A Semiparametric Bayesian Model for Repeatedly Repeated Binary Outcomes

Fernando A. Quintana

Departamento de Estadística, Facultad de Matemáticas,

Pontificia Universidad Católica de Chile, Casilla 306, Correo 22, Santiago, CHILE.

Peter Müller and Gary L. Rosner

Department of Biostatistics, The University of Texas, M. D. Anderson Cancer Center,

Box 447, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.

Mary V. Relling

Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital,

332 N. Lauderdale, Memphis, TN 38105-2794 USA

Department of Pharmaceutical Research

St. Jude Children's Research Hospital

Memphis, TN 38105

Summary. We discuss the analysis of data from single nucleotide polymorphism (SNP) arrays comparing tumor and normal tissues. The data consist of sequences of indicators for loss of heterozygosity (LOH) and involve three nested levels of repetition: chromosomes for a given patient, regions within chromosomes, and SNPs nested within regions. We propose to analyze these data using a semiparametric model for multi-level repeated binary data. At the top level of the hierarchy we assume a sampling model for the observed binary LOH sequences that arises from a partial exchangeability argument. This implies a mixture of Markov chains model. The mixture is defined with respect to the Markov transition probabilities. We assume a nonparametric prior for the random mixing measure. The resulting model takes the form of a semiparametric random effects model with the matrix of transition probabilities being the random effects. The model includes appropriate dependence assumptions for the two remaining levels of the hierarchy, i.e., for regions within chromosomes and for chromosomes within patient.

1. Introduction

In many biomedical studies, investigators collect data on a number of repeat experiments for a given set of patients or subjects, often involving multiple levels of repetition. We specifically

consider the case where a sequence of binary responses is collected from each experiment. The discussion is motivated by inference about regions of increased loss of heterozygosity (LOH) from single nucleotide polymorphism (SNP) arrays comparing tumor and normal tissue samples from a group of children who experience treatment-related leukemia. The data include three nested levels of repetition: chromosomes within a patient, regions within a chromosome, and SNPs within regions. We denote the number of patients by n and the collected data as $y_{icjk} = 1$ if LOH was recorded (and zero otherwise) in the kth SNP from region j within chromosome c for patient i, where $i = 1, \ldots, n$; $c = 1, \ldots, 22$; $j = 1, \ldots, n_{ic}$; and $k = 1, \ldots, n_{icj}$. In short, the data consist of binary sequences nested within multiple levels of repeat experiments recorded for each individual. We generally describe this data structure as "repeatedly repeated binary measurements."

Newton et al. (1998) and Newton and Lee (2000) proposed the "instability-selection" model for the analysis of LOH data. The model assumes that the observed losses (deletion of genetic material) occur in chromosomes according to a single binary Markov process. An extension of the instability-selection model to pooled analysis of LOH from several experiments is described in Miller et al. (2003). Lin et al. (2004) consider permutation-based methods with windows and hidden Markov models to assess LOH. A related model can be found in Beroukhim et al. (2006).

Our modeling approach is based on similar assumptions. We deviate from these approaches in that we consider Markov processes that are specific to regions within chromosomes rather than one Markov chain for the entire chromosome, and we model the dependence of these Markov chains across regions and chromosomes. See Section 2 for a description of how regions are constructed. The use of parameters that are specific to regions allows us to define region-specific rates of LOH and enables us to report the desired inference about regions of increased LOH. The instability-selection model was developed for a different inference goal, namely the mapping of tumor suppression genes.

In summary, we define a model structure with three nested levels of repetition: sequences of binary indicators for LOH within each region, consecutive regions within each chromosome, and chromosomes within a patient. We define a hierarchical model over the three levels. For the first level of repeated measurements we assume that the binary LOH sequence within each region is partially exchangeable of order ℓ. A probability model for a binary sequence y_k, $k = 1, \ldots, n$, is order-ℓ exchangeable if the joint distribution is invariant under permutations that leave the initial ℓ values and all order-ℓ transition counts unaltered (Quintana and Newton, 2000; Quintana and Müller, 2004). For example, let $t(0,0) = \sum_{k=2}^{n} I\{y_{k-1} = 0, y_k = 0\}$ and similarly

for $t(0,1)$, $t(1,0)$ and $t(1,1)$, denote the order-1 transition counts. An order-1 exchangeable probability model for the sequence $(y_k,\ k = 1, \ldots, n)$, is a probability measure that is invariant under any permutation that leaves $t(0,0)$, $t(0,1)$, $t(1,0)$, $t(1,1)$ and the initial response y_1 unchanged. Quintana and Newton (1998) show that such sequences can be represented as mixtures of order-ℓ Markov chains. The mixture is with respect to the Markov transition probabilities. This assumption is in agreement with the instability-selection model, but we will allow higher orders of dependence, thus extending the modeling scope.

We complete the description of the mixture of Markov chains model with a nonparametric model on the mixing measure for the corresponding transition matrices (TM). By including latent parameters, the mixture model can be written as a hierarchical model, with the latent TMs interpreted as random effects. Finally, we assume a parametric structure to link the latent parameters across consecutive regions in a chromosome and across chromosomes within a patient. See Section 3 for details.

Fully parametric versions of such models are successfully used for Bayesian inference in multi-level repeated measurement data. Related multi-level models for discrete data are reviewed in Goldstein (2003). Heagerty and Zeger (2000) discuss maximum likelihood inference using a marginal models approach, i.e., regressing the marginal mean, rather than the conditional mean given the random effects, on covariates. Mixture models with a nonparametric mixing measure to define semiparametric random effects distributions are extensively used in nonparametric Bayesian inference, including, for example, Müller and Rosner (1997), Mukhopadhyay and Gelfand (1997) and Kleinman and Ibrahim (1998). The special case of binary outcomes has been discussed, among many others, by Basu and Mukhopadhyay (2000). An advantage of the model specification with such mixtures is that models with no random effects and fully parametric models with a parametric random effects distribution can be seen as special versions of the nonparametric case.

The nonparametric component in the proposed model is based on the Dirichlet process (DP) model introduced in Ferguson (1973). The main reasons for choosing the DP model are the intuitive nature of the prior predictive distributions and computational ease of posterior simulation. Recent reviews of semiparametric Bayesian inference appear in Walker et al. (1999) and Müller and Quintana (2004).

