2-28-2007

A HIDDEN MARKOV MODEL FOR JOINT ESTIMATION OF GENOTYPE AND COPY NUMBER IN HIGH-THROUGHPUT SNP CHIPS

Robert B. Scharpf
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, rscharpf@jhsph.edu

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Jonathan Pevnser
Johns Hopkins University School of Medicine, Department of Neuroscience

Ingo Ruczinski
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

A hidden Markov model for joint estimation of genotype and copy number in high-throughput SNP chips

Robert B. Scharpf, Giovanni Parmigiani, Ingo Ruczinski

Abstract

Amplifications and deletions of chromosomal DNA, as well as copy-neutral loss of heterozygosity have been associated with diseases processes. High-throughput single nucleotide polymorphism (SNP) arrays are useful for making genome-wide estimates of copy number and genotype calls. Because neighboring SNPs in high throughput SNP arrays are likely to have dependent copy number and genotype due to the underlying haplotype structure and linkage disequilibrium, hidden Markov models (HMM) may be useful for improving genotype calls and copy number estimates that do not incorporate information from nearby SNPs. We improve previous approaches that utilize a HMM framework for inference in high throughput SNP arrays by integrating copy number, genotype calls, and the corresponding confidence scores when available. Using simulated data, we demonstrate how confidence scores control smoothing in a probabilistic framework. Software for fitting HMMs to SNP array data is available in the R package *ICE*.

1 Introduction

Affymetrix SNP chips were originally described as a high-throughput assay for calling genotypes at thousands of SNPs (Kennedy et al., 2003; Di et al., 2005). Several recent algorithmic improvements to genotyping have been described (Rabbee and Speed, 2006; Affymetrix, 2006; Hua et al., 2007; Carvalho et al., 2006). Recently, SNP chips have been used for simultaneous estimates of copy number variants (CN) and genotype calls (GT) (Zhao et al., 2004). Because both CN and loss of heterozygosity (LOH) may be associated with disease (Dutt and Beroukhim, 2007; Cavenee et al., 1983; Shaw-Smith et al., 2004; Aggarwal et al., 2005; Aguirre et al., 2004), SNP arrays provide a convenient tool for exploring large regions of the genome for chromosomal anomalies. This paper builds on a modular approach for the analysis of Affymetrix SNP chip data (Carvalho et al., 2006). In particular, we view the analysis of SNP chip data as having the following four tiers: (i) appropriate adjustment and pre-processing of probe-level data, (ii) estimation of SNP-level summmaries of probe-level data such as GT and/or CN, (iii) statistical models that borrow strength from neighboring SNPs to infer regions of loss of heterozygosity, amplification, and deletion from SNP-level summaries, and (iv) statistical models for between sample variation to learn about abnormalities associated with disease processes. This paper describes two contributions to level (iii) of this hierarchy: first we propose a HMM that incorporates information from both genotypes and copy number and secondly we describe how confidence estimates of SNP-level summaries can improve inference and control smoothing within a probabilistic framework.

We assume that tier (i) has been addressed by most statistical methods that provide SNP-level summaries of probe-level data. A detailed description of the design of high-throughput Affymetrix SNP chips is described here (Kennedy et al., 2003). We caution that, as with gene expression technologies, pre-processing of probe-level data is an important consideration and several recent papers have described fragment-length and sequence effects that may be introduced by the polymerase chain reaction (PCR) used to amplify the DNA (Nannya et al., 2005; Carvalho et al., 2006). We assume that SNP-level

summaries for each interrogated SNP have been adjusted for probe-specific biases to the extent possible. Statistical models such as BRLMM and CRLMM that use Hapmap data for training have been shown to provide better genotype calls when the centers of the bivariate scatterplots for the A and B allele intensities are less well-defined. Genotype calls for most genotyping algorithms are concordant for over 99.9% of the measured SNPs in the Affymetrix 100k and 500k chips. Statistical methods that provide an indication of the uncertainty of the genotype call, such as the single to noise ratio (SNR) and log likelihood ratio (LLR) defined by CRLMM, can be useful for algorithms in tier (iii). Specifically, statistical models that borrow strength from neighboring SNPs to infer loss or retention of heterozygosity should incorporate the uncertainty of the genotype call, giving less weight to genotype calls that are measured with high uncertainty and more weight to well-estimated genotypes. Figure 1 illustrates how estimates of uncertainty can be helpful for genotype calls.

