Cytogenet Genome Res 123:343–351 (2008)
DOI: 10.1159/000184726

Cytogenetic and Genome Research

Computational methods for identification of recurrent copy number alteration patterns by array CGH

S.P. Shah

Department of Computer Science, University of British Columbia, and Department of Pathology (UBC), British Columbia Cancer Agency, Vancouver, B.C. (Canada)

Accepted in revised form for publication by H. Kehrer-Sawatzki and D.N. Cooper, 20 August 2008.

Abstract. Recurrent DNA copy number alterations (CNA) are widely studied in diagnostic and cytogenetic cancer research. CNAs reveal locations that may alter gene dosage and thus expression of the genes contained within. Array comparative genomic hybridization has emerged as a popular high-throughput, genome-wide technique to interrogate tumor genomes for copy number alterations. When studying a group of tumors derived from a patient cohort, it is of great interest to detect the copy number alterations that are common across the population and thus assumed to be potential diagnostic markers and/or predictors of clinical outcome. In this paper, we review extant and available computational approaches for detecting such recurrent copy number alterations from array comparative genomic hybridization (aCGH) data. This is a challenging computational problem due to various sources of noise in the data that can obscure the recurrent copy number signals or induce false positives in their prediction. In this paper, we qualitatively evaluate methods designed to detect recurrent copy number alterations for aCGH data based on their analytical strengths and limitations, and discuss expected future directions in this important area of cancer research.

1 Introduction

DNA copy number alterations (CNAs) are present in nearly all tumor genomes (Hanahan and Weinberg, 2000). CNAs are intervals of a chromosome ranging in size from a few kilobases to whole chromosome arms where genetic material is deleted or amplified. Such alterations can indicate the genomic instability of a tumor and are a result of acquired somatic mutations in the evolution of the tumor cells from a normal state to a neoplastic state. Numerous studies have revealed that the gene dosage alterations induced by CNAs result in differential expression of the genes contained within them (Pollack et al., 2002; Aguirre et al., 2004; Heidenblad et al., 2005; Chin et al., 2007; Lee et al., 2008). As such, identification of genomic amplifications (or gains) harboring oncogenes and deletions (losses) harboring tumor suppressor genes are of particular interest (Balmain et al., 2003).

Recent work has revealed previously undescribed recurrent CNAs that are implicated in cancer (Hosoya et al., 2006; Chin et al., 2007), demonstrating that the catalogue of clinically relevant CNAs is far from complete. It is routine to study cohorts of patients with similar phenotype or cancer (sub)type to determine population-level patterns of CNAs. Generally it is assumed that recurrent CNAs, amplifications or deletions found in the same location across the set of patient tumors, are evidence for so-called 'driver' alterations, or alterations that are symptomatic and/or causative of the disease (Pollack et al., 2002). Examples of clini-

S.P.S. is supported by the Michael Smith Foundation for Health Research.

Request reprints from Sohrab P. Shah
British Columbia Cancer Agency, 3427-600 W 10th Ave
Vancouver, B.C. (Canada)
telephone: +1 604 877 6000 x2589; fax: +1 604 877 6089
e-mail: sshah@bccrc.ca

1424–8581/08/1234–0343$24.50/0

Accessible online at:
www.karger.com/cgr

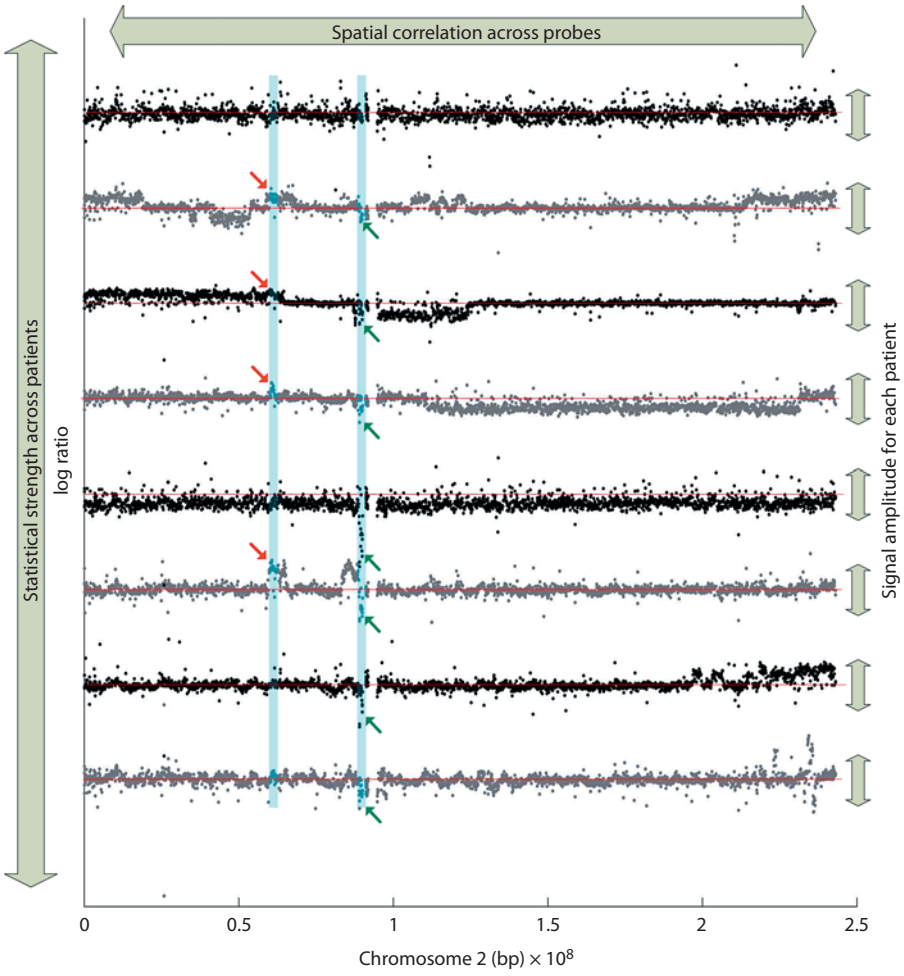

Fig. 1. Example aCGH data from eight mantle cell lymphoma cell lines (de Leeuw et al., 2004) showing two examples of recurrent CNAs (shaded in blue) found on chromosome 2. Each horizontal set of dots represents the log ratios of a given cell line (or patient). The red dotted lines indicate the 0 log ratio (or expected neutral value). The probes that lie in the blue shaded areas (recurrent loss on the right, and recurrent gain on the left in four cell lines depicted by red arrows) comprise the desired output of an algorithm to detect recurrent CNAs. Note that for the recurrent CNAs, statistical strength across patients can be leveraged to detect them. Also note that CNAs tend to span regions of contiguous probes, thus spatial correlation across the chromosome should be leveraged. Finally the amplitude of the signal for each patient should be considered in the analysis.

