Bayesian Random Segmentation Models to Identify Shared Copy Number Aberrations for Array CGH Data

Veerabhadran Baladandayuthapani, Yuan Ji, Rajesh Talluri, Luis E. Nieto-Barajas, and Jeffrey S. Morris

1. INTRODUCTION

1.1 Detection of Shared Aberrant Genetic Regions in Cancer

Genomic abnormalities in the number of DNA copies in a cell have been associated with cancer development and progression (Pinkel and Albertson 2005). During cell replication, various types of errors can occur that lead to either insertion of an extra copy or deletion of part of a DNA sequence in the genome. Left unchecked, these errors can silence important genes or amplify their expression, in either case leading to an improperly functioning cell. If they involve amplification of proto-oncogenes that regulate cell division or deletion of tumor-suppressor genes that prevent unwanted cell division or induce programmed cell death, these errors can be contributing factors in the initiation stage of carcinogenesis. Accumulation of particular combinations of these genetic errors can cause a group of cells to cross the threshold to cancer, at which point the cells’ increased genetic instability and high replication rate will lead to even more errors, possibly leading to progression or metastasis. Thus we can expect that shared genomic regions with common DNA copy alterations in a particular population of cancer patients may contain genes that are crucial in characterizing this population, be it a group of patients with a common cancer type, a set of patients who metastasize versus those who do not, or a subset of patients responding to a particular biological therapy. The detection of these shared regions of aberration and assessment of differential alterations between groups have the potential to impact the basic knowledge and treatment of many types of cancers and can play a role in the discovery and development of molecular-based personalized cancer therapies.

1.2 Array CGH

Comparative genomic hybridization (CGH) methods were developed to survey DNA copy number variations across a whole genome in a single experiment (Kallioniemi et al. 1992). With CGH, differentially labeled test (e.g., tumor) and reference (e.g., normal individual) genomic DNAs are cohybridized to normal metaphase chromosomes, and fluorescence ratios along the length of chromosomes provide a cytogenetic representation of the relative DNA copy number variation. Because chromosomal CGH resolution is limited to 10–20 Mb, any aberration smaller than that will not be detected. Array-based CGH (aCGH) is a recent modification of CGH that provides greater resolution by using microarrays of DNA fragments rather than metaphase chromosomes (Pinkel et al. 1998; Snijders et al. 2001). These arrays can be generated with different types of DNA preparations. One method uses bacterial artificial chromosomes (BACs), each of which consists of a 100- to 200-kb DNA segment. Other arrays are based on complementary DNA (cDNA; Pollack et al. 1999, 2002) or oligonucleotide fragments (Lucito et al. 2000). As in CGH analysis, the resultant map of gains and losses is obtained by calculating fluorescence ratios measured via image analysis tools. An alternative high-resolution technique for detecting copy number variation is accorded by single nucleotide polymorphism (SNP) genotyping methods (Zhao et al. 2004; Herr et al. 2005).
genotyping large numbers of DNA sequences, one might possibly use aCGH to determine gains and losses with high resolution across the entire genome. The broad goal of determining such genomic patterns of gains and losses can be subsequently applied in possible cancer diagnosis and management. For example, for a group of patients diagnosed with the same pathological type of cancer, genetic subtyping can predict markedly different responses to chemotherapies and offer powerful prognostic information.

Like most microarray analyses, normalization of the intensity ratios (or the corresponding log-ratios) is conducted before any downstream analysis, to adjust for sources of systematic variation not attributable to biological variation. The most common normalization techniques are of a global nature, such as centering the data about the sample mean or median for a given hybridization (Fridlyand et al. 2004). (See Khojasteh et al. 2005 for further discussion on normalization methods for aCGH data.) For the present analysis, we assume that the data have been appropriately normalized to adjust for experimental artifacts.

In an idealized scenario where all of the cells in a tumor sample have the same genomic alterations and are uncontaminated by cells from surrounding normal tissue, the log2 ratio of normal probes is $\log_2(2/2) = 0$, that of single copy losses is $\log_2(1/2) = -1$, and that of single copy gains is $\log_2(3/2) = 0.58$. Multiple copy gains have values of $\log_2(4/2), \log_2(5/2)$, and so on. Loss of both copies theoretically would correspond to a value of $-\infty$. In this idealized situation, all copy number alterations could be promptly observed from the data, obviating the need for statistical techniques. In real applications, however, the log2 ratios differ considerably from these expected values, for various reasons. First, aCGH data are characterized by high noise levels that add random measurement errors to the observations. Second, a given tumor sample is not completely homogeneous, because there may be contamination with neighboring normal cells and considerable genomic variability among the individual tumor cells. This heterogeneity means that we actually measure a composite copy number estimate across a mixture of cell types, which tend to result in attenuation of the ratios toward zero.

1.3 Existing Approaches

As mentioned earlier, one of the key goals in aCGH data analysis is to infer regions of gains and losses in the copy number across the genome map. A host of methods have attempted to fulfill this need with varying degrees of success. Most proposed methods fall into one of two categories: calling methods or segmentation methods. Calling methods model the aCGH profile at a probe/clone level and call the state of each probe gain, loss, or neutral. The most popular of these methods are the hidden Markov models (HMM). Guha, Li, and Neuberg (2008) proposed a Bayesian HMM to account for the dependence between neighboring clones by specifying the true copy number states as the latent states in the HMM scheme. Shah et al. (2007) extended the HMMs to detect shared aberrations by modeling the shared profile by a master sequence of states that generates the samples. Hodgson et al. (2001) proposed a three-component Gaussian mixture model corresponding to gain, loss, or neutral states, respectively. Another related approach using HMM proposed by Fridlyand et al. (2004) also shares characteristics of the segmentation approaches described below.

Segmentation methods or change point models seek to identify contiguous regions of common means, separated by breakpoints, and to estimate the means in these regions. Sen and Srivastava (1975) proposed a frequentist solution of detecting a single change point, which was subsequently extended by Olshen et al. (2004) for aCGH data as the circular binary segmentation (CBS). The CBS recursively detects pairs of change points to identify chromosomal segments with altered copy numbers. Another alternative is a penalized maximum likelihood approach in which the data likelihood is maximized for a fixed number of change points, usually with an added (heuristic) penalty to control for overfitting, such as the least squares criterion (Huang et al. 2005), penalized quantile smoothing (Eilers and de Menezes 2005), and fused lasso penalty (Tibshirani and Wang 2008).