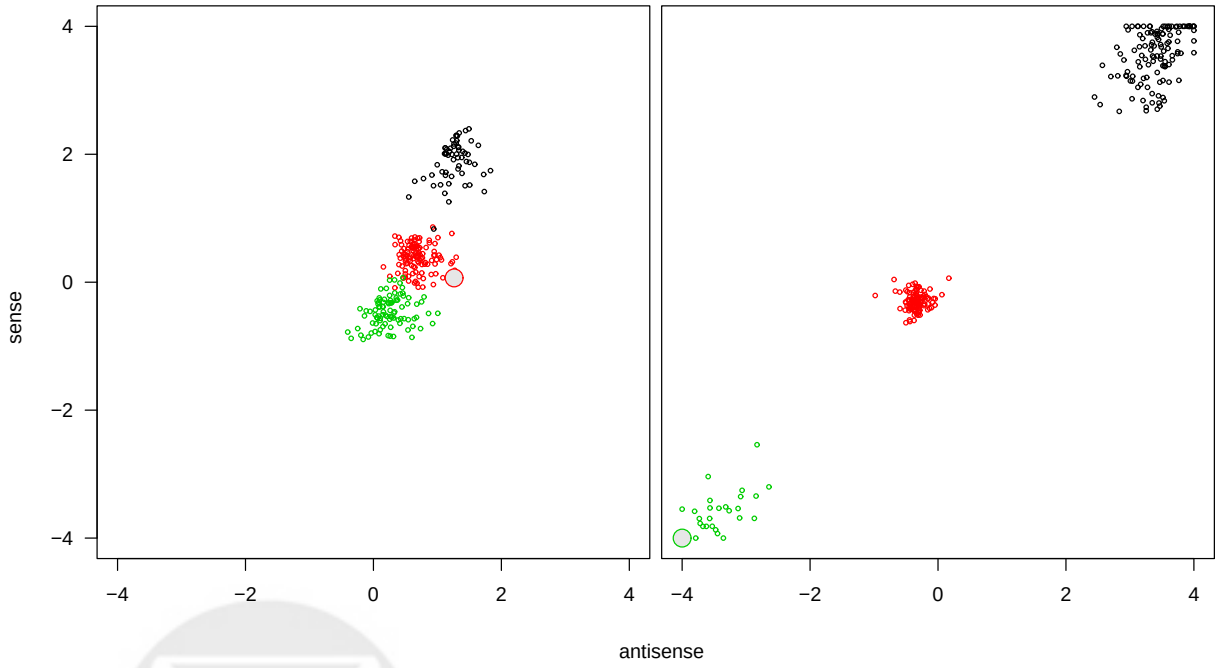

Figure 1: Hapmap genotype calls (the gold standard) for a *bad* SNP (left) and a *good* SNP (right) for 269 samples measured on Affymetrix 100k SNP chips. Genotype calls are indicated by color: AA (green), AB (red), and BB (black). Within any given sample, there is typically a mixture of separation in the genotype clusters. For instance, the large circles with gray background in each panel are the genotype estimates from the same sample. This figure motivates an approach that incorporates uncertainty estimates to control smoothing.

Methods developed for estimation of CN in tier (ii) using SNP chips are even more susceptible to the probe-specific biases described above, as CN can vary markedly. Note also that methodologies for copy number estimation developed for aCGH and ROMA platforms are relevant, particularly if SNP-level summaries of CN are viewed as the data for higher-level models that smooth noisy estimates.

Procedures for CN estimation described for aCGH and ROMA, including circular binary segmentation (Olshen et al., 2004) and hidden Markov models (HMM) for the latent number of chromosomal copies (typically an integer) (Fridlyand et al., 2004; Shah et al., 2006), least squares regression (Huang et al., 2005), and others, are therefore applicable.

SNP chip data differ from array CGH data in two imporant ways: a) SNP chips also provide imformation for the genotype and b) provide a more dense coverage, currently generating genotype information and copy number estimates at locations in excess of 500,000 SNPs. The correlation structure between those estimates has to be an essential part of any statistical modeling approach. The most promising methods currently available are based on hidden Markov models. In particular, to infer LOH regions and to estimate CN changes, the dChip software and methods are the most widely used in the literature for the analysis of SNP chip data. The dChip methods are based on separate HMMs for copy number ((Zhao et al., 2004)) and genotype analysis ((Lin et al., 2004; Beroukhim et al., 2006)). Correlation to CN estimates are considered in (Beroukhim et al., 2006), but not explicitly modeled. Other appoaches for CN estimation with SNP chips have been implemented in CNAG (Nannya et al., 2005), CARAT (Huang et al., 2006), PLASQ (Laframboise et al., 2006), and CNRLMM (Wang et al., 2006). Notably, (Laframboise et al., 2006) and (Wang et al., 2006) provide allele-specific estimates of CN. To our knowledge, no statistical approach has been reported yet that allows for simultaneous analysis of LOH and CN changes. In addition, algorithms in tier (iii) that integrate information from surrounding SNPs as well as confidence estimates of CN can augment such approaches.

We develop an HMM for the paired sequence of CN and GT observations, placing simultaneous inference on LOH and CN within the probabilistic framework of the HMM. Additionally, we illustrate how integrating confidence scores of the SNP-level summaries in the HMM can further improve inference for the underlying hidden states using simulated data. These ideas are implemented in the R package *ICE*.

2 Results

SNP-level summaries were obtained using a combination of real and simulated data for 9200 SNPs measured on chromosome 1 of an Affymetrix 100k SNP chip. Because the hidden states of the Markov model are determined by whether $\widehat{\text{GT}}$ (example 1), $\widehat{\text{CN}}$ (example 2), or both (example 3) are available, we organize the results accordingly. See Section 4.5 for a description of the examples.

The simulated $\widehat{\text{GT}}$ in example 1 are plotted in the top panel of Figure 2. Features A - C are simulated as described in Section 4.5. Features D and E represent changes in chromosomal copy number that can not be detected without estimates of copy number. We fit two HMM models to the data as described in Section 4. The predicted states from the vanilla and ICE HMMs are similar, but the vanilla HMM smooths over the two heterozygous SNPs in A. ICE only smooths over the two heterozygous SNPs in B where the uncertainty of the $\widehat{\text{GT}}$ is greater.

Figure 3 (top) plots $\widehat{\text{CN}}$ in example 2 with a simulated amplification (D) and deletions (B and C). In panel 2 of Figure 3 we plot the predicted states from a vanilla HMM that does not use confidence scores of the $\widehat{\text{CN}}$. Panel 3 plots the predicted states from ICE. The vanilla HMM smooths over the two SNPs with $\widehat{\text{CN}}$ near 2 in regions D and A, calling the entire region an amplification ($\nearrow$) or deletion ($\searrow$), respectively. By contrast, ICE transitions from $\searrow$ to $\rightarrow$ and from $\nearrow$ to $\rightarrow$ only when the confidence scores of the $\widehat{\text{CN}}$ estimates are high. When the confidences scores for the $\widehat{\text{CN}}$ are low, ICE uses the information from neighboring SNPs and smooths over these regions providing predictions that are comparable to vanilla. Additionally, ICE detects the microamplification (5 SNPs) and microdeletion (3 SNPs) in region E that each had high confidence scores.

The data plotted in Figure 4 (top) was obtained by superimposing the $\widehat{\text{GT}}$ in example 1 on the

$\widehat{CN}$ in example 2. Note that by considering both GT and CN simultaneously, both HMMs correctly distinguish deletion (Del) from copy-neutral LOH. In addition, ICE detects small aberrations in E that are well-estimated, and smooths noisy estimates by borrowing information from neighboring SNPs.