The rest of this article is organized as follows. Section 2 describes the LOH dataset. In section 3, we describe the main features of the proposed model, with emphasis on how we model dependence of the binary LOH sequence within each region, and dependence across regions and chromosomes. Posterior simulation schemes are discussed in Section 4. Section

5 reports the resulting inference, including a comparison with the instability-selection model. Section 6 concludes with a final discussion.

2. The Data

The motivating dataset comes from a study conducted at the St. Jude Children's Research Hospital in Memphis, Tennessee (SJCRH). A full description and discussion of these data can be found in Hartford et al. (2006). We briefly summarize the study. Some children develop second cancers after successful treatment for their initial cancer diagnosis. Of particular concern to the investigators are therapy-related leukemias. Several studies have identified characteristics of particular anti-cancer therapies that may affect the child's risk of developing a later secondary leukemia (Pedersen-Bjergaard, 2005). Some of these treatments may cause genetic alterations that lead to the patient's subsequent secondary cancer.

Investigators at SJCRH carried out genome-wide studies of children with secondary leukemia, in order to learn about genetic factors that may be associated with a patient's risk of a secondary leukemia. A genomic alteration of particular interest to the investigators was LOH. Heterozygosity refers to the presence of two different alleles of a gene at corresponding loci of a pair of chromosomes (i.e., being heterozygous). LOH is when an allele at a particular locus is missing. LOH can occur if part or all of one of the paired chromosomes is missing or if there is a deletion or mutation of part of one of the chromosome pairs. LOH can lead to cancer, for example, if it occurs at the site of a normal tumor suppressor gene that was keeping a cancer-susceptibility gene in check.

The investigators used SNP arrays to compare germline (normal) and tumor (secondary leukemia blasts) samples. The arrays interrogated more than 100,000 SNPs in samples from 13 patients with a diagnosis of treatment-related leukemia. These patients had enrolled in SJCRH protocols for treatment of their initial diagnosis of acute lymphoblastic leukemia (ALL) (Relling et al., 2003). Specifically, they amplified, labeled, and hybridized 500 ng of DNA from each sample to the Affymetrix GeneChip®Human Mapping100K Set. After scanning the chips, they applied the GeneChip®DNA Analysis Software to make the genotype calls for the data. With the genotype calls, the investigators declared each SNP investigated by the array LOH, retention, or indeterminate, following the approach of Lin et al. (2004). The germline samples for these patients came from DNA that the investigators had extracted from normal leukocytes (white blood cells) at the time the child achieved his or her initial remission. Bone marrow at the time of diagnosis of secondary leukemia was the source of the leukemic blast samples. The data consist of $n = 13$ binary sequences with an outcome $y = 1$ for a recorded LOH at a given

SNP, and a zero otherwise. Each sequence is of length 116,204.

The primary objective of this study is the identification of regions of increased LOH, i.e., the main event of interest is a property of regions of SNPs. Consequently, we divide the LOH sequences into regions. A fully model-based approach could consider the choice of regions as a random element itself, and define a prior probability model for region boundaries. However, as we will later show the final inference is relatively robust with respect to detail choices in the definition of the regions. We therefore proceed with an arbitrary definition of regions based on only the following practical considerations. If regions were too long, then inference about increased LOH would be too coarse to be practically useful. Also, since overall less than 1% of the SNPs report LOH we would almost universally report no increased LOH if regions were to span too many SNPs. On the other hand, if regions were too short, then the SNP sequence within each region would not provide much evidence about increased LOH within a region. Also, for many chromosomes and subjects the data records essentially no LOH. Inference about increased LOH for these portions of the SNP sequence does not require small scale regions. Based on these considerations we use the following definition. For chromosomes with more than 0.5% recorded LOH we used regions of length 55 SNPs each. For all other chromosomes we used regions of length 835. This resulted in a total of 874 regions of lengths either 55 (for chromosomes 5 through 9 and 15, 16 and 17) or 835 (for all other chromosomes). The first two nested levels of repeated measurements are thus given by regions within chromosomes and the sequence of recorded indicators for LOH within each region. Besides a general notion of dependence, little is known about appropriate probability models for such data. In the process of meiosis, chromosomes cross and genetic material gets shuffled. Nucleotides closer to each other are less likely to get separated than those that are farther away. This linkage disequilibrium phenomenon suggests a Markovian dependence, in line with the sampling model to be described in Section 3.

The context of the data that we analyze here differs from that in the aforementioned papers. Our data concern patients who were previously treated for cancer and subsequently developed treatment-related leukemia. The investigators hypothesize that the chemotherapy causes chromosomal damage by mechanisms that may be different from those leading to spontaneous (new) cancers. We therefore propose a model that is developed for the goal of identifying consistent *regions* of high LOH without relying on any assumed mechanism.

3. A Model for Repeatedly Repeated Binary LOH Measurements

3.1. The Sampling Model

Recall that y_{icjk} represents the binary indicator of LOH for SNP k in region j of chromosome c for patient i. Let $\boldsymbol{y}_{icj} = (y_{icjk},\, 1 \leq k \leq n_{icj})$ be the entire LOH sequence from the j-th region for chromosome c of the i-th patient. We model correlation at the level of the observed binary outcomes by assuming the sequences $\boldsymbol{y}_{icj}$ to be partially exchangeable of some order. We assume a fixed maximum order ℓ that is common to all sequences. It can be shown (Quintana and Newton, 1998) that order-ℓ exchangeability implies that $p(\boldsymbol{y}_{icj})$ can be expressed as a mixture of homogeneous order-ℓ Markov chains. The mixture is with respect to the order-ℓ transition probabilities.

Let $\boldsymbol{\theta}_{icj}$ denote the transition matrix (TM) that defines the Markov model for subject i, chromosome c and region j. The transition probabilities $\boldsymbol{\theta}_{icj}$ can be represented as a 2^{ℓ}-dimensional vector of transition probabilities $\theta_{icj,m_\ell,\ldots,m_1,0}$ from state $(m_\ell, m_{\ell-1}, \ldots m_1)$ to $(m_{\ell-1}, \ldots, m_1, 0)$, where $m_k \in \{0,1\}$ for all k. Denote by $t_{icj}(m_\ell, \ldots m_1, 0) = \sum_{k=\ell+1}^{n_{icj}} I(y_{icjk} = 0, y_{icj,k-1} = m_1, \ldots, y_{icj,k-\ell} = m_\ell)$, the count of transitions from state $(m_\ell, \ldots, m_1)$ to state $(m_{\ell-1}, \ldots, m_1, 0)$, for region j in chromosome c of patient i. The likelihood is given by

$$p(\boldsymbol{y}_{icj} \mid \boldsymbol{\theta}_{icj}) = \prod_{m_\ell,\ldots,m_1 \in \{0,1\}} \Big\{ \theta_{icj}(m_\ell, \ldots, m_1, 0)^{t_{icj}(m_\ell,\ldots,m_1,0)} \times \big[1 - \theta_{icj}(m_\ell, \ldots, m_1, 0)\big]^{t_{icj}(m_\ell,\ldots,m_1,1)} \Big\}. \quad (1)$$