cally relevant recurrent CNAs such as *HER2 (ERBB2)* amplification in breast cancer (Slamon et al., 1987), *EGFR* in non-small cell lung cancer (Hirsch et al., 2003), and *MYCN* in neuroblastoma (Iehara et al., 2006) provide precedent for this assumption, although we caution the reader that this assumption does not always hold (see Section 4).

Indeed, some of these alterations are used for prognostic testing (Yaziji et al., 2004) and the development of diagnostic tools (Schwaenen et al., 2004). Furthermore, recurrent CNAs are thought to be selected for in the clonal evolution of a tumor (Weinberg, 2007) and their study can suggest the presence of genes involved in disrupted cancer-related biochemical pathways (Coe et al., 2006). We note parenthetically that the ability to detect driver alterations depends on the molecular homogeneity and composition of the patient cohort. If the cohort is heterogeneous (composed of several distinct molecular subtypes), important driver alterations of a rare subtype could be obscured by patterns from the remainder of the population (see Section 4).

The 'passenger' CNAs, in contrast, are those that are patient specific and are generally not shared across the population. These alterations may result from acquired genomic instability, non-pathological copy number variations (Redon et al., 2006; Wong et al., 2007) or other mechanisms that are not well-understood. Thus, separating driver CNAs from passenger CNAs is critical to reveal potential diagnostic/prognostic markers as well as therapeutic targets for improved clinical care and management of the disease (Chin and Gray, 2008).

A powerful experimental tool for studying CNAs is high-resolution genome-wide profiling with array comparative genomic hybridization (aCGH) (Ishkanian et al., 2004; Pinkel and Albertson, 2005). (Detailed discussion of aCGH is beyond the scope of this contribution. We refer the reader to a comprehensive review by Pinkel and Albertson (2005) for further reading.) Current examples of aCGH profiling of tumors include large-scale studies to characterize lung (squamous cell), brain (glioblastoma multiforme) and ovarian serous cancer as part of the Cancer Genome Atlas project (Collins and Barker, 2007). aCGH allows CNAs to be interrogated on a genome-wide level, and their boundaries to be mapped to within a few kb. Due to the high dimensionality of aCGH (25,000–1,000,000 probes depending on the platform), it is essential to turn to computational meth-

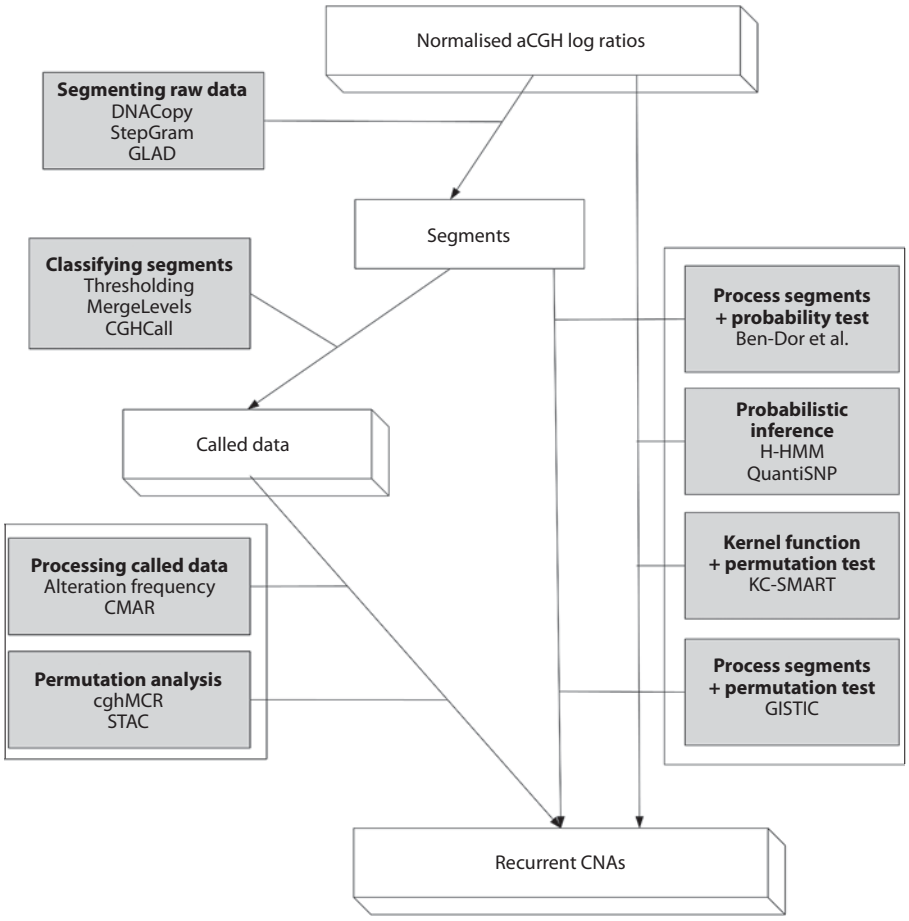

Fig. 2. Workflows for inferring recurrent CNAs from aCGH data. We show the various steps used in predicting recurrent CNAs from aCGH data. The top part of the diagram shows the preprocessing steps some algorithms use to map raw data to called data. The algorithms in the white box on the left then process the called data to infer recurrent CNAs while those in the white box on the right process either continuous segmented data or the raw data directly. Also shown are which algorithms use probabilistic inference or permutation testing to produce their results. Please refer to Table 1 for availability of software.

ods to interpret CNAs from aCGH data. The main problem we focus on in this paper is the task of detecting recurrent CNAs given aCGH data from a cohort of patients using computational approaches, under the assumption that they represent putative driver alterations.