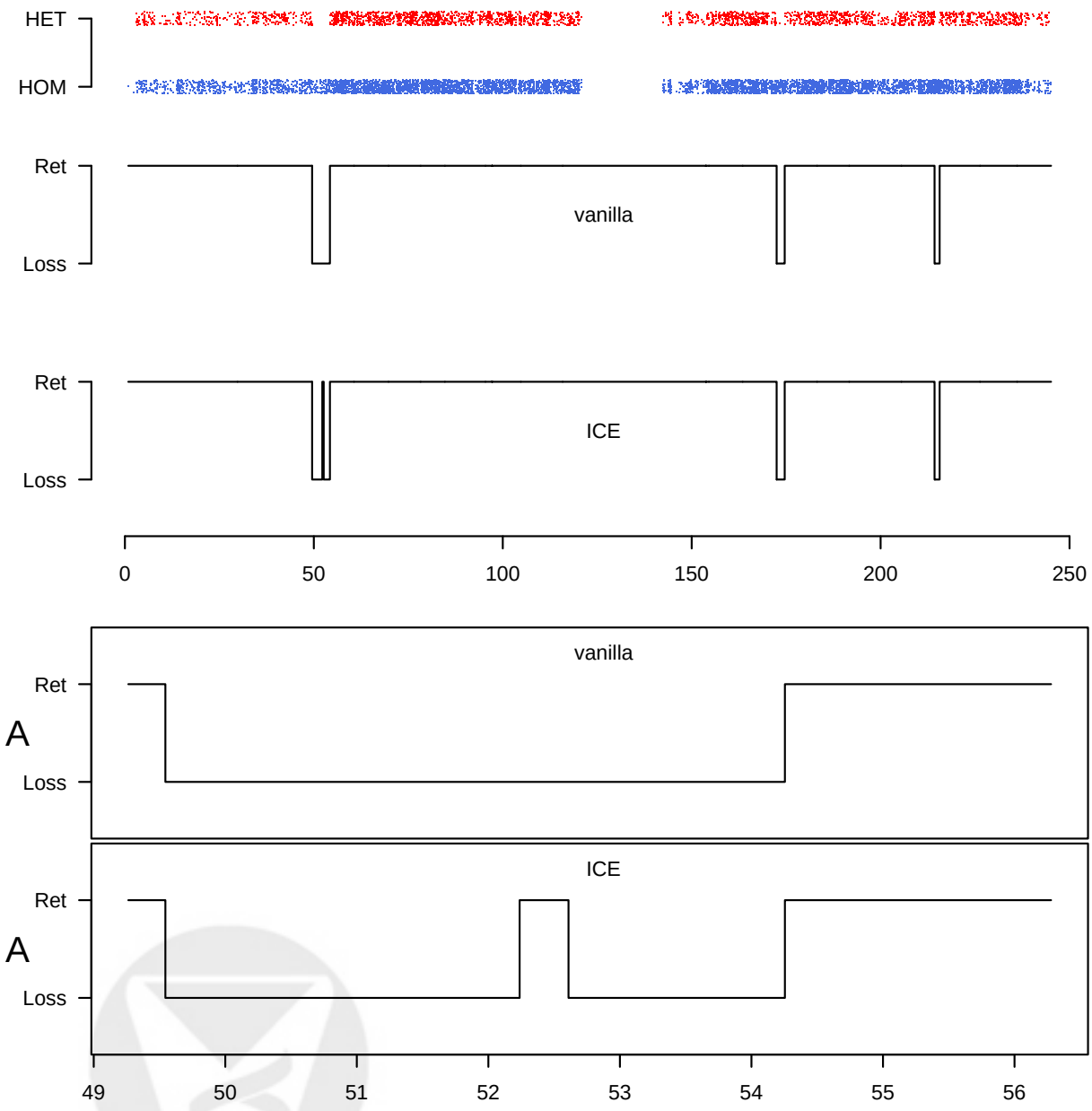

Figure 2: Top: simulated $\widehat{\text{GT}}$ for chromosome 1 with four *Loss* features. Region A contains two *Loss* features separated by a short chromosomal segment in state *Rentention* (Ret) containing two heterozygous genotype calls. Region B is a single *Loss* region with two false calls (two SNPs called heterozygous). A spans 200 SNPs, whereas B and C both span 100 SNPs. We assigned a high probability of a correct classification for the heterozygous calls in A and a lower probability of correct classification for the heterozygous calls in B. Panels 2 and 3 plot the predicted underlying states obtained from the vanilla and ICE HMMs, respectively.