Bayesian approaches for change point models typically involve a joint prior on the configuration of possible change point(s) and the associated parameters. Carlin, Gelfand, and Smith (1992) proposed a hierarchical Bayesian analysis of change point models in the context of a single change point, and Inclan (1993) and Stephens (1994) considered models for multiple change points. Barry and Hartigan (1993) discussed alternative formulations of the change points in terms of product partition distributions, which Erdman and Emerson (2008) subsequently tailored for aCGH data. Chib (1998) proposed a formulation of the multiple change point model in terms of latent discrete state variables that indicated the regime from which a particular observation was drawn. Hutter (2007) developed an exact Bayesian regression algorithm for piecewise constant functions using unknown segment number and boundary locations. Denison, Mallick, and Smith (1998) proposed an approach to the variable change point problem in a different context, using reversible-jump techniques, but considered only the single function case. Our approach generalizes the approach of Denison, Mallick, and Smith (1998) to multiple functions in functional regression framework.

In the aCGH context, all of these segmentation methods provide breakpoint locations and corresponding means, but do not discern whether the corresponding segments represent a true aberration. That is, once the change points are identified and associated mean log ratio levels are estimated, it is not at all clear which segments of common mean represent real genetic aberrations (i.e., copy number gain or loss) and which are simply due to experimental variability. Thus an additional post-model-fitting procedure is implemented to call the segments as gains or losses, often based on ad hoc thresholding criteria, such as the median of the median absolute deviations (Rossi et al. 2005). Other approaches include clustering-based approaches to combining similar segments (Hupe et al. 2004) and combining segments based on their distributions (Willenbrock and Fridlyand 2005). Thus the final inference is highly dependent on the performance of the segmentation procedure, which is usually based on user-defined parameters. Moreover, because the subsequent calling procedures are not part of the natural model building scheme, the variability in estimation of the segments is inherently ignored in the subsequent inference.
All of the foregoing calling and segmentation methods are formulated for single-array CGH profiles and do not explicitly address the problem of detecting shared patterns of aberration within a common group of patients. Shared copy number aberrations (CNAs) define patterns that provide a molecular characterization of a common group phenotype, potentially detecting a disrupted genetic process. A common strategy for detecting CNAs involves a two-step approach, first making gain/loss calls on individual arrays/samples using single-array approaches and then inferring common regions of alteration in which the frequency of alteration exceeds some specified threshold (Aguirre et al. 2004; Diskin et al. 2006). This two-stage approach has two key drawbacks. First, preprocessing each sequence separately might remove information by smoothing over short or low-frequency signals that characterize the population (Shah et al. 2007), and consequently some shared CNAs might be misdiagnosed as experimental noise. Second, it underutilizes the information in the data because it fails to borrow strength across samples when determining regions of copy number change. By modeling the samples together, it is possible to gain power for detecting shared regions of alteration when using a model that effectively reduces the noise level while reinforcing shared signals across samples. This increase in power may yield improved sensitivity for detecting shared CNAs, especially the changes of small magnitude present in a high proportion of samples in the population or changes in the presence of high noise levels.

As an illustration, Figure 1 plots aCGH samples from a real lung cancer dataset analyzed in Section 6. The log2 ratios are plotted against their genomic location from 1 to 50 Mb in the p-arm on chromosome 1 for six samples from a subtype of lung cancer. To exemplify our approach, we focus our attention on two areas of the genome, 2–3 Mb and 38–40 Mb, marked by two parallel vertical dashed gray lines toward the left and right of the $x$-axis, respectively, which appear to exhibit CNAs (mostly gains). Only four samples seem to exhibit a clear gain in copy number around the 2–3 Mb location (samples 1, 4, 5, and 6, with the samples numbered 1–6 from top to bottom), whereas only three samples seem to exhibit a clear gain around the 38–40 Mb location (samples 2, 3, and 6). Combined with this variable frequency of aberrations in samples is the fact that the size of the aberrations (i.e., vertical height) differs from...
sample to sample. We aimed to borrow strength across samples in a statistically coherent manner, to detect such patterns of shared aberrations. Figure 9 in Section 6 shows the corresponding plot of posterior probabilities of shared aberrations (gains and losses) plotted as a function of the genome location, corresponding plot of posterior probabilities of shared aberrations (gains and losses) plotted as a function of the genome location, and the subject-specific curves \( \alpha_{gij}(\cdot) \) are the random-effect functions, representing the patterns of subject-to-subject variation.

Before fitting this model, we need to consider representations for the group mean, random-effect, and residual error functions. Here we use a basis function approach (Ramsay and Silverman 2005), representing each of the functions as a sum of basis coefficients and basis functions. For ease of exposition, we drop the subscript \( g \) from our ensuing discussion and concentrate on a single group analysis. We model (1) via low-dimensional basis function projection as (ignoring group ordering \( g \))

\[
Y_{ij} = \sum_{k=1}^{K} b_k(X_{ij})\beta_k + \sum_{k=1}^{K} b_k(X_{ij})b_{ik} + \epsilon_{ij},
\]

where the following definitions and model assumptions are made: \( b_k(\cdot), k = 1, \ldots, K \), are the basis functions used to represent both the group mean function \( \mu_g(\cdot) \) and random-effect functions \( \alpha_{gij}(\cdot) \) in (1), with corresponding basis coefficients \( \beta_k \) and \( b_{ik} \). The measurement errors \( \epsilon_{ij} \) are assumed to be Normal(0, \( \sigma^2_{\epsilon_i} \)) and uncorrelated with \( \beta_k \) and \( b_{ik} \). Thus we assume that the error variance \( \sigma^2_{\epsilon_i} \) varies from patient to patient, accounting for the between-patient variability. Our model can easily accommodate different kinds of error structures, such as autoregressive (AR) errors and robust estimation via t-distributed errors, but we do not pursue these structures here. We assume that the random-effects coefficients, \( b_i = (b_{i1}, \ldots, b_{ik})^T \), are normally distributed with mean 0 and covariance \( \sigma^2_{\epsilon_i} \Sigma_b \). This admits the following covariance structure on the within-array functional observations: \( \text{cov}(Y_i) = \sigma^2_{\epsilon_i} (B_i(X_i)\Sigma_b B_i^T(X_i) + I) \), where \( Y_i = (Y_{ij1}, \ldots, Y_{ijn})^T \) and \( B_i(X_i) \) is the \( n \times K \) basis matrix corresponding to subject \( i \).

General classes of basis functions, including splines (Guo 2002) and wavelets (Morris and Carroll 2006), have been used with functional mixed models, but here we take a different approach. Motivated by the underlying biology of the data, we use piecewise-constant basis functions with endpoints stochastically determined by the data. Note that the use of Haar wavelets (Hsu et al. 2005) induces piecewise-constant basis functions, but the corresponding change points are constrained to occur at specific locations that involve splitting the genomic domain by a factor of two. Our choice is more flexible, allowing change points at arbitrary locations as suggested by the data.
Suppose that the ordered genome locations $X \in A$ are bounded by lower and upper elements $A^{[0]}, A^{[w]}$. Suppose that $A$ is partitioned into $K$ disjoint sets, such that $A = \bigcup_{i=1}^{K} \Delta_i$ and $\Delta_i \cap \Delta_j = \emptyset$ for $i \neq j$. The partition is determined by $(K-1)$ ordered change points $\hat{c} = \{c_1, \ldots, c_{K-1}\}$, where $c_1 < c_2 < \cdots < c_{K-1}$, such that $\Delta_1 = [A^{[0]}, c_1], \Delta_2 = (c_1, c_2], \ldots, \Delta_K = (c_{K-1}, A^{[w]}].$ The basis function $B_i(X_{ij})$ is then defined to be 1 if $X_{ij} \in \Delta_i$ and 0 otherwise. This is a “zero-order basis” that strongly smooths or borrows strength across all observations within a common segment.

