Statistical analysis of Single Nucleotide Polymorphism microarrays in cancer studies

Pierre Neuvial, Henrik Bengtsson and Terence P. Speed

Abstract In this chapter, we focus on statistical questions raised by the identification of copy number alterations in tumor samples using genotyping microarrays, also known as Single Nucleotide Polymorphism (SNP) arrays. We define the copy number states formally, and show how they are assessed by SNP arrays. We identify and discuss general and cancer-specific challenges for SNP array data preprocessing, and how they are addressed by existing methods. We review existing statistical methods for the detection of copy number changes along the genome. We describe the influence of two biological parameters — the proportion of normal cells in the sample, and the ploidy of the tumor — on observed data. Finally, we discuss existing approaches for the detection and calling of copy number aberrations in the particular context of cancer studies, and identify statistical challenges that remain to be addressed.

Pierre Neuvial
Department of Statistics, University of California, Berkeley, USA
e-mail: pierre@stat.berkeley.edu

Henrik Bengtsson
Department of Statistics, University of California, Berkeley, USA and
Department of Epidemiology & Biostatistics, University of California, San Francisco, USA
e-mail: hb@stat.berkeley.edu

Terence P. Speed
Department of Statistics, University of California, Berkeley, USA and
Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research, Parkville, Australia
e-mail: terry@stat.berkeley.edu

Contents

1 From biological questions to statistical challenges

Each normal human cell has 23 pairs of chromosomes. For each of them, one chromosome has been inherited from each biological parent. Tumor cells harbor numerous structural alterations of their DNA including point mutations, translocations, small insertion or deletion events, larger scale copy number changes, amplifications, and loss of heterozygosity (LOH), which corresponds to the loss of the contribution of one parent in a genomic region. These alterations can affect genes and regulatory transcripts, which may result in cellular modifications including angiogenesis, immune evasion, metastasis, and altered cell growth, death and metabolism [1]. They are thought to be associated with diagnostic and prognostic factors [2].

An immediate goal of copy number studies in cancer research is to estimate the underlying *copy number state* (to be defined more formally in the next section) at each position along the genome of a tumor sample. Microarray-based technologies have been used for more than a decade to quantify copy numbers at a large number of genomic loci [2, 3, 4]. In particular, genotyping microarrays (SNP arrays) are a technology of choice because they combine a high-density of markers along the genome (in the order of millions for the current generation) with the ability to assess both changes in total copy number and loss of heterozygosity in a single assay. This is what make them particularly relevant to cancer studies, where both pieces of information are needed to understand the underlying copy number state of the tumor.

In this chapter, we review statistical challenges raised by the analysis of SNP array data in cancer studies. We focus on the analysis of *one* tumor sample. Identifying copy number states from a tumor sample requires *detecting* changes in copy number signals, and *calling* regions, that is, assigning a copy number state to each region detected. The main ingredient for the detection part is the fact that DNA copy number is *locally constant* along the genome: *locus-level* estimates can thus be combined in *region-level* estimates. However, for this property of local constancy to be fully exploited, SNP array data first have to be pre-processed so that locus-level estimates for a given sample are comparable across loci. For the calling step to be performed satisfactorily, biological factors that influence the estimated copy number levels — tumor ploidy and normal contamination — have to be understood and acknowledged for.

Outline

We begin by defining the copy number states of interest in cancer studies, and showing how estimates can be obtained from preprocessed SNP array data for each locus (Section 2). We then describe current methods for SNP array data preprocessing, with a focus on specific challenges for copy number studies in

cancers (Section 3). In Section 4 we review statistical methods that have been proposed to combine locus-level copy number estimates (as obtained after preprocessing) to detect copy number changes along the genome. In Section 5 we describe the influence of tumor ploidy and normal contamination on observed signals and their interpretation. In Section 6 we show how the methods described in Section 4 have been applied to SNP array data in cancer studies, by accounting for or taking advantage of the characteristics of the data described in Section 5. We conclude by identifying ongoing challenges for the statistical analysis of SNP array data in cancer studies (Section 7).

2 Minor and major copy numbers in cancer studies

We define the *copy number state* of a tumor at a given genomic locus j as a pair of numbers $(\underline{\gamma}_j, \overline{\gamma}_j)$, where $\underline{\gamma}_j \geq 0$ and $\overline{\gamma}_j \geq 0$ are respectively the smaller and the larger of the two parental copy numbers at this locus. By definition we have $\underline{\gamma}_j \leq \overline{\gamma}_j$, and $\gamma_j = \underline{\gamma}_j + \overline{\gamma}_j$ is the total copy number. The quantities $\underline{\gamma}_j$ and $\overline{\gamma}_j$ are called minor and major copy numbers, respectively. Note that $\underline{\gamma}_j$, $\overline{\gamma}_j$, and γ_j need not be whole numbers, especially because of the possible presence of normal cells in the tumor sample. This point is explained in detail in Section 5.

The two-dimensional vector $(\underline{\gamma}_j, \overline{\gamma}_j)$ does not characterize parental copy numbers at locus j in the tumor. Indeed, the information of which of minor or major copy numbers corresponds to the maternal chromosome at locus j, and which one corresponds to the paternal chromosome is missing from $(\underline{\gamma}_j, \overline{\gamma}_j)$, and it may change across loci. In short, because of the constraint $\underline{\gamma}_j \leq \overline{\gamma}_j$, minor and major copy numbers (CNs) are not *phased* in terms of parental copy numbers.

The remainder of this section is organized as follows. In Section 2.1 we focus on *true* copy number signals, that is, the actual copy numbers in the biological samples. We demonstrate that knowing true minor and major copy numbers is enough to characterize copy number events of interest in cancer studies. In Section 2.1 we show that true copy numbers, including minor and major copy numbers, can be *estimated* from SNP array data at the locus level. Notation used in the chapter is summarized in Section 2.3.

2.1 Information relevant to copy number studies in cancers

Table 1 summarizes the copy number states relevant to cancer studies in terms of minor and major copy numbers. They are described as the conjunction of information regarding total copy numbers and (loss of) heterozygosity. For

example, knowing the total copy number in a region of LOH ($\underline{\gamma} = 0$) allows us to distinguish between hemizygous deletions $(\underline{\gamma}, \overline{\gamma}) = (0, 1)$, that is, single copy deletions, from LOH when the total copy number is two $(0, 2)$, so-called copy-neutral LOH or acquired uniparental disomy. Conversely, among regions of neutral copy number ($\gamma = 2$), regions of copy-neutral LOH $(0, 2)$ can be distinguished from normal regions $(1, 1)$ based on the LOH status of the region. This is distinction is important for data interpretation, as copy-neutral LOH is a known mechanism through which a recessive tumor suppressor gene can be expressed with no apparent change in total copy number [5].

	Deletion	Neutral	Gain
Loss of Heterozygosity	$(0, 1)$	$(0, 2)$	$(0, \overline{\gamma})$ with $\overline{\gamma} \geq 3$
Heterozygosity	$(0, 0)$	$(1, 1)$	$(\underline{\gamma}, \overline{\gamma})$ with $1 \leq \underline{\gamma} \leq \overline{\gamma}$

Table 1 *Minor and major copy number states of interest for cancer studies, presented as the conjunction of information regarding total copy number (columns) and heterozygosity status (rows).*

Regions of LOH are characterized by the absence of one of the two parental chromosomes, that is, by a null minor copy number: $\underline{\gamma}_j = 0$. However, (loss of) heterozygosity is a binary concept which can be insufficient (even when combined with total copy numbers) to fully characterize subtle copy number events such as complex gains, as in the lower right cell of Table 1 which corresponds to a copy number gain with retention of heterozygosity. For example, $(1, 3)$ and $(2, 2)$ are two states that fall into this category, with the same total copy number. However, the biological interpretation of these two states can be quite different: $(2, 2)$ is a balanced duplication of a chromosomal region, while $(1, 3)$ corresponds to an allele-specific amplification, which can typically pinpoint regions containing oncogenes.

This example illustrates the need for a quantitative measure to characterize *allelic imbalance* between parental copy numbers at a given locus, rather than a binary variable (retention or loss of heterozygosity). Several closely related measures have been proposed to quantify allelic imbalance in cancers [6, 7, 8]. These measures can be written in terms of minor and major copy numbers and quantify the distance tho the heterozygous status. In this chapter, we denote the *allelic imbalance* at locus j by $\delta_j \in [0, 1]$, and use the following definition:

$$\delta_j = \frac{\overline{\gamma}_j - \underline{\gamma}_j}{\overline{\gamma}_j + \underline{\gamma}_j} \,. \tag{1}$$

In the above example, $(\underline{\gamma}, \overline{\gamma}) = (2, 2)$ yields $\delta = 0$ (allelic balance or heterozygosity), while $(1, 3)$ yields $\delta = 1/2$ (*partial* loss of heterozygosity). Note how a hemizygous deletion $(0, 1)$ and a copy-neutral LOH $(0, 2)$ both yield $\delta = 1$.

2.2 What can be estimated from SNP array data

Single Nucleotide Polymorphisms (SNPs) are genomic positions where the DNA sequence varies at a substantial rate across individuals of some population. For most SNPs only two (out of four) variants are observed. These variants are called *alleles* and arbitrarily denoted by A and B. SNP arrays are a microarray-based technology which targets both alleles of a large number of SNPs. Although they were originally developed for genotyping studies, they have also been proved quite useful for copy number studies, especially in cancers.

Current generations of SNP arrays (Affymetrix GenomeWideSNP_6 and Illumina Human1M-Duo) interrogate approximately one million SNPs, that is, of the order of 10% of the total number of known human SNPs. They also incorporate copy number probes, which measure total copy numbers at non-polymorphic loci for increased resolution of copy number studies. We refer to [9] for a more comprehensive review on SNP array technologies. Specific characteristics of SNP array assays that are relevant to the data analysis and particularly to data preprocessing are explained in more detail in Section 3.

For the present section it is sufficient to note that SNP array data (after preprocessing as explained in Section 3) can be summarized by a two-dimensional vector $(c_j, b_j)_{j\in\mathcal{J}}$ of *locus-level estimates*, where $\mathcal{J}$ denotes the set of J loci targeted by the microarray. When j is a SNP, c_j is the sum of the contribution of the two alleles at j called allele-specific copy numbers, and b_j is the corresponding fraction of signal coming from allele B at j. Following [10, 11, 6], b_j will be called *allele B fraction*. The corresponding allele A fraction is $a_j = 1 - b_j$. The corresponding allele-specific copy numbers A and B can therefore be written as $(a_j c_j, b_j c_j)_{j\in\mathcal{J}}$. When j is a copy number probe, c_j is the total intensity signal at j, while b_j and a_j are not defined.

