Web-based Supplementary Materials for

Segmentation and Estimation for SNP Microarrays: a Bayesian Multiple Change Point Approach

by Yu Chuan Tai, Mark N. Kvale, John S. Witte

Web Appendix A

This section extends the familial-sample model to any size pedigree. Suppose there are G generations for a pedigree, each with n_1,...,n_G members. Members at generation g are denoted by i_{g1},...,i_{gn_g}. From generations 2 to G, each member has 0, 1, or 2 parents available. The observed log_2 ratios for the k^{th} member i_{gk} are $s_{i_{gkl}}$ and $c_{i_{gkl}}$ are modeled by $s_{i_{gkl}}|\lambda_{i_{gkl}}, \sigma^2_{si_{gkl}} \sim N(\lambda_{i_{gkl}}, \sigma^2_{si_{gkl}})$ and $c_{i_{gkl}}|\lambda_{i_{gkl}}, \sigma^2_{ci_{gkl}} \sim N(\lambda_{i_{gkl}}, \sigma^2_{ci_{gkl}})$. The posterior has the same form as equation (5) in the main article. The differences are all the products over f, m, o ($\prod_{i=f,m,o}$) are replaced by $\prod_{g=1,...,G}\prod_{i=i_{g1},...,i_{gn_g}}$, and the term $\prod_{i=f,m} P(\boldsymbol{d}_i|\boldsymbol{d}_{i0})$ is replaced by the product over all members without any parent (e.g. first generation, marry-ins). The same for $\prod_{i=f,m} P(\boldsymbol{d}_{i0})$. Finally, the term $P(\boldsymbol{d}_o|\boldsymbol{d}_f, \boldsymbol{d}_m)$ is replaced by the product of conditional probabilities of all members with at least one parent. Here, we assume that the copy number of an individual is conditional independent from those ancestors at least one generation apart, given his/her parents' copy numbers.

Web Appendix B

This section describes the parameter inference for BMCP. The method proposed here is based on those in Suchard et al. (2003) and Tai et al. (2009). The algorithm comprises four move types: birth step, death step, location update, and parameter update. Their corresponding proposal probabilities are b_K, o_K, η_K, λ_K respectively, where $b_K = c \times min\{1, P(K+1)/P(K)\}$, $o_K = c \times min\{1, P(K-1)/P(K)\}$, and $\eta_K = \lambda_K = (1 - b_K - o_K)/2$. The boundary conditions are $b_{K_{max}} = 0$ and $o_0 = 0$. The constant c is set up to be as large as possible with the constraint that $b_K + o_K \leq 0.9$. We describe how we implement these four move types based on the one-sample problem. The implementation for the multi-sample and familial-sample problems are similar so the details are omitted.

The parameter and location updates are done using the standard Metropolis-Hastings (MH) algorithm. If the proposed move is a parameter update, then we randomly select one of the segment-specific copy number difference d_ls, say d_j, and draw its new value d_j^* from an interval centered at d_j of truncated normal $N_{-\alpha_{jR}}(d_j, u^2)$, where u is a tuning constant. Similarly, we randomly select one of the $\boldsymbol{d}_0$, say, $d_{j'0}$, and draw its new value $d^*_{j'0}$ from an interval centered at $d_{j'0}$ of $N_{-\alpha_{j'R}}(d_{j'0}, u^2)$. Next, we update $\boldsymbol{\beta}$. To satisfy the identifiability constraint, we identify the segment with the smallest absolute copy number difference which is least likely to be a CNV and randomly select one of the remaining segments and draw its new β from truncated normal distribution with mean equaling the current value and truncation upper and lower bounds 1 and 0, respectively. The β for the segment with the smallest absolute copy number difference is also adjusted such that the identifiability constraint is met. Since this segment is most likely to be the one whose copy number difference is 0, adjusting its β value to meet the identifiability constraint has the least impact on the likelihood. Next we update $\boldsymbol{\sigma}_s^2$ and $\boldsymbol{\sigma}_c^2$. Again, we randomly select one of the $\boldsymbol{\sigma}_s^2$ and $\boldsymbol{\sigma}_c^2$ and draw their new values from uniform distribution centered at their current values within a small window size w, with the constraint that the new values are positive. The δ^2 value is fixed and assigned as $(0.3 \times \kappa)^2$, where κ is the SD for all the data points. The factor 0.3 could be replaced by any value between 0 and 1, to reflect the belief that true copy number variances should be smaller than observed in the data. This is a reasonable assumption, since SNP microarray data are typically noisy.

If the proposed move is a location update, then we randomly select one change point, say τ_j, and sample its new location τ_j^* from a set of candidate locations within a fixed window centered at the current location, i.e. $[\tau_j - W_j,\, \tau_j + W_j]$, where the window sizes W_j are chosen such that the order of the new location is the same as that of the current one, and the new location is integer valued. The set of candidate locations is defined based on the absolute differences in log_2 ratios between consective data points. We select the top $x\%$ locations with the largest absolute differences as the candidate locations to sample from. When $x = 100$, the algorithm searches through all locations. When $0 < x < 100$, we only sample from these candidate locations according to probabilities proportional to these absolute differences, since the larger the absolute differences, the more likely these locations have copy number difference changes. This subsampling scheme is done to reduce computational burden, and its effect on the ability to detect change points is investigated in section 3 of the main article.

