Supplementary Material for 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 MCMC conditionals

Our model (2) can be written in matrix notation as

$$\mathbf{Y}_i = \boldsymbol{B}_i(\mathbf{X}_i)\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in})$ denotes the $n \times 1$ vector of measurements of all probes in array i (from a given chromosome) and $\boldsymbol{B}_i(\mathbf{X})$ is an $n \times K$ basis matrix of step functions with corresponding coefficients $\boldsymbol{b}_i$ of size $K \times 1$. We assume that the random effects $\boldsymbol{b}_i$ are normally distributed with mean $\boldsymbol{\beta}$ and covariance $\sigma^2_{\epsilon i}\Sigma_b$.

The MCMC sampling shall proceed in the following steps.

1. Sampling the regression coefficients, $(\boldsymbol{b}_i, \boldsymbol{\beta})$: Since we are working in a Gaussian prior and likelihood framework, our conditional distributions for the regression coefficients are Gaussian. Specifically, b_i for each array i are sampled from their conditional distribution $N(\mu^*_{b_i}, \Sigma^*_{b_i})$, where $\boldsymbol{B}_i$ is the basis matrix in (1), $\mu^*_{b_i} = \sigma^{-2}_{\epsilon i}\Sigma^*_{b_i}\{\boldsymbol{B}_i^T\mathbf{Y}_i + \Sigma_b^{-1}\boldsymbol{\beta}\}$ and $\Sigma^*_{b_i} = \sigma^2_{\epsilon i}\{\boldsymbol{B}_i^T\boldsymbol{B}_i + \Sigma_b^{-1}\}^{-1}$. Suppose $\lambda^* = (\lambda_{1*}, \ldots, \lambda_{K*})$, denote the current vector of latent variables for the current model K, i.e., a vector of 1's indicating to which of the three mixtures the elements of β belong. The conditional distribution of $\boldsymbol{\beta}$ after integrating out $\boldsymbol{b}$ (as $\boldsymbol{b}$ has a normal kernel) is dependent on the mixture to which the corresponding segment for β_k belongs:

- If it belongs to a neutral mixture the conditional is still a normal $\beta_k|\lambda_k^0 = 1, \text{rest} \sim N(\mu^*_{\beta_{i0}}, \Sigma^*_{\beta_{i0}})$
- If it belongs to a loss mixture the conditional is a truncted normal $\beta_k|\lambda_k^- = 1, \text{rest} \sim N(\mu^*_{\beta_i}, \Sigma^*_{\beta_i})U(-\kappa_-, \epsilon_-)$
- If it belongs to a gain mixture the conditional is a truncted normal $\beta_k|\lambda_k^+ = 1, \text{rest} \sim N(\mu^*_{\beta_i}, \Sigma^*_{\beta_i})U(\epsilon_+, \kappa_+)$,

where $\mu^*_\beta = \Sigma^*_\beta\{(\sigma^2_{\epsilon i}\Sigma_b)^{-1}\sum_i A_i B_i^T Y_i\}$ and $\Sigma^*_\beta = \{\sigma^{-2}_{\epsilon i} M\Sigma_b^{-1} - \sigma^{-2}_{\epsilon i}\Sigma_b^{-1}\sum_i A_i\Sigma_b^{-1}\}^{-1}$ where $A_i = (\Sigma_b^{-1} + B_i^T B_i)^{-1}$ $\mu^*_{\beta_i}$ and $\Sigma^*_{\beta_i}$ are diagonal elements of μ^*_β and Σ^*_β respectively. $\mu^*_{\beta i0} = \mu^*_{\beta_i} * \frac{\tau^2}{\tau^2+\Sigma^*_{\beta_i}}$ and $\Sigma^*_{\beta i0} = \frac{\tau^2\Sigma^*_{\beta_i}}{\tau^2+\Sigma^*_{\beta_i}}$