We adopt (1) with fixed order $\ell = 2$. The implied data reduction by sufficiency to a set of $2^{\ell+1} = 8$ transition counts is critical to facilitate fast likelihood evaluation. The assumption $\ell = 2$ implies that 4 parameters are required to represent each of the $11,362$ TMs (874 per patient) involved in the likelihood model. The choice of $\ell = 2$ generalizes the order-1 Markov models used in Newton and Lee (2000) and Lin et al. (2004). It is also supported by exploratory analysis based on the permutation Monte Carlo tests described in Quintana and Newton (1998). To this effect, we randomly chose one region within each chromosome and considered the corresponding binary sequence for every patient. This amounts to a total of 286 sequences of lengths either 55 or 835. For each sequence we carried out the Monte Carlo test, using a 5% significance level. For 276 of the sequences we find a selected order of $\ell = 0$, for 7 sequences we find $\ell = 1$, and for the remaining 3 sequences we find $\ell = 2$. We conclude that $\ell = 2$ is adequate for all sequences.

3.2. Random Effects Model

We complete the definition of the order-ℓ exchangeable model (with $\ell = 2$) as a mixture of Markov chains by adding a probability model for the TMs $\boldsymbol{\theta}_{icj}$ in (1). In words, we use a non-parametric prior to define the joint distribution of subject-specific effects, a hierarchical normal model to define dependence across chromosomes, and a normal autoregression model to define dependence across regions. The main features of the proposed model are the use of flexible nonparametric priors for the implied marginal distribution of the random effects at all three levels, i.e., regions, chromosomes and subjects, and the use of parsimonious parametric models to define the dependence structure across regions and chromosomes.

We start with a model for the TM $\boldsymbol{\theta}_{ic1}$ corresponding to the first region, i.e. $j = 1$ of chromosome c and patient i. Focusing on only one region ($n_{ic} = 1$) for the moment, we reduce notation to $\boldsymbol{\theta}_{ic} \equiv \boldsymbol{\theta}_{ic1}$. We assume a normal hierarchical model across chromosomes, $p(\text{logit}(\boldsymbol{\theta}_{ic}) \mid \boldsymbol{\mu}_{ic}) = N(\boldsymbol{\mu}_{ic}, \boldsymbol{S})$ and $p(\boldsymbol{\mu}_{ic} \mid \boldsymbol{\mu}_i) = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_1)$, independently across $c = 1, \ldots, 22$. In other words, we define the dependence across chromosomes by assuming an exchangeable normal model for the TMs on a logit scale. We complete the model, still restricting to the first region only, by assuming an exchangeable prior on $\boldsymbol{\mu}_i$ across patients, $\boldsymbol{\mu}_i \sim F$. Instead of specific parametric assumptions for F we use a non-parametric prior. Formally, we define F to be a random probability measure and write $F \sim p(F)$. We define the specific choice of RPM below. In summary, we assume

$$\text{logit}(\boldsymbol{\theta}_{ic}) \sim N(\boldsymbol{\mu}_{ic}, \boldsymbol{S}), \quad \boldsymbol{\mu}_{ic} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_1), \quad \boldsymbol{\mu}_i \sim F \quad \text{and} \quad F \sim p(F). \tag{2}$$

The most commonly used prior $p(F)$ for non-parametric Bayesian inference is the Dirichlet process (DP) prior (Ferguson, 1973). For a review of the DP model see, for example, Müller and Quintana (2004). We only briefly review the main implications of the use of the DP prior in (2). The DP prior is indexed with two parameters, the base measure F_0 and the total mass parameter M, and we write $F \sim DP(M, F_0)$. The base measure defines the prior expectation, $E(F) = F_0$, and the total mass parameter is a precision parameter. The baseline F_0 may itself depend on additional hyper-parameters. A key feature of the DP is that F is almost surely discrete. The discrete nature of F implies a positive probability for ties among the $\boldsymbol{\mu}_i$ values. The groups of patients sharing a common value can be interpreted as *clusters*. Let $\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_L^*$ be the unique values among $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$. We define latent indicators s_i for *cluster memberships*, such that $s_i = h$ if $\boldsymbol{\mu}_i = \boldsymbol{\mu}_h^*$, and let m_h represent the *size* of the hth cluster, i.e. the number of $\boldsymbol{\mu}_i$s equal to $\boldsymbol{\mu}_h^*$. The DP model can be best understood by considering the prior predictive distribution for a sample $\boldsymbol{\mu}_i \sim F$, $i = 1, \ldots, n$, generated from a random probability model with a DP prior. Marginalizing with respect to F, the following prior predictive probabilities

apply:

$$p(\boldsymbol{\mu}_{n+1} \mid \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n) = \begin{cases} \delta_{\boldsymbol{\mu}_h^*}(\boldsymbol{\mu}_{n+1}), & \text{with pr. } m_h/(M+n),\ h = 1, \ldots, L \\ F_0(\boldsymbol{\mu}_{n+1}), & \text{with pr. } M/(M+n) \end{cases} \tag{3}$$

The predictive rule (3) implies that with some probability the TMs for a new patient mimic some of the previous ones (up to residual variation); with the remaining probability, the latent parameters controlling the TMs are drawn from the baseline distribution F_0. A computationally convenient choice for the baseline measure F_0 is a conjugate prior to the kernel in (2). In the case of the normal kernel we use $F_0(\boldsymbol{\mu}) = N(\boldsymbol{\mu}; \boldsymbol{m}, \boldsymbol{V})$.