Because our goal in this analysis is to detect common regions of aberrations across samples characterizing the population, we use the same set of basis functions, and thus the same change points, in both the population average profile and the array-specific profiles. This provides computational feasibility for these high-dimensional data and effectively allows the model to borrow strength across samples in determining the shared regions. As we discuss in Section 2.2, we start with a large number of potential change points and allow the data to determine which to include in the modeling. Although the arrays from the same group share the same segmentation, the segmentation differs among groups to account for the heterogeneity in the aberrations over different types (and/or stages) of cancer. We also assess which genome regions have different aberrations among the groups by comparing group-level mean alterations.

2.2 Prior Specification

2.2.1 Distribution of the Population Profile. In the previous section we outlined a segmentation model for a sample of aCGH profiles. Here we introduce the calling of aberrations in the population or “master” profile via a hierarchical formulation. Because the aCGH profiles can be thought of as a mixture of three generic copy number states—copy number deletion. Because the aCGH profiles can be thought of as a mixture of three generic copy number states—copy number deletion. Because the aCGH profiles can be thought of as a mixture of three generic copy number states—copy number deletion.

Here we introduce the calling of aberrations in the population-level mean encompassing the expected maximum range of the log ratios. The lower limits, $\epsilon_-$ and $\epsilon_+$, are set to small constants to determine the boundaries, such that the mean of segments, $\beta_k$, are classified as a gain or a loss. Following Guha, Li, and Neuberg (2008), who suggested values of $(\epsilon_-, \epsilon_+)$ between [0.05, 0.15], we set the values to be 0.1 for all of the analyses presented here.

The variance of the normal distribution in the mixture, $\delta^2$, controls the spread of normal distribution and can be either fixed or estimated. Letting $\delta^2 \rightarrow 0$ leads to a point mass at 0 that does not overlap the adjoining components. In this article we estimate $\delta^2$ by specifying an inverse-gamma prior with a (small) mean set to $E(\delta^2) = \max(\epsilon_-, \epsilon_+)$. This normal/uniform mixture, depending on the values of the $\kappa$’s and $\delta^2$, can lead to lighter or heavier tails compared to the normal distribution. Because in this application we are interested in heavier tails than normal, we impose the constraint that $(\kappa_-, \kappa_+) > \delta^2$.

The mixture parameters $\lambda_k = \{\lambda_k^-, \lambda_k^0, \lambda_k^+\}$ (for each segment $k$) follow an independent multinomial distribution as $\lambda_k \sim \text{Multi}(1, \pi_-, \pi_0, \pi_+)$. The associated vector of probabilities $\pi = (\pi_-, \pi_0, \pi_+)$ follows a Dirichlet distribution (in saturated form) as $\text{Dir}(\pi_{10}, \pi_{20}, \pi_{30})$. A plot of this mixture prior is shown in Figure 2, where the individual components are $\lambda(-4, -0.1)$ in red, $N(0, 0.10)$ in gray, and $\lambda(0.1, 4)$ in green. The convolved prior using equal weights is plotted as a solid black line.

Such normal/uniform mixtures have been used in clustering (Fraley and Raftery 2002) and for gene expression data (Parmigiani et al. 2002). In essence, the mixture distribution on the population-level coefficients $\beta_k$ implies that the master sequence of array CGH profiles arise from a discrete mixtures of
gains, losses, or neutral states. Essentially the calls, as an aber-
ration or not, are done at the population level rather than at the
sample level, because our interest lies in the detection of shared
regions of aberration across samples.

2.2.2 Prior Specification of Change Point Configurations.
As alluded to in Section 1, one of the key goals of aCGH analy-
sis is to find the number and location of change points, and thus
the shared regions of common CNAs. With \( n \) probes per chro-
mosome, the possible number of change points is \( n - 1 \), which
leads to \( 2^{(n-1)} \) possible change points configurations. Modern
aCGH arrays typically have on the order of thousands of probes
on a given chromosome, and thus an exhaustive search for the
optimal configuration is typically not computationally feasible.
There are basically two approaches to this problem. One
approach is to start with a large (fixed) number of segments
and control overfitting via an explicit penalty added to the like-
lihood. The optimal configuration of change points is then de-
termined using an empirical criterion. In practice, a heuristic
methods (Tibshirani and Wang 2005) or complex (Picard 2005)
models. Alternatively, a more exact approach is to treat the number and locations of the
change points/knots as random variables and conduct a Markov
chain Monte Carlo (MCMC)-based stochastic search over the
posterior space to discover configurations with high posterior
probabilities. We take this latter approach here.

Depending on the resolution of the array used and other
information, we may have a prior expectation of the distribution of the number of segments, which can be used to set the hyper-
parameters of the prior on \( K \), such as Poisson(\( K \mid \gamma \)), where \( \gamma \)
is the prior expectation of the number of segments, and density
\( \gamma^K \text{exp}(-\gamma)/K! \). This prior was originally adopted by Green
(1995) on the number of model components in a different con-
text. Another option is the negative binomial distribution, a
gamma mixture of Poisson that is more flexible than the Poisson
distribution, which has only one parameter that controls both
the mean and the variance. However, in the absence of such
information, we can set a flat prior on \( K \), such as the discrete
uniform prior \( U(0, \ldots, K_{\text{max}}) \), where \( K_{\text{max}} \) is an upper limit on the
number of change points expected in the data. We found that
our posterior inference is insensitive to the choice of the prior on \( K \) (see Supplementary Figure 1), and thus we use it as a
default specification in all of our analyses.

