

**Stochastic segmentation models for array-based comparative
genomic hybridization data analysis
SUPPLEMENTARY APPENDIX**

APPENDIX A

Proof of (6), (12), (14), (18), (21), (22), (23) and (25)

To prove (6), we make use of

$$\phi_{\mu_1, v_1}(\theta)\phi_{\mu_2, v_2}(\theta) = \phi_{\bar{\mu}, \bar{v}}(\theta) \sqrt{\frac{\bar{v}}{v_1 v_2}} \exp \left\{ \frac{1}{2} \left[\frac{\bar{\mu}^2}{\bar{v}} - \frac{\mu_1^2}{v_1} - \frac{\mu_2^2}{v_2} \right] \right\} = \frac{\phi_{\mu_1}(0)\phi_{\mu_2}(0)}{\phi_{\bar{\mu}, \bar{v}}(0)} \phi_{\bar{\mu}, \bar{v}}(\theta), \quad (\text{A.1})$$

where $\bar{v} = (v_1^{-1} + v_2^{-1})^{-1}$ and $\bar{\mu} = \bar{v}(\mu_1/v_1 + \mu_2/v_2)$, as can be shown by completing the squares. From (5), it follows that

$$P(\theta_t = 0 | \mathcal{Y}_t) \propto \{(1-p)p_{t-1} + cq_{t-1}\} \phi_{0, \sigma^2}(y_t), \quad (\text{A.2})$$

and the density function of the absolutely continuous component of θ_t is proportional to

$$(pp_{t-1} + bq_{t-1})\phi_{\mu, v}(\theta)\phi_{\theta, \sigma^2}(y_t) + a \sum_{i=1}^{t-1} q_{i, t-1} \phi_{\mu_i, v_i, t-1}(\theta)\phi_{\theta, \sigma^2}(y_t), \quad (\text{A.3})$$

with the constant of proportionality equal to the reciprocal of the conditional density function of y_t given \mathcal{Y}_{t-1} . From (A.1), it follows that

$$\begin{aligned} \phi_{\mu, v}(\theta)\phi_{\theta, \sigma^2}(y_t) &= \phi_{\mu, v}(\theta)\phi_{y_t, \sigma^2}(\theta) = \phi_{\mu_{t, t}, v_{t, t}}(\theta) \{ \phi_{\mu, v}(0)\phi_{y_t, \sigma^2}(0) / \phi_{\mu_{t, t}, v_{t, t}}(0) \} \\ &= \phi_{\mu_{t, t}, v_{t, t}}(\theta) (\psi / \psi_{t, t}) \phi_{0, \sigma^2}(y_t), \end{aligned} \quad (\text{A.4})$$

$$\phi_{\mu_{i, t-1}, v_{i, t-1}}(\theta)\phi_{\theta, \sigma^2}(y_t) = \phi_{\mu_{i, t}, v_{i, t}}(\theta) (\psi_{i, t-1} / \psi_{i, t}) \phi_{0, \sigma^2}(y_t), \quad (\text{A.5})$$

Putting (A.4) and (A.5) into (A.2) and (A.3) then yields (6).

The formula for α_t in (12) has already been proved in (11). Let $f_t(\cdot | \mathcal{Y}_n)$, $f_t(\cdot | \mathcal{Y}_t)$ and $f_t(\cdot | \mathcal{Y}_{t+1, n})$ denote the density functions of the absolutely continuous components of θ_t given \mathcal{Y}_n , \mathcal{Y}_t , $\mathcal{Y}_{t+1, n}$, respectively, and let $\dot{\pi}$ denote the density function of the absolutely continuous component of π . Then applying Bayes' theorem as in (11),

$$f_t(\theta | \mathcal{Y}_n) \propto f_t(\theta | \mathcal{Y}_t) f_t(\theta | \mathcal{Y}_{t+1, n}) / \dot{\pi}(\theta). \quad (\text{A.6})$$