Figure 1 shows Affymetrix GenomeWideSNP_6 data 50Mb-long genomic region on Chromosome 2 of an ovarian tumor sample from the Cancer Genome Atlas (TCGA). TCGA is a collaborative initiative to provide a high-throughput molecular characterization of a large number of tumors from different cancer types, with the goal to improve biological understanding and clinical treatment of these cancers [12, 13]. These data have been preprocessed using an allele-specific version of the CRMAv2 method [14], called AS-CRMAv2, followed by the TumorBoost method [15] for normalization of raw allele-specific copy numbers.

Previous copy number analyses led by TCGA have shown that this tumor has two copy number transitions in this region. The first one occurs at ∼124.2Mb, between a normal region: $(\underline{\gamma}, \overline{\gamma}) = (1, 1)$ and a region of single chromosome gain: $(\underline{\gamma}, \overline{\gamma}) = (1, 2)$. The second transition occurs at ∼140.9Mb, between a region of single gain and a region of copy-neutral LOH: $(\underline{\gamma}, \overline{\gamma}) = (0, 2)$.

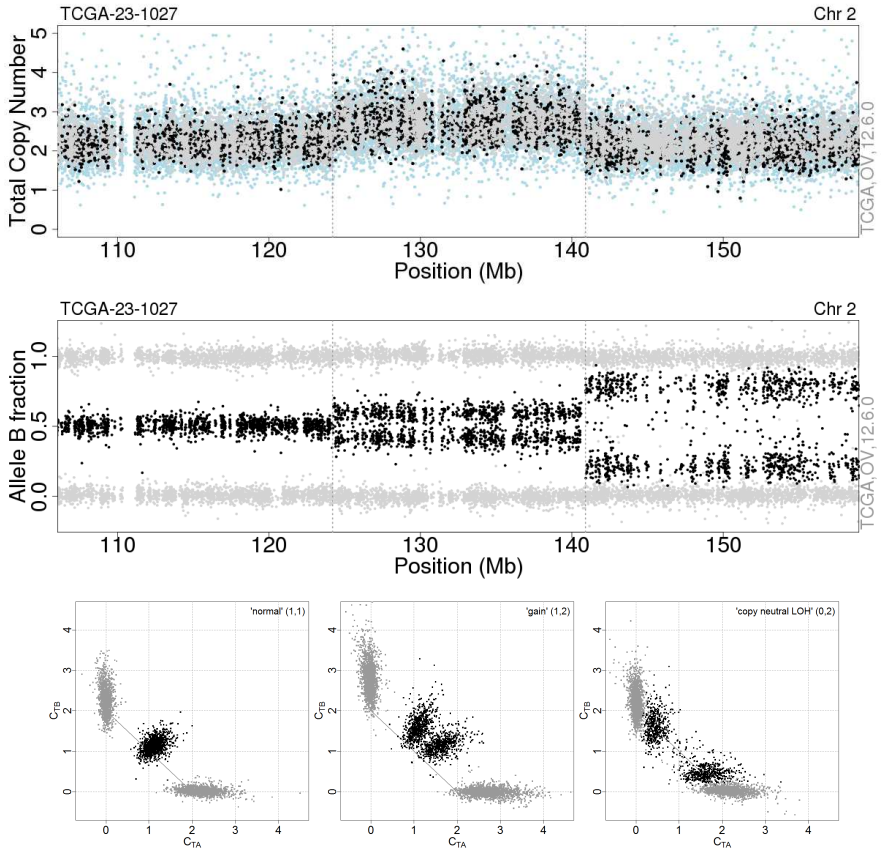

Fig. 1 Locus-level estimates from Affymetrix GenomeWideSNP_6 data in three copy number regions on chromosome 2 of a TCGA ovarian tumor sample: normal $(1,1)$, gain $(1,2)$ and copy-neutral LOH $(0,2)$. Top panel, total copy numbers (c_j) along chromosome 2. Middle panel, allelic ratios (b_j) along chromosome 2. Transitions between the three copy number states are indicated by dashed gray vertical lines in the top and middle panels. Bottom panels, allele-specific copy numbers: $(a_j c_j, b_j c_j)$ in each of the three regions. Light blue: copy number probes; gray: SNPs called homozygous in the paired normal sample; black: SNPs called heterozygous in a paired normal sample (not shown). The data were preprocessed using AS-CRMAv2 [14] followed by TumorBoost [15].

Obtaining locus-level estimates of minor and major copy numbers, and allelic imbalances.

For any configuration of the paternal and maternal genotypes at SNP j, true allelic ratios α_j and β_j satisfy

$$(\alpha_j, \beta_j) \in \left\{0, \underline{\gamma}_j/\gamma_j, \overline{\gamma}_j/\gamma_j, 1\right\} , \tag{2}$$

with the constraint $\alpha_j + \beta_j = 1$. In particular, if SNP j is heterozygous in the germline, then by definition the alleles inherited from the two parents at this locus differ, and the minimum and maximum allelic ratios satisfy $\min(\alpha_j, \beta_j) = \underline{\gamma}_j/\gamma_j$ and $\max(\alpha_j, \beta_j) = \overline{\gamma}_j/\gamma_j$. Therefore, minor and major copy numbers may be estimated as the locus level by

$$\begin{cases} \underline{c}_j & = c_j \cdot \min(a_j, b_j) \\ \overline{c}_j & = c_j \cdot \max(a_j, b_j) \end{cases} . \tag{3}$$

The true allelic imbalance as defined in Equation (1) may then be written as $\delta_j = 1 - 2 \cdot \min(\alpha_j, \beta_j)$, and the corresponding locus-level estimate becomes

$$d_j = 1 - 2 \cdot \min(a_j, b_j) \, . \tag{4}$$

2.3 Notation

The notation used in this chapter for true copy number signals (Greek letters) and the corresponding *locus-level* estimates (Roman letters) is gathered in Table 2.

	True	Locus-level estimate	Locus type
Total copy number	γ_j	c_j	SNP and CN
Allele A fraction	α_j	a_j	SNP
Allele B fraction	$\beta_j = 1 - \alpha_j$	$b_j = 1 - a_j$	
Minor copy number	$\underline{\gamma}_j = \gamma_j \cdot \min(\alpha_j, \beta_j)$	$\underline{c}_j = c_j \cdot \min(a_j, b_j)$	Heterozygous SNP
Major copy number	$\overline{\gamma}_j = \gamma_j \cdot \max(\alpha_j, \beta_j)$	$\overline{c}_j = c_j \cdot \max(a_j, b_j)$	
Allelic imbalance	$\delta_j = (\overline{\gamma}_j - \underline{\gamma}_j)/\gamma_j$	$d_j = 1 - 2 \cdot \min(a_j, b_j)$	

Table 2 *Notation: true copy numbers and corresponding locus-level estimates from SNP arrays.*

3 Preprocessing

The goal of this section is to explain how *locus-level estimates* for total copy numbers (c_j) and allelic ratios (a_j and $b_j = 1 - a_j$), as defined in Section 2, can be obtained from the observed signal intensities retrieved from SNP array experiments. We focus on the two main SNP array platforms, which are manufactured by Affymetrix [16, 17] and Illumina [18, 19, 20]. The first steps that have to be carried out for low-level analysis of microarray data consist in correcting data for sources of unwanted variation, in order to make ob-

served signals *comparable across samples for a given locus*. These steps are described in Section 3.1. We note that the methods described in this section are generally technology-specific, but not specific to cancer studies — they are relevant to any SNP array data analysis. In cancer studies however, the observed signals also need to be *comparable across loci for a given sample*, so that downstream analysis methods can take advantage of the local constancy of the signal along the genome to combine locus-level estimates into region-level estimates. This question is addressed in Section 3.2. Note that because the methods developed in Section 3.2 rely on *reference samples* for the estimation of copy numbers, their application requires making signal intensities comparable across samples (as explained in Section 3.1) in the first place. Section 3.2 is not technology-specific; however it is only relevant to copy number studies in cancers.

3.1 Making signals comparable across samples

SNP arrays were originally developed and used for genotyping purposes in genome-wide association studies (GWAS). Genotype calls are generally estimated independently for each SNP, by comparing the distribution of allelic signals across samples. Necessarily, preprocessing methods for SNP arrays were initially focused on making signals comparable across samples. In this section, we briefly review the design principles of existing methods addressing this point. As Affymetrix and Illumina assays are quite different, these methods are mostly platform-specific.

3.1.1 Affymetrix

A variety of preprocessing methods have been suggested for Affymetrix SNP arrays, e.g. (implicit or explicit) background correction, allelic-crosstalk calibration, probe-sequence normalization, PCR fragment-length normalization, several distribution-based normalization methods, and various methods summarizing probe-level signals into locus-level estimates.

Correction of PCR and sequence related effects. Affymetrix genotyping assays involve a Polymerase Chain Reaction (PCR) amplification step [16, 17]. In the assays where restriction enzymes are used to fragment the target DNA, the locus-specific copy number estimates may be correlated with the fragment length. Since the fragments are known from the genome annotation it is straightforward to estimate and correct for such effects [21, 22, 23, 24, 14]. Moreover, it has been reported that observed intensities are also correlated to the GC content [21, 24], but also more complex relationships such as the nucleotide sequences of the probes [23, 14].

As these parameters may vary across assays and between hybridizations, they need to be corrected for in order to make comparisons across samples meaningful and more precise. Existing approaches typically involve non-linear regression of signal intensities on PCR fragment length, GC content and nucleotide position [21, 22, 23, 24, 14].

Generic probe-level normalization is a crucial step of microarray preprocessing which aims at making probe signals comparable between samples. For Affymetrix data, methods originally developed for the preprocessing of expression microarray data — lowess normalization [25], invariant-set normalization [26] or quantile normalization [27], have been successfully applied to SNP array data. These approaches explicitly constrain probe-levels signals to be comparable across arrays.

Correction for allelic crosstalk. However, It has been recently shown that most of the non-biological differences between the distribution of probe-level signals across samples could be attributed to *allelic crosstalk* (including an offset correction), that is, cross-hybridization between probes targeting the two alleles of a SNP [24, 28]. One advantage of allelic crosstalk calibration is that it effectively makes probe-level signals comparable across samples without imposing constraints on intensity distributions [14, 24]. It can also be applied to each array separately.

Summarization of probe-level signals. Summarization combines normalized probe-level signals into locus-level estimates by fitting a log-additive or multiplicative model of the intensities. These models were first developed for the analysis of oligonucleotide expression microarrays [26, 27] and later adapted to SNP arrays [29, 30, 23]. Related multi-array models that explicitly models the allelic crosstalk at the summarization step have also been suggested [31, 28].

A common feature of Affymetrix SNP arrays is that each SNP is associated with a set of 25 nucleotide-long probe sequences. Half of these *probe sets* target allele A and the other half targeting allele B. However, the technology has evolved substantially across generations of SNP arrays, as a result of an effort from both the manufacturer and the scientific community [9]. With the latest generation of SNP arrays (GenomeWideSNP_5 and 6), all probes targeting a given allele-specific or total copy number locus are technical replicates. With this simplified probe set design, using the median of replicated probes within an array as a summary has been shown [14] to perform as good as or better than previously proposed summarization models that required several arrays to be used.