2. Sampling of the variance components, $(\sigma^2_{\epsilon i}, \Sigma_b)$: With an $\mathcal{IG}(a_\epsilon, b_\epsilon)$ prior for all i, the conditional distribution of $\sigma^2_{\epsilon i}$ is $\mathcal{IG}[(M+K)/2 + a_\epsilon, \{(SSY_i + SSb_i)/2 + 1/b_\epsilon\}^{-1}]$ where $SSY_i = (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{b}_i)^T(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{b}_i)$ and $SSb_i = (\boldsymbol{b}_i - \boldsymbol{\beta})^T\Sigma_b^{-1}(\boldsymbol{b}_i - \boldsymbol{\beta})$. We assume that the diagonal elements of $\Sigma_b = \text{diag}(\sigma^2_{\text{b1}}, \ldots, \sigma^2_{\text{bK}})$ have a (conjugate) i.i.d. $\mathcal{IG}(a_\sigma, b_\sigma)$ prior with a corresponding posterior, $\sigma^2_{bk} \sim \mathcal{IG}[M/2 + a_\sigma, \{\sum_i((b_{ik} - \beta_k)^2/2\sigma^2_{\epsilon i}) + 1/b_\sigma\}^{-1}]$. We set the hyperparameters of the inverse-gamma distribution to be $(1, 1)$ to impart little information.

3. Sampling the mixture parameters: $(\boldsymbol{\lambda}_k, \boldsymbol{\pi}_k)$ have multinomial and Dirichlet posteriors, $\boldsymbol{\lambda}_k \sim Multi\left(1; \frac{\pi_- f_-}{\sum_h \pi_h\phi_h}, \frac{\pi_0 f_0}{\sum_h \pi_h\phi_h}, \frac{\pi_+ f_+}{\sum_h \pi_h\phi_h}\right)$ and $\boldsymbol{\pi} \sim Dir(\pi_{10} + s_-, \pi_{20} + s_0, \pi_{30} + s_+)$, respectively, where $f_\bullet$ are the corresponding mixture density and $s_\bullet = \sum_{k=1}^K \lambda_{k\bullet}$

4. We use the following RJMCMC algorithm to sample the number and location of break points. Following Green (1995) and Denison *et al.* (1998) we include three types of moves in our algorithm: BIRTH, in which we add a new segment location,; DEATH in which we

delete a segment location; and MOVE, in which we relocate a segment location, with corresponding prior probabilities (p_B, p_D, p_M) where $p_M = 1-(p_B+p_D)$. The challenging aspect of the algorithm is comparing models in the RJMCMC samples. The acceptance probability for a proposed move from model $\mathcal{M}$ (of say, dimension K) to $\mathcal{M}^*$ (if dimension K^*) for a reversible jump algorithm in its most general form is

$$\min\left\{1, \frac{\mathrm{f}(\mathbf{Y}|\mathcal{M}^*)}{\mathrm{f}(\mathbf{Y}|\mathcal{M})}\frac{\mathrm{p}(\mathcal{M}^*)}{\mathrm{p}(\mathcal{M})}\frac{\mathrm{q}(\mathcal{M}|\mathcal{M}^*)}{\mathrm{q}(\mathcal{M}^*|\mathcal{M})}|\mathrm{J}|\right\},$$

the ratio of marginal likelihoods, prior distributions and proposal distributions together with a Jacobian term $|J|$. We focus on each of the ratios one by one. Here the likelihood ratio is based on the marginal likelihood $f(\mathbf{Y}|\mathcal{M})$, where $\mathcal{M} = (\boldsymbol{\beta}, \vec{c}, K, \boldsymbol{\lambda}, \Sigma_b, \boldsymbol{\sigma}_\epsilon^2)$ which has a closed form since we can integrate out the random effect coefficients $\boldsymbol{b}_i$ using its Gaussian specification. We further integrate out the population level coefficients $\boldsymbol{\beta}$ using the Gaussian-ronrod quadrature method, because the Gaussian kernel is preserved due to the normal/uniform mixture construction. Note that the proposed model $\mathcal{M}^*$ depends on the variance components Σ_b and $\boldsymbol{\sigma}_\epsilon^2$. In order to accept/reject a model based on its segments, we need to minimize the effects of these variances on the acceptance probability. Specifically, we will assume $\boldsymbol{\sigma}_\epsilon^2 = \boldsymbol{\sigma}_\epsilon^{2*}$ at the current sampled value. Because Σ_b and Σ_b^* may be of different dimensions, we cannot assume that they are equal. However, we can assume that they are equal in the elements corresponding to the segments common to both models and condition only on those elements. Following Bigelow and Dunson (2008), we can integrate out the additional variance parameter from the marginal likelihood via a numerical integration procedure such as a Laplace approximation, yielding our approximate marginal likelihood as $f(\mathbf{Y}|\mathcal{M})$, where $\mathcal{M} = (\vec{c}, K, \boldsymbol{\lambda})$. Now suppose a BIRTH step is proposed to add another