The constant of proportionality in (11) and (A.6) is $g(\mathcal{Y}_t)g_*(\mathcal{Y}_{t+1, n})/g^*(\mathcal{Y}_n)$, where g , g_* and g^* denote the respective joint density functions. As shown in Section 2.1,

$$\dot{\pi}(\theta) = \phi_{\mu, v}(\theta)p / (p + c). \quad (\text{A.7})$$

Simple algebra that involves completing squares as in (A.1) can be used to show that if $v^{-1} < v_1^{-1} + v_2^{-1}$, $\tilde{v} = (v_1^{-1} + v_2^{-1} - v^{-1})^{-1}$ and $\tilde{\mu} = \tilde{v}(\mu_1/v_1 + \mu_2/v_2 - \mu/v)$, then

$$\frac{\phi_{\mu_1, v_1}(\theta)\phi_{\mu_2, v_2}(\theta)}{\phi_{\mu, v}(\theta)} = \phi_{\tilde{\mu}, \tilde{v}}(\theta) \sqrt{\frac{v\tilde{v}}{v_1 v_2}} \exp\left\{\frac{1}{2}\left[\frac{\tilde{\mu}^2}{\tilde{v}} + \frac{\mu^2}{v} - \frac{\mu_1^2}{v_1} - \frac{\mu_2^2}{v_2}\right]\right\} = \frac{\phi_{\mu_1, v_1}(0)\phi_{\mu_2, v_2}(0)}{\phi_{\tilde{\mu}, \tilde{v}}(0)\phi_{\mu, v}(0)} \phi_{\tilde{\mu}, \tilde{v}}(\theta). \quad (\text{A.8})$$

Combining (A.6), (A.7) with (5) and (9), and making use of (A.8) in the case $t < j$, we obtain β_{ijt}^* in (12).

Let $\tilde{K}_t = \min\{s \geq t : \theta_t = \dots = \theta_s \neq \theta_{s+1}\}$ be the counterpart of K_t (defined at the beginning of Section 2.2) for the time-reversed chain. In view of the preceding argument and (10),

$$\beta_{ijt} = P\{\theta_{K_t} \neq 0, K_t = i, \tilde{K}_t = j | \mathcal{Y}_n\}. \quad (\text{A.9})$$

From the definitions of K_t and \tilde{K}_t , it follows that the event in (A.9) is the same as C_{ij} defined in (14). Hence (14) holds.

To prove (18), note that

$$P(\theta_t = 0, y_t \in dt | \mathcal{Y}_{t-1}) = P(\theta_t = 0 | \mathcal{Y}_{t-1}) \phi_{0, \sigma^2}(y_t) dt = p_t^* \phi_{0, \sigma^2}(y_t) dt, \quad (\text{A.10})$$

$$\begin{aligned} P(\theta_t \neq 0, y_t \in dt | \mathcal{Y}_{t-1}) &= \int f(\theta_t = \theta \neq 0 | \mathcal{Y}_{t-1}) \phi_{\theta, \sigma^2}(y_t) d\theta dt \\ &= \int \left\{ (pp_{t-1} + bq_{t-1}) \phi_{\mu, v}(\theta) + a \sum_{i=1}^{t-1} q_{i, t-1} \phi_{\mu_i, v_i, v_{i, t-1}}(\theta) \right\} \phi_{\theta, \sigma^2}(y_t) d\theta dt \\ &= \sum_{i=1}^t q_{i, t}^* \phi_{0, \sigma^2}(y_t) dt, \end{aligned} \quad (\text{A.11})$$

by (A.4), (A.5) and (6). From (A.10) and (A.11), (18) follows.

To prove (21), we modify (A.2) as

$$P(\theta_t = 0, \theta_{t-1} = 0 | \mathcal{Y}_t) \propto (1-p)p_{t-1} \phi_{0, \sigma^2}(y_t), \quad P(\theta_t = 0, \theta_{t-1} \neq 0 | \mathcal{Y}_t) \propto cq_{t-1} \phi_{0, \sigma^2}(y_t).$$