3.1.2 Illumina

Almost all studies performed using Illumina data use the preprocessing method provided by Illumina's BeadStudio software [32, 10], which is an affine transformation of the original data that corrects for offset and signal

compression (or allelic crosstalk), and scales the data based on control points. The parameters for this affine transformation are estimated independently for each sample, for each *sub-bead pool*. As the Infinium assay does not involve PCR amplification, correction for sequence effects is not needed for Illumina SNP arrays.

Recent works demonstrated that the signals after BeadStudio normalization suffer from a dye bias [33]: the distribution of normalized signals differ substantially between the two types of fluorescent dyes (Cy3 and Cy5) that are used in the Infinium II assay [34]. The correction method proposed by [33] consists in applying Quantile Normalization [27] to the normalize the two dyes. Importantly, this is still done independently for each array.

3.2 Making signals comparable across probes

Signal intensities at a given locus j can be assumed to be proportional to the corresponding true copy numbers, but the proportionality coefficient is unfortunately locus specific and unknown [26, 27, 35]. These coefficients are known as *locus affinities*. In copy number studies, true copy numbers are expected to be locally constant along the genome. This property is exploited by downstream segmentation methods to detect copy number changes along the genome, as explained in Sections 4 and 6. It is therefore fundamental for these downstream analyses that these locus affinities be canceled beforehand, in order to make copy number signals comparable across neighboring loci. This section describes how existing methods address this question for total copy number and allelic signals.

3.2.1 Total copy numbers

As locus affinities are not sample-specific, they can be effectively canceled from total signals by dividing the observed(summarized) signal intensity y_j at locus j by an observed *reference* signal intensity, $y_j^{(R)}$, at the same locus, which is obtained from a sample or a pool of samples for which the true copy number at locus j, $\gamma_j^{(R)}$, is known:

$$c_j = \gamma_j^{(R)} \frac{y_j}{y_j^{(R)}} . \tag{5}$$

In general the reference is chosen to be copy-number neutral ("copy neutral"), that is, so that $\gamma_j^{(R)} = 2$ for $j \in \mathcal{J}$. There are several choices of total reference signal $y_j^{(R)}$, depending on the study design [36, 37]. For instance, in a paired tumor-normal study, the reference signal at a given locus may be the

corresponding total signal from a matched normal tissue sample or normal blood sample, whereas in a tumor study without matched normals, it may be the corresponding robust average (e.g. a median) of all samples in the study. If some of the samples in the study are normal samples, their robust average may be used as a reference instead.

It is in general better to use a reference from the same lab as the test sample, and possibly from the same batch of arrays. This is illustrated by Figure 2, where three different sets of cytogenetically normal samples were used as references for the same tumor SNP array. The tumor SNP array is from a breast cancer cell line hybridized at the Lawrence Berkeley National Laboratory (LBNL). All samples were hybridized on the Affymetrix GenomeWideSNP_6 platform, normalized using CRMAv2 [14]. Copy number profiles were segmented using the Circular Binary Segmentation (CBS) method [38]. The figures were generated using ChromosomeExplorer within the aroma.affymetrix framework [39].

The signals in the three panels of Figure 2 are of similar amplitude: the difference between copy number estimates (black segments) between two successive copy number regions is comparable across panels. Therefore, signal to noise ratios can be compared on the basis of the corresponding noise levels. We quantified the noise level (along the whole genome) for each choice of a reference using a robust first-order standard deviation estimator [40, 41]:

$$\widehat{\sigma}_{\Delta} = \frac{1}{\sqrt{2}} \cdot \Phi^{-1}(3/4) \cdot \underset{j}{\text{median}} \left(\left| z_j - \underset{j'}{\text{median}}(z_{j'}) \right| \right) , \qquad (6)$$

where $z_j = c_{j+1} - c_j$ for $j = 1, \ldots, J-1$. The scaling factors $1/\sqrt{2} \approx 0.7071$ and $\Phi^{-1}(3/4) \approx 1.4826$ make $\widehat{\sigma}_{\Delta}$ a consistent estimator of the c_j under the assumption that c_j, and hence z_j, is Gaussian and i.i.d. Because this estimator relies on the first order differences z_j, it is robust against change points and can therefore be used without knowing where the true change points are.

The noise level is high when samples from a different lab are used as references (top panel: $\widehat{\sigma}_{\Delta} = 0.60$), even when the number of samples is large (197). It is substantially smaller when references from the same lab (LBNL in this particular example) are used (middle panel: $\widehat{\sigma}_{\Delta} = 0.44$), even in a much smaller number (36). It is even lower when references from the same *batch* of arrays (bottom panel: $\widehat{\sigma}_{\Delta} = 0.37$): in this example, the reference set consisted of only 22 arrays hybridized on the same day as the tumor sample.

3.2.2 Allelic ratios

Allelic ratios for a given SNP j are usually estimated as the ratio of the signal intensity of one allele relative to the total signal intensity. For B allele fractions, this yields

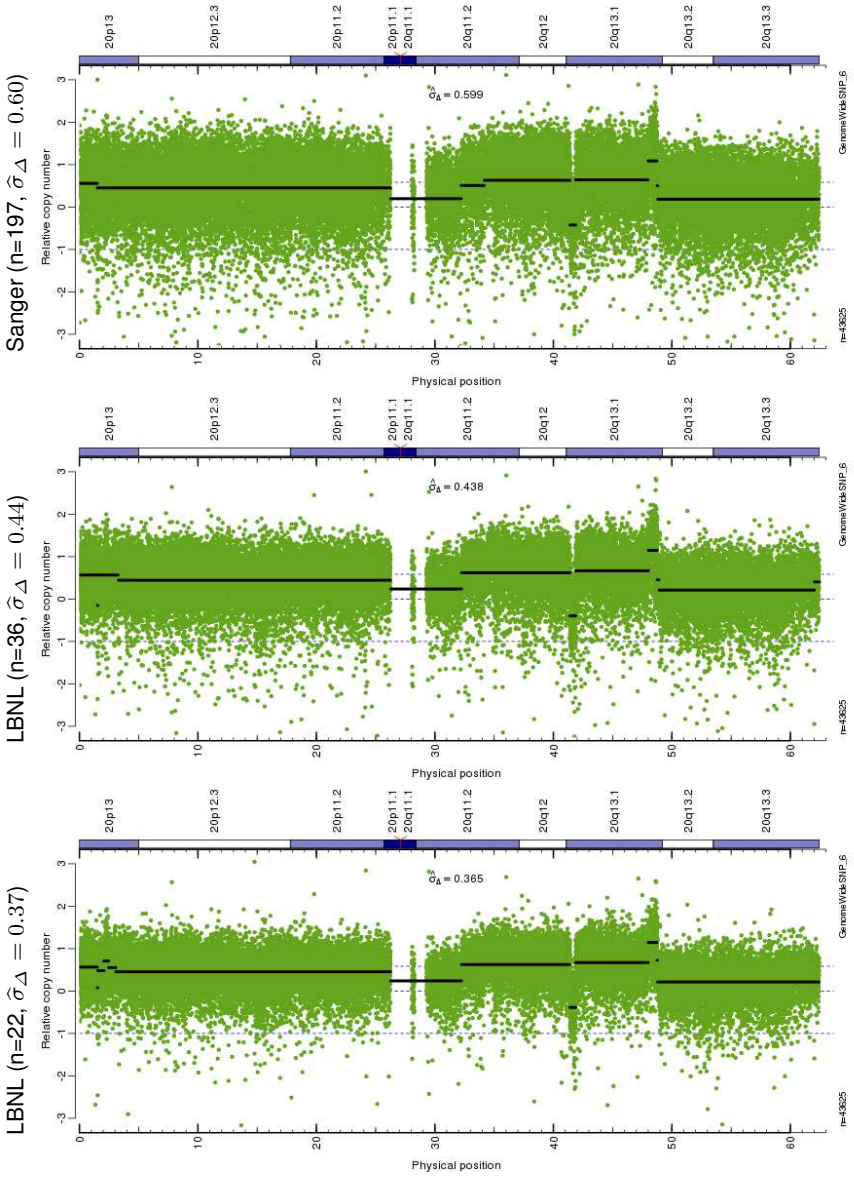

Fig. 2 *Influence of the choice of a reference sample on the signal to noise ratio in total copy number signals. Three different sets of normal references are used to estimate total copy numbers for the same tumor SNP array hybridized at the Lawrence Berkeley National Laboratory (LBNL): 197 samples from another lab (top panel), 36 arrays from LBNL (middle panel), and 22 arrays from LBNL, and the same batch as the tumor sample (bottom panel). Green dots: locus-level estimates; black segments: region-level estimates after segmentation by CBS [42].*

$$b_j = \frac{y_{jB}}{y_j}\,, \tag{7}$$

where $y_j = y_{jA} + y_{jB}$, and y_{jA} and y_{jB} are the observed signal intensities for allele A and B, respectively. Note that contrary to total signals, no external reference is needed at this stage: allelic ratios can be estimated from a single hybridization. However, these estimates have been reported to suffer from systematic deviations from their corresponding true values [20, 10, 43, 15]. One possible explanation for this effect is that locus affinities are not only locus-specific but *allele-specific*, so that they may not be adequately canceled by the ratio in Equation (7). Several approaches have been developed to normalize raw allelic ratios based on paired or unpaired normal reference hybridizations, greatly improving the signal to noise ratio for downstream analyses for Illumina data [20, 10], or both Affymetrix and Illumina data [15]. This is illustrated by Figure 3 for the TumorBoost method [15].

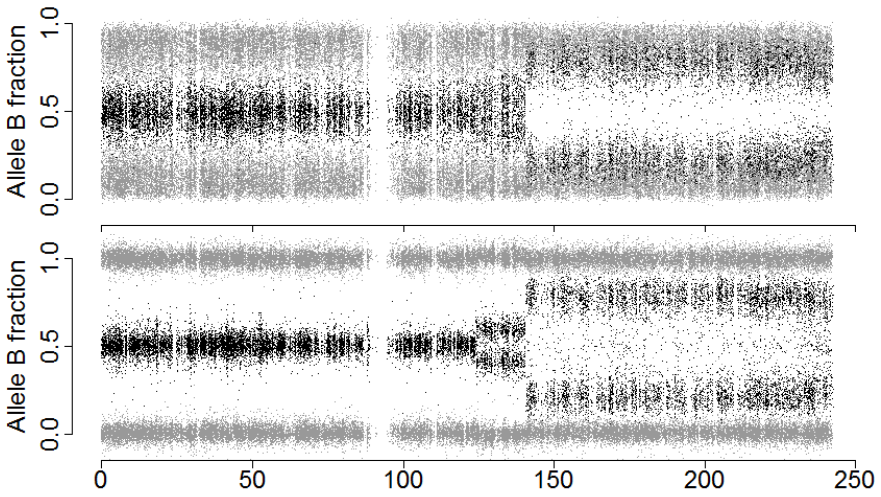

Fig. 3 *Improved signal to noise ratio after normalization by the TumorBoost method [15]. Top: raw allelic ratios as in Equation* (7)*. Bottom: TumorBoost-normalized allelic ratios, using allelic ratios from of a paired normal hybridization. Data is taken from the same tumor sample and chromosome as in Figure 1.*

4 Copy number change detection: from locus-level to region-level estimates

Copy number profiles in tumors are consequences of genomic events at the regional scale, such as small or large deletions or gains. Therefore, true copy

number signals can safely be modeled as locally constant in tumor samples. This assumption is one of the bases of all the algorithms that have been proposed for detecting copy number changes from microarray data. The goal of this section is to explain how locus-level copy number estimates (obtained after preprocessing as described in Section 3) can be combined to *detect copy number changes* along the genome.