In Fig. 1 we show aCGH data from a set of eight mantle cell lymphoma cell lines, originally published in de Leeuw et al. (2004). (Note that routine studies often consist of 10s or 100s of cases, but we use this example for illustrative purposes.) Recurrent CNAs, identified by visual inspection, are shown in the blue shaded areas. The problem of computationally detecting such recurrent CNAs is relatively underrepresented in the bioinformatics literature. As such, the limitations of current algorithms in practice are not yet fully understood. The purpose of this paper is to survey existing published algorithms designed for this purpose, point out their relative strengths and weaknesses, and suggest how this field can move forward to fully meet this important need in aCGH data analysis. We aim to provide the clinical or molecular biology investigator with a synopsis of the current analytical approaches for this problem and the computational scientist with an overview of the current approaches, and identify possibilities for further advancement. (Note that in general, the theories and assumptions of the algorithms presented herein should apply to the analysis of CNVs in population genetics studies, though this is not the focus of this contribution.)

This paper is organized as follows. In Section 2, we introduce notation and formalize the computational problem at hand. In Section 3, we will survey the available methods for detection of recurrent CNAs and discuss their relative strengths and weaknesses. In Section 4, we discuss future research directions for this problem and how the problem can be placed within the context of emerging technologies for investigation of molecular alterations in cancer.

2 Notation and computational problem

The concept of separating driver alterations from passenger alterations is mirrored by the notion of computational dimensionality reduction and/or feature selection. If we consider the set of probes in the array as features, the task is to select a small number of features that are likely to represent recurrent CNAs, and thus a molecular profile of the disease. To help formalize this problem, we introduce notation in this section and define the computational problem of inferring recurrent CNAs from aCGH data. A schematic diagram of computational workflows to aid the reader is shown in Fig. 2. The algorithms we will discuss in Section

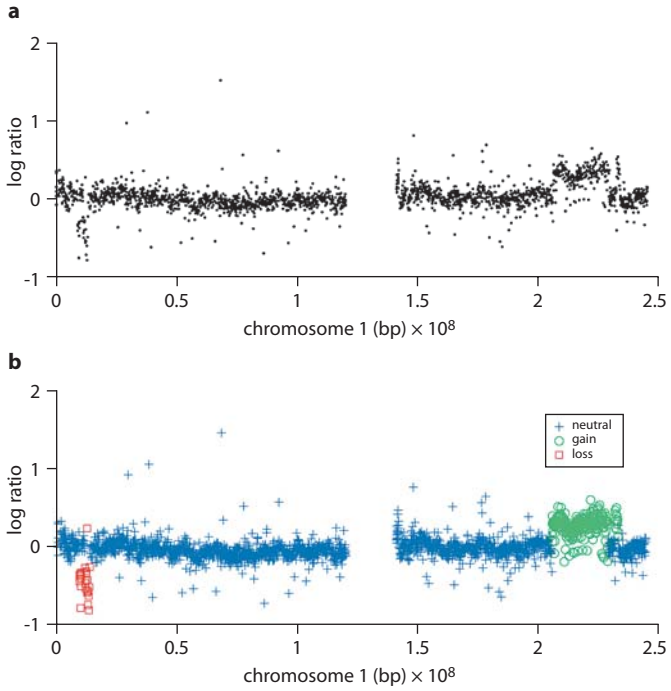

Fig. 3. (**a**) Example aCGH data from de Leeuw et al. (2004) for chromosome 1 of mantle cell lymphoma cell line HBL2. The horizontal axis is the physical chromosomal location of the probe and the vertical axis is the corresponding log ratio. (**b**) Same data as **a** with CNAs labeled by an expert. Blue crosses indicate neutral probes, green circles are gains and red squares are losses.

3 differ in their steps taken to traverse this diagram, starting at the raw aCGH data in ending with recurrent CNAs. We will see how the paths through the workflow diagram confer certain advantages/disadvantages in prediction of recurrent CNAs. The algorithms in the white box on the left operate on called or discrete data, while the algorithms in the white box on the right operate on raw data (see next section).

The aCGH data consist of a log ratio $Y_t^p \in \mathcal{R}$ of hybridization intensity of tumor DNA versus normal DNA for each probe $t \in (1, ..., T)$ in the array and for each patient $p \in (1, ..., P)$ in the population. $Y_{1:T}^{1:P}$ thus represents the full data matrix and represents a noisy measurement of actual copy number. Note that we assume that $Y_{1:T}^{1:P}$ is derived from a normalization step that has been adjusted for technical biases and artifacts from the experimental protocol (see Khojasteh et al., 2005; Neuvial et al., 2006; Marioni et al., 2007 for details). A common step in analyzing aCGH data is to find a mapping $Y_{1:T}^P \rightarrow Z_{1:T}^P$ for each patient p where $Z_t^p = k$ represents a discrete copy number state, $k \in$ {*loss, neutral, gain*}, to denoise the data. These states correspond to regions of genomic deletion, no change, and amplification, thus interpreting each probe's continuous log ratio with a biologically meaningful discrete label. We show this schematically in Fig. 3 with the raw data for a single aCGH sample taken from chromosome 1 and the associated discrete labels indicating the location of CNAs.