In summary, we have defined dependence across chromosomes c by the hierarchical model (2) and dependence across subjects i by (3). A critical feature of the proposed model is that the subject specific random effects $\boldsymbol{\mu}_i$ are of dimension 2^ℓ. A non-parametric prior for the joint vector of all $\boldsymbol{\mu}_{ic}$ would be of prohibitive dimension. Therefore, we use a parametric model to specify dependence across c in (2), and a non-parametric prior to define dependence of $\boldsymbol{\mu}_i$ across i in (3). Note that the marginal model for each $\boldsymbol{\mu}_{ic}$, marginalizing w.r.t. $\boldsymbol{\mu}_i$, is also a semiparametric mixture of normals

$$p(\boldsymbol{\mu}_{ic} \mid F) \sim \int N(\boldsymbol{\mu}, \boldsymbol{S} + \boldsymbol{\Sigma}_1)\, dF(\boldsymbol{\mu}).$$

We now complete the model by defining dependence across the second level of repeat experiments in the data structure, i.e., dependence across regions, using a similar modeling strategy. The only difference is that an appropriate model for dependence across regions is based on spatial dependence rather than exchangeability. We return to the general case with multiple regions, $j = 1, \ldots, n_{ic}$, for each chromosome and patient, as described in Section 2. Let $\boldsymbol{\theta}_{ic} = (\boldsymbol{\theta}_{ic1}, \ldots, \boldsymbol{\theta}_{icn_{ic}})$, $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}, \ldots, \boldsymbol{\theta}_{i,22})$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$. Let also $\boldsymbol{\mu}_{ic}$, $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}$ be analogously defined. The likelihood remains as in (1). As in (2) we introduce latent variables $\boldsymbol{\mu}_{icj}$ with $\text{logit}(\boldsymbol{\theta}_{icj}) \mid \boldsymbol{\mu}_{icj} \sim N(\boldsymbol{\mu}_{icj}, \boldsymbol{S})$, and replace (2) by an extended model to include dependence across regions. We introduce a vector of autoregressive coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{2^\ell})$. The coefficient α_h is used to characterize the change of the h-th coefficient of the transition matrix μ_{icj} across regions. Let $D(\boldsymbol{\alpha}) = \text{diag}(\boldsymbol{\alpha})$ denote the $2^\ell \times 2^\ell$ diagonal matrix with $\boldsymbol{\alpha}$ on the diagonal. We replace (2) by

$$\boldsymbol{\mu}_{i0} \sim F,\ \ \boldsymbol{\mu}_{ic0} = \boldsymbol{\mu}_{i0} + \epsilon_{ic0},\ \ \boldsymbol{\mu}_{ic1} = \boldsymbol{\mu}_{ic0} + \epsilon_{ic1} \quad \text{and} \quad \boldsymbol{\mu}_{icj} = \boldsymbol{\mu}_{ic0} + D(\boldsymbol{\alpha})(\boldsymbol{\mu}_{ic,j-1} - \boldsymbol{\mu}_{ic0}) + \boldsymbol{\epsilon}_{icj}, \tag{4}$$

where $\boldsymbol{\epsilon}_{icj}$ and $\boldsymbol{\epsilon}_{ic0}$ are independent normal residuals with $\boldsymbol{\epsilon}_{ic0} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_1)$, and $\boldsymbol{\epsilon}_{icj} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_2)$ for $j \geq 1$. The model assumes first-order stationarity, and includes a strictly stationary model

as a special case by appropriate choices of $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\alpha}$. Marginally for each region j, the model still implies the nonparametric mixture of normals model for $\text{logit}(\boldsymbol{\theta}_{icj})$, as before, now with the kernel $N(\text{logit}(\boldsymbol{\theta}_{ic});\ \boldsymbol{\mu}_{ic}, \boldsymbol{S})$ replaced by a normal kernel $p(\text{logit}(\boldsymbol{\theta}_{icj}) \mid \boldsymbol{\mu}_{ic0}, \boldsymbol{S}, \boldsymbol{\alpha}) = N(\boldsymbol{\mu}_{ic0}, V(\boldsymbol{S}, \boldsymbol{\alpha}))$ with the variance-covariance matrix $V(\boldsymbol{S}, \boldsymbol{\alpha})$ implied by marginalizing (4) with respect to $\boldsymbol{\mu}_{ic1}, \ldots, \boldsymbol{\mu}_{ic,j-1}$. The cluster structure on patients remains determined by the SSM assumption for F. The desired learning about regions of increased LOH is then accomplished by examining the posterior distributions of an appropriate function of the $\boldsymbol{\theta}_{icj}$ parameters. See Section 5 below for details on how we do this.

The structure in (4) highlights how the proposed semi-parametric model relates to a simpler parametric model. If we were to assume a parametric model for $\boldsymbol{\mu}_{i0}$, for example the base measure of the DP, $\boldsymbol{\mu}_{i0} \sim N(\boldsymbol{m}, \boldsymbol{V})$, the model would reduce to a fully parametric hierarchical model. Besides increased flexibility the advantage of the semi-parametric extension is that it remains more faithful to the prior judgement about the binary sequences by building on only the order ℓ exchangeability assumption.

The model specification is completed by defining hyperpriors on all remaining parameters. Let $\boldsymbol{\eta}$ denote the set of all other hyper-parameters. These include the regression coefficients $\boldsymbol{\alpha}$, the covariance matrices $\boldsymbol{S}$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, and hyper-parameters from the baseline distribution F_0, $\boldsymbol{m}$ and $\boldsymbol{V}$. For $\boldsymbol{\alpha}$ we use a normal prior, $p(\boldsymbol{\alpha}) = N(\boldsymbol{\alpha};\ \boldsymbol{a}_0, \boldsymbol{A}_0)$. Next, for $\boldsymbol{S}$ we choose an inverse-Wishart prior, $p(\boldsymbol{S}) = IW(\boldsymbol{S};\ \boldsymbol{S}_0, \nu_S)$. We also assume independent conjugate inverse-Wishart priors for the residual variances: $p(\boldsymbol{\Sigma}_1) = IW(\boldsymbol{\Sigma}_1;\ \boldsymbol{\Sigma}_{01}, \nu_1)$ and $p(\boldsymbol{\Sigma}_2) = IW(\boldsymbol{\Sigma}_2;\ \boldsymbol{\Sigma}_{02}, \nu_2)$. Finally, for the hyper-parameters in F_0 we use $p(\boldsymbol{m}, \boldsymbol{V}) = N(\boldsymbol{m};\ \boldsymbol{m}_0, \boldsymbol{V}/\lambda_0) \times IW(\boldsymbol{V};\ \boldsymbol{V}_0, \nu_V)$. In the earlier definitions we assume $\boldsymbol{a}_0$, $\boldsymbol{A}_0$, ν_A, $\boldsymbol{S}_0$, ν_S, ν_1, $\boldsymbol{\Sigma}_{01}$, ν_2, $\boldsymbol{\Sigma}_{02}$, $\boldsymbol{m}_0$, $\boldsymbol{V}_0$, λ_0 and ν_V to be known.