The methods described in this section can be used to segment total, minor and major copy numbers, or allelic imbalances: these applications are discussed in Section 6. For simplicity of notation and vocabulary, we will loosely refer to *copy numbers* and use the notation c for locus-level estimates, and γ for the corresponding true values.

Two main types of methods have been developed and are used in practice: change-point models and Hidden Markov Models (HMM). In the context of copy number analyses, they were initially applied to microarray technologies that only assess *total* signals, in particular array Comparative Genomic Hybridization (array-CGH) [3]. The practical performance of these methods has been reviewed in [44, 45]. The present section provides an up to date statistical review of currently available methods for copy number segmentation using change point or Hidden Markov Models.

For simplicity, we will only use genomic positions $j = 1, 2, \ldots, J$ corresponding to the ordering of loci, rather than the physical location (in base-pairs) of the loci, in the following discussion and equations. This is also the most commonly used approach in existing methods. Incorporating physical locations as well introduces another level of complexity to the notation and the models that is unnecessary for the overview presented here.

4.1 Change-point models

We assume that there exists a partition of the genome into K segments, $k = 1, 2, \ldots, K$, such that true copy numbers are constant in each segment. Specifically, there exists an index vector of $K + 1$ loci $\mathbf{t}(K) = (t_k)_{0 \leq k \leq K}$ called *change points*, such that $1 = t_0 < t_1 < \cdots < t_{K-1} < t_K = J$, and an associated vector of K *region-level true copy numbers* $\mathbf{\Gamma} = (\Gamma_k)_{1 \leq k \leq K}$ such that true copy numbers $\boldsymbol{\gamma} = (\gamma_j)_{j \in \mathcal{J}}$ are constant equal to Γ_k in the interval $[t_{k-1}, t_k)$. That is,

$$\gamma_j = \Gamma_k \,;\ \forall j \in [t_{k-1}, t_k), \forall k \in \{1, \ldots, K\}\,. \tag{8}$$

Letting $k(j)$ be the largest index k such that $t_k \leq j$, the observation $\mathbf{c} = (c_j)_{j \in \mathcal{J}}$ may then be modeled as

$$c_j = \Gamma_{k(j)} + \varepsilon_j\,, \tag{9}$$

where the errors $(\varepsilon_j)_{j\in\mathcal{J}}$ are independent and identically distributed (i.i.d.), and generally assumed to be Gaussian ($\mathcal{N}(0,\sigma^2)$). When the number K of segments and the vector $\mathbf{t}(K) = (t_0,\ldots,t_K)$ of change point locations are known, the log-likelihood $\ell(K,\mathbf{t}(K),\mathbf{\Gamma};\mathbf{c})$ of the model described by Equation (9) is additive in each segment:

$$\ell(K,\mathbf{t}(K),\mathbf{\Gamma};\mathbf{c}) = J\log(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{k=1}^{K}\sum_{j\in[t_{k-1},t_k)}\left(c_j - \Gamma_{k(j)}\right)^2 . \quad (10)$$

In this idealized situation, the maximum likelihood estimator of each Γ_k is the empirical mean of the observed signals within the k^{th} segment. In practice though, both K and $\mathbf{t}(K)$ are unknown, which gives rise to a model selection problem (choosing K), and a combinatorial problem: choosing $\mathbf{t}(K)$ for a given K. Indeed, the number of possible configurations for $\mathbf{t}(K)$ is $\binom{K-1}{J-1}$, that is, $O(J^{K-1})$, which is prohibitively large in realistic situations where K is in the dozens and J is currently of the order of 10^5 to 10^6.

Heuristics. The first approach taken to address these issues has been to combine a Bayesian Information Criterion (BIC) penalization for the choice of K with a genetic programming algorithm for the choice of $\mathbf{t}(K)$ [46]. Three main directions have been explored to improve on this early attempt. The most widely used method in practice, known as Circular Binary Segmentation, implements a greedy approach which recursively looks for the best partition of the data into two (or three) segments [38]. The depth of the recursion is determined by the significance of the change points, which implicitly determines K. This step has been made faster using permutation techniques in a second version of the method [42], which has been used to produce the segmentation obtained in Figure 4. A modified BIC criterion has also been proposed to estimate K directly [47].

Exact solutions. Second, several methods have been proposed that solve the original problem exactly. First, one can take advantage of the additivity of the log-likelihood in the segments and use dynamic programming to reduce the complexity of the exhaustive search for the best $\mathbf{t}(K)$ for a given K from $O(J^{K-1})$ to $O(K\cdot J^2)$. This idea has been combined with an adaptive penalization method [48] to build a quadratic ($O(K\cdot J^2)$) change point detection algorithm [49]. Such a method cannot be used to segment DNA copy number profiles from the latest generations of microarrays, for which more than 10^6 loci can be interrogated. A pruned dynamic programming algorithm has been proposed recently that recovers the optimal solution much faster [50]. Although its worst case complexity is still $O(K\cdot J^2)$, in practical situations it is almost linear in J, which makes it quite appealing for current copy number segmentation problems.

Convex relaxations. A third direction uses *convex relaxation*, which is a classical approach in statistical machine learning. It consists in replacing a non-convex optimization problem by a slightly different, but convex, version

of the problem, which can be solved efficiently. Two regression methods based on Lasso-type penalties have been applied to the problem of detecting changes in DNA copy number signals [51, 52].

The first method [51] is an adaptation of the Fused Lasso [53], which solves the constrained optimization problem

$$\min_{(\gamma_j)_{1\le j\le J}} \sum_{j=1}^{J} (c_j - \gamma_j)^2 \quad \text{s.t.} \quad \sum_{j=1}^{J-1} |\gamma_{j+1} - \gamma_j| \le v \text{ and } \sum_{j=1}^{J} |\gamma_j - 2| \le u. \tag{11}$$

Formally, this method constrains the ℓ_1 norm of the jumps in $\boldsymbol{\gamma}$, which can be seen as a convex relaxation of constraining the number of jumps (that is, the ℓ_0 norm of the jumps). In words, the mean amplitude of the changes in estimated copy-number levels ($|\gamma_{j+1} - \gamma_j|$) is not allowed to be too large. Moreover, this model incorporates a sparsity constraint on $|\gamma_j - 2|$ enforcing that most loci correspond to the copy-neutral state, where 2 represents the copy number of the copy-neutral state. For non-diploid copy-neutral state, this copy-number level should be adjusted accordingly. The complexity of the algorithm proposed in [51] is (at best) quadratic in the number of data points, that is $O(J^2)$, which is too expensive for recent data sets.

The second method [52] is a relaxed version of Equation (11) where only the amplitudes of the changes are constrained, resulting in the constrained optimization problem

$$\min_{(\gamma_j)_{1\le j\le J}} \sum_{j=1}^{J} (c_j - \gamma_j)^2 \quad \text{s.t.} \quad \sum_{j=1}^{J-1} |\gamma_{j+1} - \gamma_j| \le v\,. \tag{12}$$

This optimization problem can be written as a Lasso-type regression problem and can therefore be solved in $O(K^3 + J{\cdot}K^2)$ using a Least Angle Regression (LARS) algorithm [54] to select the first K change points. The authors of [52] suggest to prune the obtained set $\mathbf{t}(K)$ of candidate change points by running the aforementioned dynamic programming algorithm on the set of partitions consisting of subsets of $K' < K$ points in $\mathbf{t}(K)$. Because this set is much smaller than the original searching space, this pruning step has a low complexity of $O(K^3)$. Finally, they define a heuristic for choosing K based on the magnitude of the increments of the empirical risk when a change point is added.

4.2 Hidden Markov Models

Hidden Markov Models (HMMs) assume that the observed copy numbers $\mathbf{c} = (c_j)_{j\in\mathcal{J}}$ are emitted by an underlying Markov chain according to H hidden region-level true copy number states $\boldsymbol{\Gamma} = \{\Gamma_1, \ldots \Gamma_H\}$. A HMM of order 1

is defined by a specific set $\mathbf{\Gamma}$ of hidden states, and transition probabilities $(p(u,v))$ for $(u,v) \in \{1,\ldots,H\}^2$, such that

$$\mathbb{P}(\gamma_{j+1} = \Gamma_v | \gamma_j = \Gamma_u) = p(u,v)\,;\ \forall j \in \{1,\ldots,J\}\,. \tag{13}$$

HMM naturally incorporate and take advantage of the fact that different regions can have the same true copy number, which is not the case of change-point models as the one described by Equation (9). Several HMM-based methods have been proposed for estimating *total* copy numbers. These methods mainly differ in the assumptions that are made for the dynamics of the underlying Markov chain, and the approaches used for the estimation of the hidden states.

The earliest approach assumes that the state sequence is a discrete Markov chain [55]. The number of hidden states is estimated using model selection. More recently, a Bayesian HMM approach with four ($H = 4$) hidden states has been proposed [56]. Because it relies on Bayesian estimation procedures, it provides not only a segmentation of the original observations but also confidence intervals for (the index location of) each copy number change point. However, because the posterior distribution is analytically intractable, posterior inference in this model is performed using simulation-based methods.

The underlying copy number process can also be modeled as a continuous-valued Markov jump process [57]. This type of model is appealing for applications to tumor samples as it does not require the number of of hidden states (H) to be specified in advance. Moreover, contrary to [56], the posterior distribution of the hidden variables in [57] can be computed explicitly, which implies that posterior estimates, including confidence assessment of a given segmentation, are available without simulations.

In contrast to change-point methods, HMM-based approaches rely on assumptions on the distribution of the underlying copy number state sequence, and the distribution of the size of copy number regions. Although such assumptions may be unrealistic in the context of cancer studies, a number of state of the art methods for estimating copy numbers from SNP array data use HMMs, as will be explained in Section 6.

5 Purity and ploidy

Figure 4 displays the same data as in Figure 1 after segmentation of total copy numbers by the Circular Binary Segmentation algorithm [38, 42], and estimation of total, minor and major copy numbers as well as allelic ratios in regions of constant total copy numbers.