If the proposed move is the birth step, then we randomly sample a new change point from the set of candidate locations. Suppose that the newly sampled location τ^* falls into the segment $[\tau_j,\, \tau_{j+1} - 1]$. This segment is then split into two new segments: $[\tau_j,\, \tau^* - 1]$ and $[\tau^*,\, \tau_{j+1} - 1]$. The location-specific parameters for these two new segments are:

$$\begin{array}{ll}
d_r^* = (d_j + 2)e^{-w_l \sigma_d z_d} - 2 & d_l^* = (d_j + 2)e^{w_r \sigma_d z_d} - 2, \\
d_{0r}^* = (d_{0j} + 2)e^{-w_l \sigma_{d0} z_{d0}} - 2 & d_{0l}^* = (d_{0j} + 2)e^{w_r \sigma_{d0} z_{d0}} - 2, \\
\sigma_{sr}^{2*} = \sigma_{sj}^2 e^{-w_l \sigma_{\sigma_s} z_{\sigma_s}} & \sigma_{sl}^{2*} = \sigma_{sj}^2 e^{w_r \sigma_{\sigma_s} z_{\sigma_s}}, \\
\sigma_{cr}^{2*} = \sigma_{cj}^2 e^{-w_l \sigma_{\sigma_c} z_{\sigma_c}} & \sigma_{cl}^{2*} = \sigma_{cj}^2 e^{w_r \sigma_{\sigma_c} z_{\sigma_c}}, \\
\beta_r^* = \dfrac{\beta_j e^{-w_l^{'} \sigma_\beta z_\beta}}{(1 - \beta_j + \beta_j e^{-w_l^{'} \sigma_\beta z_\beta})} & \beta_l^* = \dfrac{\beta_j e^{w_r^{'} \sigma_\beta z_\beta}}{(1 - \beta_j + \beta_j e^{w_r^{'} \sigma_\beta z_\beta})},
\end{array} \qquad (1)$$

where $w_l = C_j^*/(C_j^* + C_{j+1}^*)$, $w_r = C_{j+1}^*/(C_j^* + C_{j+1}^*)$, $w_l^{'} = C_j^*/T$, $w_r^{'} = C_{j+1}^*/T$, and σ_d, σ_{d0}, σ_{σ_s}, σ_{σ_c} and σ_β are tuning constants, z_d, z_{d0}, z_{σ_s}, z_{σ_c}, z_β are random variables drawn from uniform distribution $U(-u, u)$, C_j^*, C_{j+1}^* and T are the total number of data points in the new segments $[\tau_j, \tau^* - 1]$, $[\tau^*, \tau_{j+1} - 1]$, and in the entire chromosome, respectively. The Jacobian J_B for the birth step mapping is thus equal to

$$\sigma_d \frac{(d_r^* + 2)(d_l^* + 2)}{d_j + 2} \sigma_{d0} \frac{(d_{0r}^* + 2)(d_{0l}^* + 2)}{d_{0j} + 2} \frac{\sigma_{\sigma_s} \sigma_{sr}^{2*} \sigma_{sl}^{2*}}{\sigma_{sj}^2} \frac{\sigma_{\sigma_c} \sigma_{cr}^{2*} \sigma_{cl}^{2*}}{\sigma_{cj}^2} \sigma_\beta (w_l^{'} + w_r^{'}) \frac{\beta_l^* (1 - \beta_l^*) \beta_r^* (1 - \beta_r^*)}{\beta_j (1 - \beta_j)}.$$

The acceptance probability for the birth step is $min(1, A_B)$, where

$$A_B = \frac{P(\boldsymbol{d}^*, \boldsymbol{\rho}^*, \boldsymbol{\sigma}_s^{2*}, \boldsymbol{\sigma}_c^{2*}, \boldsymbol{d}_0^* | data)}{P(\boldsymbol{d}, \boldsymbol{\rho}, \boldsymbol{\sigma}_s^2, \boldsymbol{\sigma}_c^2, \boldsymbol{d}_0 | data)} \times \frac{P(K|K+1)}{P(K+1|K) f(z_d) f(z_{d0}) f(z_{\sigma_s}) f(z_{\sigma_c}) f(z_\beta)} \times J_B, \quad (2)$$

and $\boldsymbol{d}^*$, $\boldsymbol{\rho}^*$, $\boldsymbol{\sigma}_s^{2*}$, $\boldsymbol{\sigma}_c^{2*}$ and $\boldsymbol{d}_0^*$ are the proposed parameters and partition for the birth. The proposal ratio can be written as

$$\frac{o_{K+1} q_{\tilde{\tau}}}{b_K p_{\tau^*}^{'} f(z_d) f(z_{d0}) f(z_{\sigma_s}) f(z_{\sigma_c}) f(z_\beta)}, \quad (3)$$

where $f(z)$ is the density for $U(-u, u)$ at z, and $q_{\tilde{\tau}}$ is the probability (inversely proportional to the absolute difference in log_2 ratio at $\tilde{\tau}$) of the change point $\tilde{\tau}$, being removed out of the $K+1$ existing change points after the birth, and $p^{'}$ is the prior probability that τ^* is a change point (proportional to the absolute difference in log_2 ratio at $\tilde{\tau}$).