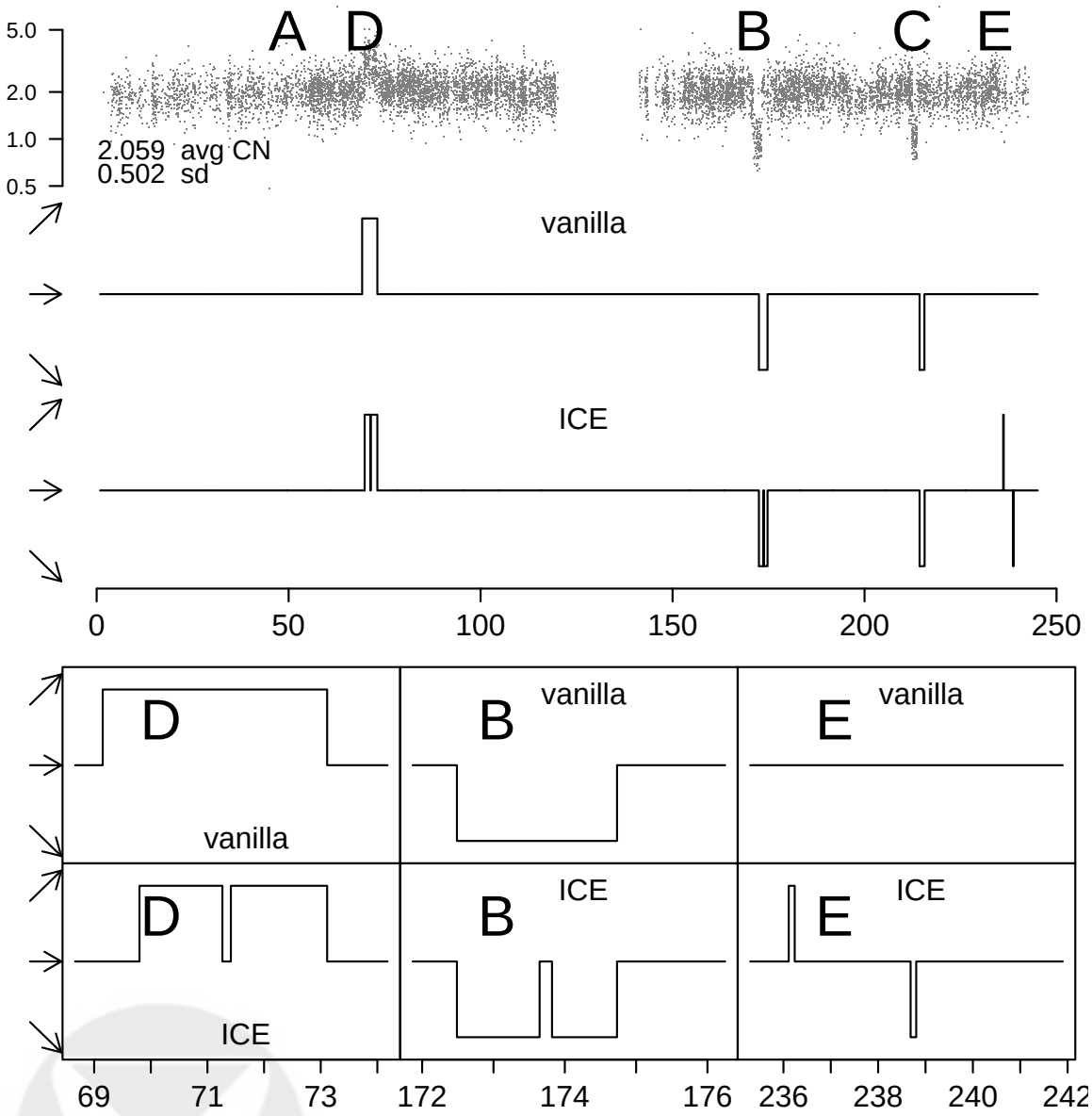

Figure 3: Top: we used $\widehat{\text{CN}}$ from the Affymetrix CNAT tool for chromosome 1 of a normal Hapmap sample. In features D (200 SNPs), B (100 SNPs), and C (100 SNPs), we replace the CNAT estimates with simulated data to mimic an amplification and two deletions, respectively. For three SNP pairs in the center of D, B and C, we assigned a $\widehat{\text{CN}}$ of ≈ 2. For regions D and B, the confidence score was high (low standard error) and for region C the confidence score was low (high standard error). Standard errors for the remaining $\widehat{\text{CN}}$ were simulated from a Gamma distribution as described in Section 4. Rows 2 and 3 plot the predictions for the hidden states deletion ($\searrow$), normal ($\rightarrow$), and amplification ($\nearrow$) for the vanilla and ICE HMMs. Rows 5 and 6 magnify the predictions for D and B where these algorithms differ.