Combining this with $P(\theta_t = 0 | \mathcal{Y}_{t+1, n}) / \pi(0)$ as in (11) yields (21). A similar argument applied to the time reverse chain yield (22). To prove (23), we use a similar argument to obtain

$$\begin{aligned} P(0 = \theta_{t-1} \neq \theta_t \in d\theta | \mathcal{Y}_t) &\propto pp_{t-1} \phi_{\mu, v}(\theta) \phi_{\theta, \sigma^2}(y_t) d\theta, \\ P(0 \neq \theta_{t-1} \neq \theta_t \in d\theta | \mathcal{Y}_t) &\propto bq_{t-1} \phi_{\mu, v}(\theta) \phi_{\theta, \sigma^2}(y_t) d\theta. \end{aligned}$$

We obtain (23) by combining this with $f_t(\theta | \mathcal{Y}_{t+1, n}) / \tilde{\pi}(\theta)$ and

$$P(\theta_{t-1} \neq \theta_t \neq 0 | \mathcal{Y}_t) = \sum_{t \leq j \leq n} P(\theta_{t-1} \neq \theta_t = \dots = \theta_j \neq 0, \theta_j \neq \theta_{j+1} | \mathcal{Y}_t) = \sum_{t \leq j \leq n} \beta_{tjt}.$$

The first equation in (25) follows from

$$E(\theta_t \mathbf{1}_{\{0 \neq \theta_t \neq \theta_{t-1}\}} | \mathcal{Y}_n) = \sum_{t \leq j \leq n} E(\theta_t | K_t, \tilde{K}_t = j, \theta_t \neq 0, \mathcal{Y}_n) P(K_t = t, \tilde{K}_t = j, \theta_t \neq 0 | \mathcal{Y}_n),$$

noting that $E(\theta_t | K_t, \tilde{K}_t = j, \theta_t \neq 0, \mathcal{Y}_n) = \mu_{t,j}$ in view of (10). The second equation also follows similarly since $(\theta_t - \mu)^2 = \theta_t^2 - 2\mu\theta_t + \mu^2$.

APPENDIX B.

Proof of (16)

Applying Bayes' theorem as in (11) yields

$$\begin{aligned} a_t &= P(\theta_t = 0 | \theta_{t-1}, \mathcal{Y}_n) \propto P(\theta_t = 0 | \theta_{t-1}, y_t) P(\theta_t = 0 | \mathcal{Y}_{t+1,n}) / \pi(0) \\ &\propto \left[(1-p) \mathbf{1}_{\{\theta_{t-1}=0\}} + c \mathbf{1}_{\{\theta_{t-1} \neq 0\}} \right] \phi_{0,\sigma^2}(y_t) P(\theta_t = 0 | \mathcal{Y}_{t+1,n}) / \pi(0), \end{aligned} \quad (\text{B.1})$$

as in (A.2). In the case $\theta_{t-1} \neq 0$ (and therefore, unlike 0, θ_{t-1} is not an atom of the stationary distribution π), a similar argument yields

$$c_t = P(\theta_t = \theta_{t-1} | \theta_{t-1}, \mathcal{Y}_n) \propto a \phi_{\theta_{t-1}, \sigma^2}(y_t) f_t(\theta_{t-1} | \mathcal{Y}_{t+1,n}) / \dot{\pi}(\theta_{t-1}), \quad (\text{B.2})$$

where $f_t(\cdot | \mathcal{Y}_{t+1,n})$ and $\dot{\pi}$ are the same as in (A.1) and (A.2). Moreover, the absolutely continuous component of the conditional distribution of θ_t given $(\theta_{t-1}, \mathcal{Y}_n)$ has density function proportional to

$$\begin{aligned} &\left[p \mathbf{1}_{\{\theta_{t-1}=0\}} + b \mathbf{1}_{\{\theta_{t-1} \neq 0\}} \right] \phi_{\mu,v}(\theta) \phi_{\theta,\sigma^2}(y_t) f_t(\theta | \mathcal{Y}_{t+1,n}) / \dot{\pi}(\theta) \\ &= \left[p \mathbf{1}_{\{\theta_{t-1}=0\}} + b \mathbf{1}_{\{\theta_{t-1} \neq 0\}} \right] \phi_{0,\sigma^2}(y_t) (\psi / \psi_{t,t}) \phi_{\mu_{t,t}, v_{t,t}}(\theta) f_t(\theta | \mathcal{Y}_{t+1,n}) / \dot{\pi}(\theta), \end{aligned} \quad (\text{B.3})$$

by (A.4). We can then apply (9) and (A.8) to derive (16) from (B.1) - (B.3).