segment/cutpoint, thus going from dimension $K \rightarrow K+1$. Our prior ratio is given by

$$\frac{p(\mathcal{M}^*)}{p(\mathcal{M})} = \frac{p(K+1)p(\vec{c}_{K+1}|K+1)p(\boldsymbol{\lambda}_{K+1})|K+1)}{p(K)p(\vec{c}_K|K)p(\boldsymbol{\lambda}_K|K)}.$$

Note that the ratio of the priors on $\boldsymbol{\lambda}$ cancels out because the multinomial prior does not depend on K. Using the prior in the ratio of prior on cutpoints turns out to be $K+1/T-K$, while the first ratio depends on the prior we adopt for K.

To identify candidate change point locations, we followed the following algorithm. We preprocessed the individual samples via the CBS algorithm (Olshen *et al.* 2004) and obtained a candidate set of change points for each sample. We further added a few probe locations to the right and left of the location of the change points by drawing a Poisson random variable, with a mean depending on the length of the segment, to obtain $\mathcal{T}_i$. We set α, the tuning parameter in the CBS algorithm, such that we obtain a large number of segments, for instance $\alpha = 0.01$. We then set our candidate change point set to $\mathcal{T} = \bigcup_i \mathcal{T}_i$.

The proposal ratio involves both the BIRTH and the DEATH steps. While adding a new segment location we use the proposal density $q(\mathcal{M}^*|\mathcal{M}) = (p_B/T-K)q(\boldsymbol{\lambda}^*)$. This is made up of the probability of actually attempting a BIRTH step together with that of choosing the particular new value c_{K+1} given $\vec{c}_K$. This can be done in $T-K$ ways, as c_{K+1} must be distinct to the K elements of $\vec{c}_K$ and there are only T possibilities in total. The probability of proposing the reverse move is $q(\mathcal{M}|\mathcal{M}^*) = (p_D/K+1)q(\boldsymbol{\lambda})$, which is the probability of proposing a DEATH step and then choosing the proposed change point to remove.

The key step here is specifying the proposal $q(\boldsymbol{\lambda}^*|\boldsymbol{\lambda})$, since the RJMCMC algorithm suffers from slow convergence if the proposal rarely leads to parameter values with high likelihood. To this effect, we choose our proposal based on our likelihood considerations, which can improve convergence in practice. In a BIRTH step we break a given segment into two seg-

ments but choose a change point, say, c_p such that a new segment is formed between genome locations (c_p, c_q) with $c_p < c_q$, to the right of the proposed change point. Denote the mixture indicator for this *new* segment to be $\boldsymbol{\lambda}^*$, which can take three possible values (corresponding to a three component mixture) with probability $p_h^* = p(\lambda_h^* = 1), h = 1, 2, 3$. We define $p_h^* = f(\mathbf{Y}|\lambda_h^* = 1)/\sum_l f(\mathbf{Y}|\lambda_l^* = 1)$ where f is the (approximate) marginal likelihood as we described above. Thus our proposal density $q(\boldsymbol{\lambda}^*|\boldsymbol{\lambda})$ is just a multinomial density with probabilities p_h^*. The reverse proposal $q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^*)$ is computed similarly, conditional on the current marginal model likelihood. The MOVE step is easy since it does not involve a dimension jump; hence we can sample the parameters using a Metropolis Hastings algorithm or an independence sampler. Finally, note that since our reversible jump step involves only jumps on discrete parameters we do not need a Jacobian term to match the dimensions.

2 MCMC specifics

2.1 Sensitivity to priors

To study the effect of prior specification on the number of changepoints K, we re-ran our analyses for different priors on K, specifically using negative binomial and Poisson distributions, as an alternative to our default uniform prior. Figure 1 shows the MCMC trace plots for the aCGH data from chromosome 1 for the SC group (see Section 6 in manuscript). The priors we used were,

1. Discrete Uniform prior: $\pi(K) \sim \text{U}(0, \ldots, K_{\text{max}})$, with $K_{\text{max}} = 196$, which was the cardinality of the candidate change point set $\mathcal{T}$. This distribution has a mean of 98 segments.