4. Posterior Simulation

Model (2) has the great advantage of conditional independence at various levels. This conditional independence facilitates implementation of a Gibbs sampling algorithm. In particular, the transition probabilities $\boldsymbol{\theta}_{icj}$ are conditionally independent across i, c and j, given all other parameters. As a consequence, the $\boldsymbol{\theta}_{icj}$ can be updated one at a time. Sampling from the corresponding full conditionals can be accomplished using standard methods for logistic regression, as discussed, e.g., in Carlin and Louis (1996).

Next, consider updating $\boldsymbol{\mu}_{i0}$ in model (4). Details on updating the configurations of ties among the $\{\boldsymbol{\mu}_{i0},\ i = 1, \ldots, n\}$, and the unique values $\boldsymbol{\mu}_h^*$ are described, among others, in MacEachern and Müller (2000) and Neal (2000). The complete conditional posterior for $\boldsymbol{\mu}_{icj}$,

$j \geq 1$, including conditioning on logit($\boldsymbol{\theta}_{icj}$) and $\{\boldsymbol{\mu}_{ics}\}$, $s \neq j$, is a straightforward normal linear regression.

Updating the autoregression parameters in (4) proceeds by draws from the complete conditional posterior distribution. Model (4) is linear in $\boldsymbol{\alpha}$. The conjugate normal prior assumption for $\boldsymbol{\alpha}$ allows for straightforward posterior simulation for $\boldsymbol{\alpha}$ conditional on imputed values for all other parameters. Finally, the remaining parameters are easily updated from the corresponding conjugate-style conditionals. See further details in Quintana and Müller (2004) and in Müller et al. (2007).

5. Identifying Regions of Increased LOH

We assume model (1) with $\ell = 2$, that is, $\boldsymbol{\theta}_{icj}$ is of dimension $2^\ell = 4$ and represents the full order-2 TM for the jth region of chromosome c of the ith patient. For the random effects distribution $p(\boldsymbol{\theta}_{icj}, j = 1, \ldots, n_{ic})$, we use model (4) with a $DP(M, F_0)$ prior with $M = 1$. The model treats the responses from different regions as conditionally independent given region-specific parameters.

Figure 1 shows the estimated marginal posterior means plus and minus one posterior standard deviation for the components of the $\boldsymbol{\alpha}$ and $\boldsymbol{m}$ coefficients. Recall that $\boldsymbol{\alpha}$ parametrizes the AR model in (4), and $\boldsymbol{m}$ is the mean of the DP base measure F_0. The posterior means for α_h are well away from zero, suggesting a significant autoregression effect, except possibly for α_3, which corresponds to the logit of transition probabilities from $(1, 0)$ to $(0, 1)$ (i.e., LOH skipping a SNP). Also, they are significantly away from 1, except possibly for α_1, i.e the coefficient for $(0, 0)$ to $(0, 1)$ transitions. In other words, the data suggests an autoregressive effect that is not a random walk-type process. On the other hand, the $\boldsymbol{m}$ coefficients, which control the center of the baseline measure (on the logit scale), are all well negative. This reflects the fact that over 99% of all the responses are zero, and so the baseline values for transition probabilities from any previous two values to an LOH response are very low for any given region. This is further reflected in Figure 2a which shows the estimated posterior means for the (0,0) to (0,0) transition probabilities (on the logit scale). The estimated transition probabilities are very high for almost all regions and patients.

If one wishes to evaluate LOH in any given region, we recommend using the long-run proportion of LOH in that region. This is based on the fact that chromosome regions are large enough to justify approximating (ergodic) LOH averages by their corresponding limits. From

elementary Markov chain theory (Ross, 2002), we find that these are given by

$$\lim_{k\to\infty} P(y_{icjk}=1) =$$
$$\lim_{k\to\infty} \sum_{j_1,j_2\in\{0,1\}} P(y_{icjk}=1|y_{icj,k-1}=j_1, y_{icj,k-2}=j_2)P(y_{icj,k-1}=j_1, y_{icj,k-2}=j_2) =$$
$$\sum_{j_1,j_2\in\{0,1\}} P(y_{icj2}=1|y_{icj1}=j_1, y_{icj0}=j_2) \lim_{k\to\infty} P(y_{icj,k-1}=j_1, y_{icj,k-2}=j_2).$$

The first term in the last summation is the transition probability from $(j_2 j_1)$ to $(j_1 1)$, while the limit probabilities in the second part are given by the stationary distribution corresponding to the appropriate TM, and therefore, easily identified as functions of $\boldsymbol{\theta}_{icj}$.

Figure 2b shows the estimated marginal posterior long- run probabilities of LOH on the logit scale, computed as indicated above. Patients 2, 6, 8, 9, and 10 show some regions with higher probability of LOH than the other patients. In general we find an overall low percentage of observed LOH sites.

The estimated marginal probabilities do not yet address the original inference goal of identifying regions of increased LOH. We address this goal by defining an indicator of "increased LOH" for patient i and region j as

$$I_{icj} = \begin{cases} 1 & \text{if } \lim_{k\to\infty} P(y_{icjk}=1) > p_0 \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $0 < p_0 < 1$ is a fixed threshold. In other words, we say that a given region has increased LOH if the marginal long-run probability of LOH is greater than p_0. As noted earlier, $\lim_{k\to\infty} P(y_{icjk}=1)$ is a function of the transition probabilities $\boldsymbol{\theta}_{icj}$ only. Figure 3a reports the posterior expectations of I_{icj} on the logit scale, using $p_0 = 1\%$. Patients 2, 6, 8, 9, and 10 stand out again. These five patients have longer stretches of neighboring loci that have high posterior probability of increased LOH relative to the threshold of 1%.