Therefore, the joint prior on \((\hat{c}, K)\) is given by

\[
p(\hat{c}, K) = \frac{1}{K_{\text{max}} + 1} \times \frac{1}{K}
\]

for \( K = 0, \ldots, K_{\text{max}} \), where \( K = \text{dim}(\hat{c}) \) is the number of ele-
ments in \( \hat{c} \) and \( T = |T| \), is the size of the candidate set of change
point locations \( T \). The first term in the prior ensures that
each configuration of change points of dimension \( K \) has equal
weight. Thus we assume that given \( K \), any set of change points
is found by sampling \( K \) items from a candidate set \( T \) without
replacement. This in turn ensures that the elements of \( \hat{c} \) are dis-
tinct and \( K_{\text{max}} \leq T \). The second term assumes that each possi-
ble dimension \( K \) is equally likely. Although in theory the num-
ber of candidate change points can be set equal to the number
of probes on the chromosome, this might not be computa-
tionally feasible given the resolution of current aCGH arrays. Thus,
from a practical standpoint, we need to restrict the set of can-
didate change points while maintaining flexibility in estimat-
ing the segmented profiles. In our implementation, we obtain a
candidate set of change points by first applying a segmentation
method to each individual profile \( i \) to obtain a set of individual-
level segments \( T_i \), and then taking the union of these individual
segments to obtain our candidate change point set \( T = \bigcup T_i \).
In the individual segmentations, we recommend choosing the
tuning parameters conservatively, erring on the side of over-
segmenting rather than undersegmenting. Further details of the
implementation procedure are provided in the Supplementary
Materials.

2.2.3 Priors on Variance Components. One of the key is-
ues from both a practical and a methodological standpoint is the modeling of the random-effect variance \( \Sigma_b \), which is
of dimension \( K \) and thus involves estimation of \( K(K + 1)/2 \) parameters if left unstructured. We assume a diagonal struc-
ture, \( \Sigma_b = \text{diag}(\sigma_{b1}^2, \ldots, \sigma_{bk}^2) \). While assuming independence
between segments, this structure also accounts for the corre-
lation between markers within the same segment and allows sep-
ate subject-to-subject or array-to-array variances for different
segments. To aid conjugacy, we assume independent inverse-
gamma priors for the individual elements of \( \Sigma_b \) and the error
variances \( \sigma^2 = [\sigma_{e1}^2, \ldots, \sigma_{eM}^2] \) with the hyperparameters set to
(1, 1) to impart little information.

3. POSTERIOR COMPUTATION VIA MCMC

We fit this fully specified Bayesian model using MCMC tech-
niques (Gilks, Richardson, and Spiegelhalter 1996). Because we
are allowing the number of change points to be random, the
dimension of our parameter space varies in each MCMC
iteration; thus we use the reversible-jump MCMC (RJMCMC; Green
1995). Our RJMCMC sampler involves three kinds of moves: BIRTH, in which we add a new segment; DEATH, in which
we delete a segment location; and MOVE, in which we relo-
cate a segment location, with corresponding prior probabil-
ities \( (p_B, p_D, p_M) \), where \( p_M = 1 - (p_B + p_D) \). Our RJMCMC
algorithm proceeds by iterating among the following steps:

(a) Initially, select \( K \) change points and location parameters
\( \hat{c}_K \).

(b) Generate a Uniform(0, 1) random number \( U \).

(i) If \( U < p_B \), perform the BIRTH step.

(ii) If \( p_B < U < p_B + p_D \), perform the DEATH step.

(iii) Otherwise perform the MOVE step.

(c) Update other model parameters, \( M = \{b, \beta, \lambda, \Sigma_b, \sigma_e^2\} \),
from their full conditionals.

Due to the conjugacy in our model, it is possible to integrate
out the random-effect parameters when updating the segment
parameters in the reversible step of the algorithm, resulting in
fast calculations and a MCMC sampler with good mixing prop-
erties. The full MCMC scheme and the full conditionals are
included in the Supplemental Materials. Usual convergence di-
agnostic methods, such as that of Gelman and Rubin (1992),
do not apply here because we are moving within a (potentially)
infinite model space and the parameters are not common to all
models. Instead we assess MCMC convergence via trace plots of $K$ and the log-likelihoods, which have a coherent interpretation throughout the model space (Brooks and Giudici 2000). Detailed information on implementation and convergence assessment of our MCMC algorithm can be accessed via the Supplemental Materials.

4. FDR–BASED DETERMINATION OF SHARED ABERRATIONS

The MCMC samples explore the distribution of possible change point configurations suggested by the data, with each configuration leading to a different segmentation of the population-level aCGH profile. Some change points that are strongly supported by the data may appear in most of the MCMC samples, whereas others with less evidence may appear less often. There are different ways to summarize this information in the samples. One could choose the most likely (posterior mode) change point configuration and conduct conditional inference on this particular segmentation. The benefit of this approach would be the resulting single set of defined segments, but the drawback is that the most likely configuration might still appear in only a very small proportion of MCMC samples. Alternatively, one could use all of the MCMC samples and, using Bayesian model averaging (BMA) (Hoeting et al. 1999), mix the inference over the various configurations visited by the sampler. This approach better accounts for the segmentation uncertainty in the data, leads to estimators of the mean population aCGH profile $\mu(x)$ with the smallest mean square error, and should lead to better predictive performance if class prediction is of interest (Raftery, Madigan, and Hoeting 1997). We use this Bayesian model averaging approach.

To summarize the overall population-level mean aCGH profiles $\mu(x)$, we compute the posterior mean of each $\mu(x)$ across all samples. Recall that by our prior structure (3), for each iteration of the MCMC a certain number of markers $x \in \Delta_{k}$, for which $\lambda_{k} = 1$ are considered copy number neutral and will have $\mu(x) = 0$. We can define a marker-based indicator of gain, loss, or neutral state, $\lambda(x) = 1$, $-1$, or 0 if $x \not{\in} \Delta_{k}$ and $\lambda_{k} = 1$, $\lambda_{k} = 1$, or $\lambda_{k} = 1$, respectively. From this, we compute $p_{-} = P(\lambda(x) = 0 | Y)$, $p_{+} = P(\lambda(x) = 1 | Y)$, and $p_{0} = P(\lambda(x) = 0 | Y)$ $= 1 - p_{-} - p_{+}$ to summarize the probability of the population average copy number state being a loss, a gain, or neutral, respectively, for the marker at position $x$. These can be displayed as probability plots as a function of genomic location $x$, with the vertical axis plotted on a logit scale to more clearly show the endpoints of the [0, 1] interval; see Figure 7 in Section 6, which plots $p_{+}$ and $p_{-}$ in green and red, respectively.

We then assume that any marker at genomic location $x$ with $p_{0}(x) < \phi$, for some threshold $\phi$, contains a true shared alteration in the population of interest. Let $\mathcal{X}_{\phi} = \{x : p_{0}(x) < \phi\}$ represent the set of all genomic locations considered shared aberrations. Note that $p_{0}(x)$ summarizes the posterior probability that marker $x$ is in fact not a shared aberration and thus is a Bayesian $q$-value, or an estimate of the local FDR (Storey 2003; Newton et al. 2004) and is appropriate for correlated data, as shown by Morris et al. (2008) and Ji et al. (2007). The significance threshold $\phi$ can be determined based on classical Bayesian utility considerations (such as those in Müller et al. 2004), based on the elicited relative costs of false-positive and false-negative errors, or can be set to control the average Bayesian FDR. For example, suppose that we are interested in finding the value $\phi_{a}$ that controls the overall average FDR at some level $\alpha$, meaning that we expect only $100\alpha\%$ of the markers declared as shared aberrations to actually be false-positive.