If the proposed move is the death step, then we delete one of the current change points, say τ_j, according to q_{τ_j}. Prior to the death, this point divides two consective segments $[\tau_{j-1}, \tau_j - 1]$ and $[\tau_j, \tau_{j+1} - 1]$. After τ_j is deleted, these two segments are joined together into $[\tau_{j-1}, \tau_{j+1} - 1]$. The new parameters for this new segment come from the inverse mapping of the birth step, which are

$$\begin{aligned} d^* &= e^{w_l log(d_l + 2) + w_r log(d_r + 2)} - 2, \\ d_0^* &= e^{w_l log(d_{0l} + 2) + w_r log(d_{0r} + 2)} - 2, \\ \sigma_s^{2*} &= e^{w_l log(\sigma_{sl}^2) + w_r log(\sigma_{sr}^2)}, \\ \sigma_c^{2*} &= e^{w_l log(\sigma_{cl}^2) + w_r log(\sigma_{cr}^2)}, \\ \beta^* &= \frac{e^{w_l^{'} logit(\beta_l) + w_r^{'} logit(\beta_r)}}{1 + e^{w_l^{'} logit(\beta_l) + w_r^{'} logit(\beta_r)}}. \end{aligned} \quad (4)$$

The acceptance ratio for the death step is just $min(1, A_B^{-1})$ with the proposed model modified as a death and K is replaced by $K - 1$ in A_B.

The estimated number of change points is the mode of the posterior number of change points. Once the number of change points $\hat{K}$ is determined, the change points are estimated by selecting the top $\hat{K}$ locations with the highest posterior probabilities being change points. All chains start with no change points with these starting values: $d = d_0 = 0$, $\sigma_s^2 = s_s^2$ (sample SNP variance), $\sigma_s^2 = s_c^2$ (sample CNV variance), and $\beta = 0.5$.

Web Appendix C

This section describes our simulation method in detail. The simulation data were generated based on Affymetrix SNP 6.0 chromosome 21 arrays from HapMap samples and a case study on skin cancer. It is known that marker density (number of markers in a region*dividedby*length of a region) varies across genome locations, so any region could be classified into one of these four categories: 1) length < 5 kb, above average density; 2) length < 5 kb, on or below average density; 3) length ≥ 5 kb, above average density; 4) length ≥ 5 kb, on or below average density. Regions of type 2) are the hardest to detect, and there may not be a clear distinction between these regions and outliers. It is important to see how the presence of outliers affect the performance of a segmentation algorithm. Therefore, we mainly focused on BMCP's performance on 1), while incorporating 2) in outlier analysis. Intuitively, regions of types 3) and 4) (in particular, type 3) should be easier to detect by any algorithm. Table 1 in the main article gives the details of each simulation model. To see how the algorithm performs under different situations, we varied the values of $\boldsymbol{\sigma}_s^2$, $\boldsymbol{\sigma}_c^2$, $\boldsymbol{\beta}$ and x. For each model, we simulated both the samples with and without outliers. For the latter, the outliers were added at six locations (two within each normal region) that are at least 111 kb apart from any other outlier or change point location so that they are indeed outliers.

Model 1 (M_1) represents the situation when data are of the same variability as typically seen in real data and the degree of contamination is moderated. Under this model, the algorithm searches through the top 20% of candidate locations. M_2 is the same as M_1, except that the degree of contamination is larger. M_3 is the same as M_1 but with larger data variability. M_4 differs in the priors for $\boldsymbol{d}$ by assuming perfect correlation among true copy number differences. M_5 and M_6 test the effect of the percentage of locations being searched by the algorithm on the results. Under M_5, it only searches through the top 5% of candidate locations, by contrast, M_6 searches through the top 95% of candidate locations. M_7 tests the performance of the algorithm when copy numbers are mosaic. M_8 is the same as M_1, except that the CNV regions are longer.