To infer this mapping computationally is non-trivial (mainly due to sources of noise in the data) and has been extensively studied. The techniques for mapping $Y_{1:T}^p \rightarrow Z_{1:T}^p$ are reviewed/benchmarked in Lai et al. (2005), Willenbrock and Fridlyand (2005) and Pique-Regi et al. (2008). The mapping is often done by segmenting the data into piecewise constant sets of probes with the same assumed underlying copy number by model-based approaches such as hidden Markov models (HMM) (Fridlyand et al., 2004; Shah et al., 2006; Marioni et al., 2006; Rueda and Diaz-Uriarte, 2007), a Markov random field method proposed by Broet and Richardson (2006) and a pseudolikelihood approach by Engler et al. (2006). The segments in these approaches are assigned a discrete copy number state during segmentation. Alternatively, the data is segmented into a piecewise constant signal where the segments are assigned a continuous value (empirical mean log ratio) (Hupe et al., 2004; Olshen et al., 2004; Venkatraman and Olshen, 2007). These methods operate on finding segments that maximize between segment variance and minimize within segment variance. We call this approach continuous segmentation. The segments are often post-processed and classified as either loss, neutral or gain by so-called segment classification or merging algorithms (MergeLevels, Willenbrock and Fridlyand, 2005; CGHCall, van de Wiel et al., 2007).

We refer to the $Z_{1:T}^{1:P}$ matrix as called data and the $Y_{1:T}^{1:P}$ matrix as raw data. The algorithms for inferring recurrent CNAs can be grouped to a large extent on whether they accept called or raw data as input. We will discuss the relative merits of the extant approaches and how the called or raw data can affect results. The output of all algorithms for detecting recurrent CNAs is a profile, which we represent by $\phi_{1:T}$. In some cases $\phi_{1:T}$ will represent a statistic or probability that a probe is recurrently altered, in other cases $\phi_{1:T}$ is simply a binary representation indicating presence or absence of a recurrently altered probe. Generally it is assumed by most algorithms that recurrent CNAs span a relatively small fraction of the probes in the array.

3 Computational approaches for inferring recurrent CNAs

We divide the current approaches for inferring recurrent alterations into two categories: those that input the called data matrix $Z_{1:T}^{1:P}$ and those that input the raw data matrix $Y_{1:T}^{1:P}$. In general, there are three axes or dimensions across which the various approaches operate: i) the actual amplitude of the log-ratio signal contained in $Y_{1:T}^{1:P}$, ii) spatial correlation across probes (i.e. across columns in the data matrix) and iii) concurrence across the population (i.e. across rows of the data matrix). These dimensions are depicted with double-ended grey arrows on Fig. 1. We will see how different algorithms exploit these characteristics. Note that an underlying assumption of some algorithms is that the patient cohort is relatively homogeneous. Preprocessing the data to separate the patients into subgroups with shared mo-

Table 1. List of algorithms for recurrent CNAs, their data input, whether or not probabilistic/statistical output is provided, their availability if applicable

Algorithm	Input data	Probabilistic/ statistical output	Availability
cghMCR (Aguirre et al., 2004)	Called	Y	http://www.bioconductor.org
CMAR (Rouveirol et al., 2006)	Called	N	from author
STAC (Diskin et al., 2006)	Called	N	http://cbil.upenn.edu/STAC
CoCoA (Ben-Dor et al., 2007)	Raw	Y	n/a
H-HMM (Shah et al., 2007)	Raw	Y	http://www.cs.ubc.ca/~sshah/acgh
KC-SMART (Klijn et al., 2008)	Raw	Y	n/a
QuantiSNP (Colella et al., 2007)	SNP	Y	http://www.well.ox.ac.uk/QuantiSNP
GISTIC (Beroukhim et al., 2007)	SNP	Y	http://www.broad.mit.edu

lecular patterns, or by known clinical subtype should be done if possible. We will discuss some methodological progress in this area in Section 4.

3.1. Algorithms for called data

The simplest algorithm for inferring recurrent CNAs from $Z_{1:T}^{1:P}$ is to compute a frequency of alterations for each probe such that

$$\phi_t(k) = \frac{1}{P}\sum_{p=1}^{P} I\left(Z_t^p = k\right),$$

where $I(Z_t^p = k)$ is a function indicating that Z_t^p is in state k. Sets of probes from ϕ_t (loss) and ϕ_t (gain) are then selected based on frequency thresholding (for a recent example, see Idbaih et al., 2008). We refer to this method as alteration frequency (AF). While AF may be effective in some cases, it is limited in that it does not directly output meaningful statistics or probabilities to the investigator to quantify and thus compare the observed results. Many authors treat the recurrent CNA problem as finding regions in the $Z_{1:T}^{1:P}$ matrix spanning contiguous set of probes in CNAs that maximally overlap across the patients. An example of this approach is reported in Aguirre et al. (2004) and available in the Bioconductor (Gentleman et al., 2004) package cghMCR (see Table 1 for availability). This method uses a step-wise approach and a permutation test to find recurrent CNAs based on a statistical score. The data are first segmented with DNACopy (Olshen et al., 2004; Venkatraman and Olshen, 2007). Segments above an upper and lower user-settable threshold are labeled as CNAs resulting in the previously discussed $Y_{1:T}^p \rightarrow Z_{1:T}^p$ mapping. Highly altered CNAs are retained as important regions that define discrete locus boundaries. These regions are compared across patients to identify overlapping groups of positive or negative value segments. Minimal common regions (MCRs) are defined as regions having at least a user-defined recurrence rate across samples and where the median log ratio for the probes with the segment across patients is above the 95% percentile in a permutation test. This method was the first to suggest a computational approach for identifying recurrent CNAs. Although it was shown to be effective in Aguirre et al. (2004), it is somewhat ad-hoc and depends on a number of user-settable thresholds. One does not typically know how to correctly choose these thresholds, and a particular setting may not generalize well to other data sets.

A more mathematically motivated approach is presented by Rouveirol et al. (2006). They present a formal definitional framework based on a binarized representation of the data (treating losses and gains separately and independently). The framework is used to develop two algorithms for discovering recurrent CNAs termed minimal altered region (MAR) and constrained minimal altered region (CMAR). Both are data mining methods based on finding close constrained itemsets (sequences) in the binary matrix restricted to sequential data – a necessary extension due to the spatially ordered nature of the aCGH probes in the genome. A simplistic summary of the CMAR algorithm is that it searches for small rectangles of 1's in the input binary matrix, similar in concept to biclustering. The CMAR algorithm has a worst case running time proportional to the square of the number of probes, which may limit its use to aCGH platforms with smaller numbers of probes.