We still need to relate the reported probabilities to the desired decision of identifying regions of increased LOH. In carrying out this decision we face a massive multiplicity problem. In the context of high throughput gene expression data, several procedures have been proposed to address such decision problems based on the notion of false discovery rates (Benjamini and Hochberg, 1995; Storey, 2002). Most discussions are for the stylized setup of a two-group comparison microarray experiment. For each of a large number of genes recorded on the microarrays, we wish to make a decision about differential expression. Under the same setup and using a Bayesian decision-theoretic perspective, Müller et al. (2004) show that under a variety of loss functions the optimal decision is characterized by flagging all those comparisons

with marginal probability of differential expression beyond a certain threshold. The conclusion is valid for any probability model. In particular, the probability model can include dependence, as in the proposed model for LOH indicators. Thus the same solution applies. The optimal inference about regions of increased LOH is achieved by marking all regions with marginal probability of increased LOH beyond a threshold. Let $\delta_{icj} \in \{0, 1\}$ denote an indicator for reporting increased LOH for region (icj). Let $D = \sum_{icj} \delta_{icj}$ denote the total number of reported regions. Let $I_{icj} = I_{icj}(\boldsymbol{\theta}_{icj})$ denote the unknown truth, as in (5) and define the false discovery proportion (FDP) as

$$\text{FDP} = \sum_{i,j} (1 - I_{icj})\, \delta_{icj} / D.$$

The FDP is a function of the unknown parameters I_{icj} and the data, implicitly through δ_{icj}. Let $v_{icj} = E(I_{icj} \mid \text{data})$ denote the marginal posterior probability of increased LOH in region (icj). These probabilities are reported in Figure 3a. The posterior expected FDP, $\overline{\text{FDR}} = E(\text{FDP} \mid \text{data})$, is evaluated as $\overline{\text{FDR}} = \sum_{i,j} (1 - v_{icj})\, \delta_{icj} / D$. Using a threshold of 0.29, i.e., $\delta_{icj} = I(v_{icj} > 0.29)$, we find $\overline{\text{FDR}} = 10\%$. This inference is reported in Figure 3b, with black bars indicating the decision to flag a region as exhibiting "increased LOH." For two patients (2 and 6) we see uniformly increased LOH across all regions. Note the double thresholding that is implicit in the definition of δ by thresholding the posterior expectation of I_{icj}, which in turn is defined by a threshold on the limiting probabilities. This arises because we are interested in regions of increased LOH, rather than regions of LOH (the latter would be very few for a comparable FDP).

For comparison we considered three alternative approaches for the desired inference and compare these with our proposed scheme. We used the methods proposed in Newton et al. (1998) and Lin et al. (2004) and a fully parametric version of our proposed model. We used a set of 100 equally spaced loci for the method of Newton et al. (1998). We found all the log of odds (LOD) scores to be zero, i.e. no region of high LOH was detected. This somewhat surprising conclusion is explained by the fact that the binary sequences are very long and the proportion of recorded LOH per sequence is very low. Thus a model that assumes a single Markovian process across all regions, may not be flexible enough to capture local behavior as our approach does. In this case, the MLEs required to implement the model in Newton et al. (1998) are essentially driven by the overwhelming proportion of observed "no loss" and thus the result.

Results for the approach proposed in Lin et al. (2004) are shown in Figure 4. We used the implementation in dChip, the public domain software that is introduced in Lin et al.

(2004). Figure 4a plots the probability of LOH using the hidden Markov model score defined in dChip. Compare the inference with Figure 2b. While the general patterns are similar under both methods, inference under the proposed model includes more extensive borrowing strength within the hierarchical model. Also, the reported inference provides region-specific probabilities under a coherent joint probability model across regions, chromosomes and samples. This allows the investigator to report summaries like the inference about regions of increased LOH based on joint probability models. On the other hand, an important advantage of the approach in Lin et al. (2004) is the highly reduced reliance on a specific model. In fact the approach includes an option to compute simple entirely model-independent LOH scores. Their method judges significance by a permutation test with appropriate multiplicity control.

Finally, we implemented a fully parametric model by replacing the nonparametric model $\boldsymbol{\mu}_{i0} \sim F$ in (4) with a parametric model $\boldsymbol{\mu}_{i0} \sim N(\boldsymbol{m}, \boldsymbol{V})$. All other model choices are left unchanged. Figure 4b shows the resulting probability of LOH. Compared with Figure 2b we see less smoothing across regions in the reported probabilities but otherwise similar results. The additional smoothing in the semi-parametric model arises from the clustering that is implied by the DP prior. In contrast, the parametric model can be described as assuming all singleton clusters, i.e., all random effects $\boldsymbol{\mu}_{i0}$ in (4) are distinct. The main argument for the semi-parametric model is that it naturally follows from the prior judgement about partial exchangeability of the binary sequences.

Finally, we assess the sensitivity with respect to the arbitrary definition of regions. We consider two alternative schemes. First, we construct regions as before, but using lengths of either 100 or 1000, rather than 55 or 835. In the second alternative scheme that we consider, the number of base pairs in each region is kept constant and thus the number of SNPs per region is variable. We chose the number of base pairs such that the number of regions per chromosome remains the same as before (i.e., when regions were defined by 55 and 835 SNPs, respectively). The resulting plots (not shown) of probability of LOH, probability of increased LOH, and the decisions to flag regions for increased LOH remain almost unchanged from Figures 2ab and 3a.

6. Conclusion

Motivated by the analysis of LOH data, we have introduced a semiparametric Bayesian model for binary measurements with multiple nested levels of repetition. The top-level repetition is modeled as a mixture of Markov chains. The mixture is defined with respect to the transition matrix for a given order of dependence ℓ for SNPs within a given region. Marginally, for each second level repeated measurement unit (chromosome region), a nonparametric model charac-

terizes the random effects related to that unit. The proposed approach completes the model by defining dependence across regions and across chromosomes, using a parametric hierarchical model.

The data analysis carried out can be extended and complemented in several ways. The goal of the inference in our motivating application was to detect regions of increased LOH. In other applications with LOH data, one might be interested in modeling for loss of tumor suppression genes that are hypothesized to be associated with the observed LOH. Newton and Lee (2000) propose an instability-selection model that facilitates such inference. The instability-selection model could be used to replace the partially exchangeable sampling model in our approach, while still keeping dependence across SNPs as in (4).

For other applications it might be useful to extend the proposed model to include location-specific covariates in (1). This could be achieved by using a log-linear model for the transition probabilities. The log-linear model would include the regression on the lagged outcomes as well as additional location-specific covariates. The regression coefficients in the log-linear model would then replace $\text{logit}(\boldsymbol{\theta}_i)$ in (2).

Acknowledgments

This research was supported, in part, by grants CA075981 and GM061393 from the U.S. National Cancer Institute, and by grants FONDECYT 1060729 and Laboratorio de Análisis Estocástico PBCT-ACT13. We thank Dr. Wenjian Yang for help with the data files and advice about the analysis in dChip. We also thank the Associate Editor and Referees for their valuable comments and suggestions.