As explained in Section 2, TCGA has shown that the copy number states observed in the genomic region displayed in Figure 1 are a normal diploid region $(1,1)$, a single gain $(1,2)$ and a copy-neutral LOH $(0,2)$.

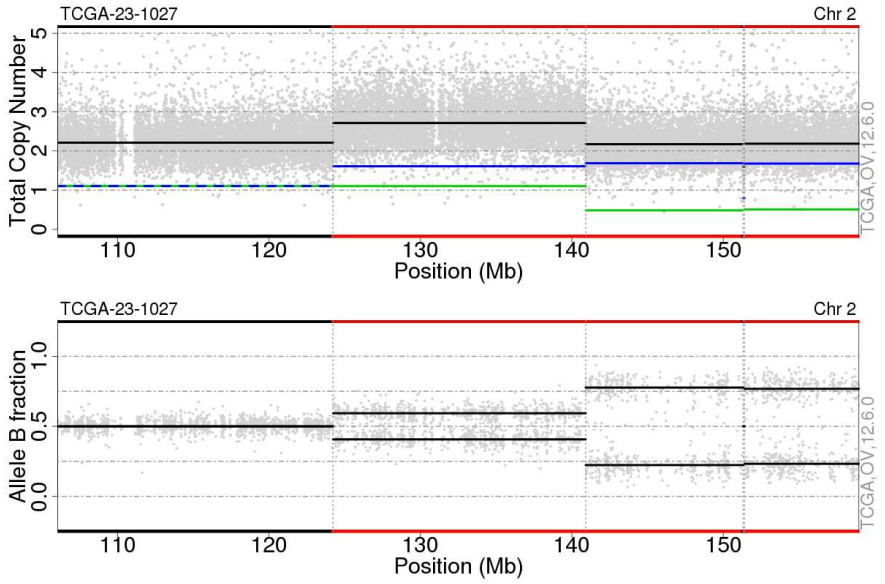

Fig. 4 Locus and region-level estimates. Input data is the same as in the top panels of Figure 1. Two main change points in total copy numbers (top panel) have been detected by the CBS algorithm [38, 42], and are reported in both panels as dashed gray vertical lines. Top panel: locus-level total copy number estimates (gray dots), and total (black), major (blue) and minor (green) region-level copy number estimates after change point detection. Bottom panel: locus-level (gray dots) and region-level allele B fractions estimates after change point detection (black lines) for heterozygous SNPs. Regions of allelic imbalance (unequal parental copy numbers) are highlighted in red.

However, it is not straightforward to infer these copy number states only by looking at Figure 4: the observed region-level copy number estimates do not reflect the true copy numbers in the tumor cells of the sample. First, the total copy number is slightly greater than 2 in the normal diploid region. Then, the difference between successive region-level total copy numbers is substantially smaller than the true difference (one copy number unit). Even more strikingly, allele B fractions in the region of copy-neutral LOH (rightmost region) are far from the expected values of 0 or 1.

In this section we explain that these observations are not due to imperfections of the preprocessing method or the microarray assay itself, as they reflect two *biological features* of the data: the ploidy of the tumor, and the presence of normal cells (and possibility of several cytogenetically distinct kinds of tumor cells) in what is usually called a *tumor sample*. For simplicity we will assume that the reference used in the estimation of locus-specific copy numbers (as explained in Section 3.2) is a cytogenetically normal sample (either normal tissue, or normal blood extract) from the same individual

as the tumor. Methods that take these biological parameters into account in the estimation of copy number states are discussed in Section 6.

5.1 Pure tumor samples

In Figure 5 we have represented the true copy numbers in a sample assumed to contain only one kind of tumor cells and having the same copy number states as those observed in Figure 4: a normal $(1,1)$ region, followed by a region of gain of a single copy $(1,2)$, and by a region of copy-neutral LOH $(0,2)$.

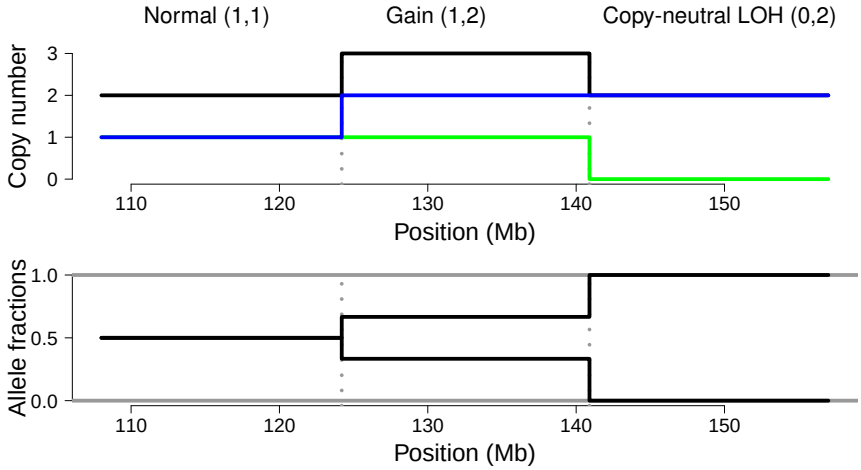

Fig. 5 Assumed true total copy numbers and allelic rations in the tumor cells depicted in Figure 4. Top panel: total (black), major (blue) and minor (green) copy numbers. Bottom panel: allele B fractions: homozygous SNPs (gray) and heterozygous SNPs (black).

By Equation (2), true allele B fractions satisfy $\beta_j \in \{0, \underline{\gamma}_j/\gamma_j, \overline{\gamma}_j/\gamma_j, 1\}$, and the pattern of allelic ratios observed in Figure 5 can be interpreted as follows. In a region of allelic balance (left region), where the two parental copy numbers are identical (and not zero), the two heterozygous states merge into $\beta_j = 1/2$ and there are three distinct states: $\beta_j \in \{0, 1/2, 1\}$. In a region of allelic imbalance with retention of heterozygosity (middle region) where the two parental copy numbers are different and neither are zero, β_j can take four distinct values. In a region of LOH (right region), where the minor copy number is 0, heterozygous states disappear and we observe two distinct states: $\beta_j \in \{0, 1\}$. The only type of scenario not represented in Figure 5 is

the case of homozygous deletions, where both parental copy numbers are null and true allele B fractions are not defined.

5.2 Contamination by normal cells

In practice however, "tumor samples" are generally a mixture of a tumor cells and a normal cells. In this situation, Equation (2) still holds, but the observed parental copy numbers need not be whole numbers anymore. They are a mixture of the unknown parental copy numbers in the tumor, and the parental copy numbers in normal cells, which are typically but not always $(1, 1)$. The exceptions are so-called copy number polymorphisms (CNPs) [58, 59, 60]. For simplicity, we will in what follows only consider SNPs that are diploid in the normal cells.

Assuming that normal cells are diploid, and denoting by $\kappa \in [0, 1]$ the proportion of normal cells in the sample, then the true minor and major copy numbers in the sample are given by

$$\begin{cases} \underline{\gamma}_j &= (1-\kappa)\underline{\gamma}^\star_j + \kappa \\ \overline{\gamma}_j &= (1-\kappa)\overline{\gamma}^\star_j + \kappa \end{cases} \tag{14}$$

where $\underline{\gamma}^\star_j$ and $\overline{\gamma}^\star_j$ are the true minor and major copy numbers *of the tumor cells from the sample* at locus j, as if there were no normal cells. Note that these true copy numbers neither need not be whole numbers as the tumor cells of a DNA sample may themselves be a mixture of several tumoral populations (or clones), each with distinct whole-number copy number profiles. The corresponding total copy numbers and allelic imbalances (when j is a heterozygous SNP) are given by

$$\begin{cases} \gamma_j &= (1-\kappa)\gamma^\star_j + 2\kappa \\ \delta_j &= \frac{\overline{\gamma}^\star_j - \underline{\gamma}^\star_j}{\gamma^\star_j + 2\kappa/(1-\kappa)} \end{cases} \tag{15}$$

True allele B fractions satisfy $\beta_j \in \{0, 1/2 - \delta_j/2, 1/2 + \delta_j/2, 1\}$. The influence of normal contamination on true total copy numbers and allelic ratios is shown in Figure 6. Normal contamination moves the observed allelic ratios towards those of the corresponding normal genotypes, and the observed total copy numbers towards the copy number of normal cells. A major difference with the case of no normal contamination is that one still observes heterozygous states in regions of LOH in the tumor: indeed, in regions of LOH, the minor copy number is 0 in tumor cells ($\underline{\gamma}^\star = 0$), and we have

$$\beta \in \left\{ 0; \frac{\kappa}{(1-\kappa)\gamma^\star + 2\kappa}; \frac{(1-\kappa)\gamma^\star + \kappa}{(1-\kappa)\gamma^\star + 2\kappa}; 1 \right\}, \tag{16}$$

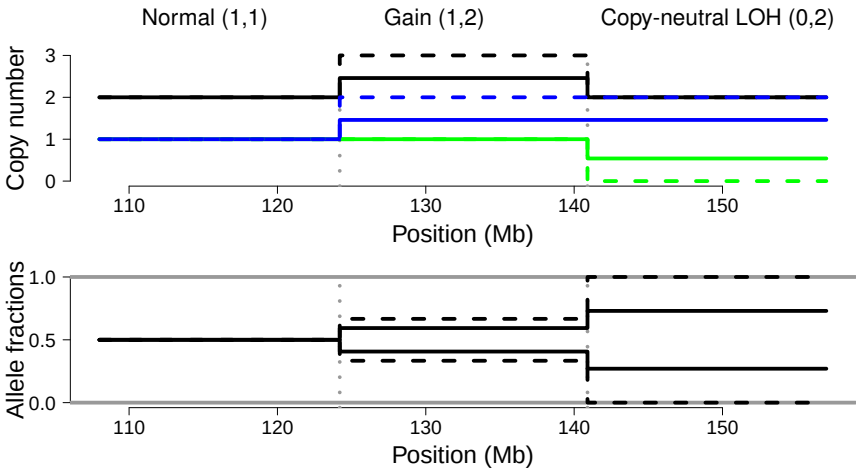

Fig. 6 Influence of contamination by normal cells on true copy numbers: comparing 0% contamination (pure tumor as in Figure 5, dashed lines) with 54% contamination (solid lines). Top panel: true total (black), major (blue) and minor (green) copy numbers. Bottom panel: true allele B fractions: homozygous SNPs (gray) and heterozygous SNPs (black).

which corresponds to four distinct modes for allelic ratios. This is illustrated by Figure 6 (right) in the particular situation of copy-neutral LOH, where $\gamma^\star = 2$, leading to $\beta \in \{0; \kappa/2; 1-\kappa/2; 1\}$.

From a modeling point of view, it is worth noting that normal cell contamination is a particular case of contamination, because it may be estimated and corrected for based on either diploid assumptions or explicit measurements of a matched normal (germline) sample. Simply speaking, it is in many cases possible to remove the normal component in the tumor-normal mixture. This is rarely possible for other type of cell contaminations, as they are generally not directly measured. In particular, the problem of identifying different tumor clones from one heterogeneous tumor sample is a harder one.