Diskin et al. (2006) also input a binary matrix similar to Rouveirol et al. (2006). Their method, STAC, computes two complementary statistics for quantifying the likelihood of observed recurrent CNAs. The first estimates how often the observed frequency of an alteration would occur by chance. This is expected to uncover highly frequent alterations. The second is termed a footprint statistic, and is computed on the results of a greedy search strategy to find overlapping 'stacks' of alterations in the population. This is expected to detect recurrent CNAs that are low-frequency, yet possibly of clinical importance. In both cases, permutation analysis is performed to assess the statistical significance of what is observed. The statistical output allows the prioritization for experimental follow-up not possible in Rouveirol et al. (2006) which produces binary output.

cghMCR, CMAR and STAC all input called data. Working with called data has its advantages in that with respect to some characteristics, the data are assumed to be denoised. Thus the specificity of predictions is expected to be high. However, in the next section we discuss how working with called data may limit the sensitivity of predictions in certain circumstances.

3.2. Algorithms for raw data

Thus far we have considered algorithms that require discrete called data, or the $Z_{1:T}^{1:P}$ matrix, described above as input. Several authors: Lipson et al. (2006), Shah et al. (2007), Klijn et al. (2008) and Ben-Dor et al. (2007) assert that inputting the raw data as input has advantages over called data.

Ben-Dor et al. (2007) argue that the amplitude of aberration should contribute to the inference of recurrent CNAs. Consider the case where the data are discretized into a small number of states e.g. {*loss, neutral, gain*}, then (for example) high-level amplicons, which arguably provide stronger evidence of being selected in the clonal evolution of the tumor, would contribute equally as a single copy gain to discovering recurrent CNAs. Furthermore, important high-level amplicons may be infrequently targeted, making them harder to detect by methods discussed thus far. To leverage the amplitude of the signal, Ben-Dor et al. use a statistical framework based on the concept of measuring probe penetrance. The approach begins by segmenting the raw data using continuous segmentation (using StepGram, Lipson et al., 2006), thus producing an intermediate form of the data that preserves amplitude, but is piecewise constant. Depending on the amplitude and the relative abundance of CNAs in the sample, a statistic is computed to quantify the significance of each CNA. More formally, given a region R spanning a putative CNA in a given patient p, the algorithm computes how many other regions of R's size in the continuous segmented data of p have at least the same average amplitude across patients. This computes a patient-specific score $s(p,R)$. Given P patients, the statistical significance of observing the data spanned by R in the whole population is given by an adjusted probability density function based on the Binomial distribution. This approach therefore differs from the methods described in Section 3.1 in two important ways: it provides probabilistic output, and it models the signal amplitude across patients.

In Shah et al. (2007), we introduced a hierarchical HMM designed to explicitly model shared alterations (or putative driver alterations) and patient-specific alterations. Our approach extends HMMs for single patient analysis to the multiple patient case. The hierarchical HMM (H-HMM) is capable of switching the generative model, so the copy number state of each probe in the data is generated by a 'master' process representing putative driver alterations, or a patient-specific process representing passenger alterations. Two key ideas are presented. First, the 'master' process is inferred by jointly considering the raw data from all the patients. Thus the algorithm can borrow statistical strength across patients and detect 'subtle' shared alterations that may get smoothed over when making the 'calls' required by the algorithms in Section 3.1. Second, in regions where there is little evidence for a recurrent CNA, the algorithm allows for the data to be modeled by a patient-specific Markov process representing the passenger alterations. Both processes model the spatial correlation known to exist in the data. The resulting output is a probabilistic estimate of the uncertainty of each probe representing a canonical loss, neutral, gain, or a location predominately modeled by the passenger alteration process. The output is probabilistic and takes advantage of signals present across all three dimensions of the data (recall Fig. 1). A potential limitation we have noticed is that it does not readily detect infrequent alterations.

Klijn et al. (2008) suggest a method called KC-SMART, a locally weighted regression algorithm based on kernel convolution to compute a smoothed estimate of the recurrent CNAs. The method also considers all three dimensions of the data: amplitude, spatial correlation and frequency of alteration. Using $Y_{1:T}^{1:P}$ as input, the data are separated into positive and negative log-ratios. The positive ratios are summed across patients and the negative ratios are summed across patients. These sums are used in computing the amplitude of a Gaussian kernel convolution function whose values are then smoothed providing a single estimate for the combined log ratios across the population at an arbitrary genomic position. Thus the profile is a smooth, continuous representation of the raw data matrix. Statistical significance of the amplitude of the peaks is assessed using permutation analysis. A key feature of the algorithm is that the width of the kernel is defined by the user and thus can be tuned to find large recurrent CNAs and small recurrent CNAs.

3.3. Related algorithms for SNP arrays

Genotyping technology is also commonly used for copy number analysis. Single nucleotide polymorphism (SNP) chips can interrogate more than one million loci in the human genome in one experiment and consequently robust computational approaches have been developed for their analysis. Although not explicitly designed for aCGH, the approaches described in this section are easily modified to use with aCGH data and thus are very relevant to the discussion. Colella et al. (2007) suggest an HMM approach, QuantiSNP, where the log ratios and allele frequencies are combined to form the emission model. The hidden states in the model represent the combined copy number and genotype for each probe. The transition matrix between these states is non-stationary and is computed using a distance based prior, accounting for unequal genomic spacing of the probes. Importantly, this method introduces Bayes factors for assessing significance levels for altered regions. These significance measures are computed on segments, whereas most HMMs for aCGH output likelihood of the best sequence, or probe-specific probabilities. For population-level analysis, the authors suggest placing a transition matrix at each probe that is jointly updated across patients. Thus the non-stationary transition matrix models recurrent CNAs by leveraging statistical strength across patients. We have recently adapted this approach for aCGH data. Preliminary results indicate it performs similarly to the H-HMM from Shah et al. (2007) in a simulation study (data not shown).