2. Negative Binomial Distribution: $\pi(K) \sim \text{NB}(r, p)$ with $r = 1, p = 0.01$ which has a mean at 99, similar to the discrete uniform distribution.

3. Poisson prior: $\pi(K) \sim$ Pois(λ) with $\lambda = 98$, with mean equal to 98, same as the uniform distribution.

The left panels show the trace plots of 10000 MCMC iterations for K, with uniform, negative binomial and Poisson priors (top to bottom) All seemed to converge around a value of 160 change points for this data, indicating relative insensitivity to the prior distribution. Plotted on the right panels are the corresponding kernel density estimates of the post burn-in samples, discarding the first 5000 samples. We use a thinning factor of 5, thus basing our inference on 1000 samples. The mean and standard deviation of K for the three priors were: 158.29 (1.95), 158.69 (2.10) and 157.90 (1.98) respectively - no significant difference in the number of segments.

2.2 MCMC convergence and diagnostics

Usual convergence diagnostic methods, such as Gelman and Rubin (1992) do not apply in our case since 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). We also ran multiple chains with diverse starting values and all seemed to converge reasonably. Figure 2 shows the trace plots of K (top panel) and log-likelihood (bottom panel) using two diverse starting values plotted in red and blue colors for the data from chromosome 1 for the SC group (see Section 6 in manuscript). We observe reasonable convergence and mixing on both fronts.

3 Additional simulations results

3.1 Comparison of sensitivities

Table 1 contains the sensitivities of the three methods split by prevalence and noise level for various cutoff values of the false positive rate (1-specificity=0.05, 0.10, 0.20). The BDSAcgh performed much better than cghMCR for all values of τ and H-HMM for $\tau = 0.2, 0.3$. H-HMM performed remarkably well for $\tau = 0.1$ but the performance degraded with increasing τ. For higher prevalences, the BDSAcgh was competitive to the cghMCR for $\tau = 0.1, 0.2$ and performed much better at the high noise levels $\tau = 0.3$.

3.2 Performance a function of CNV length

We also assessed the performance of BDSAcgh as a function of CNV length. In our simulations the widths of shared aberrations were generated from a gamma distribution with parameters (a, b) and is parameterized such that the mean is a/b. We set $(a, b) = (2.5, 0.05)$ such that the mean of the distribution is 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 define the following: $[1, 10]$ as "fine" segments, $(10, 25]$ probes as short segments, $(25, 65]$ probes as medium length and > 65 probes as long segments Note that the 25th and 75th quantile of the gamma distribution is 25 and 65 respectively. We then calculated the sensitivity of detecting a true aberration as a function of segment length and the results are plotted in Figure 3. We used a threshold value corresponding to a fixed specificity of 0.90, for calling an aberration for all the methods: BDSAcgh, cghMCR with recurrence rate of 0.5, and cghMCR with recurrence rate of 0.2 shown in white, light gray and dark gray bars respectively.

Overall, we noticed that our method performed better than cghMCR in detecting aberrations,

across all segment lengths and noise scenarios. In particular, BDSAcgh has higher sensitivity in detecting short and medium length segments as compared to cghMCR, especially in the high noise scenario. In addition, BDSAcgh has marginally higher sensitivity in detecting longer segments than shorter, which is as expected since we have more probes to estimate copy number changes in those segments. With a recurrence rate set of 0.2, cghMCR does really well in the low noise scenario as compared to 0.5 recurrence rate, and marginally well for fine and short segments. However, its performance marginally decreased for medium and long segments, especially for $\tau = 0.3$.

Also, shown in Figure 4 is the histogram of the segment lengths determined by BDSAcgh for chromosome 1 for SC group. The minimum segment length found was 1 and maximum was 295 probes long, with a median length was 5 probes. This shows that BDSAcgh is flexible in finding small as well long segments.