For all markers $x_{j}, j = 1, \ldots, n$, we first sort $p_{i} = p_{0}(x_{i})$ in ascending order to yield $p_{(j)}, j = 1, \ldots, n$. Then $\phi_{a} = \phi_{(j)}$, where $\xi = \max\{j : j < n \sum_{j=1}^{j} p_{(j)} \leq \alpha\}$. The set of regions $\mathcal{X}_{\phi_{a}}$ can then be claimed to be shared aberrations based on an average Bayesian FDR of $\alpha$. These regions can be marked on the probability plot in a different color to set them apart from the neutral regions.

The posterior samples also can be used to perform FDR-based inference to determine differential aberrations between different populations by using the FDR-based pointwise functional inference approach set to control the average Bayesian FDR (as in Morris et al. 2008). Suppose that we are interested in detecting regions of the genome with differences in average genomic alteration of at least $15\%$ between two groups. After running separate Bayesian hierarchical segmentation models for each group, we take the posterior samples for the mean aCGH profiles for the two groups, say $\mu_{1}(x)$ and $\mu_{2}(x)$, and at each position $x$ containing a marker to compute the posterior probabilities of at least a $1.15$-fold difference between the means, which is $p_{12}(x) = P(|\mu_{1}(x) - \mu_{2}(x)| > \log_{2}(\delta))$ for $\delta = 1.15$. These quantities measure the probability that the two groups have mean aCGH profiles that differ by at least $15\%$ at position $x$ in the genome. The quantities $1 - p_{12}(x)$ are then $q$-values for assessing differential aberrations between the two populations because they measure the probability of a false-positive if position $x$ is called a “discovery,” defined as a region with at least a $1.15$-fold difference in the population aCGH profiles. A threshold $\phi_{12,\alpha}$ on the posterior probabilities can be determined so that markers with $1 - p_{12}(x) < \phi_{12,\alpha}$ are flagged while controlling the expected Bayesian FDR at level $\alpha$, as described earlier. Probability plots can be generated for each group comparison to highlight the probability of each genomic region having different aberrations (see Figure 8 in Section 6).

5. SIMULATIONS

We performed simulation studies to evaluate the operating characteristics of our method under various scenarios and for comparison with other approaches in the literature. We generated a series of aCGH data sets with prespecified known regions of aberration of various sizes and prevalences and white noise added. We generated 20 aCGH profiles consisting of 2000 markers, with 10 regions of aberration. These 10 regions included one region of loss and one region of gain for each of five prevalence levels $\omega \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, representing the proportion of samples in the population containing this aberration. For example, an aberration with $\omega = 0.2$ would appear in $20\%$ of the samples, whereas an aberration with $\omega = 1.0$ would appear in all of the samples. We generated the widths of shared aberrations from a gamma distribution with parameters $(a,b)$ and mean $a/b$. We set $(a,b) = (2.5, 0.05)$ such that the mean of the distribution was 50 and the 99% interval corresponded to $(5, 168)$ rounded to the nearest integer. Thus the range of shared
aberrations could vary substantially, accommodating both large and short segments.

We then added white noise to these noiseless aCGH profiles. The estimated value of the noise variance, $\tau$, in the real aCGH data introduced in the next section was around 0.2. To investigate scenarios with higher or lower noise, we varied the noise variance within the range $\tau = \{0.1, 0.2, 0.3\}$, corresponding to the low, medium, and high levels of noise in the log2 ratios. We drew the effect sizes for individual gains and losses uniformly from $[0.1, 0.25]$. Considering these effect size distributions, this yielded signal-to-noise ratios (SNR) in the ranges of $[1, 2.5]$ for the low-noise scenarios, $[0.5, 1.25]$ for the medium-noise scenarios, and $[0.33, 0.83]$ for the high-noise scenarios. Because our effect sizes were not the same across the genome, our SNR varied across individual profiles, as is typical in aCGH data. Figure 3 shows three aCGH profiles for each of the noise scenarios. The signal is increasingly blurred with increasing noise variance, and the aberrations in the test cases appear realistic and nontrivial. We generated 10 datasets for each value of $\tau$, for a total of 30 datasets with 20 profiles each.

We fit the BDSAcgh model with default priors and parameterizations as described in Section 2.2, except that we set $\sigma_{\epsilon_i}^2 = \sigma^2$ for all $i$, that is, common variance across all arrays. We compared our method with two approaches for estimating copy numbers for multiple samples, the cghMCR algorithm of Aguirre et al. (2004) and the hierarchical hidden Markov model (H-HMM) of Shah et al. (2007).

The cghMCR algorithm locates the minimum common regions (MCRs), regions of a chromosome showing common gains/losses across array CGH profiles derived from different samples. In this algorithm, the profiles are first segmented individually, and highly altered segments are then compared across samples to identify positive or negative valued segments. MCRs are then defined as contiguous spans having at least a recurrence rate defined by a parameter (recurrence) across samples that is calculated by counting the occurrence of highly altered segments. This is an example of a two-step approach where segmentation is done at the sample level independent of the calling. For our analysis, we used the R package—cghMCR, available from the Bioconductor project at http://www.bioconductor.

![Figure 3](image-url)
segmentation for cghMCR using the CBS algorithm of Olshen et al. (2004) for the individual samples, with tuning parameter \( \alpha = 0.01 \). The cghMCR is controlled by three user-defined parameters: (1) upper and lower threshold values of percentile above and below which the segments are identified as altered, (2) the number of base pairs that separate two adjacent segments (gap parameter), and (3) the rate of recurrence for a gain or loss that is observed across samples. We fix the gap parameter and rate of occurrence and vary the threshold to compute the sensitivities and specificities. We set the gap parameter to 50, which was the mean length of the altered segments that we considered in our simulation. We set the rate of recurrence to 50%, which corresponds to a central location in our range of aberration prevalences, \( \omega \).

The H-HMM model extends the single-sample HMM to multiple samples to infer shared aberrations, by modeling the shared profile by a master sequence of states that generates the samples. The H-HMM is closer in spirit to BDSAcgh in terms of borrowing strength across samples to infer shared regions of aberrations, but is based on different probability model: the hidden Markov model. The H-HMM model assumes that the samples are conditionally independent given an underlying hidden state and follows a Gaussian observation model. The hidden states in the H-HMM are loss, gain, neutral, and undefined, and the probability of being at particular state is estimated by pooling information across samples. The model parameters are estimated using an MCMC algorithm. For the H-HMM model, we used the MATLAB implementation of the method provided at http://people.cs.ubc.ca/~sshah/acgh/index.html. Implementation details of the H-HMM method are provided in the Supplemental Materials.