5.3 Tumor ploidy

As explained in Section 3.2, the total copy number at locus j is generally estimated relative to a reference as in Equation (5), in order to cancel locus-specific affinities. We can actually interpret c as an estimator of the true copy number in the tumor sample if *the same number of cells were hybridized to the microarray in the tumor and in the normal assay.*

This assumption does not necessarily hold, because of copy number alterations in the tumor. Indeed, the experimental protocol constrains the amount of DNA, not the number of cells, to be the same for each sample

assayed [36, 11]. For example, a purely tetraploid tumor with two copies of the genome and no other chromosomal alteration could not be distinguished from a cytogenetically normal (*diploid*) sample, as the genomic material hybridized on the SNP array is the same in both situations. We refer to [35, Section 4.4] for further discussion on this issue.

In this chapter we define the *ploidy* λ of a biological sample as the total amount of genomic DNA in this sample relative to that of a normal sample. Therefore, ploidy as defined here needs not be a whole number, because of chromosomal gains and losses and as the tumor sample may be a mixture of normal cells and tumor cells or one or more types of tumor cells with different patterns of genomic alteration. Figure 7 illustrates the influence of tumor ploidy on SNP array signals when using Equation (5) to estimate total copy numbers, that is, when assuming that the average true copy number in the normal is 2. Ploidy acts as a scaling factor for total, minor and major

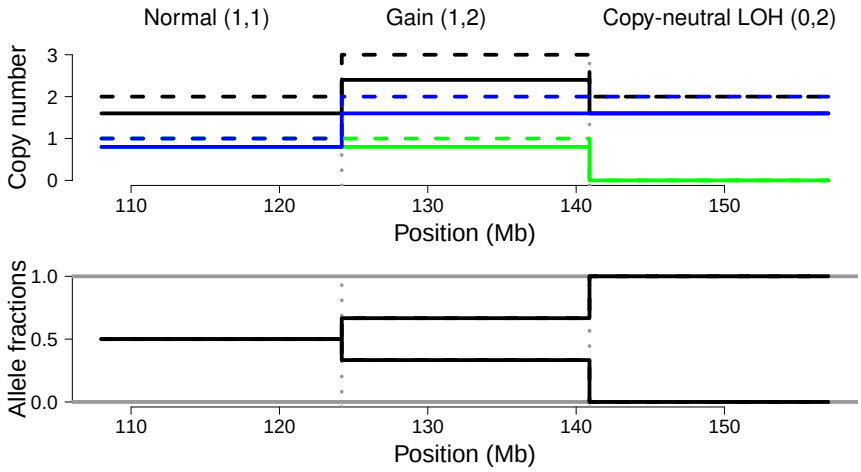

Fig. 7 Influence of tumor ploidy on true copy numbers in absence of normal contamination: comparing ploidy 2 (as in Figure 5, dashed lines) to ploidy 2.5 (solid lines). Top panel: true total (black), major (blue) and minor (green). Bottom panel: true allele B fractions: homozygous SNPs (gray) and heterozygous SNPs (black).

copy numbers. Allelic signals as defined in Equation (7) are not affected.

5.4 Combined influence of purity and ploidy

As a result of the combined influence of purity and ploidy on the actual composition of a biological sample, the true minor and major copy numbers at a SNP j may be written as

$$\underline{\gamma}_j = \frac{1}{\lambda}\left[(1-\kappa)\underline{\gamma}_j^\star + \kappa\right] \tag{17}$$

$$\overline{\gamma}_j = \frac{1}{\lambda}\left[(1-\kappa)\overline{\gamma}_j^\star + \kappa\right] \tag{18}$$

The corresponding true total copy numbers and allelic ratios are given by:

$$\gamma_j = \frac{1}{\lambda}\left[(1-\kappa)\gamma_j^\star + 2\kappa\right] \tag{19}$$

$$\delta_j = \frac{\overline{\gamma}_j^\star - \underline{\gamma}_j^\star}{\gamma_j^\star + 2\kappa/(1-\kappa)} \tag{20}$$

As explained above, we note that allelic imbalances (δ_j) are only affected by normal contamination, not by ploidy. Figure 8 illustrates the combined influence of purity and ploidy by comparing the true total copy numbers and allelic ratios for a pure tumor without normal contamination (as in Figure 5) with a non-diploid tumor with normal contamination according to Equations (19) and (20).

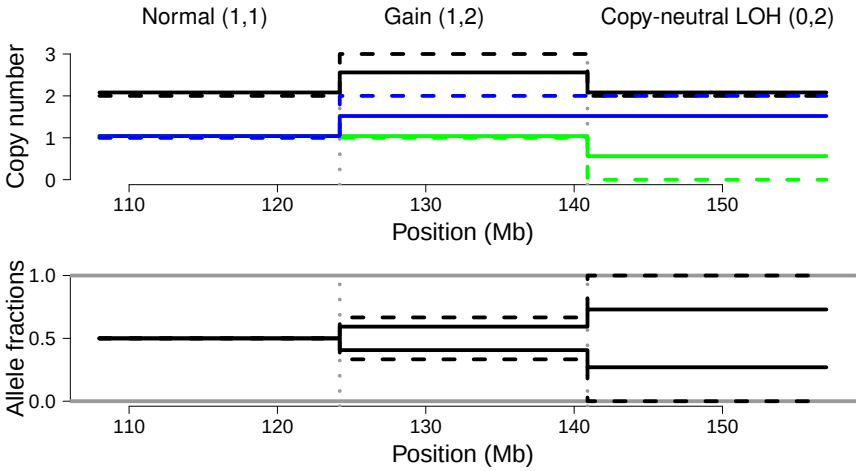

Fig. 8 Combined influence of tumor ploidy and normal cell contamination on true copy numbers: comparing ploidy 2 and no normal contamination (as in Figure 5, dashed lines) with ploidy 1.8 and 54% normal contamination (solid lines). Top panel: true total (black), major (blue) and minor (green). Bottom panel: true allele B fractions: homozygous SNPs (gray) and heterozygous SNPs (black).

When accounting for both purity and ploidy, the copy number patterns become quite similar to those observed with real data. This is illustrated by the comparison between the true copy numbers in Figure 8 and locus- and region-level copy number estimates in Figure 4.

6 Estimation of copy number states in cancer studies

Copy number studies in cancer research aim at identifying the unknown copy number state in a tumor sample, as defined in Section 2. As explained above, the word *identification* actually covers two different statistical questions: *detecting* changes in copy number signals, and *calling* regions, that is, assigning a copy number state to each region detected. Because SNP arrays interrogate allele-specific signals, they can be used for both detection and calling.

Segmentation can be performed regardless of purity and ploidy, although these two biological parameters do influence the detection power of any given segmentation method, through the distance between true region-level copy number states. However, both purity and ploidy have to be acknowledged in order to call copy number states in the tumor cells of a given sample.

6.1 Existing methods

A number of methods for analyzing SNP array data were developed in the context of Copy Number Variation (CNV) studies in normal samples: VanillaICE [61], PennCNV [62], QuantiSNP [63], and BirdSuite [64]. Most of them are based on HMMs. Because these methods are dedicated to, and well-designed for CNV studies, their model states do not adequately describe the copy number states in Table 1. More specifically, either they do not consider allele-specific amplifications [62, 63], or the distinction between normal and copy-neutral LOH [61], or they are only designed to detect rare CN aberrations [64]. Moreover, their states generally do not account for possible tumor heterogeneity or contamination by normal cells.

Table 3 lists methods that actually combine total and allele-specific signals in order to call copy number states (as defined in Table 1) in cancer studies. They are described in terms of the type of information they take into account and the type of method they use for *detecting* copy number changes, whether their application requires the availability of a paired normal reference, and whether they explicitly account for tumor purity and ploidy as discussed in Section 5.

We have shown in Section 2 that SNP array signals were two-dimensional by nature, and that both dimensions were needed to *call* copy number states as defined in Table 1. All methods cited in Table 3 indeed make use of both dimensions at the calling step, but not necessarily at the detection step. These methods can be classified in terms of the type of input data they are using at the detection step, as indicated by the horizontal lines in Table 3. We note here that although raw allelic signals typically have several modes in a region of constant copy number (as explained in Section 2), direct segmentation methods can be used to detect changes in allelic signals from SNPs that are heterozygous in the germline [6, 7, 67, 66].

Name	Settings	Detection method	Ploidy	Purity
MCP [8]	unpaired	HMM on δ	no	no
Gardina [11]	unpaired	HMM on "genotypes"	yes	no
BAFsegmentation [6]	paired or unpaired	segmentation of δ	no	yes
SOMATICS [7]	unpaired	segmentation of δ	no	yes
AsCNAR/CNAG [36]	unpaired	HMM on δ	no	yes
OverUnder [65]	unpaired	$2\times$ 1d smoothing	yes	no
psCBS [66]	paired	two-way segmentation	no	no
GAP [67]	unpaired	$2\times$ 1d segmentation	yes	yes
Lamy [68]	paired	HMM on (γ, δ)	no	yes
PSCN [69]	unpaired	HMM on (γ, δ)	no	no
PICNIC [43]	unpaired	HMM on (γ, δ)	yes	no
genoCNA [70]	unpaired	HMM on (γ, δ)	no	yes

Table 3 *Existing methods for copy number studies in cancers using SNP arrays. Settings: has the method been developed for studies with available paired normal samples, or not ? Last two columns: does the method explicitly account for ploidy, purity ?.*

Methods from the first group use only one piece of information for detection [8, 71, 7, 36, 6]. As these methods are mostly interested in loss of heterozygosity, they all allelic imbalances (or genotypes) and not total copy numbers, as an input for the detection step. Methods from the second group combine both pieces of information, either by independent smoothing [65] or segmentation [67] of each piece of information, or by segmentation of total signals followed by segmentation of allelic signals [66]. In particular, GAP [67] is to our knowledge the only method that explicitly accounts for both purity and ploidy. Finally, methods from the third group perform truly joint detection of copy number changes [68, 69, 43, 70]. In the next section, we show that such joint approaches can be more powerful to detect copy number changes.

6.2 Joint detection provides more power to detect copy number changes

In this section, we demonstrate that there is substantial statistical power to gain by considering both pieces of information for the detection step. Figure 9 shows that total and allelic signals have comparable power to detect the two change points studied throughout this chapter: a transition between a diploid normal state $(1, 1)$ and a gain $(1, 2)$ (left panel), and a transition between a gain $(1, 2)$ and a region of copy-neutral LOH $(0, 2)$ (right panel). For each type of signal studied in Figure 9 is a ROC curve used to measure the separation between two copy number states at a change point of known location based on this signal. We refer to [15, 41, 14] for a comprehensive description of this evaluation.