We assumed the reference sample has 2 copies across all locations. For each model, we simulated 50 independent samples for the cases with outliers and without outliers, separately. For M_1 to M_7, we assigned 4 true change points at locations $21,855,876$, $21,858,980$, $39,870,416$, and $39,874,612$ bps giving CNVs of sizes 3.1 and 4.2 kbs. These locations were chosen to assess the algorithm's ability to detect regions of type 1). Since there were only fewer than 5 markers within each region, we randomly added more markers to increase the marker densities of both regions. For M_8, the true change points were assigned to be $21,541,564$, $22,173,087$, $39,555,928$, and $40,187,781$ bps giving CNV sizes both about 632 kbs. This model was used to test the algorithm's ability to detect regions of type

4). For each sample, we first draw $\boldsymbol{d}$ from a truncated multivariate normal distribution $N_{-2}(\boldsymbol{d}_0, \mathbf{V})$. The $\boldsymbol{d}_0$ was generated by assigning these copy numbers from the start to the end segments: 2, 3, 2, 1, 2 for M_1 to M_7. For M_8, they are 2, 2.5, 2, 1.5, 2. The ij-th element of $\mathbf{V}$ is $\delta^2 r^{|i-j|}$. We set $r > 0$ in all the simulation models so that we can assess the effect of ignoring potential correlations among copy numbers. Then $\boldsymbol{\lambda}$ was calculated as $log_2(\boldsymbol{\beta} \times \boldsymbol{d} + 2) - 1$. Conditional on $\boldsymbol{\lambda}$, $\boldsymbol{\sigma}_{\boldsymbol{s}}^2$ and $\boldsymbol{\sigma}_{\boldsymbol{c}}^2$, we next draw s_l and c_l from $N(\lambda_l, \sigma_{sl}^2)$ and $N(\lambda_l, \sigma_{cl}^2)$, respectively. Each chain was run $100,000$ iterations. The burn-ins were the first $50,000$ iterations and were thrown away, and we took every 10 of the remaining iterations for the final analyses. The upper bound for the number of change points (K_{max}) was set to 20 for all simulations.

We compared the ability of change point detection with those of the two commonly used algorithms: Circular Binary Segmentation (CBS) and Hidden Markov Model (HMM) implemented in the Bioconductor softwares DNAcopy (Olshen et al., 2004; Venkatraman and Olshen, 2007) and aCGH (Fridlyand et al., 2004), respectively. For models without outliers, we did not do any outlier smoothing, undo split (DNAcopy), or merge step (aCGH). Nor did we do any other ad hoc adjustment on the results (e.g. combination or removal of very short segments). BIC was used as the model selection criterion for aCGH analysis. For models with outliers, we did outlier smoothing first prior to segmentation using DNAcopy, to remove the potential effects from outliers. For aCGH, we merged states whose predicted values are less than 0.25 apart. The comparisons were made based on the numbers of false positives (FP) and false negatives (FN). Since each method gives estimated change points, we were able to determine the number of detected true change points. By this, we mean the number of true change points located within distance of 100 markers from the subsampling of at least one estimated change point. The number of false negatives is the total number of true change points minus the number of true change points detected by at least one estimated change point. Similarly, we define the number of false positives as the total number of estimated change points minus the number of estimated change points located within distance of 100 markers from subsampling of at least one true change point.

Web Table 1

Parameter	True Value	Estimate	95% CI
τ_1	21,855,876	21,855,900	[21,160,407, 21,856,237]
τ_2	21,858,980	21,858,980	[21,858,765, 22,152,248]
τ_3	39,870,416	39,870,416	[39,870,416, 39,870,416]
τ_4	39,874,612	39,874,612	[39,874,611, 39,874,612]
d_2	0.82	0.66	[0.44,0.96]
d_4	-0.92	-0.72	[-0.98,-0.60]
β_2	0.70	0.72	[0.50,0.90]
β_4	0.70	0.90	[0.66,0.99]

Table 1: Estimated parameters of interest from BMCP approach of 1 simulation with outliers from M_1.

Web Table 2

Rank	Location (bp)	Posterior Prob.
1	56129270	0.1200
1	79865244	0.1200
1	81856486	0.1200
4	75212671	0.1100
4	63287151	0.1100
6	98319393	0.0730
7	98312568	0.0400
8	69333663	0.0180
9	69335653	0.0120
10	63294085	0.0100
11	72623668	0.0089
12	72433414	0.0076
13	125296227	0.0073
14	98319560	0.0067
15	125918133	0.0066
16	72473859	0.0065
17	65997458	0.0063
18	72505882	0.0056
19	125121004	0.0053
20	125573852	0.0047

Table 2: Posterior probabilities of the top 20 locations for the tumor sample by BMCP.

Web Table 3

Parameter	Estimate	95% CI
τ_1	56,129,270	[56,129,270, 56,129,270]
τ_2	63,287,151	[63,287,151, 63,294,085]
τ_3	69,333,663	[69,027,272, 73,244,833]
τ_4	75,212,671	[74,902,548, 75,212,671]
τ_5	79,865,244	[79,865,244, 79,865,244]
τ_6	81,856,486	[81,856,486, 81,856,486]
τ_7	98,312,568	[98,312,568, 98,319,560]
τ_8	98,319,393	[124,772,608, 130,247,840]
d_1	0.15	[0.054, 0.28]
d_2	-0.46	[-0.5, -0.4]
d_3	0.13	[-0.018, 0.29]
d_4	-0.16	[-0.35, 0.065]
d_5	0.46	[0.36, 0.57]
d_6	-0.33	[-0.39, -0.18]
d_7	0.64	[0.62, 0.68]
d_8	0.45	[0.059, 0.68]
d_9	0.031	[-0.02, 0.086]
β_1	0.49	[0.16, 0.96]
β_2	0.49	[0.47, 0.54]
β_3	0.51	[0.48, 0.59]
β_4	0.5	[0.48, 0.53]
β_5	0.43	[0.38, 0.5]
β_6	0.59	[0.55, 0.69]
β_7	0.51	[0.5, 0.55]
β_8	0.51	[0.49, 0.55]
β_9	0.5	[0.47, 0.53]

Table 3: Estimated parameters of interest for the tumor sample by BMCP.