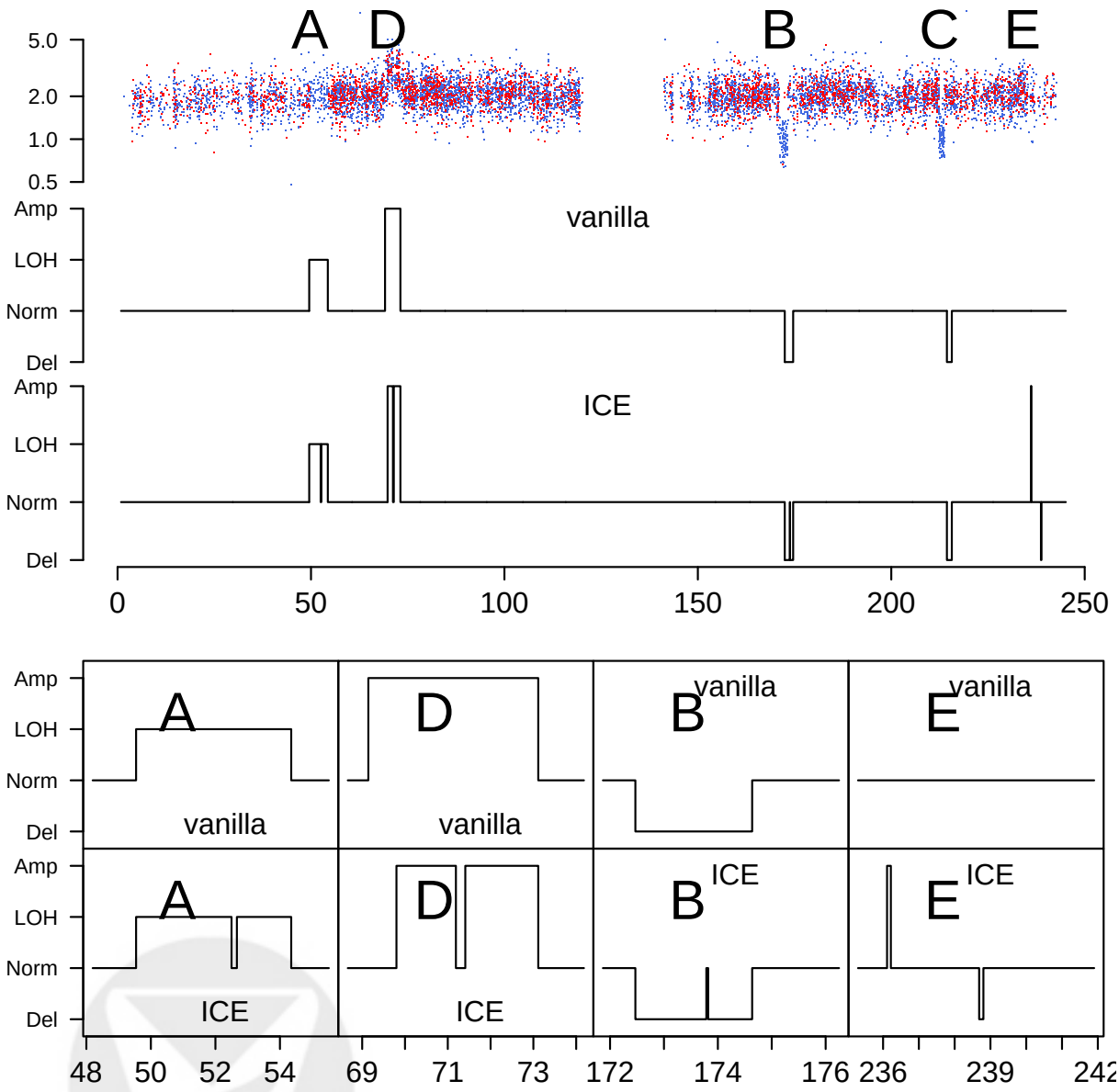

Figure 4: We superimpose the $\widehat{CN}$ in Figure 3 onto the $\widehat{GT}$ in Figure 2. We fit HMMs to the joint observation sequence of $\widehat{CN}$ and $\widehat{GT}$ without (vanilla) and with (ICE) confidence scores of the SNP-level summaries. Simulated in this figure are the genotype calls, copy number, and confidence scores for the features in A-E. For the remaining SNPs, we used a combination of real (CNAT estimates of CN) and simulated (genotype calls and confidence scores) data. The predicted states are deletion (Del), normal (Norm), LOH, and amplification (Amp).

3 Discussion

Our paper builds on a modular approach for analysing SNP chip data, extending the functionality of statistical algorithms that pre-process probe-level data to produce SNP-level summaries of GT and CN. As SNP-level summaries of GT and CN are often correlated, HMMs that use information from neighboring SNPs have the potential to improve noisy estimates. Previous approaches using HMMs have considered GT and CN separately, but not simultaneously in a single unifying statistical model. We develop a HMM model for SNP chips that use the joint observation sequence of $\widehat{\text{CN}}$ and GT as input. Using simulated data, we demonstrate that a HMM model that uses both $\widehat{\text{CN}}$ and $\widehat{\text{GT}}$ can distinguish copy-neutral LOH from deletion-induced LOH. Additionally, the joint HMM should improve the ability to identify small deletions, where both LOH and reduced $\widehat{\text{CN}}$ are present. Finally, we demonstrate how pre-processing algorithms that provide confidence scores of SNP-level summaries can be integrated into the emission probabilities of the HMM to control smoothing in a probabilistic framework. In particular, confidence estimates may help identify very small, well-estimated features that may otherwise be missed.

Our results are a proof of principle that HMMs can (i) effectively incorporate multiple sources of SNP-level summaries and (ii) that confidence estimates of SNP-level summaries may help identify small regions of gain in copy number, deletion, or LOH. We used CRLMM estimates of confidence scores for $\widehat{\text{GT}}$ in our HMM, but confidence scores from other algorithms could easily be incorporated.

The tools we provide are helfpul for identifying chromosomal alterations, including deletions and long stretches of homozygosity. We do not currently prescribe a method for ranking alterations detected by the HMM. In part, this reflects our view that LOH and copy number variation are prevalent in many phenotypically normal individuals. Identifying features that may be associated with a phenotype are better handled by statistical models for between-sample variation in studies with phenotypically normal and diseased populations. Such models reside in the next tier of our modular approach to the analysis of SNP chip data and are an extension of the ideas presented here. Finally, while our analysis focuses on methods for Affymetrix SNP chip data, these ideas are portable to other high throughput platforms such as Illumina.

4 Methods

4.1 Hidden Markov Models

The hidden states are governed by the available data, which may depend on the pre-processing software. The hidden states for copy number HMM are hemizygous or homozygous deletion (↘), normal (→), and amplification (↗). For GT, our interest lies in whether a SNP is in a neighborhood where heterozygosity is retained (◐) or not (○). Specifically, we define the state ○ as an unusually long sequence of homozygous calls. The HMM for the joint GT and CN observation sequence has the following four hidden states: Deletion (⊘), Normal (◒), copy-neutral LOH (⊖), and Amplification (◪). Again, we define state ⊖as an unsually long sequence of homozygous genotype calls with no evidence of decreased $\widehat{\text{CN}}$. Note that whether the absence of heterozygosity in a region is induced by a deletion or a copy-neutral mechanism is distinguished by the hidden states ⊘and ⊖, respectively. For each of the three examples (see Section 4.5), we fit two HMMs: a *vanilla* HMM that uses only the SNP-summary estimates and allows options for updating parameters by EM and a HMM that integrates confidence estimates (*ICE*) of GT and/or CN.

4.2 Genotype calls

The observation sequence for the genotype HMM is a string of homozygous (AA or BB) and heterozygous (AB) genotypes. The vanilla HMM for the hidden states Retention ($\CIRCLEARROW$) and Loss ($\CIRCLE$) (of heterozygosity) require specification of the initial state probability distribution, the emission probabilities, and the transition probability between the true states. Commonly employed in the literaure for the latter is the "instability-selection" model for LOH analysis (Newton et al., 1998; Beroukhim et al., 2006) that describes the dependencies between the underlying states of adjacent SNPs as a function of distance. The model assumption is as follows. For any two adjacent SNPs, let θ be the probability that the state of the first marker is not informative (denoted by I^c) for the state of the second marker. As the distance between SNPs affects this probability, it is modeled as $\theta(d) = (1 - \exp(-2d))$, where d is the physical distance (usually in 100Mb units) between the SNPs. We assume that with probability $1-\theta(d)$, $\text{SNP}_{(i)}$ is informative (denoted by I) for $\text{SNP}_{(i+1)}$ and that no change in state occurs between the adjacent SNPs. For example, this leads to

$$\begin{aligned}
\tau_{\CIRCLE/\CIRCLE}(d) &= P(\CIRCLE_{i+1}|\CIRCLE_i, d) \\
&= P(\CIRCLE_{i+1}, I|\CIRCLE_i, d) + P(\CIRCLE_{i+1}, I^c|\CIRCLE_i, d) \\
&= P(\CIRCLE_{i+1}|I, \CIRCLE_i, d) \times P(I|\CIRCLE_i, d) + P(\CIRCLE_{i+1}|I^c, \CIRCLE_i, d) \times P(I^c|\CIRCLE_i, d) \\
&= 1 - \theta(d) + P(\CIRCLE) \times \theta(d) \qquad (4.1)
\end{aligned}$$

as the probability that the state of $\text{SNP}_{(i+1)}$ is *Loss*, given that the state of $\text{SNP}_{(i)}$ with distance d was *Loss*. It follows that

$$\tau_{\CIRCLEARROW/\CIRCLE}(d) = P(\CIRCLEARROW_{i+1}|\CIRCLE_i, d) = 1 - P(\CIRCLE_{i+1}|\CIRCLE_i, d) = \theta(d) \times P(\CIRCLEARROW). \qquad (4.2)$$