Beroukhim et al. (2007) suggest a method called 'Genomic Identification of Significant Targets in Cancer' (GISTIC) based upon a step-wise workflow similar in spirit to KC-SMART, albeit with some notable differences. In Ber-

houkhim et al., the amplitudes of the CNAs are summed across patients to compute a probe-level score. The probes are repeatedly permuted and scores are recalculated for each permutation. The probes with scores in the original data that occur rarely by chance are selected by thresholding. A novel contribution is that the data in significant 'regions' are post-processed to characterize 'peaks' as focal alterations (for example that span single genes), broad alterations (for example that span entire chromosomes), or overlapping peaks of both types. In contrast to Klijn et al. (2008) where small and large peaks are found by iterative runs with difference parameter settings, GISTIC explicitly classify large and small peaks in a single run.

4 Discussion

4.1. The need for a benchmark study and data sets

We have surveyed and reviewed available methods (listed in Table 1) for detecting recurrent CNAs from aCGH data. We have discussed the merits and weaknesses of these methods; however, we are currently limited to qualitative arguments in our assessments. In contrast to the single aCGH mapping problem ($Y^p_{1:T} \rightarrow Z^p_{1:T}$) where there has been a convergence in the literature towards using HMMs (Fridlyand et al., 2004; Guha et al., 2006; Marioni et al., 2006; Shah et al., 2006; Rueda and Diaz-Uriarte, 2007; Stjernqvist et al., 2007; Andersson et al., 2008), the algorithms for inferring recurrent CNAs are diverse, to say the least. What is not known is how each approach confers accuracy advantages over its competitors, and under what characteristics of the data. How are the approaches affected by molecular heterogeneity of the population? What is the comparative sensitivity to detecting rare but important high level amplicons? Is there an associated cost to specificity? Analogous systematic studies to those performed by Lai et al. (2005) and Willenbrock and Fridlyand (2005) are needed to address these and other important questions in the context of detecting recurrent CNAs. Although we previously reported benchmark results on synthetic data (Shah et al., 2007), and Klijn et al. (2008) showed some qualitative comparisons to another method, quantitative benchmarking on real-data sets with experimentally validated recurrent CNAs is required in order to definitively assess where the current weaknesses lie in the current compendium of algorithms. Such data sets will be a valuable resource for computational scientists working in this field to improve on the current state-of-the-art and will better position biological scientists to make an informed methodological choice when analyzing newly generated data.

4.2. Recurrence as an indicator of importance

All of the algorithms we surveyed assume that recurrence or frequency of a CNA in the population has a relationship with its molecular importance, and thus indicates its potential as a candidate driver mutation. This assumption is a strong one that may not always hold. In particular, there may be cases where passenger alterations frequently occur in the same location due to fragile sites in genomically unstable tumours. When inferring recurrent alterations computationally from data alone, it would be difficult to distinguish these events from potential driver CNAs. Careful cross-referencing with the literature related to fragile sites would be required in order to rank the relevance of predictions. In addition, it is imperative to consider computational predictions as 'hypothesis generators' that guide experimental follow-up. In general, computational analyses should be thought of as a mechanism to prioritize functional assays that can determine the role of the recurrent CNAs.

4.3. Clinical and molecular subtypes

The problem of computationally stratifying a patient cohort into molecularly distinct subgroups based on CNA patterns is under-represented in the literature. Van Wieringen et al. (2007) suggest distance metrics for hierarchical clustering of the patients; however, this appears to be the only method developed specifically for aCGH data. It is essential that the issue of heterogeneity be appropriately addressed as the number, characteristics and distribution of molecular subtypes is unknown in many diseases. For some diseases, molecular cytogenetic subgroups have been identified (i.e. for follicular lymphoma, Hoglund et al., 2004), yet some subtypes occur with low frequency in the population (~10%). The CNAs that characterize such subtypes might be ignored if the heterogeneity of the population is not considered in the analysis.

4.4. Copy number variation (CNV) integration

There is potential for naturally occurring structural variations in the form of CNVs (Redon et al., 2006; Wong et al., 2007) to be mistaken for somatic recurrent CNAs. When reporting CNAs in tumor genomes, it is essential to consider CNVs in assessing the clinical relevance of the CNA. This problem could be addressed experimentally using matched normal DNA as the reference; however, in cases where a pooled reference is used, CNVs must be considered. The database of genomic variants (Iafrate et al., 2004) is a useful resource for cross-referencing putative CNVs in post-processing of predicted recurrent CNAs. Numerous large-scale studies have revealed a large collection of non-overlapping sets of CNVs that are stored in this database. Alternatively, these previously reported CNVs could be integrated directly into algorithms for recurrent CNAs, perhaps as Bayesian priors in model-based approaches (for example Shah et al., 2006), in order to distinguish them from somatic alterations. Furthermore, although the algorithms discussed herein are primarily described for the problem of detecting somatic alterations in cancer, many of the algorithms might be applied to the detection of CNVs in population genetics studies based on aCGH data with little or no change.

4.5. Multiple molecular tests and emerging technologies

The extent to which CNAs alter the expression of the genes contained within them is highly variable and com-

plex. Recent work by Lee et al. (2008) proposed an analytical approach for integrating gene expression and aCGH data. The authors found both direct and indirect effects of CNAs on gene expression demonstrating that the relationship between the two data types is more complex than originally proposed. Sophisticated models will need to be developed in order to understand and exploit the relationship that exists between CNA and gene expression data. Authors such as Lipson et al. (2004), Berger et al. (2006), and van Wieringen and van de Wiel (2008) have shown promising results for this problem, but the literature for this topic is far from saturated. Furthermore, other alterations such as epigenetic modifications play a role in altering gene expression in cancer. As data are produced by genome-wide assays for promoter hypermethylation (Esteller, 2007), they will need to be robustly integrated with CNA data and gene expression data in order to discern the combined effects of genomic and epigenomic alterations on gene expression. This will require the development of new analytical methodologies built on top of current CNA algorithms.