Web Table 4

Individual	Parameter	Estimate	95% CI
All	τ_1	105,803,129	[105,771,742, 105,814,900]
All	τ_2	105,841,589	[105,823,899, 105,841,721]
All	τ_3	111,179,089	[111,179,089, 111,179,089]
All	τ_4	111,185,496	[111,185,496, 111,185,496]
NA18852	d_1	-0.028	[-0.076, 0]
NA18852	d_2	-0.24	[-0.58, 0]
NA18852	d_3	-0.019	[-0.081, 0.037]
NA18852	d_4	-1.1	[-1.7, 0]
NA18852	d_5	-0.0082	[-0.041, 0.003]
NA18853	d_1	-0.013	[-0.11, 3.00E-04]
NA18853	d_2	-0.22	[-0.74, 0.36]
NA18853	d_3	0.0044	[-0.036, 0.051]
NA18853	d_4	-0.48	[-0.9, 0]
NA18853	d_5	0.0012	[-0.028, 0.12]
NA18854	d_1	-0.015	[-0.033, 0]
NA18854	d_2	-0.32	[-0.71, 0]
NA18854	d_3	-0.0067	[-0.032, 0.004]
NA18854	d_4	-0.41	[-0.88, 0]
NA18854	d_5	0.00064	[-0.065, 0.019]
NA18852	β_1	0.26	[0,0.53]
NA18852	β_2	0.46	[0,0.98]
NA18852	β_3	0.38	[0, 1]
NA18852	β_4	0.65	[0,1]
NA18852	β_5	0.48	[0,1]
NA18853	β_1	0.3	[0,0.61]
NA18853	β_2	0.5	[0,0.98]
NA18853	β_3	0.44	[0,1]
NA18853	β_4	0.071	[0, 0.24]
NA18853	β_5	0.26	[0,0.57]
NA18854	β_1	0.29	[0,0.5]
NA18854	β_2	0.33	[0,0.84]
NA18854	β_3	0.56	[0,1]
NA18854	β_4	0.53	[0,0.98]
NA18854	β_5	0.4	[0,1]

Table 4: Estimated parameters of interest for the trio by BMCP familial-sample model.

Web Table 5

ID	Log2 Ratio	Estimated Copy Number	Absolute Difference
Multi-Sample Model			
CNV1			
NA18852	-1.5	1.68	0.97
NA18853	-1.4	1.81	1.1
NA18854	-1.3	1.84	1
CNV2			
NA18852	-1.5	1.62	0.91
NA18853	0.087	2.06	0.064
NA18854	-0.28	1.91	0.26
Total Abs. Diff.			4.3
Familial-Sample Model			
CNV1			
NA18852	-0.72	1.76	0.55
NA18853	-0.68	1.78	0.53
NA18854	-0.66	1.69	0.42
CNV2			
NA18852	-2.80	0.91	0.62
NA18853	0.19	1.53	0.75
NA18854	-0.48	1.6	0.17
Total Abs. Diff.			3.04

Table 5: Absolute Differences in Copy Number Estimates and True Copy Numbers for both CNVs in the Trio by BMCP multi-sample and faimilial-sample models.

