistics in Medicine* **17**, 1425–1445.
- NEWTON, M. & LEE, Y. (2000). Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data. *Biometrics* **56**, 1088–1097.
- 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.
- PEIFFER, D. A., LE, J. M., STEEMERS, F. J., CHANG, W., JENNIGES, T., GARCIA, F., HADEN, K., LI, J., SHAW, C. A., BELMONT, J., CHEUNG, S. W., SHEN, R. M., BARKER, D. L. & GUNDERSON, K. L. (2006). High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome Research* **16**, 1136–1148.
- PENG, J., WANG, P., ZHOU, N. & ZHU, J. (2009a). Partial correlation estimation by joint sparse regression model. *Journal of the American Statistical Association* **104**, 735–746.
- PENG, J., ZHU, J., BERGAMASCHI, A., HAN, W., NOH, D., POLLACK, J. & WANG, P. (2009b). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics*, in press. .

- PICARD, F., ROBIN, S., LAVIELLE, M., VAISSE, C. & DAUDIN, J. (2005). A statistical approach for array cgh data analysis. *BMC Bioinformatics* **6**, 27.
- PINKEL, D., SEGRAVES, R., SUDAR, D., CLARK, S., POOLE, I., KOWBEL, D., COLLINS, C., KUO, W. L., CHEN, C., ZHAI, Y., DAIRKEE, S. H., LJUNG, B. M., GRAY, J. W. & ALBERTSON, D. G. (1998). High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**, 207–11.
- POLLACK, J., PEROU, C., ALIZADEH, A., EISEN, M., PERGAMENSCHIKOV, A., WILLIAMS, C., JEFFREY, S., BOTSTEIN, D. & BROWN, P. (1999). Genome-wide analysis of dna copy-number changes using cdna microarrays. *Nature Genetics* **23**, 41–46.
- REDON, R., ISHIKAWA, S., FITCH, K. R., FEUK, L., PERRY, G. H., ANDREWS, D. T., FIEGLER, H., SHAPERO, M. H., CARSON, A. R., CHEN, W., CHO, E. K., DALLAIRE, S., FREEMAN, J. L., GONZALEZ, J. R., GRATACOS, M., HUANG, J., KALAITZOPOULOS, D., KOMURA, D., MACDONALD, J. R., MARSHALL, C. R., MEI, R., MONTGOMERY, L., NISHIMURA, K., OKAMURA, K., SHEN, F., SOMERVILLE, M. J., TCHINDA, J., VALSESIA, A., WOODWARK, C., YANG, F., ZHANG, J., ZERJAL, T., ZHANG, J., ARMENGOL, L., CONRAD, D. F., ESTIVILL, X., TYLER-SMITH, C., CARTER, N. P., ABURATANI, H., LEE, C., JONES, K. W., SCHERER, S. W. & HURLES, M. E. (2006). Global variation in copy number in the human genome. *Nature* **444**, 444–454.
- ROUVEIROL, C., STRANSKY, N., HUPÉ, P., LA ROSA, P., VIARA, E., BARRILLOT, E. & RADVANYI, F. (2006). Computation of recurrent minimal genomic alterations from array-cgh data. *Bioinformatics* **22**, 849–856.
- SCHIFFMAN, J. D., WANG, Y., MCPHERSON, L. A., WELCH, K., ZHANG, N., DAVIS, R., LACAYO, N. J., DAHL, G. V., FAHAM, M. & FORD, J. M. (2009). Molecular inversion probes reveal patterns of 9p21 deletion and copy number aberrations in childhood leukemia. *Cancer Genetics and Cytogenetics* **193**, 9–18.
- SHAH, S. P., LAM, W. L., NG, R. T. & MURPHY, K. P. (2007). Modeling recurrent dna copy number alterations in array cgh data. *Bioinformatics* **23**, 450–458.
- SHENDURE, J., MITRA, R. D., VARMA, C. & CHURCH, G. M. (2004). Advanced sequencing technologies: methods and goals. *Nature reviews. Genetics* **5**, 335–344.
- SNIJDERS, A. M., NOWAK, N., SEGRAVES, R., BLACKWOOD, S., BROWN, N., CONROY, J., HAMILTON, G., HINDLE, A. K., HUEY, B., KIMURA, K., LAW, S., MYAMBO, K., PALMER, J., YLSTRA, B., YUE, J. P., GRAY, J. W., JAIN, A. N., PINKEL, D. & ALBERTSON, D. G. (2001). Assembly of microarrays for genome-wide measurement of dna copy number. *Nature genetics*. **29**, 263–264.
- STAUF, J., LINDGREN, D., VALLON-CHRISTERSSON, J., ISAKSSON, A., GORANSSON, H., JULIUSSON, G., ROSENQUIST, R., HOGLUND, M., BORG,

- A. & RINGNER, M. (2008). Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome snp arrays. *Genome Biology* **9**, R136+.
- TAYLOR, B. S., BARRETINA, J., SOCCI, N. D., DECAROLIS, P., LADANYI, M., MEYERSON, M., SINGER, S. & SANDER, C. (2008). Functional copy-number alterations in cancer. *PLoS ONE* **3**, e3179+.
- TIBSHIRANI, R. & WANG, P. (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics* **9**, 18–29.
- VENKATRAMAN, E. & OLSHEN, A. (2007). A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics* **23**, 657–663.
- WANG, H., VELDINK, J. H., OPHOFF, R. A. & SABATTI, C. (2008). Markov models for inferring copy number variations from genotype data on illumina platforms. *Technical Report, Dept. of Statistics, University of California at Los Angeles*.
- WANG, K., LI, M., HADLEY, D., LIU, R., GLESSNER, J., GRANT, S. F. A., HAKONARSON, H. & BUCAN, M. (2007). Penncnv: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome Research* **17**, 1665–1674.
- 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.
- WEN, C., WU, Y., HUANG, Y., CHEN, W., LIU, S., JIANG, S., JUANG, J., LIN, C., FANG, W., HSIUNG, C. & CHANG, I. (2006). A bayes regression approach to array-cgh data. *Statistical Applications in Molecular Biology* **5**.
- WILLENBROCK, H. & FRIDLAND, J. (2005). A comparison study: applying segmentation to arraycgh data for downstream analyses. *Bioinformatics* **21**, 4084–4091.
- XING, B., GREENWOOD, C. M. T. M. & BULL, S. B. B. (2007). A hierarchical clustering method for estimating copy number variation. *Biostatistics* **8**, 632–653.
- ZHANG, N., SENBABAOGU, Y. & LI, J. Z. (2009a). Joint estimation of dna copy number from multiple platforms. *Technical Report, Department of Statistics, Stanford University*.
- ZHANG, N. & 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. & LI, J. Z. (2009b). Detecting simultaneous change- points in multiple sequences. *Biometrika* **in press**.