3.3 Effect of tuning parameters on cghMCR algorithm

We also assessed the performance of the cghMCR algorithm on the tuning parameters: rate of recurrence and α (the parameter that controls the number of segments in CBS - higher α more number of segments). The varied the rate of recurrence across 5 levels (0.2,0.4,0.5,0.8,1) and α across 5 levels as (0.01,0.05,0.2,0.5,0.9). The corresponding AUC are shown in Table 2. The performance of the cghMCR algorithm is somewhat robust to specification of α but drastically changes with recurrence rate, especially for higher values (0.8,1). Thus, the cghMCR algorithm is not robust to mis-specification of recurrance rate, while in contrast, our proposed method does not require specification of any such parameters.

3.4 Performance at lower prevalances

We re-ran our simulations, for both BDSAcgh and cghMCR, with 100 samples and at lower prevalences (0.05,0.10) and the results are shown in Figure 5. For prevalence of 0.05 neither method does well in detecting shared aberrations for all noise scenarios (AUC $\approx$ 0.5). For prevalence of 0.10, BDSAcgh does better than cghMCR at all noise scenarios but with much less power compared to those at prevalences of 0.2 and higher. The performance of BDSAcgh across prevalences in the range (0.2-1) is similar to the results presented in Section 5 in the paper.

4 H-HMM implementation

To compare our method to the H-HMM method of Shah et. al. (2007) we used the MATLAB implementation of the method provided by the authors at `http://people.cs.ubc.ca/~sshah/acgh/index.html`. We used all the default settings in the code `multiSampleHHMMSynthetic.m`, to analyze our simulated data except that we changed $\epsilon = 0.8$, which the authors used in their simulation study. The code outputs posterior probabilities corresponding to loss, neutral, gain and undefined for every probe location. We used these posterior probabilities and varied the threshold from 0 to 1 to obtain the ROC curves.

REFERENCES

Bigelow, J. and Dunson, D. B. (2008), Bayesian adaptive regression splines for hierarchical data, *Biometrics*, 63, 724-732.

Brooks, S. and Giudici, P. (2000). Markov chain monte carlo convergence assessment via two-way analysis of variance. *Journal of Computational and Graphical Statistics* , 9(2), 266285.

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, *Biostatistics*, 4, 557-572.

Gelman, A. and Rubin, D. P. (1992). Inference from iterative simualtion using mut- liple sequences. *Statistical Science*, 7, 457

Table 1: Simulation Study: The entries in the body of the table are the sensitivities for the three methods -cghMCR. H-HMM and BDSAcgh and for various levels of noise, $\tau = \{0.1, 0.2, 0.3\}$, and various cut-off values of false positive rate (FPR): $\{0.05, 0.10, 0.20\}$. Columns 4-8 correspond to the varying degrees of prevalence with the overall sensitivity in column 9.

			Prevalence					
τ	FPR	Method	0.2	0.4	0.6	0.8	1	Overall
0.1	0.05	cghMCR	0.1451	0.6707	0.9673	0.9894	0.9978	0.6984
		H-HMM	0.8779	0.9536	0.9724	0.9124	0.8266	0.9419
		BDSAcgh	0.9153	0.9310	0.8997	0.9381	0.8735	0.9492
	0.10	cghMCR	0.2029	0.6916	0.9701	0.9901	0.9978	0.7267
		H-HMM	0.9570	0.9678	0.9989	0.987	0.9683	0.9816
		BDSAcgh	0.9604	0.9484	0.9856	0.9638	1.0000	0.9636
	0.20	cghMCR	0.3186	0.7334	0.9757	0.9916	0.9978	0.7709
		H-HMM	0.9866	0.9810	1.0000	0.9932	1.0000	0.9931
		BDSAcgh	0.9648	0.9541	0.9872	0.9678	1.0000	0.9676
0.2	0.05	cghMCR	0.1247	0.4803	0.3652	0.8284	0.9954	0.4788
		H-HMM	0.1264	0.2887	0.3534	0.3878	0.3078	0.2289
		BDSAcgh	0.3133	0.6069	0.8879	0.9303	0.9263	0.6696
	0.10	cghMCR	0.1937	0.5152	0.4396	0.8405	0.9957	0.5180
		H-HMM	0.2410	0.4598	0.4478	0.5164	0.5484	0.4309
		BDSAcgh	0.4318	0.7339	0.9464	0.9387	0.9367	0.9650
	0.20	cghMCR	0.3318	0.5766	0.5059	0.8566	0.9963	0.5939
		H-HMM	0.4141	0.6473	0.5835	0.6781	0.8345	0.6263
		BDSAcgh	0.5947	0.8259	0.9523	0.9455	0.9437	0.9689
0.3	0.05	cghMCR	0.0751	0.5073	0.5354	0.4907	0.7162	0.3825
		H-HMM	0.0635	0.07023	0.0889	0.1187	0.1292	0.0843
		BDSAcgh	0.1203	0.5836	0.8678	0.8479	0.8002	0.7112
	0.10	cghMCR	0.1288	0.5413	0.5697	0.5105	0.7651	0.4243
		H-HMM	0.1279	0.1405	0.1779	0.2152	0.2505	0.1687
		BDSAcgh	0.2115	0.7221	0.9658	0.8922	0.9318	0.8982
	0.20	cghMCR	0.2363	0.6002	0.6193	0.5546	0.8263	0.5330
		H-HMM	0.2531	0.2882	0.3241	0.3851	0.4123	0.3125
		BDSAcgh	0.3023	0.7636	0.9737	0.9042	0.9393	0.9094