Web Figure 1

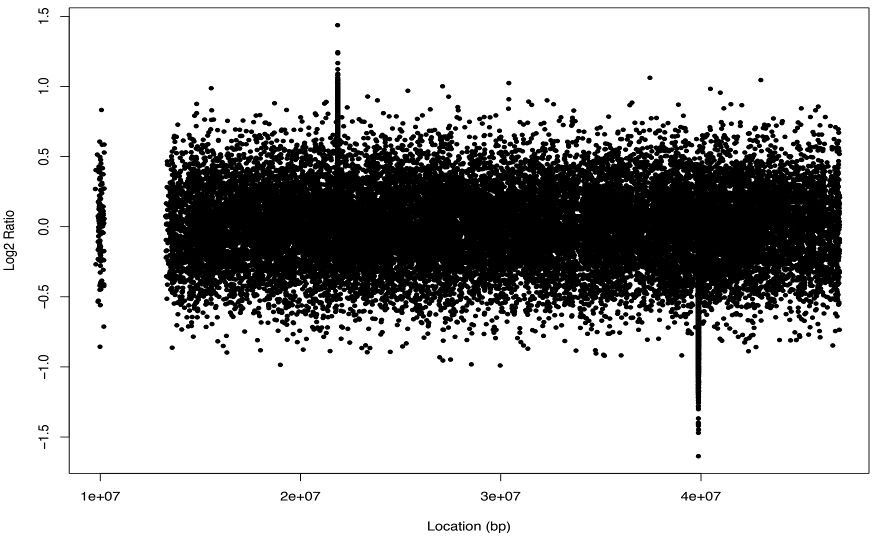


Figure 1: Plot of 1 simulated dataset from model M_1

Web Figure 2

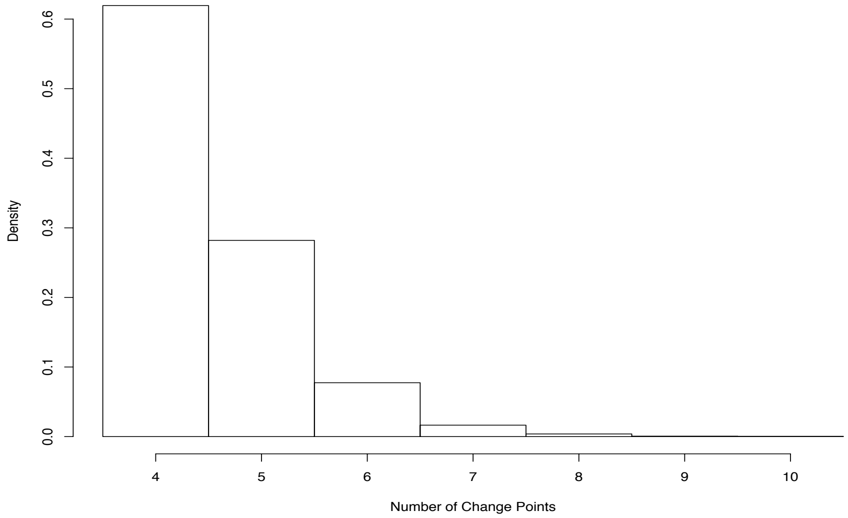
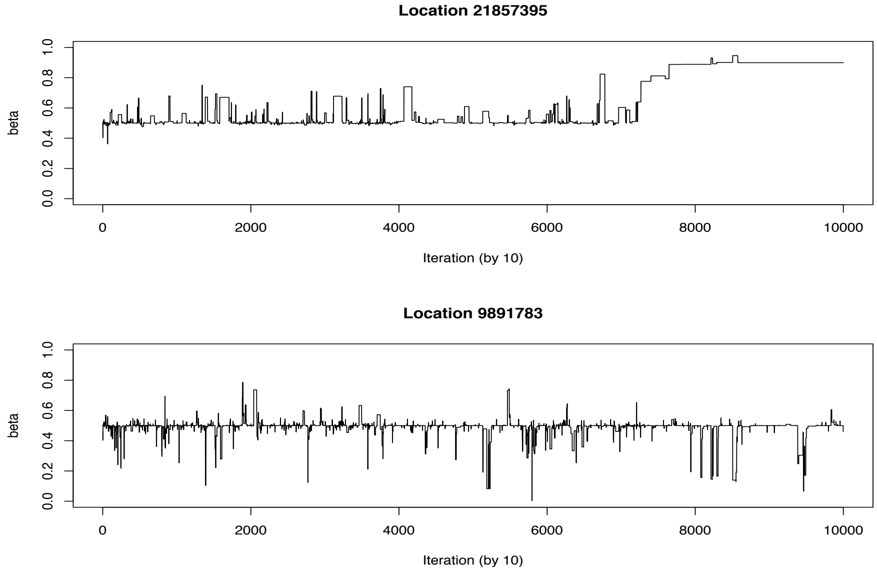


Figure 2: Distribution of number of change points of 1 simulated dataset from model M_1

Web Figure 3



Location 21857395

Location 9891783

Figure 3: Trace plots for β at two locations, one within the CNV region (top) and the other within the normal region (bottom) of 1 simulated dataset from model M_1. The traces are plotted every 10 iterations.