Campbell et al. (2008) recently showed it is possible to detect CNAs using emerging massively parallel sequencing technology. This promising approach suggests that digital measurements of copy numbers can be achieved given adequate sequence coverage of the genome. Furthermore, high-throughput paired-end sequencing technology provides a powerful ability to detect potential gene fusions induced by genomic deletions (Tomlins et al., 2005) in addition to a host of other genetic abnormalities. As sequence yields of this transformative technology increase and costs are reduced, studies that apply this technology to cohorts of patients will be possible. The algorithms described in Section 3.1 and 3.2 will need to be adapted to accept this data as input as high-throughput sequencing techniques gain traction in the clinical genomics community.

4.6. Clinical application

Bejjani and Shaffer (2006) reviewed the extent to which aCGH can be applied for clinical diagnostics. Recent activity in the literature suggest that robust classifiers based on patterns of CNAs in a patient cohort that exhibit a given phenotype can be constructed (Rapaport et al., 2008). The approaches outlined herein potentially provide valuable feature selection tools for such algorithms as many aim to reduce dimensionality of the data and output a small set of probes that are representative of a population and may therefore provide more discriminative power. Moreover, we can envision placing this problem in the context of simultaneous feature selection and classification (Krishnapuram et al., 2005). This area is under-explored in aCGH data analysis. The approaches described in this paper may serve as important groundwork for the classification problem as adoption of aCGH as a clinical tool increases. Note that for translational research, experimental validation of predicted CNAs is required and furthermore, biological interpretation of recurrent CNAs is essential to fully understand the pathophysiology of the disease (see for example Guan et al., 2007).

5 Conclusion

We have surveyed and described the current published approaches for detecting recurrent copy number alterations in aCGH data. We suggest qualitatively that algorithms that input raw data are potentially more sensitive to detection of recurrent low-level alterations and the detection of rare, but high-level amplicons over those that input called data. However, quantitative benchmarking is needed in this field to systematically evaluate the strengths and weaknesses of the various approaches in the context of robust accuracy metrics. Future directions for methods to improve on the current state-of-the-art for recurrent CNA detection include CNV integration, consideration of molecular subtypes in the analysis, integration with different data types such as gene expression and methylation, and development of classifiers for clinical use.

Acknowledgements

The author wishes to thank Drs. David Huntsman and Christian Steidl of the British Columbia Cancer Agency for helpful comments in the preparation of this manuscript.

References

Aguirre AJ, Brennan C, Bailey G, Sinha R, Feng B, et al: High-resolution characterization of the pancreatic adenocarcinoma genome. Proc Natl Acad Sci USA 101:9067–9072 (2004).

Andersson R, Bruder CE, Piotrowski A, Menzel U, Nord H, et al: A segmental maximum *a posteriori* approach to genome-wide copy number profiling. Bioinformatics 24:751–758 (2008).

Balmain A, Gray J, Ponder B: The genetics and genomics of cancer. Nat Genet 33:238–244 (2003).

Bejjani BA, Shaffer LG: Application of array-based comparative genomic hybridization to clinical diagnostics. J Mol Diagn 8:528–533 (2006).

Ben-Dor A, Lipson D, Tsalenko A, Reimers M, Baumbusch LO, et al: Framework for identifying common aberrations in DNA copy number data, in Research in Computational Molecular Biology, Volume 4453 of Lecture Notes in Computer Science, pp 122–136 (Springer, Berlin 2007).

Berger JA, Hautaniemi S, Mitra SK, Astola J: Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. IEEE/ACM Trans Comput Biol Bioinform 3:2–16 (2006).

Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, et al: Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci USA 104:20007–20012 (2007).

Broet P, Richardson S: Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. Bioinformatics 22:911–918 (2006).

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, et al: Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40:722–729 (2008).