Table 2: The entries in the body of the table are the AUC for the cghMCR over varying values of noise (τ) and tuning parameters (α and recurrance).

		Recurrance				
τ	α	0.2	0.4	0.5	0.8	1
0.1	0.01	0.9758	0.8985	0.8326	0.0084	0
	0.05	0.9756	0.8996	0.8390	0.0083	0
	0.20	0.9770	0.8958	0.8390	0.0104	0
	0.50	0.9735	0.8830	0.8324	0.0063	0
	0.90	0.9526	0.8414	0.7529	0.0022	0
0.2	0.01	0.9036	0.7796	0.7072	0.0214	0
	0.05	0.9236	0.8055	0.7518	0.0188	0
	0.20	0.9377	0.8379	0.7600	0.0112	0
	0.50	0.9283	0.8409	0.7721	0.0178	0
	0.90	0.8654	0.7718	0.6913	0.0019	0
0.3	0.01	0.7860	0.7302	0.6755	0.0242	0
	0.05	0.8461	0.7616	0.7125	0.0131	0
	0.20	0.8889	0.8107	0.7534	0.0077	0
	0.50	0.8917	0.8336	0.7687	0.0087	0
	0.90	0.8364	0.7749	0.6887	0.0006	0

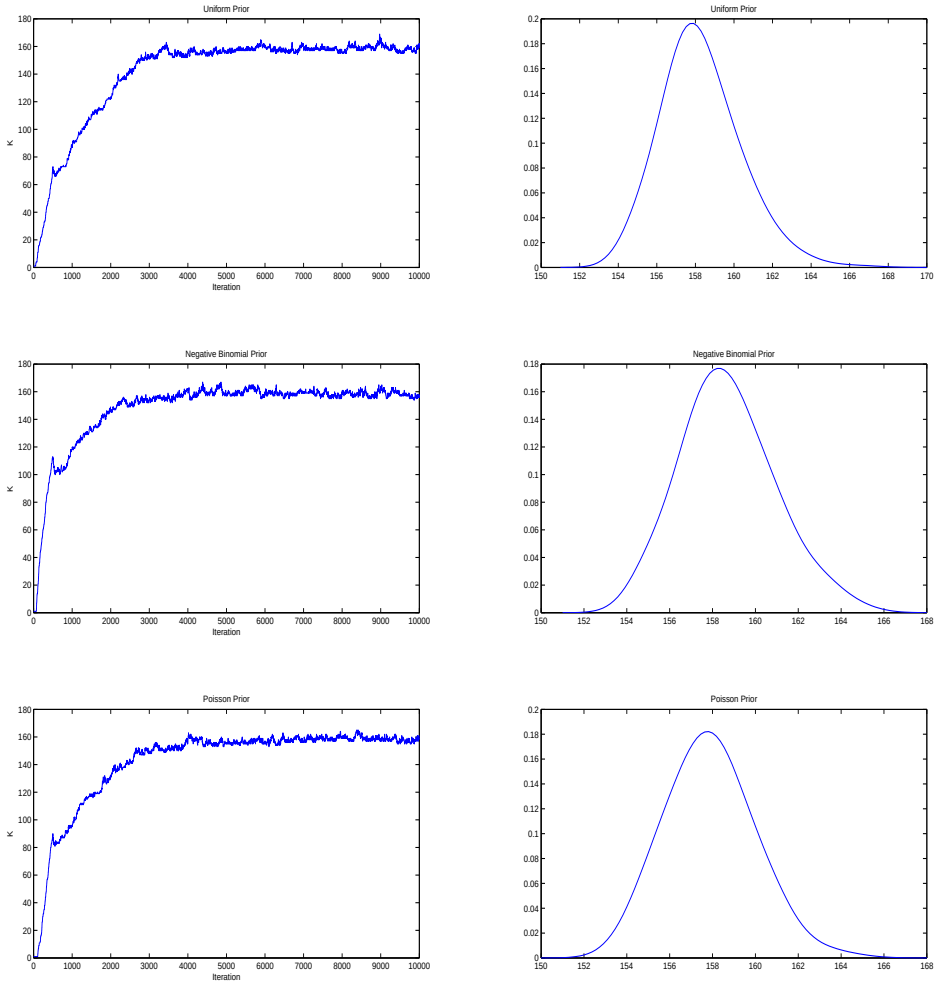

Figure 1: *Prior specification on the number of changepoints:* The left panels show the MCMC trace plots of the parameter K, the number of changepoints, with different prior specifications: uniform, negative binomial and Poisson (top to bottom). The right panel show the corresponding density estimates of K using 1000 post burn-in samples.

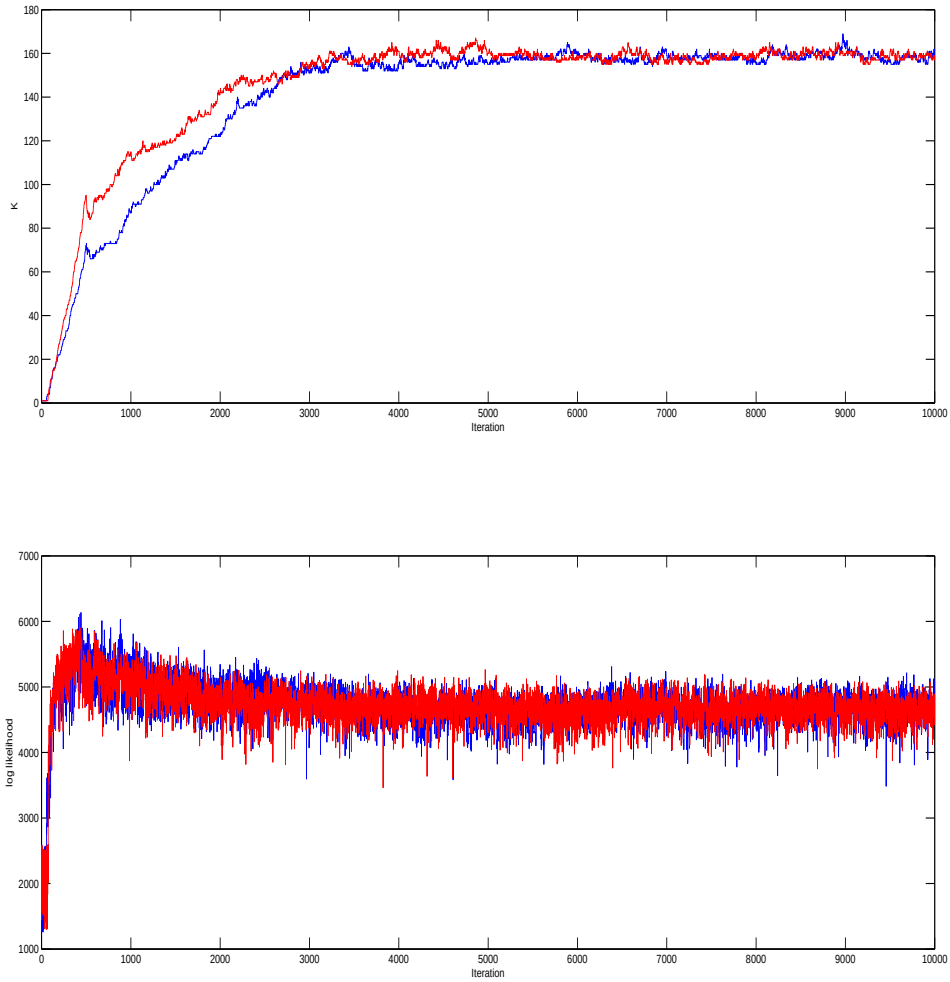

Figure 2: Top panel: shown are the MCMC traceplot of K using two diverse starting values corresponding to red and blue respectively. Bottom panel: the corresponding chains of the log-likelihood values