APPENDIX C

Estimation of hyperparameters and implementation.

It is shown in Appendix A that the conditional density function of y_t given \mathcal{Y}_{t-1} is

$$f(y_t | \mathcal{Y}_{t-1}) = (p_t^* + \sum_{i=1}^t q_{i,t}^*) \phi_{0,\sigma^2}(y_t), \quad (\text{18})$$

where p_t^* and $q_{i,t}^*$ are given by (6) and are functions of the hyperparameter vector $\Phi = (p, b, c, \mu, v, \sigma^2)$. Given Φ and the observed data \mathcal{Y}_n , the log likelihood function is

$$l(\Phi) = \sum_{t=1}^n \log f(y_t | \mathcal{Y}_{t-1}) = \sum_{t=1}^n \log \left\{ (p_t^* + \sum_{i=1}^t q_{i,t}^*) \phi_{0,\sigma^2}(y_t) \right\}, \quad (\text{19})$$

in which $f(\cdot|\cdot)$ denotes conditional density function. Maximizing (19) over Φ yields the maximum likelihood estimate $\widehat{\Phi}$.

Since Φ is a 6-dimensional vector and the functions $p_t^*(\Phi)$ and $q_{i,t}^*(\Phi)$ have to be computed recursively for $1 \leq t \leq n$, direct maximization of (19) may be computationally expensive due to the curse of dimensionality. An alternative approach is to use the EM algorithm which exploits the much simpler structure of the log likelihood $l_c(\Phi)$ of the complete data $\{(y_t, \theta_t), 1 \leq t \leq n\}$:

$$\begin{aligned} l_c(\Phi) = & -\frac{1}{2} \sum_{t=1}^n \left\{ \frac{(y_t - \theta_t)^2}{\sigma^2} + \log(2\pi\sigma^2) \right\} - \frac{1}{2} \sum_{t=1}^n \left\{ \frac{(\theta_t - \mu)^2}{v} + \log(2\pi v) \right\} \mathbf{1}_{\{0 \neq \theta_t \neq \theta_{t-1}\}} \\ & + \sum_{t=1}^n \left\{ [\log(1-p)] \mathbf{1}_{\{\theta_t = \theta_{t-1} = 0\}} + (\log p) \mathbf{1}_{\{\theta_t \neq \theta_{t-1} = 0\}} \right\} \\ & + \sum_{t=1}^n \left\{ [\log(1-b-c)] \mathbf{1}_{\{\theta_t = \theta_{t-1} \neq 0\}} + (\log c) \mathbf{1}_{\{\theta_t = 0 \neq \theta_{t-1}\}} + (\log b) \mathbf{1}_{\{0 \neq \theta_t \neq \theta_{t-1} \neq 0\}} \right\}. \end{aligned} \quad (20)$$

Since $l_c(\Phi)$ decomposes into normal and multinomial components, the E-step of the EM algorithm involves $E((\theta_t - \mu)^2 | \mathcal{Y}_n)$, $E((\theta_t - y_t)^2 | \mathcal{Y}_n)$ and the conditional probabilities

$$P(\theta_t = 0 = \theta_{t-1} | \mathcal{Y}_n) = \frac{(1-p)p_{t-1}\alpha_t}{(1-p)p_{t-1} + cq_{t-1}}, \quad P(\theta_t = 0 \neq \theta_{t-1} | \mathcal{Y}_n) = \frac{cq_{t-1}\alpha_t}{(1-p)p_{t-1} + cq_{t-1}}, \quad (21)$$

$$P(\theta_t \neq \theta_{t-1} = 0 | \mathcal{Y}_n) = c\tilde{q}_t\alpha_{t-1} / \{(1-p)\tilde{p}_t + c\tilde{q}_t\}, \quad (22)$$

$$P(0 \neq \theta_t \neq \theta_{t-1} \neq 0 | \mathcal{Y}_n) = \left(\sum_{j=t}^n \beta_{tj} \right) bq_{t-1} / \{bq_{t-1} + pp_{t-1}\}, \quad (23)$$