Web Figure 4

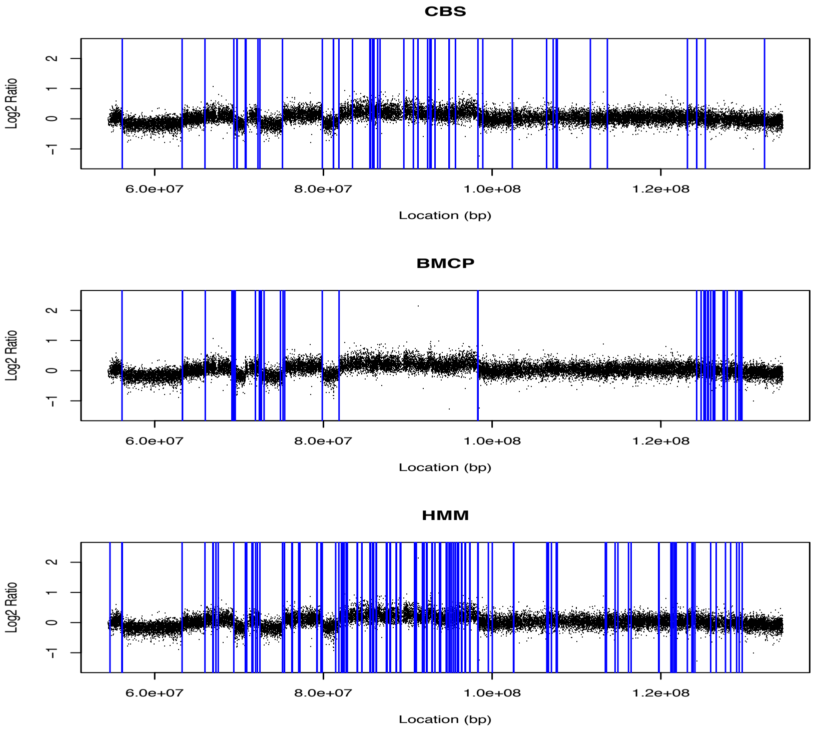


Figure 4: Estimated change point locations for the tumor sample by CBS and HMM without outlier handling, and the top 48 locations by BMCP.

Web Figure 5

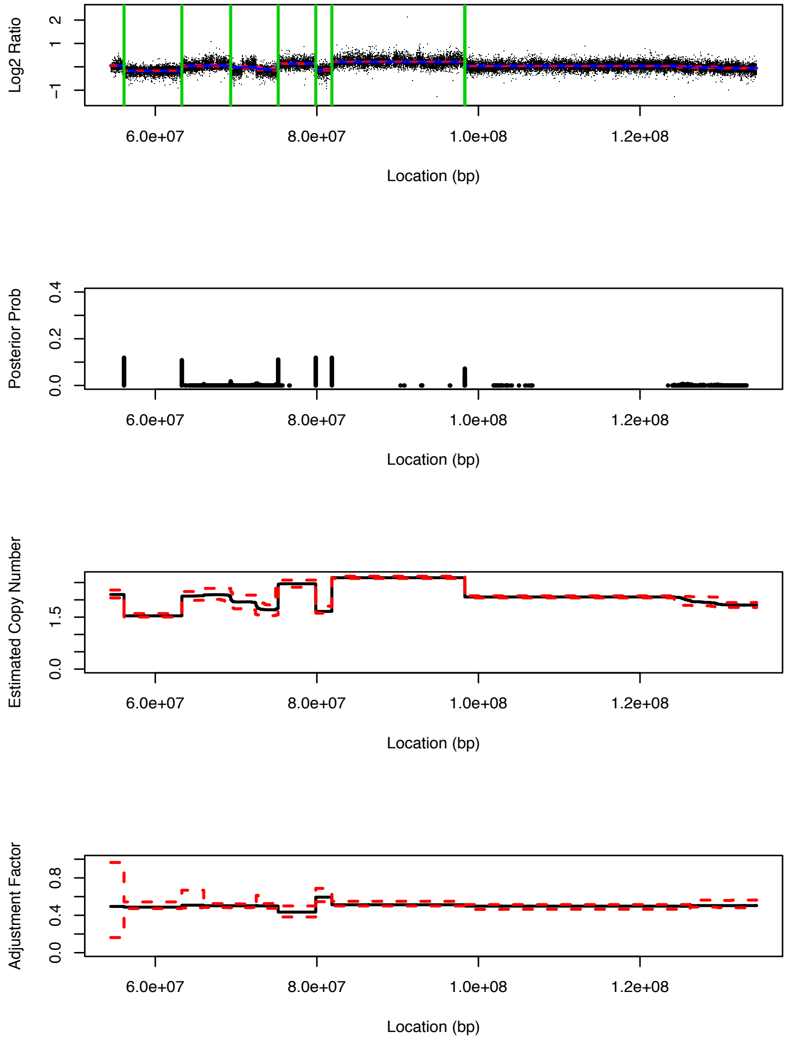


Figure 5: Estimated log_2 ratio, posterior probability, copy number, and adjustment factor for the tumor sample. Dashed lines are the 95% CIs.