References

Basu, S. and Mukhopadhyay, S. (2000) Bayesian analysis of binary regression using symmetric and asymmetric links. *Sankhyā*, **62**, 372–387.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Methodological*, **57**, 289–300.

Beroukhim, R., Lin, M., Park, Y., Hao, K., Zhao, X., Garraway, L. A., Fox, E. A., Hochberg, E. P., Mellinghoff, I. K., Hofer, M. D., Descazeaud, A., Rubin, M. A., Meyerson, M., Wong, W. H., Sellers, W. R. and Li, C. (2006) Inferring loss-of-heterozygosity from un-

paired tumors using high-density oligonucleotide snp arrays. *PLoS Computational Biology*, **2**. URLhttp://www.citebase.org/abstract?id=oai:pubmedcentral.gov:1458964.

Carlin, B. P. and Louis, T. A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.

Goldstein, H. (2003) *Multilevel Statistical Models, Third Edition*. Kendall's Library of Statistics, 3. London: Arnold Publishers.

Hartford, C., Yang, W., Cheng, C., Liu, W., Su, X., Pounds, S., Neale, G., Fan, Y., Raimondi, S. C., Bogni, A., Pui, C.-H. and Relling, M. V. (2006) Genome Scan for Therapy-Related Myeloid Leukemia. *Tech. rep.*, Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital.

Heagerty, P. J. and Zeger, S. L. (2000) Marginalized multilevel models and likelihood inference. *Statistical Science*, **15**, 1–19.

Kleinman, K. and Ibrahim, J. (1998) A semi-parametric bayesian approach to the random effects model. *Biometrics*, **54**, 921–938.

Lin, M., Wei, L. J., Sellers, W. R., Lieberfarb, M., Wong, W. H. and Li, C. (2004) dchipsnp: significance curve and clustering of snp-array-based loss-of-heterozygosity data. *Bioinformatics*, **20**, 1233–1240.

MacEachern, S. N. and Müller, P. (2000) Efficient mcmc schemes for robust model extensions using encompassing dirichelt process mixture models. In *Robust Bayesian Analysis* (eds. F. Ruggeri and D. R. Insua). New York.

Miller, B. J., Wang, D., Krahe, R. and Wright, F. A. (2003) Pooled Analysis of Loss of Heterozygosity in Breast Cancer: a Genome Scan Provides Comparative Evidence for Multiple Tumor Suppressors and Identifies Novel Candidate Regions. *American Journal of Human Genetics*, **73**, 748–767.

Mukhopadhyay, S. and Gelfand, A. E. (1997) Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association*, **92**, 633–639.

Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004) Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, **99**.

Müller, P. and Quintana, F. (2004) Nonparametric Bayesian Data Analysis. *Statistical Science*, **19**, 95–110.

Müller, P. and Rosner, G. (1997) A bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association*, **92**, 1279–1292.

Müller, P., Rosner, G. and Quintana, F. A. (2007) Semiparametric Bayesian Inference for Multilevel Repeated Measurement Data. *Biometrics*, **63**, 280–289.

Neal, R. M. (2000) Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Newton, M. A., Gould, M. N., Reznikoff, C. A. and Haag, J. D. (1998) On the statistical analysis of allelic-loss data. *Stat Med*, **17**, 1425–45.

Newton, M. A. and Lee, Y. (2000) Inferring the Location and Effect of Tumor Suppressor Genes by Instability-Selection Modeling of Allelic-Loss Data. *Biometrics*, **56**, 1088–1097.

Pedersen-Bjergaard, J. (2005) Insights into leukemogenesis from therapy-related leukemia. *New England Journal of Medicine*, **352**, 1591–1594.

Quintana, F. and Müller, P. (2004) Nonparametric bayesian assessment of the order of dependence for binary sequences. *Journal of Computational and Graphical Statistics*, **13**, 213–231.

Quintana, F. A. and Newton, M. A. (1998) Assessing the Order of Dependence for Partially Exchangeable Binary Data. *Journal of the American Statistical Association*, **93**, 194–202.

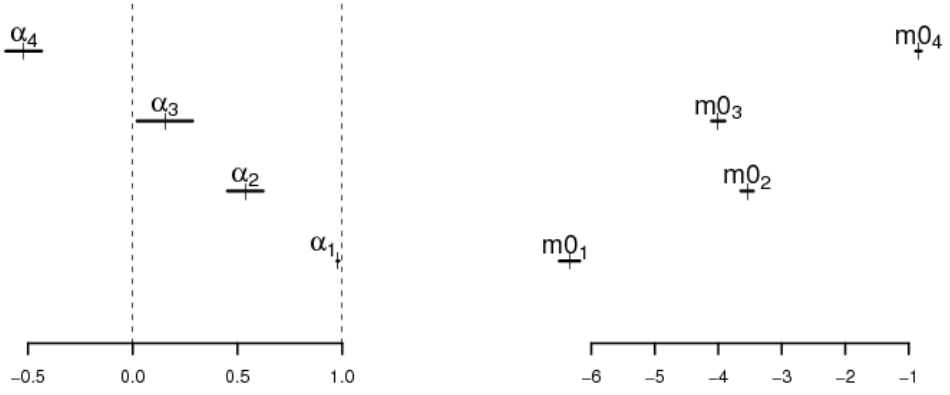
— (2000) Computational aspects of Nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences. *Journal of Computational and Graphical Statistics*, **9**, 711–737.

Relling, M.V.and Boyett, J., Blanco, J., Raimondi, S., Behm, F., Sandlund, J., Rivera, G., Kun, L., Evans, W. and Pui, C. (2003) Granulocyte colony-stimulating factor and the risk of secondary myeloid malignancy after etoposide treatment. *Blood*, **101**, 3862–3867.

Ross, S. M. (2002) *Introduction to Probability Models, Eighth Edition*. Academic Press.