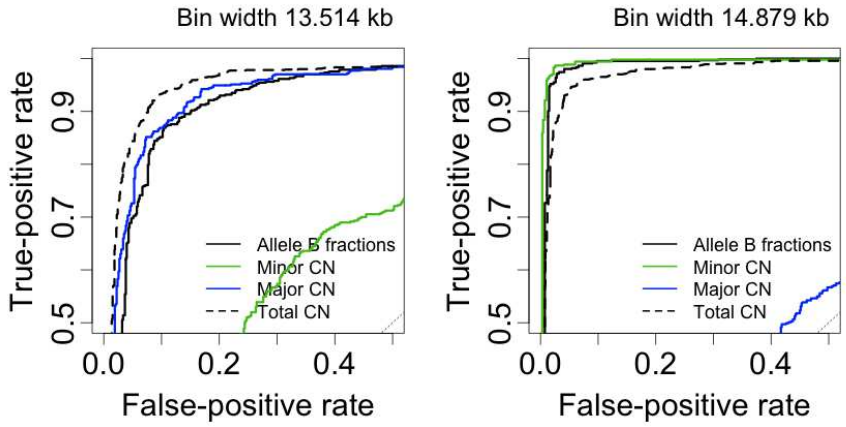

Fig. 9 *At a fixed resolution, total copy numbers and allelic signals have comparable power to detect copy number changes. ROC curves for the two copy number change points studied in Figure 1: left panel, between a normal region* $(1,1)$ *and a single gain* $(1,2)$*; right panel, between a single gain* $(1,2)$ *and a region of copy-neutral LOH* $(0,2)$*. Affymetrix GenomeWideSNP_6 data.*

Allelic signals have a lower density than total signals as copy number probes only measure total copy numbers, and because only SNPs that are heterozygous in the germline are informative in terms of allelic imbalances. However, these ROC curves can be compared across signals because the evaluation is performed at a fixed *resolution* for each change point. Each resolution corresponds to a different number of markers for allelic and total signals. The change point between states $(1,1)$ and $(1,2)$ is detected slightly better with total signals (dashed) than with allelic signals (solid), allelic imbalances (black) or major copy numbers (green). As expected, the change point is not detected by minor copy numbers (red), as there is no change in true minor copy numbers. The change point between $(1,2)$ and $(0,2)$ is detected with similar or higher power using allelic signals than using total signals. Similar patterns are observed for other types of change points, suggesting that there is substantial detection power to gain in using both total and allelic signals for the detection of copy number changes.

6.3 Comparison between existing joint methods

Four methods based on HMM perform truly joint TCN and AI analyses: PICNIC [43], PSCN [69], genoCNA [70], and the method proposed in [68]. One advantage of HMM-based methods is that they can incorporate different probe types (SNPs and copy number probes) naturally, although in practice this seems to have been done only in PICNIC [43].

As discussed in Section 4, HMMs with discrete hidden state spaces perform the detection and calling steps at the same time, and are necessarily limited in terms of number of copy number states, that is, they cannot adapt to the intrinsic number of copy number states of a given problem. To our knowledge, PSCN [69] is currently the only method for joint TCN and AI analysis which is based on a continuous hidden state space. Conversely, one drawback of this type of approach is that it does not give a hard segmentation of the data in copy number states. Instead, copy numbers are estimated at each particular location and the method has to be combined with some thresholding in order to actually provide a segmentation of the original data. Moreover, downstream analyses are needed to estimate and/or call minor and major copy numbers.

We advocate the development of a joint direct segmentation method, that could take fully advantage of the two dimensions of SNP array data, as the above HMM do, but without assuming a particular form for the distribution of the copy number states sequence or the distribution of the size of copy number regions. Such a method could rely on the same type of models as those developed for joint direct segmentation of several copy number profiles [72, 73, 74].

7 Concluding remarks

In this chapter, we have underlined key aspects the analysis of SNP array data, including the influence of purity and ploidy on the observed data, and explained how they should be accounted for in the identification of copy number states. Although existing methods adequately address several of the challenges we focused on in this chapter, a few questions remain to be solved besides the above-mentioned development of a joint direct segmentation method. For the problem of detecting copy number changes, most existing methods assume that the errors follow a Gaussian distribution, although microarray data may be more heavy tailed. Current statistical models can be extended to other types of error distribution, but the main difficulty resides in developing efficient practical implementations.

For calling copy number states, although the effects of purity and ploidy are now widely acknowledged, methods to account for them — and also for tumor heterogeneity, that is, the possible presence of several tumoral clones in the tumor sample — will probably have to be improved and adapted to different types of cancers. A critical assessment of such methods is desirable, and would require producing validation data where purity and ploidy are known.

We have focused on the identification of copy number changes for one sample from one SNP array platform. In conclusion, we indicate statistical questions that arise in more general settings: when several samples are con-

sidered at a time, when one sample has been assayed on several platforms, and with newer copy number technologies.

Identifying recurrent allele-specific events. Even though some of the preprocessing methods described in Section 3 require several microarrays, currently available methods for identifying copy number states from SNP arrays analyze each tumor sample separately. However, the joint analysis of several samples from the same tumor type should be more powerful if the same biological events can be shared by several samples, as already demonstrated for total copy numbers for array-CGH data [72, 75, 73, 74]. Extensions of such methods to allelic signals remain to be developed.

Combining allele-specific signals across platforms. When the same sample is analyzed by two different platforms, combining signals across platforms should lead to improved detection of copy number alterations. This has been demonstrated for total copy numbers [41, 76] but still has to be investigated for allelic signals.

High-throughput sequencing. Currently, high-throughput sequencing technologies are more expensive than SNP arrays for whole genome allele-specific copy number studies, because accurate estimation of allelic ratios from read count data requires high sequencing coverage. The rapid evolution of these technologies suggests that allele-specific copy number studies will be cost-effective in the near future, leading to new statistical issues that will need to be addressed.

Acknowledgments

We gratefully acknowledge the Lawrence Berkeley National Laboratory (LBNL) and The Cancer Genome Atlas (TCGA) for making data and results available. This work was support by NCI grant U24 CA126551.

References

1. D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000. 4
2. Lynda Chin and Joe W. Gray. Translating insights from the cancer genome into clinical practice. *Nature*, 452(7187):553–563, April 2008. 4
3. D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, 20:207–211, 1998. 4, 16
4. Donna G Albertson and Daniel Pinkel. Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet*, 12 Spec No 2:R145–52, Oct 2003. 4

5. Musaffe Tuna, Sakari Knuutila, and Gordon B Mills. Uniparental disomy in cancer. *Trends in Molecular Medicine*, 15(3):120–128, March 2009. PMID: 19246245. 6
6. Johan Staaf, David Lindgren, Johan Vallon-Christersson, Anders Isaksson, Hanna Goransson, Gunnar Juliusson, Richard Rosenquist, Mattias Hoglund, Ake Borg, and Markus Ringner. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biology*, 9(9):R136, 2008. 6, 7, 26, 27
7. G Assié, T LaFramboise, P Platzer, J Bertherat, CA Stratakis, and C Eng. SNP arrays in heterogeneous tissue: highly accurate collection of both germline and somatic genetic information from unpaired single tumor samples. *Am J Hum Genet*, 82:903–915, 2008. 6, 26, 27
8. Cheng Li, Rameen Beroukhim, Barbara A Weir, Wendy Winckler, Levi A Garraway, William R Sellers, and Matthew Meyerson. Major copy proportion analysis of tumor samples using SNP arrays. *BMC Bioinformatics*, 9:204, 2008. 6, 27
9. Thomas LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, 37(13):4181–4193, July 2009. PMID: 19570852. 7, 11
10. Daniel A Peiffer, Jennie M Le, Frank J Steemers, Weihua Chang, Tony Jenniges, Francisco Garcia, Kirt Haden, Jiangzhen Li, Chad A Shaw, John Belmont, Sau Wai Cheung, Richard M Shen, David L Barker, and Kevin L Gunderson. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, 16(9):1136–1148, September 2006. 7, 11, 15
11. P.J. Gardina, K.C. Lo, W. Lee, J.K. Cowell, and Y. Turpaz. Ploidy status and copy number aberrations in primary glioblastomas defined by integrated analysis of allelic ratios, signal ratios and loss of heterozygosity using 500 K SNP Mapping Arrays. *BMC genomics*, 9(1):489, 2008. 7, 24, 27
12. Francis S. Collins and Anna D. Barker. Mapping the cancer genome. *Scientific American*, 296(3):50–57, Mar 2007. 7
13. The Cancer Genome Atlas (TGCA) research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061–1068, 2008. 7
14. H. Bengtsson, P. Wirapati, and TP Speed. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix arrays including GenomeWideSNP 5 & 6. *Bioinformatics*, 27(17):2149–2156, 2009. 7, 8, 10, 11, 13, 27
15. Henrik Bengtsson, Pierre Neuvial, and Terence P Speed. TumorBoost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinformatics*, 11(1):245, 2010. 7, 8, 15, 27
16. Affymetrix Inc. Affymetrix Genome-Wide Human SNP Array 6.0. Data sheet, 2007. 9, 10
17. Affymetrix Inc. Affymetrix cytogenetics research solution. Data sheet, 2009. 9, 10
18. Kevin L Gunderson, Frank J Steemers, Grace Lee, Leo G Mendoza, and Mark S Chee. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet*, 37(5):549–554, May 2005. 9
19. Frank J. Steemers and Kevin L. Gunderson. Whole genome genotyping technologies on the BeadArray platform. *Biotechnology Journal*, 2(1):41–49, 2007. 9
20. Illumina, inc. SNP genotyping and copy number analysis. Illumina Product Guide, 2009. 9, 15
21. Yasuhito Nannya, Masashi Sanada, Kumi Nakazaki, Noriko Hosoya, Lili Wang, Akira Hangaishi, Mineo Kurokawa, Shigeru Chiba, Dione K Bailey, Giulia C

Kennedy, and Seishi Ogawa. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res*, 65(14):6071–6079, 2005 Jul 15. 10, 11