$P(\CIRCLEARROW)$ and $P(\CIRCLE)$ refer to the initial probabilities for *Retention* and *Loss*, respectively. These initial probabilities can be set as fixed constants using knowledge from previous experiments, or alternatively, learned via the EM algorithm (Dempster et al., 1977). Emission probabilities for states $\CIRCLE$ and $\CIRCLEARROW$ are estimated as

$$\beta_{\CIRCLE}(\widehat{\text{GT}}) \sim \text{Binomial}(p = 0.99) \text{ and } \beta_{\CIRCLEARROW}(\widehat{\text{GT}}) \sim \text{Binomial}(p = 0.7), \text{ where}$$

p is the probability of a homozygous genotype call. Efficient computation of the probability of the observation sequence given the model is computed using the forward algorithm as described in (Rabiner, 1989). The most probable state sequence given the model is calculated via the Viterbi algorithm (Viterbi, 1967; Rabiner, 1989).

Integrating confidence estimates When confidence estimates are available, we observe a pair of observations: the genotype call $\widehat{\text{GT}}$ and the uncertainty measure (i.e. score) $\text{S}_{\widehat{\text{GT}}}$. The joint distribution of $\widehat{\text{GT}}$ and $\text{S}_{\widehat{\text{GT}}}$ depends on the underlying state. For example, if the state for a particular SNP is *Loss*, the emission probability is

$$\beta_{\CIRCLE}\left\{ \widehat{\text{GT}}, \text{S}_{\widehat{\text{GT}}} \right\} = f\left\{ \widehat{\text{GT}} \,|\CIRCLE \right\} \times f\left\{ \text{S}_{\widehat{\text{GT}}} \mid \widehat{\text{GT}}, \CIRCLE \right\}. \qquad (4.3)$$

Note that the first of the two terms on the right hand side of equation (4.3) is simply the emission probability when estimates of uncertainty are not available. The second term can be understood as a weight for the former term that depends on the confidence with which the call is made. The second

term can be approximated using a density estimate of the $\mathsf{S}_{\widehat{\mathsf{GT}}}$ where the gold standard is available. From the 269 HapMap samples, we know the distributions of the uncertainty measures for all four possible combinations of called (HOM or HET) and true (HOM or HET) genotypes measured on the Affymetrix 100k SNP chips. We use kernel based density estimates to obtain

$$f\left\{\mathsf{S}_{\widehat{\mathsf{HOM}}} \mid \widehat{\mathsf{HOM}}, \mathsf{HOM}\right\},\ f\left\{\mathsf{S}_{\widehat{\mathsf{HOM}}} \mid \widehat{\mathsf{HOM}}, \mathsf{HET}\right\},\ f\left\{\mathsf{S}_{\widehat{\mathsf{HET}}} \mid \widehat{\mathsf{HET}}, \mathsf{HOM}\right\},\ f\left\{\mathsf{S}_{\widehat{\mathsf{HET}}} \mid \widehat{\mathsf{HET}}, \mathsf{HET}\right\}. \quad (4.4)$$

The first term in (4.4), for example, denotes the density of the scores when the genotype is correctly called homozyous, i.e. the called genotype is homozygous ($\widehat{\mathsf{HOM}}$), and the true genotype is homozygous (HOM). If the underlying state is *Loss*, then the true genotype is always HOM and we can use kernal-based estimates of the above densities to estimate the emission probabilities, assuming only that

$$f\left\{\mathsf{S}_{\widehat{\mathsf{HOM}}} \mid \widehat{\mathsf{HOM}}, \bigcirc\right\} \approx f\left\{\mathsf{S}_{\widehat{\mathsf{HOM}}} \mid \widehat{\mathsf{HOM}}, \mathsf{HOM}\right\} \quad \text{and} \quad f\left\{\mathsf{S}_{\widehat{\mathsf{HET}}} \mid \widehat{\mathsf{HET}}, \bigcirc\right\} \approx f\left\{\mathsf{S}_{\widehat{\mathsf{HET}}} \mid \widehat{\mathsf{HET}}, \mathsf{HOM}\right\}. \quad (4.5)$$

If the underlying state is *Retention* (◐), then the true genotype can be HET or HOM. We therefore estimate the emission probabilities for state ◐ as

$$\begin{aligned}
& \beta_{\text{◐}}\left\{\widehat{\mathsf{GT}}, \mathsf{S}_{\widehat{\mathsf{GT}}}\right\} \\
= \ & f\left\{\widehat{\mathsf{GT}} \mid \text{◐}\right\} f\left\{\mathsf{S}_{\widehat{\mathsf{GT}}} \mid \widehat{\mathsf{GT}}, \text{◐}\right\} \\
= \ & f\left\{\widehat{\mathsf{GT}} \mid \text{◐}\right\}\left(f\left\{\mathsf{S}_{\widehat{\mathsf{GT}}}, \mathsf{HOM} \mid \widehat{\mathsf{GT}}, \text{◐}\right\} + f\left\{\mathsf{S}_{\widehat{\mathsf{GT}}}, \mathsf{HET} \mid \widehat{\mathsf{GT}}, \text{◐}\right\}\right) \\
= \ & f\left\{\widehat{\mathsf{GT}} \mid \text{◐}\right\}\left(f\left\{\mathsf{S}_{\widehat{\mathsf{GT}}} \mid \mathsf{HOM}, \widehat{\mathsf{GT}}, \text{◐}\right\} f\left\{\mathsf{HOM} \mid \widehat{\mathsf{GT}}, \text{◐}\right\} + f\left\{\mathsf{S}_{\widehat{\mathsf{GT}}} \mid \mathsf{HET}, \widehat{\mathsf{GT}}, \text{◐}\right\} f\left\{\mathsf{HET} \mid \widehat{\mathsf{GT}}, \text{◐}\right\}\right) \\
= \ & f\left\{\widehat{\mathsf{GT}} \mid \text{◐}\right\}\left(f\left\{\mathsf{S}_{\widehat{\mathsf{GT}}} \mid \mathsf{HOM}, \widehat{\mathsf{GT}}\right\} f\left\{\mathsf{HOM} \mid \widehat{\mathsf{GT}}, \text{◐}\right\} + f\left\{\mathsf{S}_{\widehat{\mathsf{GT}}} \mid \mathsf{HET}, \widehat{\mathsf{GT}}\right\} f\left\{\mathsf{HET} \mid \widehat{\mathsf{GT}}, \text{◐}\right\}\right). \quad (4.6)
\end{aligned}$$

The unknown terms in Equation 4.6, $f\left\{\mathsf{HOM} \mid \widehat{\mathsf{GT}}, \text{◐}\right\}$ and $f\left\{\mathsf{HET} \mid \widehat{\mathsf{GT}}, \text{◐}\right\}$, are also estimated from the HapMap samples.