The cghMCR, H-HMM, and BDSAcgh methods flag regions of the genome as shared aberrations based on the chosen thresholds: upper and lower thresholds for cghMCR and posterior probabilities for the latter methods. Varying these parameters across their ranges (0.01–0.50 for cghMCR and 0–1 for H-HMM and BDSAcgh), we constructed receiver operating characteristic (ROC) curves that summarize each method’s ability to correctly detect the true shared aberrations in the simulated data sets. At each threshold parameter value, we computed the sensitivity (true positive rate) and 1-specificity (false-positive rate) of the shared aberration detection by computing the proportion of truly aberrated probes that were detected by the method and the proportion of probes that were not aberrated but were mistakenly deemed to be so by the method. Figure 4 shows the overall averaged ROC curves across all simulation runs and values of prevalences (\( \omega \)) for cghMCR (in blue), H-HMM (in red), and BDSAcgh (in black) for the three noise levels (top to bottom). As a measure of performance, we calculated the area under the curve (AUC; Fawcett 2006) for each ROC curve, as shown in Table 1, broken down for the three noise levels. The third column displays the mean overall AUCs, with standard errors in parentheses.

The BDSAcgh consistently outperformed cghMCR in all noise scenarios, with the difference in performance increasing with increasing noise level in the data. The \( p \)-value for a two-sided Student \( t \) test for the difference between the AUC for the two methods was \( < 10^{-6} \) for all noise levels. H-HMM performed marginally better than BDSAcgh in the low-noise scenario, but BDSAcgh performed consistently better in the medium- and high-noise scenarios. For \( \tau = 0.3 \) H-HMM performed worse than cghMCR in terms of AUC. To focus on the region of the ROC curve of most interest, we also compared the partial area under the AUC curve, truncated at the 1-specificity value of 0.20 and normalized to be on a \([0, 1]\) scale (AUC\(_{20}\)), as given in the fourth column of Table 1. The relative results from the three methods are similar to the overall AUC results, with BDSAcgh outperforming cghMCR for all noise scenarios and outperforming H-HMM in the medium- and high-noise scenarios.

To explore the performance of the methods as a function of the prevalence of aberration, we plotted the ROC curves for varying values of the prevalence (\( \omega \)) (Figure 5). In this figure, the columns correspond to the varying noise levels in the data, with the leftmost column for \( \tau = 0.1 \), the middle column for \( \tau = 0.2 \), and the rightmost column for \( \tau = 0.3 \). The five rows, from top to bottom, correspond to increasing values of \( \omega = (0.1, 0.2, 0.3, 0.4, 0.5) \). The corresponding mean AUC and AUC\(_{20}\) for each level of the prevalence in shown in the upper and lower panels of Figure 6, respectively. The red bars represent BDSAcgh, the orange bars represent cghMCR, and the yellow bars represent H-HMM, with the whiskers indicating the \((\pm 1)\) standard errors. Several interesting features can be deduced from this figure. First, BDSAcgh outperforms cghMCR, most strongly for aberrations with low/medium prevalences of \((0.4, 0.6)\) and in low-SNR scenarios (\( \tau = 0.2, 0.3 \)). For higher prevalences, two methods show very similar results, as expected because in those circumstances most shared aberrations are not too difficult to detect. BDSAcgh is more robust in low-SNR scenarios than cghMCR. In the lowest-prevalence group (\( \omega = 0.2 \)), both methods performed poorly for low SNR (\( \tau = 0.2, 0.3 \)); however, BDSAcgh performed quite well and remarkably better than cghMCR in the high-SNR scenario (\( \tau = 0.1 \)). For the high-SNR scenario, H-HMM performed remarkably well for all prevalences, but its performance deteriorated under low SNR (\( \tau = 0.2, 0.3 \)), with moderate increases in performance with increasing prevalence. In terms of AUC\(_{20}\), the relative performance of the methods was same as that of AUC, but cghMCR performed better than H-HMM, especially at the low SNR scenario (\( \tau = 0.3 \)).

In conclusion, our simulation studies suggest that our method outperforms the neo-stage cghMCR approach for detecting shared CNAs, yielding larger areas under the ROC curves for all noise levels studied here, with the greatest differences in higher-noise settings. Separated out by prevalence, BDSAcgh has dramatically greater sensitivity than cghMCR in lower-prevalence settings. BDSAcgh’s greater sensitivity likely stems from the fact that it jointly models all arrays together and borrows strength between arrays in detecting the shared aberrations. In contrast, cghMCR involves segmenting the individual arrays separately and then comparing the segments across samples. Furthermore, because our method is based on a unified hierarchical model, it appropriately accounts for the variability of change point configuration and segmentation, as well as array-to-array variability in our inference. Thus the probabilities summarizing our level of evidence of aberration are based on all of these sources of variability, which is another possible explanation for our method’s greater accuracy in calling the
truly aberrated regions compared with H-HMM, which models only measurement error and does not account for sample-to-sample variability.

We also assessed the performance of the cghMCR algorithm by varying the tuning parameters and found that the results were not robust to the misspecification of these parameters. In particular, we varied the rate of recurrence and \( \alpha \) (the parameter that controls the number of segments in CBS; the higher the \( \alpha \), the greater number of segments). We varied the rate of recurrence across five levels (0.2, 0.4, 0.5, 0.8, and 1) and varied \( \alpha \) across five levels (0.01, 0.05, 0.2, 0.5, and 0.9). The corresponding results are given in Table 2 in the Supplemental Materials. In brief, the performance of the cghMCR algorithm was somewhat robust to specification of \( \alpha \) and produced slightly better results at lower rates of recurrence (0.2 and 0.4) but was dramatically worse at higher values (0.8, 1), especially in the high-noise scenario. This might be due to the fact that in the cghMCR algorithm, the minimal common regions are found using a predetermined cutoff of recurrence rate across samples that are not informed by borrowing strength across samples. Thus the results are not robust to misspecification of the recurrence rate, especially in low-SNR scenarios, whereas in contrast, our proposed method does not require specification of any such parameters.

We also assessed the performance of our method as a function of

---

Figure 4. Simulation study: Overall ROC curves across all values of prevalences (\( \omega \)) for the three methods cghMCR (in blue), H-HMM (in red), and BDSAcgh (in black). The vertical axis is the sensitivity (true positive rate) and the horizontal axis is 1-specificity (false positive rate). (a) \( \tau = 0.1 \), (b) \( \tau = 0.2 \), and (c) \( \tau = 0.3 \).
Table 1. Simulation study: Shown here are the mean area under the curves (AUCs) for the three methods, cghMCR, H-HMM, and BDSAcgh for various levels of noise (τ). The 3rd column shows the full AUCs and the 4th column shows the partial area under the curves (AUC_{20}) truncated at false positive rate of 0.2. Also shown are the standard errors in parentheses.