together with $P(\theta_t = \theta_{t-1} \neq 0 | \mathcal{Y}_n)$, which is determined by the property that those five conditional probabilities have to sum up to 1. The proof of (21) – (23) is given in Appendix A. In view of (20), the M-step of the EM algorithm involves the closed-form updating formulas

$$\begin{aligned} 1 - \widehat{p}_{\text{new}} &= [\Sigma_1^n P(\theta_t = \theta_{t-1} = 0 | \mathcal{Y}_n, \widehat{\Phi}_{\text{old}})] / [\Sigma_1^n P(\theta_{t-1} = 0 | \mathcal{Y}_n, \widehat{\Phi}_{\text{old}})], \\ \widehat{a}_{\text{new}} &= [\Sigma_1^n P(\theta_t = \theta_{t-1} \neq 0 | \mathcal{Y}_n, \widehat{\Phi}_{\text{old}})] / [\Sigma_1^n P(\theta_{t-1} \neq 0 | \mathcal{Y}_n, \widehat{\Phi}_{\text{old}})], \\ \widehat{c}_{\text{new}} &= [\Sigma_1^n P(\theta_t = 0 \neq \theta_{t-1} | \mathcal{Y}_n, \widehat{\Phi}_{\text{old}})] / [\Sigma_1^n P(\theta_{t-1} \neq 0 | \mathcal{Y}_n, \widehat{\Phi}_{\text{old}})], \widehat{b}_{\text{new}} = 1 - \widehat{a}_{\text{new}} - \widehat{c}_{\text{new}}, \\ \widehat{\mu}_{\text{new}} &= [\Sigma_1^n E(\theta_t \mathbf{1}_{\{0 \neq \theta_t \neq \theta_{t-1}\}} | \mathcal{Y}_n, \widehat{\Phi}_{\text{old}})] / [\Sigma_1^n P(0 \neq \theta_t \neq \theta_{t-1} | \mathcal{Y}_n, \widehat{\Phi}_{\text{old}})], \\ \widehat{v}_{\text{new}} &= [\Sigma_1^n E\{(\theta_t - \widehat{\mu}_{\text{old}})^2 \mathbf{1}_{\{0 \neq \theta_t \neq \theta_{t-1}\}} | \mathcal{Y}_n, \widehat{\Phi}_{\text{old}}\}] / [\Sigma_1^n P(0 \neq \theta_t \neq \theta_{t-1} | \mathcal{Y}_n, \widehat{\Phi}_{\text{old}})], \\ \widehat{\sigma}_{\text{new}}^2 &= \Sigma_{t=1}^n [E((y_t - \theta_t)^2 | \mathcal{Y}_n, \widehat{\Phi}_{\text{old}})] / n. \end{aligned} \quad (24)$$

It is shown in Appendix A that

$$\begin{aligned}
E(\theta_t \mathbf{1}_{\{0 \neq \theta_t \neq \theta_{t-1}\}} | \mathcal{Y}_n) &= \sum_{t \leq j \leq n} \beta_{tjt} \mu_{t,j}, \\
E((\theta_t - \mu)^2 \mathbf{1}_{\{0 \neq \theta_t \neq \theta_{t-1}\}} | \mathcal{Y}_n) &= \sum_{t \leq j \leq n} \beta_{tjt} (\mu_{t,j}^2 + v_{t,j} - 2\mu \mu_{t,j} + \mu^2),
\end{aligned} \tag{25}$$

which can be applied to compute $\hat{\mu}_{\text{new}}$ and \hat{v}_{new} in (24). The iterative scheme (24) is carried out until convergence or until some prescribed upper bound on the number of iterations is reached.

To speed up the computations involved in the preceding EM algorithm, one can use the BCMIX approximations in Section 2.5 instead of the full recursions to determine $q_{i,t}$, $\tilde{q}_{j,t}$, etc. Moreover, one can accelerate the EM algorithm by using a hybrid approach that combines EM with some classical optimization technique, e.g., quasi-Newton methods as in Lange (1995). Applications to array-CGH data have shown that the EM estimates of μ, v, σ^2 and b typically converge quite fast. This suggests switching, after these parameter estimates stabilize, from the EM algorithm to global search for the optimizing p and c , which are particularly important as they represent relative frequencies of departures from, and returns to, the baseline state. The global search in this hybrid procedure uses (19) as a function only of p and c , with the other parameter estimates fixed at the time of switch from EM.