4.3 Copy number.

The hidden states for autosomal CN are hemi- or homozygous deletion (↘), two copies (→), and more than two copies (↗). A typical, and from practical experience, a reasonable model assumption when only copy number is considered (applied to aCGH and SNP chip data) is that the logarithm of the copy number estimate, after normalization, is roughly normally distributed around the true log copy number (see for example (Zhao et al., 2004)). However, it is assumed in general that the variability is constant across SNPs, which is not necessarily the case. If the variance is assumed to be constant, this parameter can be learned via the EM algorithm (Dempster et al., 1977), or estimated for example in a robust manner using quantiles from the observed data. In the examples presented here, we obtained a robust estimate for the standard deviation of the $\widehat{\mathsf{CN}}$ by averaging the 16th and 84th percentiles of the $\log_2$ transformed CN. For state $\mathcal{S}$, the mean $\mu_{\mathcal{S}}$ and variance $\sigma^2_{\mathcal{S}}$ of the Gaussians used to estimate the emission probabilities can be fixed at starting values or updated by EM. Here, we assumed a constant σ^2 and estimated the emission probabilities for state ↘, for instance, as

$$\begin{aligned}
\beta_{\searrow}(\widehat{\mathsf{CN}}) &\equiv f(\widehat{\mathsf{CN}}|\searrow) \\
&= N\left(\log_2(\widehat{\mathsf{CN}})|\mu_{\mathcal{S}} = 0, \sigma^2\right). \quad (4.7)
\end{aligned}$$

A distance-weighted transition probability for the copy number HMM is as described above for GT.

Integrating confidence estimates Standard errors for the $\widehat{\mathsf{CN}}$, $\mathsf{S}_{\widehat{\mathsf{CN}}}$, were simulated from a shifted Gamma: $\Gamma(2,2) + 0.4$. The emission probabilities for the HMM retains the same location parameters for the Gaussian, but with SNP-specific standard errors for the $\widehat{\mathsf{CN}}$. For instance, in the normal state the emission probability at a particular SNP (omitting subscripts for SNPs) is

$$\beta_{\rightarrow}\left\{\widehat{\mathsf{CN}}, \mathsf{S}_{\widehat{\mathsf{CN}}}\right\} \sim N\left(1, (\sigma \times \mathsf{S}_{\widehat{\mathsf{CN}}})^2\right). \tag{4.8}$$

The scalar σ can be estimated from the sample at hand, or set equal to one if $\mathsf{S}_{\widehat{\mathsf{CN}}}$ measures the actual variability of the copy number estimate around the true copy number.

4.4 Copy number and genotype

For the joint analysis of CN and GT, we again extend the transition probabilities in Equations 4.1 and 4.2 to the hidden states ◒, ◪, ⊖, and ⊘.

For the emission probabilities, we assume conditional independence between the copy number estimates and the genotype calls:

$$f(\widehat{\mathsf{CN}}, \widehat{\mathsf{GT}}|\mathcal{S}) \quad = \quad f(\widehat{\mathsf{CN}}|\mathcal{S}) \times f(\widehat{\mathsf{GT}}|\mathcal{S}).$$

This equation can be further simplified, as the copy number distribution only depends on the true copy number, and the genotype distribution only depends on the true underlying state being *Retention* or *Loss*. For example,

$$f\left\{\ \widehat{\mathsf{CN}}, \widehat{\mathsf{GT}} \mid \oslash\ \right\} = f\left\{\ \widehat{\mathsf{CN}} \mid \oslash\ \right\} \times f\left\{\ \widehat{\mathsf{GT}} \mid \oslash\ \right\} = f\left\{\ \widehat{\mathsf{CN}} \mid \searrow\ \right\} \times f\left\{\ \widehat{\mathsf{GT}} \mid \bigcirc\ \right\}. \tag{4.9}$$

The terms in Equation 4.9 can be estimated as described above for GT and CN. Emission probabilities for states Deletion ⊘, Amplification ◪ and LOH ⊖ can be obtained similarly.

4.5 Examples

The 100k Affymetrix SNP chip data for the samples used in this analysis are publicly available (`www.affymetrix.com`). The three datasets discussed in Section 2 were generated as follows.

Example 1 We simulated 9200 genotypes (comparable to the number of SNPs in the 100k SNP chip) from a binomial distribution with probability 0.7 of a homozygous genotype. For a region spanning 200 SNPs and 2 regions each spanning 100 SNPs, we inserted homozygous calls. We refer to these regions as A, B, and C. We replaced two SNPs in the center of A and B with heterozygous calls. In truth region A is comprised of two ○ features separated by a chromosomal segment in state ◐. By contrast, region B is a single LOH region with 2 falsely called heterozygotes. To add confidence scores to the simulated GC we took random draws from the empirical distributions of the probability that the $\widehat{\mathsf{GT}}$ was correct given the true genotype obtained from 269 Hapmap samples. In particular, confidence scores for heterozygous calls were sampled from the distribution of probabilities when HET was the correct call, and confidence scores for homozygous calls were sampled from the distribution of probabilities when HOM was the correct call. The false calls in B were assigned low confidence scores, whereas the true HET calls in A are assigned a high confidence score.