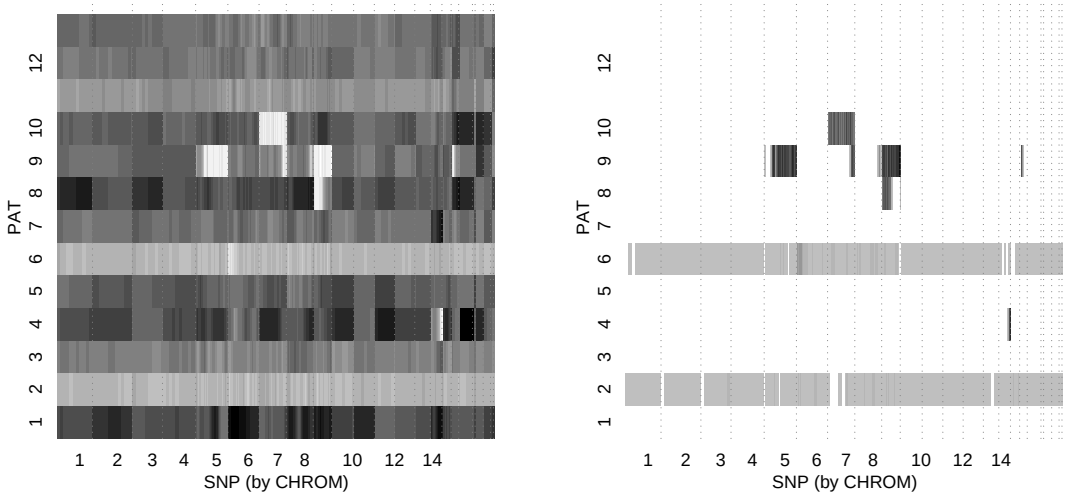
Storey, J. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B (Methodological)*, **64**, 479–498.

Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. M. (1999) Bayesian nonparametric inference for random distributions and related functions (Disc: P510-527). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **61**, 485–509.



(a) (b)

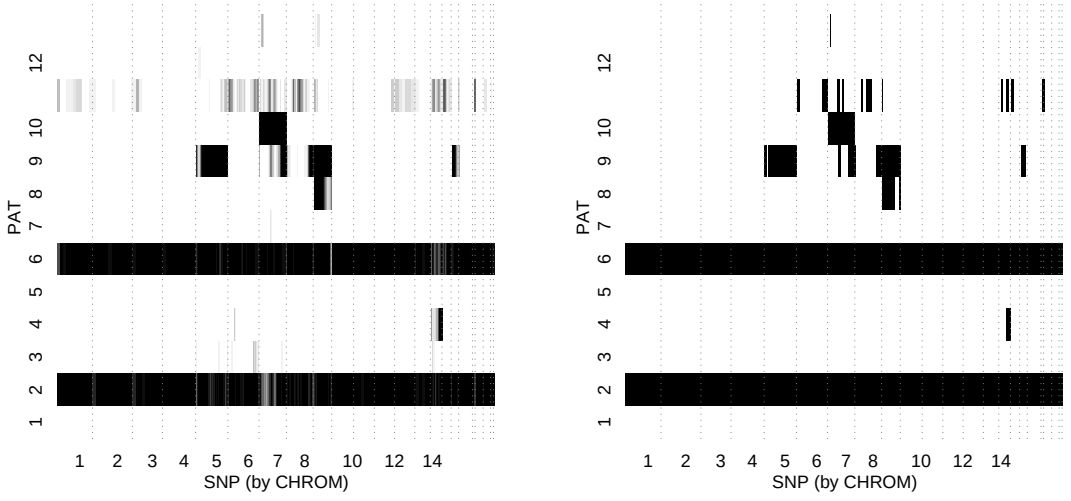
Fig. 1. *Marginal posterior means and standard deviations for (a)* $\boldsymbol{\alpha}$ *and (b)* $\boldsymbol{m}$*. The horizontal bars show the marginal posterior mean (a)* $E(\alpha_\ell \mid Y)$ *and (b)* $E(m_\ell \mid Y)$ *(marked by "| ") plus/minus one posterior standard deviation for* $\ell = 1, \ldots, 4$



(a) $(0,0) \rightarrow (0,0)$ transition probability

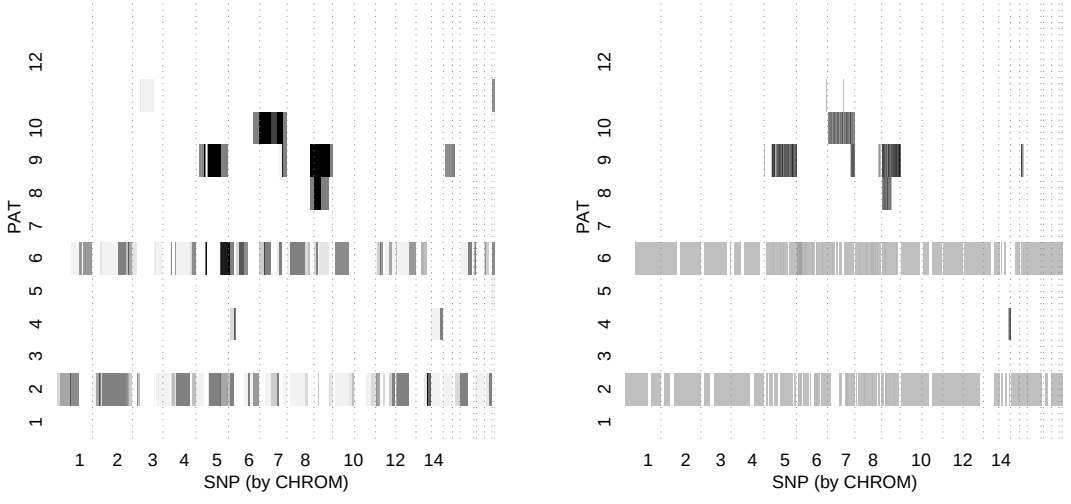
(b) probability of LOH

Fig. 2. *The left panel reports the marginal posterior means of the transition probability from state "00" to "00", on a logit scale. The right panel shows the marginal posterior probability of LOH. Values are indicated by gray shades with black corresponding to 1.0. From left to right, columns represent regions 1 through* 874*. Chromosome boundaries are indicated by dotted vertical lines, and labeled for chromosomes 1 through 14. The rows correspond to patients 1 through 13.*



(a) $Pr(I_{icj} \mid data)$ (b) δ_{icj}

Fig. 3. *The left panel reports the marginal posterior probability of increased LOH, $P(I_{icj} \mid data)$, with darker gray shades for higher probabilities. On the horizontal axis are regions arranged by chromosomes, as in Figure 2. The right panel shows the decisions δ_{icj}, with a black bar indicating that region (ij) is reported as increased LOH.*



(a) dChip (b) parametric model

Fig. 4. *Probability of LOH under two alternative methods. Panel (a) shows the inference reported by dChip using the hidden Markov model scores. Panel (b) shows inference under a fully parametric model. In both panels the horizontal axis shows regions arranged by chromosomes. Compare with Figure 2b.*