<table>
<thead>
<tr>
<th>τ</th>
<th>Method</th>
<th>AUC</th>
<th>AUC_{20}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>cghMCR</td>
<td>0.8326</td>
<td>0.7170</td>
</tr>
<tr>
<td></td>
<td>H-HMM</td>
<td>0.9746</td>
<td>0.8870</td>
</tr>
<tr>
<td></td>
<td>BDSAcgh</td>
<td>0.9599</td>
<td>0.8644</td>
</tr>
<tr>
<td>0.20</td>
<td>cghMCR</td>
<td>0.7072</td>
<td>0.5111</td>
</tr>
<tr>
<td></td>
<td>H-HMM</td>
<td>0.7571</td>
<td>0.3825</td>
</tr>
<tr>
<td></td>
<td>BDSAcgh</td>
<td>0.9450</td>
<td>0.7872</td>
</tr>
<tr>
<td>0.30</td>
<td>cghMCR</td>
<td>0.6755</td>
<td>0.4306</td>
</tr>
<tr>
<td></td>
<td>H-HMM</td>
<td>0.6160</td>
<td>0.1636</td>
</tr>
<tr>
<td></td>
<td>BDSAcgh</td>
<td>0.9149</td>
<td>0.7554</td>
</tr>
</tbody>
</table>

of the segment length and found that BDSAcgh has greater sensitivity than cghMCR in detecting short- and medium-length segments, especially in the high-noise scenario. These and additional simulation results can be assessed via the Supplemental Materials.

To assess the calibration of the Bayesian FDR threshold that we use to determine significant CNAs, for each simulated data set we computed the threshold corresponding to a nominal FDR = 0.10 and then assessed the true FDR of the regions flagged according to this rule. We found that across the 10 data sets, the median true FDR was 0.0993, with an interquartile range of [0.0789, 0.1481], suggesting that the Bayesian FDR estimates were well calibrated.

6. DATA ANALYSIS

We applied the BDSAcgh method to a lung cancer data set originally published by Coe et al. (2006) and Garnis et al. (2006), available at http://sigma.bccrc.ca/. The data consist of aCGH samples from 39 well-studied lung cancer cell lines. The samples are subdivided into four subgroups of small-cell lung cancer (SCLC) and nonsmall-cell lung cancer (NSCLC): NSCLC adenocarcinoma (NA; n = 18), NSCLC squamous cell carcinoma (NS; n = 7), SCLC classical (SC; n = 9), and SCLC variant (SV; n = 5). This dataset has been studied in depth, and shared patterns have been further validated in biological experiments across groups (Coe et al. 2006; Garnis et al. 2006). The prior and hyperprior settings that we used to analyze these data are the same as those reported in Section 3. We ran 10,000 MCMC samples after a burn-in of 5000 samples, at which point our chains had converged reasonably. We fit each of the groups separately using our proposed method and compared our results with those reported by Coe et al. (2006). We analyzed the aCGH profiles from chromosome 1 and 9 to illustrate our method.

Figure 7 shows the posterior probabilities for chromosome 9 of shared aberrations as a function of the genomic position, p_+(x) and p_-(x) for gain (green) and loss (red), respectively, for the population-level profile for the four phenotypic groups NA, NS, SC, and SV. The horizontal blue dashed line is the threshold on the posterior probabilities 1 − φ_{0.10} that controls the expected Bayesian FDR at 0.10, as described in Section 4, which were φ_{0.10} = \{0.3752, 0.3613, 0.3353, 0.2974\} for the respective phenotypic groups. Any probes with 1 − φ_{0.10} > 1 − φ_{0.10} were then flagged as significant aberrations within their group. The probes that exhibit a gain are shown in green, those that exhibit a loss are in red, and nonsignificant probes are in gray. Note that the patterns of shared aberrations differed quite significantly across the various cancer subtypes, with copy number losses seen mostly in groups NA, NS, and SV and copy number gains seen mainly in group SC. In groups NA and SV we found a loss of copy number in a significant portion of chromosome 9, as was reported by Coe et al. (2006) and Garnis et al. (2006), who used this chromosome as an example.

To follow up on these results, we constructed a list of the genes at genomic locations exhibiting a shared aberration using the MapViewer tool from the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/projects/mapview), and explored their known relationships to cancer or lung cancer using the Online Mendelian Inheritance in Man (OMIM) database and the PubMed database on the NCBI website. Table 2 summarizes the total number of genes with losses (gains) in each phenotype group with known links to lung cancer (LR) or to cancer in general (CR), or with a function that is either unknown or unrelated to cancer (NR/U). Table 3 lists the genes that we found to be directly related to lung cancer for each of the cancer phenotypes. We found a total of 34 genes within the four phenotype groups. Shah et al. (2007) identified only two genes related to lung cancer in chromosome 9 as having CNA, namely CA9 (identified as gain) and CDKN2A (identifed as loss). Here gene CA9 was identified as belonging to a region of gain only for cancer type I, and gene CDKN2A was identified as belonging to a region of loss for cancer types NA, NS, and SV.

To investigate the differences between the four lung cancer subtypes in more depth, we plotted the contrasts between the four groups in Figure 8. The left panel of the figure plots the difference between the posterior mean population mean curves, i.e., \beta_i(x) − \beta_j(x) for groups (i, j) as a function of the genomic location x, along with the 95% (pointwise) credible intervals. We are interested in detecting probes whose mean copy number profiles differ between groups by more than 15%, which we consider meaningful, from a practical standpoint. The green dashed lines indicate 15%, or 1.15-fold differences, between the groups. The right panel contains the pointwise posterior probabilities of at least a 1.15-fold difference between the phenotypic groups, with red and green marking the probes identified as differentially aberrated (exceeding a threshold 1 − φ_{0.10} to control the Bayesian FDR at 0.10, the blue dashed line). We find that the group SC is the most different from the other phenotypic groups, because a large number of probes are differentially aberrated. This is not surprising because most areas in chromosome 9 of group SC exhibit a gain in copy number while the other groups had mostly losses of copy number. These results suggest a strongly different gene copy number profile for classical small cell lung cancer and its variants and the nonsmall cell lung cancers.