Chin L, Gray J: Translating insights from the cancer genome into clinical practice. Nature 242:553–563 (2008).
Chin SF, Teschendorff AE, Marioni JC, Wang Y, Barbosa-Morais NL, et al: High resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. Genome Biol 8:R215 (2007).
Coe BP, Lockwood WW, Girard L, Chari R, Macaulay C, et al: Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer. Br J Cancer 94:1927–1935 (2006).
Colella S, Yau C, Taylor JM, Mirza G, Butler H, et al: QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35:2013–2025 (2007).
Collins FS, Barker AD: Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. Sci Am 296: 50–57 (2007).
de Leeuw RJ, Davies JJ, Rosenwald A, Bebb G, Gascoyne RD, et al: Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. Hum Mol Genet 13:1827–1837 (2004).
Diskin SJ, Eck T, Greshock J, Mosse YP, Naylor T, et al: STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. Genome Res 16: 1149–1158 (2006).
Engler DA, Mohapatra G, Louis DN, Betensky RA: A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations (aCGH). Biostatistics 7: 399–421 (2006).
Esteller M: Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet 8:286–298 (2007).
Fridlyand J, Snijders A, Pinkel D, Albertson D, Jain A: Hidden Markov Models approach to the analysis of array CGH data. J Multivariate Anal 90:132–153 (2004).
Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80 (2004).
Guan Y, Kuo WL, Stilwell JL, Takano H, Lapuk AV, et al: Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. Clin Cancer Res 13:5745–5755 (2007).
Guha S, Li Y, Neuberg D: Bayesian hidden Markov modeling of array CGH data. J Am Stat Assoc 103:485–497 (2008).
Hanahan D, Weinberg RA: The hallmarks of cancer. Cell 100:57–70 (2000).
Heidenblad M, Lindgren D, Veltman JA, Jonson T, Mahlamaki EH, et al: Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. Oncogene 24:1794–1801 (2005).
Hirsch FR, Varella-Garcia M, Bunn PA, Di Maria MV, Veve R, et al: Epidermal growth factor receptor in non-small-cell lung carcinomas: correlation between gene copy number and protein expression and impact on prognosis. J Clin Oncol 21:3798–3807 (2003).
Hoglund M, Sehn L, Connors JM, Gascoyne RD, Siebert R, et al: Identification of cytogenetic subgroups and karyotypic pathways of clonal evolution in follicular lymphomas. Genes Chromosomes Cancer 39:195–204 (2004).
Hosoya N, Sanada M, Nannya Y, Nakazaki K, Wang L, et al: Genomewide screening of DNA copy number changes in chronic myelogenous leukemia with the use of high resolution array-based comparative genomic hybridization. Genes Chromosomes Cancer 45:482–494 (2006).
Hupe P, Stransky N, Thiery J, Radvanyi F, Barillot E: Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 20:3413–3422 (2004).
Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al: Detection of large-scale variation in the human genome. Nat Genet 36:949–951 (2004).
Idbaih A, Marie Y, Lucchesi C, Pierron G, Manie E, et al: BAC array CGH distinguishes mutually exclusive alterations that define clinicogenetic subtypes of gliomas. Int J Cancer 122:1778–1786 (2008).
Iehara T, Hosoi H, Akazawa K, Matsumoto Y, Yamamoto K, et al: *MYCN* gene amplification is a powerful prognostic factor even in infantile neuroblastoma detected by mass screening. Br J Cancer 94:1510–1515 (2006).
Ishkanian A, Malloff C, Watson S, DeLeeuw R, Chi B, et al: A tiling resolution DNA microarray with complete coverage of the human genome. Nat Genet 36:299–303 (2004).
Khojasteh M, Lam WL, Ward RK, MacAulay C: A stepwise framework for the normalization of array CGH data. BMC Bioinformatics 6:274 (2005).
Klijn C, Holstege H, de Ridder J, Liu X, Reinders M, et al: Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. Nucleic Acids Res 36:e13 (2008).
Krishnapuram B, Carin L, Figueiredo MA, Hartemink AJ: Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Trans Pattern Anal Mach Intell 27:957–968 (2005).
Lai W, Johnson M, Kucherlapati R, Park P: Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics 21:3763–3770 (2005).
Lee H, Kong SW, Park PJ: Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. Bioinformatics 24:889–896 (2008).
Lipson D, Ben-Dor A, Dehan E, Yakhini Z: Joint analysis of DNA copy numbers and gene expression levels, in WABI, Lecture Notes in Computer Science (LNCS), pp 135 (Springer, Berlin 2004).
Lipson D, Aumann Y, Ben-Dor A, Linial N, Yakhini Z: Efficient calculation of interval scores for DNA copy number data analysis. J Comput Biol 13:215–228 (2006).
Marioni JC, Thorne NP, Tavare S: BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics 22:1144–1146 (2006).
Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, et al: Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol 8:R228 (2007).
Neuvial P, Hupe P, Brito I, Liva S, Manie E, et al: Spatial normalization of array-CGH data. BMC Bioinformatics 7:264 (2006).
Olshen A, Venkatraman E, Lucito R, Wigler M: Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5:557–572 (2004).
Pinkel D, Albertson D: Array comparative genomic hybridization and its applications in cancer. Nat Genet 37 Suppl:11–17 (2005).
Pique-Regi R, Monso-Varona J, Ortega A, Seeger RC, Triche TJ, Asgharzadeh S: Sparse representation and Bayesian detection of genome copy number alterations from microarray data. Bioinformatics 24:309–318 (2008).
Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, et al: Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci USA 99:12963–12968 (2002).
Rapaport F, Barillot E, Vert JP: Classification of array CGH data using a fused SVM. Bioinformatics 24:i375–i382 (2008).
Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al: Global variation in copy number in the human genome. Nature 444:444–454 (2006).
Rouveirol C, Stransky N, Hupe P, Rosa PL, Viara E, et al: Computation of recurrent minimal genomic alterations from array-CGH data. Bioinformatics 22:849–856 (2006).
Rueda OM, Diaz-Uriarte R: Flexible and accurate detection of genomic copy-number changes from aCGH. PLoS Comput Biol 3:e122 (2007).
Schwaenen C, Nessling M, Wessendorf S, Salvi T, Wrobel G, et al: Automated array-based genomic profiling in chronic lymphocytic leukemia: development of a clinical tool and discovery of recurrent genomic alterations. Proc Natl Acad Sci USA 101:1039–1044 (2004).
Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, et al: Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics 22:431–439 (2006).
Shah SP, Lam WL, Ng RT, Murphy KP: Modeling recurrent DNA copy number alterations in array CGH data. Bioinformatics 23:450–458 (2007).
Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL: Human breast cancer: correlation of relapse and survival with amplification of the *HER-2/neu* oncogene. Science 235:177–182 (1987).
Stjernqvist S, Ryden T, Skold M, Staaf J: Continuous-index hidden Markov modeling of array CGH copy number data. Bioinformatics 23: 1006–1014 (2007).
Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, et al: Recurrent fusion of *TMPRSS2* and *ETS* transcription factor genes in prostate cancer. Science 310:644–648 (2005).
van de Wiel MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, Ylstra B: CGHcall: calling aberrations for array CGH tumor profiles. Bioinformatics 23:892–894 (2007).
van Wieringen WN, van de Wiel MA: Nonparametric testing for DNA copy number induced differential mRNA gene expression. Biometrics 64:1–25 (2008).
van Wieringen WN, Van De Wiel MA, Ylstra B: Weighted clustering of called array CGH data. Biostatistics 9:484–500 (2007).
Venkatraman ES, Olshen AB: A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23:657–663 (2007).
Weinberg RA: The Biology of Cancer (Garland Science, Taylor and Francis Group, New York 2007).
Willenbrock H, Fridlyand J: A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics 21:4084–4091 (2005).
Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, et al: A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 80:91–104 (2007).
Yaziji H, Goldstein LC, Barry TS, Werling R, Hwang H, et al: *HER-2* testing in breast cancer using parallel tissue-based methods. JAMA 291: 1972–1977 (2004).