APPENDIX D: Supplementary figures.

Figure 1. BAC array CGH profile for chromosome 20 in cell line BT474. The lines are the signal levels estimated using SCP (top plot), HMM (middle plot), and CBS (bottom plot). Also shown in the top plot are the 2.5% and 97.5% quantiles (gray lines) of the posterior distribution of θ_t estimated by SCP, and the locations A, B, and C analyzed in Section 4.2.

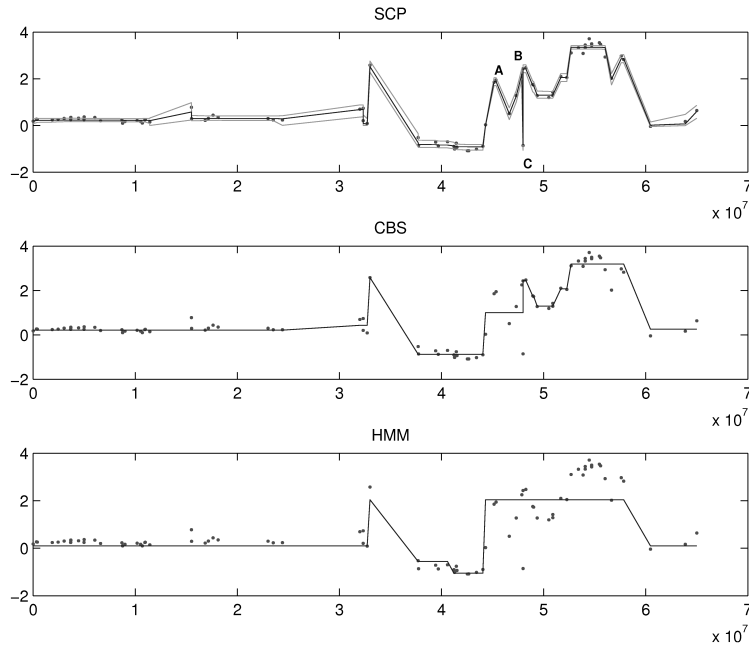


Figure 2. Histogram of number of segments in 5000 signal sequences simulated from the posterior distribution for cell line BT474.

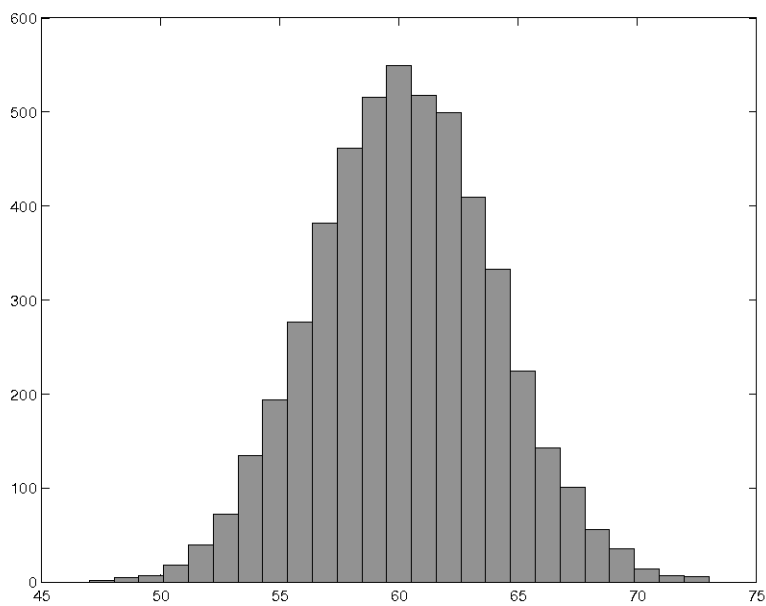


Figure 3. A simulation sequence generated from the HMM model (top plot), the stochastic change-point (SCP) model (middle plot), and the frequentist (CBS) model (bottom plot).