Web Figure 6

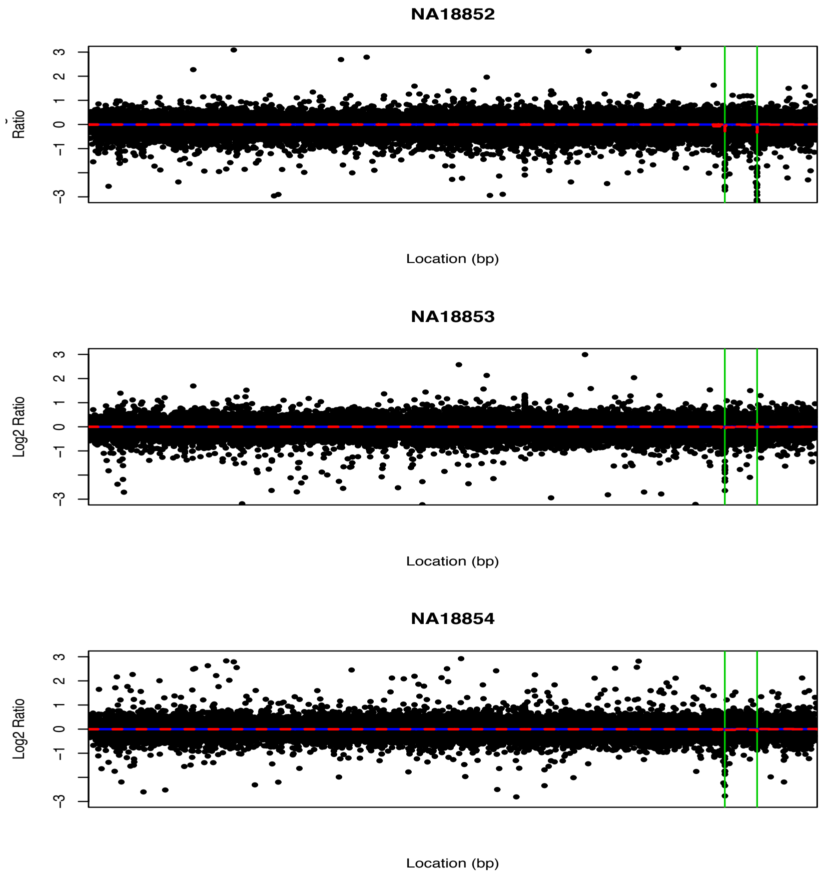


Figure 6: Estimated common change points (green vertical lines) for the trio (NA18852, NA18853, NA18854) by BMCP multi-sample model. The blue solid and red dashed lines are the estimated log_2 ratios and the corresponding 95% CIs, respectively.

Web Figure 7

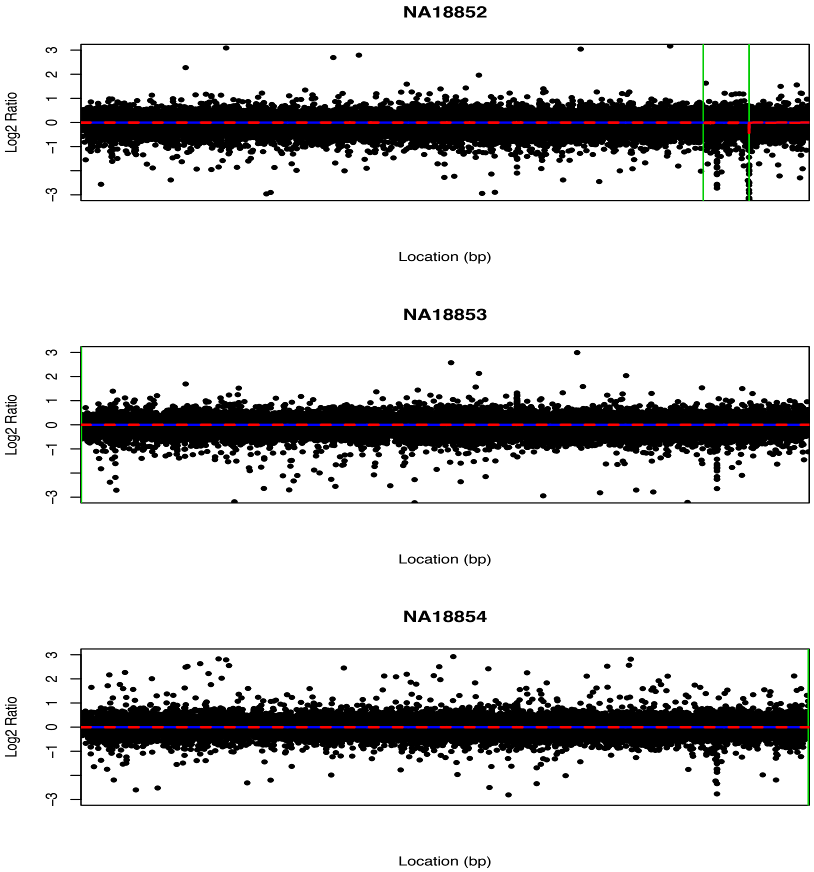


Figure 7: Estimated common change points (green vertical lines) for the trio (NA18852, NA18853, NA18854) by BMCP single-sample model. The blue solid and red dashed lines are the estimated log_2 ratios and the corresponding 95% CIs, respectively

References

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004). Hidden Markov Models Approach to the Analysis of Array CGH Data. *J. Multivar. Anal.*, 90(1):132–153.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data. *Biostat*, 5(4):557–572.

Suchard, M., Weiss, R., Dorman, K., and Sinsheimer, J. (2003). Inferring Spatial Phylogenetic Variation Along Nucleotide Sequences: A Multiple Changepoint Model. *Journal of the American Statistical Association*, 98:427–437.

Tai, Y. C., Iversen, E. S., and Parmigiani, G. (2009). Modeling Cancer Risk Variation by Mutational Spectra. *In preparation.*

Venkatraman, E. S. and Olshen, A. B. (2007). A Faster Circular Binary Segmentation Algorithm for the Analysis of Array CGH Data. *Bioinformatics*, 23(6):657–663.