22. Shumpei Ishikawa, Daisuke Komura, Shingo Tsuji, Kunihiro Nishimura, Shogo Yamamoto, Binaya Panda, Jing Huang, Masashi Fukayama, Keith W Jones, and Hiroyuki Aburatani. Allelic dosage analysis with genotyping microarrays. *Biochem Biophys Res Commun*, 333(4):1309–1314, 2005 Aug 12. 10, 11
23. Benilton Carvalho, Henrik Bengtsson, Terence P Speed, and Rafael A Irizarry. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, 8(2):485–499, 2007 Apr. 10, 11
24. H Bengtsson, R Irizarry, B Carvalho, and T P Speed. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, 24(6):759–767, 2008 Mar 15. 10, 11
25. Yee H Yang, Sandrine Dudoit, Percy Luu, David M Lin, Vivian Peng, John Ngai, and Terence P Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002. 11
26. C Li and W H Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98(1):31–36, 2001 Jan 2. 11, 12
27. B M Bolstad, R A Irizarry, M Astrand, and T P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–93, Jan 2003. 11, 12
28. Maria Ortiz-Estevez, Henrik Bengtsson, and Angel Rubio. ACNE: a summarization method to estimate allele-specific copy numbers for Affymetrix SNP arrays. *Bioinformatics*, Jun 2010. 11
29. Nusrat Rabbee and Terence P Speed. A genotype calling algorithm for Affymetrix SNP arrays. *Bioinformatics*, 22(1):7–12, Jan 2006. 11
30. Affymetrix Inc. BRLMM: an improved genotype calling method for the GeneChip Human Mapping 500K Array Set, Apr 2006. 11
31. Thomas LaFramboise, David Harrington, and Barbara A Weir. PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data. *Biostatistics*, 8(2):323–336, Apr 2007. 11
32. Illumina, inc. Illumina's genotyping data normalization methods. White paper, 2006. 11
33. Johan Staaf, Johan Vallon-Christersson, David Lindgren, Gunnar Juliusson, Richard Rosenquist, Mattias Hoglund, Ake Borg, and Markus Ringner. Normalization of illumina infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. *BMC Bioinformatics*, 9(1):409, 2008. 12
34. Frank J Steemers, Weihua Chang, Grace Lee, David L Barker, Richard Shen, and Kevin L Gunderson. Whole-genome genotyping with the single-base extension assay. *Nature Methods*, 3(1):31–33, 2006. PMID: 16369550. 12
35. Henrik Bengtsson. *Low-level analysis of microarray data*. PhD thesis, Centre for Mathematical Sciences, Division of Mathematical Statistics, Lund University, oct 2004. 12, 24
36. Go Yamamoto, Yasuhito Nannya, Motohiro Kato, Masashi Sanada, Ross L Levine, Norihiko Kawamata, Akira Hangaishi, Mineo Kurokawa, Shigeru Chiba, D Gary Gilliland, H Phillip Koeffler, and Seishi Ogawa. Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of affymetrix single-nucleotide-polymorphism genotyping microarrays. *Am J Hum Genet*, 81(1):114–126, 2007 Jul. 12, 24, 27
37. S. Pounds, C. Cheng, C. Mullighan, S.C. Raimondi, S. Shurtleff, and J.R. Downing. Reference alignment of SNP microarray signals for copy number analysis of tumors. *Bioinformatics*, 25(3):315, 2009. 12

38. A. B. Olshen, E. S. Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004. 13, 17, 19, 20
39. H. Bengtsson, K. Simpson, J. Bullard, and K. Hansen. aroma. affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. Technical report, Technical Report 745, Department of Statistics, University of California, Berkeley, 2008. 13
40. J. von Neumann, R. H. Kent, H. R. Bellinson, and B. I. Hart. The mean square successive difference. *The Annals of Mathematical Statistics*, 12(2):153–162, 1941. 13
41. Henrik Bengtsson, Amrita Ray, Paul T Spellman, and Terence P Speed. A single-sample method for normalizing and combining full-resolution copy numbers from multiple sources and technologies. *Bioinformatics*, 25(7):861—867, 2009. 13, 27, 30
42. E S Venkatraman and Adam B Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663, March 2007. 14, 17, 19, 20
43. Chris D. Greenman, Graham Bignell, Adam Butler, Sarah Edkins, Jon Hinton, Dave Beare, Sajani Swamy, Thomas Santarius, Lina Chen, Sara Widaa, P. Andy Futreal, and Michael R. Stratton. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostat*, 11(1):164–175, 2010. 15, 27, 28
44. Weil R Lai, Mark D Johnson, Raju Kucherlapati, and Peter J Park. Comparative analysis of algorithms for identifying amplifications and deletions in array-CGH data. *Bioinformatics*, 21(19):3763–3770, 2005 Oct 1. 16
45. Hanni Willenbrock and Jane Fridlyand. A comparison study: applying segmentation to array-CGH data for downstream analyses. *Bioinformatics*, 21(22):4084–4091, 2005 Nov 15. 16
46. Kees Jong, Elena Marchiori, Aad van der Vaart, Bauke Ylstra, Marjan Weiss, and Gerrit Meijer. Chromosomal breakpoint detection in human cancer. In Günther R. Raidl, Stefano Cagnoni, Juan Jesús Romero Cardalda, David W. Corne, Jens Gottlieb, Agnès Guillot, Emma Hart, Colin G. Johnson, Elena Marchiori, Jean-Arcady Meyer, and Martin Middendorf, editors, *Applications of Evolutionary Computing, EvoWorkshops2003: EvoBIO, EvoCOP, EvoIASP, EvoMUSART, EvoROB, EvoSTIM*, volume 2611 of *LNCS*, pages 54–65, University of Essex, England, UK, 14-16 April 2003. Springer-Verlag. 17
47. Nancy R Zhang and David O Siegmund. A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007. 17
48. M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, 2005. 17
49. F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.J. Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6(1):27–27, 2005. 17
50. Guillem Rigaill. Pruned dynamic programming for optimal multiple change-point detection. Arxiv preprint arXiv:1004.0887, April 2010. 17
51. R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 2007. 18
52. Z. Harchaoui and C. Lévy-Leduc. Catching change-points with lasso. *Advances in Neural Information Processing Systems*, 20:161–168, 2008. 18
53. R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 91–108, 2005. 18
54. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of statistics*, 32(2):407–451, 2004. 18

55. J. Fridlyand, A. Snijders, D. Pinkel, D. G. Albertson, and A. N. Jain. Application of hidden markov models to the analysis of the array CGH data. *Journal of Multivariate Analysis*, 90:132–153, 2004. Special Issue on Multivariate Methods in Genomic Data Analysis. 19
56. S. Guha, Y. Li, and D. Neuberg. Bayesian hidden Markov modeling of array CGH data. *Journal of the American Statistical Association*, 103(482):485–497, 2008. 19
57. Tze Leung Lai, Haipeng Xing, and Nancy Zhang. Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostat*, 9(2):290–307, April 2008. 19
58. Jonathan Sebat, B. Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Pr Lundin, Susanne Mnr, Hillary Massa, Megan Walker, Maoyen Chi, Nicholas Navin, Robert Lucito, John Healy, James Hicks, Kenny Ye, Andrew Reiner, T. Conrad Gilliam, Barbara Trask, Nick Patterson, Anders Zetterberg, and Michael Wigler. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–8, Jul 2004. 22
59. A. John Iafrate, Lars Feuk, Miguel N Rivera, Marc L Listewnik, Patricia K Donahoe, Ying Qi, Stephen W Scherer, and Charles Lee. Detection of large-scale variation in the human genome. *Nature Genet.*, 36(9):949–951, Sep 2004. 22
60. Richard Redon, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, and ... Global variation in copy number in the human genome. *Nature*, 444:444–454, 2006. 22
61. R.B. Scharpf, G. Parmigiani, J. Pevsner, and I. Ruczinski. Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *Ann. Appl. Statist.*, 2(2):687–713, 2008. 26
62. K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S.F.A. Grant, H. Hakonarson, and M. Bucan. PennCNV: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, 17(11):1665, November 2007. 26
63. Stefano Colella, Christopher Yau, Jennifer M. Taylor, Ghazala Mirza, Helen Butler, Penny Clouston, Anne S. Bassett, Anneke Seller, Christopher C. Holmes, and Jiannis Ragoussis. QuantiSNP: an objective bayes Hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucl. Acids Res.*, 35(6):2013–2025, March 2007. 26
64. Joshua M Korn, Finny G Kuruvilla, Steven A McCarroll, Alec Wysoker, James Nemesh, Simon Cawley, Earl Hubbell, Jim Veitch, Patrick J Collins, Katayoon Darvishi, Charles Lee, Marcia M Nizzari, Stacey B Gabriel, Shaun Purcell, Mark J Daly, and David Altshuler. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*, 40(10):1253–1260, 2008 Oct. 26
65. Edward F Attiyeh, Sharon J Diskin, Marc A Attiyeh, Yaël P Mossé, Cuiping Hou, Eric M Jackson, Cecilia Kim, Joseph Glessner, Hakon Hakonarson, Jaclyn A Biegel, and John M Maris. Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res*, 19(2):276–283, February 2009. 27
66. Adam B Olshen, Richard A Olshen, Henrik Bengtsson, Pierre Neuvial, Paul T Spellman, and Venkatraman E Seshan. Extension of circular binary segmentation to parent-specific copy number. Submitted, May 2010. 26, 27
67. Tatiana Popova, Élodie Manié, Dominique Stoppa-Lyonnet, Guillem Rigaill, Emmanuel Barillot, and Marc-Henri Stern. Genome alteration print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biology*, 10(11):R128, 2009. 26, 27
68. P. Lamy, C.L. Andersen, L. Dyrskjot, N. Torring, and C. Wiuf. A Hidden Markov Model to estimate population mixture and allelic copy-numbers in cancers using Affymetrix SNP arrays. *BMC bioinformatics*, 8(1):434, 2007. 27, 28

69. Hao Chen, Haipeng Xing, and Nancy R Zhang. Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays. Technical report, Stanford University, 2009. 27, 28, 29
70. Wei Sun, Fred A. Wright, Zhengzheng Tang, Silje H. Nordgard, Peter Van Loo, Tianwei Yu, Vessela N. Kristensen, and Charles M. Perou. Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucl. Acids Res.*, 37(16):5365–5377, September 2009. 27, 28
71. Rameen Beroukhim, Ming Lin, Yuhyun Park, Ke Hao, Xiaojun Zhao, Levi A Garraway, Edward A Fox, Ephraim P Hochberg, Ingo K Mellinghoff, Matthias D Hofer, Aurelien Descazeaud, Mark A Rubin, Matthew Meyerson, Wing Hung Wong, William R Sellers, and Cheng Li. Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. *PLoS Comput Biol*, 2(5):e41, 2006 May. 27
72. Nancy R Zhang, David O Siegmund, Hanlee Ji, and Jun Z Li. Detecting simultaneous change-points in multiple sequences. *Biometrika*, 2010. 29, 30
73. Kevin Bleakley and Jean-Philippe Vert. Joint segmentation of many aCGH profiles using fast group LARS. http://hal.archives-ouvertes.fr/hal-00422430/en/, October 2009. 29, 30
74. Franck Picard, Émilie Lebarbier, Eva Budinaská, and Stéphane Robin. Joint segmentation of multivariate Gaussian Processes using mixed linear models. Technical report, Statistics for Systems Biology Group, 2007. 29, 30
75. Sohrab P Shah, Wan L Lam, Raymond T Ng, and Kevin P Murphy. Modeling recurrent DNA copy number alterations in array-CGH data. *Bioinformatics*, 23(13):i450–8, July 2007. 30
76. Nancy R. Zhang, Yasin Senbabaoglu, and Jun Z. Li. Joint estimation of DNA copy number from multiple platforms. *Bioinformatics*, 26(2):153–160, November 2009. 30