To qualitatively compare the relative performance of our method with the performance of the cghMCR, we investigated the aCGH profiles from chromosome 1 for the SC group. The posterior probability plot of the shared aberration for chromosome 1 for the SC group, presented in Figure 9, reveals...
Figure 5. Simulation study: ROC curves for the three methods cghMCR (in blue), H-HMM (in red), and BDSAcgh (in black) broken down by prevalence ($\omega$). The vertical axis in each plot is the sensitivity (true positive rate) and the horizontal axis is 1-specificity (false positive rate). The columns correspond to the varying noise levels in the data with (a) column for $\tau = 0.1$, (b) column for $\tau = 0.2$, and (c) column for $\tau = 0.3$. The 5 rows from top to bottom correspond to increasing $\omega = (0.1, 0.2, 0.3, 0.4, 0.5)$.

a number of regions of high posterior probability of gains at the genomic locations corresponding to the following genes: TNFRSF4, TP73, TNFRSF9, E2F2, FGR, DMAP1, RAB13. This agrees with the findings by Shah et al. (2007), who also found that the expression level of these genes were known to be altered in lung cancer. We also ran the cghMCR method on this data using the default specifications (Aguirre et al. 2004); the aberration regions obtained are plotted in yellow at the top of each panel of Figure 9. For this case study, the cghMCR determined all the aberrated regions as gains with no regions being determined as losses. The cghMCR failed to detect the small frequency aberrations near the p-arm of chromosome 1 (0–40 Mb), thus failing to detect most of the genes at that location that were determined using our method. We also investigated the lengths of the segments found by BDSAcgh and a plot of the histogram of segment lengths can be found in the Supplementary Materials. We found that the minimum segment length detected was 1 and the maximum was 295 probes long, with a median length of 5 probes, showing that BDSAcgh is flexible in detecting short as well as large segments.

7. DISCUSSION AND CONCLUSIONS

We propose a novel method, the BDSAcgh, based on a Bayesian segmentation model for detecting shared aberrations in aCGH data. The model moves beyond the classical approach of segmenting individual arrays by introducing a functional mixed effects model to borrow strength between samples, to infer shared regions of aberration. Our method yields mean aberration profiles for different specified groups that can be individually analyzed using FDR-based methods to detect CNAs characterizing the population; the specified groups can be formally compared to each other to detect group differences while controlling the FDR. The results can be presented using posterior
Figure 6. Simulation study: top panel are the bar graphs of the mean area under the curves (AUCs) with the standard error bars for the BDSAcgh (in red), cghMCR (in orange), and HHMM (in yellow). The vertical axes are the mean AUC with the horizontal axes sorted by increasing prevalence. (a) $\tau = 0.1$ (low noise), (b) $\tau = 0.2$ (medium noise), and (c) $\tau = 0.3$ (high noise). Bottom panel: similar to top panel with the vertical bar being the mean AUC.

Probability plots that are highly interpretable to a practitioner because the shared regions of aberration or group differences are summarized in terms of probabilities rather than segmented means.

Our simulation studies suggest that our method outperforms the cghMCR, a two-stage approach for detecting shared CNAs, yielding larger areas under the ROC curves for all the noise levels studied here, with the greatest differences seen in higher noise settings. Separated out by prevalence, or percentage of individuals in the population with the aberration, we see that the BDSAcgh method has dramatically greater sensitivity than the cghMCR in lower prevalence settings. This is important, because the genetic heterogeneity of cancer suggests considerable variability, even within a well-defined population, and because there may be alterations with moderate to low abundance that could still be said to characterize the population.

The increased sensitivity of the BDSAcgh likely comes from the fact that it jointly models all arrays together and borrows strength between arrays in detecting the shared aberrations, while the commonly-used two-step approaches like the cghMCR involve segmenting the individual arrays separately and then comparing the segments across samples. BDSAcgh estimates the underlying mean aCGH profiles for the population using a hierarchical model, which automatically reinforces
Figure 7. Group profiles for chromosome 9: shown are the pointwise posterior probabilities of shared regions of aberration as a function of the genomic position, \( p_+ (x) \) and \( p_- (x) \), for the four phenotypic groups: (a) NA, (b) NS, (c) SC, and (d) SV. The blue dotted lines indicate the threshold for flagging a location as significant, controlling the expected Bayesian FDR to be less than 0.10. The significant locations that show a gain in copy number are shown in green, losses are shown in red and nonsignificant locations are shown in gray.

Our underlying prior structure partitions the mean profile into regions of gain, loss, and no change, automatically yielding a...
straightforward and intuitive measure from which we can infer which genomic regions are CNAs in the population while controlling the FDR. Further, the posterior samples from the Bayesian method can be used to compare different populations to assess which genomic regions are differentially aberrated between the populations.

Since our primary goal is to detect common regions of aberrations across samples characterizing the populations, we assume the same segmentation, in terms of number and change points, at both the sample and population level. This further engenders computational feasibility for such high-dimensional aCGH data and allows borrowing strength across samples since we effectively estimate one set of change points via the RJMCMC algorithm. Our method can be easily extended to the case where different samples have different number and location of change points although this would lead to substantial added computational burden since effectively we will be estimating $N + 1$ sets of change points and hence a further $N + 1$ RJMCMC steps in our MCMC algorithm, $N$ being the number of samples. This scenario would be especially useful if the goal is to cluster samples based on their aCGH profiles and we leave this task for future consideration.

Some aspects of our model could benefit from further development. We assume a Gaussian distribution for our random effects, which might be suspect, especially in the presence of outliers. Robust specifications of distribution on the random effect via parametric distributions, such as a $t$-distribution or scale mixture of normals, might be a viable alternative. Another attractive approach would be to specify a completely nonparametric distribution on the random effects and/or the overall mean levels like the Dirichlet process prior. Another advantage of our hierarchical Bayesian model is that it can easily be embedded into a larger modeling scheme involving other types of data. For example, one useful and natural extension of our Bayesian model is to jointly model gene expression data, copy number aberrations, and their relationships. These models can be used to integrate data across various sources to draw a systems-based biological inference.

**SUPPLEMENTAL MATERIALS**

**Results:** The supplemental materials contain detailed information regarding our MCMC computations and additional simulation results. In particular, we provide full conditional distributions for all model parameters, sensitivity to prior specifications, MCMC convergence and diagnostics as well as additional simulations results to supplement the results in Section 5 (supplementary_material_revised.pdf)

[Received April 2009. Revised May 2010.]

**REFERENCES**


Figure 8. Contrast of group mean profiles: the left column shows the pointwise differences of the population mean functions of the 4 cancer phenotypes: NA, NS, SC, and SV along with the 95% credible intervals. The green lines indicate the 1.15-fold difference between the mean profiles. The right column is corresponding pointwise posterior probability of a 1.15-fold difference. The blue dotted lines indicate the threshold for flagging a location as significant, controlling the expected Bayesian FDR to be less than 0.10. The significant locations that show a gain in copy number are shown in green, losses are shown in red, and nonsignificant locations are shown in gray.
Figure 9. Group profile for group SC: shown are the pointwise posterior probabilities of shared regions of aberration as a function of the genomic position, $p_+(x)$ and $p_-(x)$, for the phenotypic group SC. The blue dotted lines indicate the threshold for flagging a location as significant, controlling the expected Bayesian FDR to be less than 0.10. The significant locations that show a gain in copy number are shown in green, losses are shown in red, and nonsignificant locations are shown in gray. Also shown are significant genes identified to related to lung cancer. Marked in yellow are the corresponding locations identified as shared aberrations by the cghMCR algorithm.