Example 2 The Affymetrix CNAT tool was used to obtain $\widehat{\text{CN}}$ for chromosome 1 of a sample in the CEPH trios dataset. Features B, C, and D in Figure 3 were simulated as described in Section 4. For D, two amplified segments are separated by a chromosomal segment with $\widehat{\text{CN}} \approx 2$ that contains two SNPs. For B, two deletions are separated by a chromosomal segment with $\widehat{\text{CN}} \approx 2$ that contains two SNPs. The two SNPs in B and D with $\widehat{\text{CN}} \approx 2$ both have high confidence scores (low standard errros). A microdeletion of 3 SNPs and a microamplification of 5 SNPs were simulated in the distal q-arm. The $\widehat{\text{CN}}$ for the microdeletions and amplification were assigned high confidence scores. Confidence scores for the remaining SNPs were simulated from a Gamma distribution as described above.

Example 3 For example 3, we superimpose the $\widehat{\text{GT}}$ in Example 1 onto the $\widehat{\text{CN}}$ in Example 2.

5 Acknowledgments

RBS was supported by grant DMS034211 from the National Science Foundation (P. I. Giovanni Parmigiani).

References

Affymetrix (2006) Brlmm: an improved genotype calling method for the genechip human mapping 500k array set. Tech. rep., Affymetrix, Inc. White Paper

Aggarwal A, Leong SH, Lee C, Kon OL, Tan P (2005) Wavelet transformations of tumor expression profiles reveals a pervasive genome-wide imprinting of aneuploidy on the cancer transcriptome. Cancer Res 65(1):186–94

Aguirre AJ, Brennan C, Bailey G, Sinha R, Feng B, Leo C, Zhang Y, Zhang J, Gans JD, Bardeesy N, Cauwels C, Cordon-Cardo C, Redston MS, DePinho RA, Chin L (2004) High-resolution characterization of the pancreatic adenocarcinoma genome. Proc Natl Acad Sci U S A 101(24):9067–72
URL http://dx.doi.org/10.1073/pnas.0402932101

Beroukhim R, Lin M, Park Y, Hao K, Zhao X, Garraway LA, Fox EA, Hochberg EP, Mellinghoff IK, Hofer MD, Descazeaud A, Rubin MA, Meyerson M, Wong WH, Sellers WR, Li C (2006) Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. PLoS Comput Biol 2(5):e41. 2
URL http://dx.doi.org/10.1371/journal.pcbi.0020041

Carvalho B, Speed TP, Irizarry RA (2006) Exploration, normalization, and genotype calls of high density oligonucleotide SNP array data. Tech. rep., Johns Hopkins University

Cavenee WK, Dryja TP, Phillips RA, Benedict WF, Godbout R, Gallie BL, Murphree AL, Strong LC, White RL (1983) Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. Nature 305(5937):779–784

Dempster A, Laird D, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39:1–38

Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, mei Shen M, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S (2005) Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. Bioinformatics

21(9):1958–1963
URL http://dx.doi.org/10.1093/bioinformatics/bti275

Dutt A, Beroukhim R (2007) Single nucleotide polymorphism array analysis of cancer. Curr Opin Oncol 19(1):43–49
URL http://dx.doi.org/10.1097/CCO.0b013e328011a8c1

Fridlyand J, Snijders A, Pinkel D, Albertson D, Jain A (2004) Hidden Markov models approach to the analysis of array CGH data. Journal of Multivariate Analysis 90:132–153

Hua J, Craig DW, Brun M, Webster J, Zismann V, Tembe W, Joshipura K, Huentelman MJ, Dougherty ER, Stephan DA (2007) SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. Bioinformatics 23(1):57–63
URL http://dx.doi.org/10.1093/bioinformatics/btl536

Huang J, Wei W, Chen J, Zhang J, Liu G, Di X, Mei R, Ishikawa S, Aburatani H, Jones KW, Shapero MH (2006) CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. BMC Bioinformatics 7:83
URL http://dx.doi.org/10.1186/1471-2105-7-83

Huang T, Wu B, Lizardi P, Zhao H (2005) Detection of DNA copy number alterations using penalized least squares regression. Bioinformatics 21(20):3811–7
URL http://dx.doi.org/10.1093/bioinformatics/bti646

Kennedy GC, Matsuzaki H, Dong S, min Liu W, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips MS, Boyce-Jacino MT, Fodor SPA, Jones KW (2003) Large-scale genotyping of complex DNA. Nat Biotechnol 21(10):1233–1237
URL http://dx.doi.org/10.1038/nbt869

Laframboise T, Harrington D, Weir BA (2006) PLASQ: A Generalized Linear Model-Based Procedure to Determine Allelic Dosage in Cancer Cells from SNP Array Data. Biostatistics
URL http://dx.doi.org/10.1093/biostatistics/kxl012

Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH, Li C (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 20(8):1233–40
URL http://dx.doi.org/10.1093/bioinformatics/bth069

Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, Kurokawa M, Chiba S, Bailey DK, Kennedy GC, Ogawa S (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 65(14):6071–6079
URL http://dx.doi.org/10.1158/0008-5472.CAN-05-0465

Newton MA, Gould MN, Reznikoff CA, Haag JD (1998) On the statistical analysis of allelic-loss data. Stat Med 17(13):1425–1445

Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5(4):557–72
URL http://dx.doi.org/10.1093/biostatistics/kxh008

Rabbee N, Speed TP (2006) A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22(1):7–12
URL http://dx.doi.org/10.1093/bioinformatics/bti741

Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77:257–286. 1

Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, Ng R, Murphy KP (2006) Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics 22(14):e431–e439
URL http://dx.doi.org/10.1093/bioinformatics/btl238

Shaw-Smith C, Redon R, Rickman L, Rio M, Willatt L, Fiegler H, Firth H, Sanlaville D, Winter R, Colleaux L, Bobrow M, Carter NP (2004) Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. J Med Genet 41(4):241–248

Viterbi A (1967) Error bounds for convolution codes and an asymptotically optimal decoding algorithm. IEEE Transactions on Information Theory 13(2):260–269

Wang W, Carvalho B, Miller N, Pevsner J, Chakravarti A, Irizarry RA (2006) Estimating genome-wide copy number using allele specific mixture models. Tech. rep., Johns Hopkins University

Zhao X, Li C, Paez JG, Chin K, Jnne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C, Gray JW, Sellers WR, Meyerson M (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 64(9):3060–71