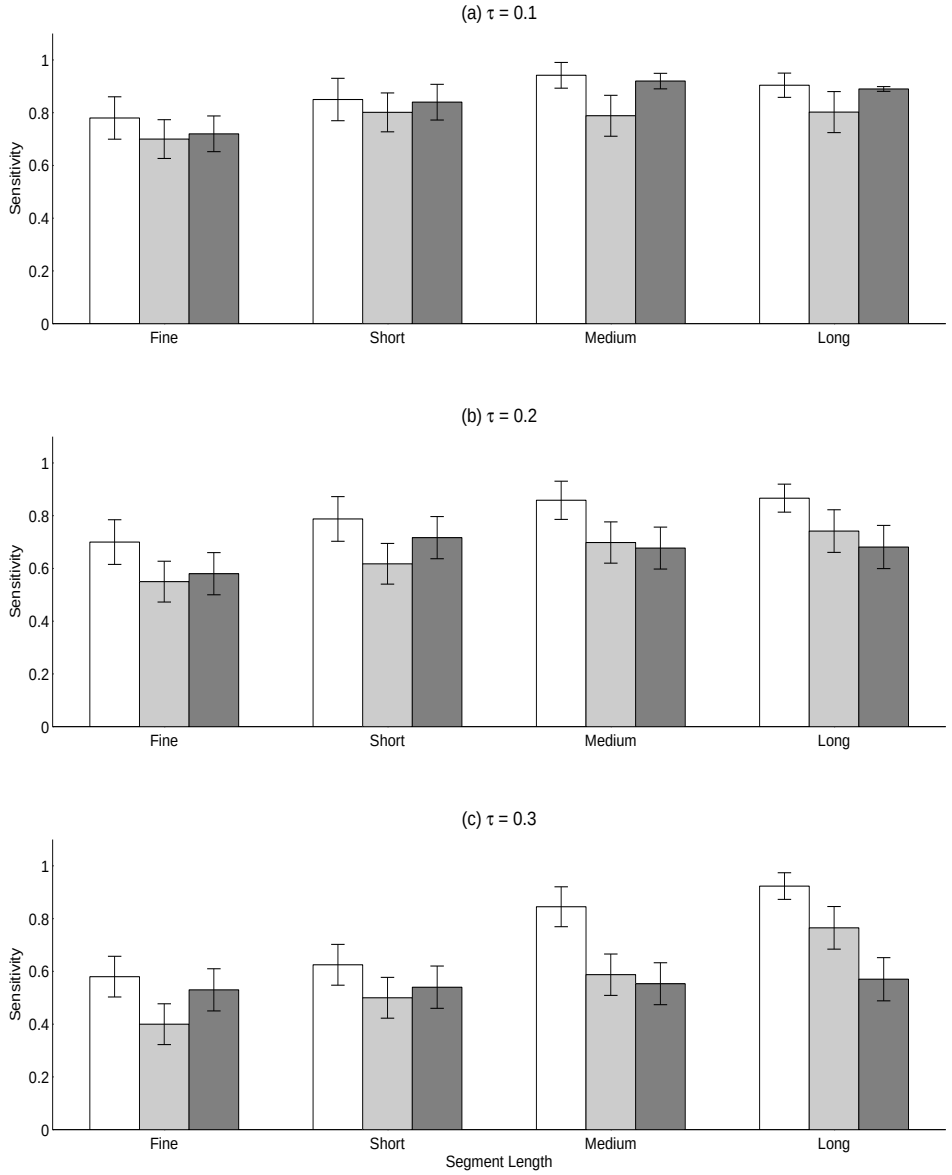

Figure 3: *Performance as function of segment length:* Shown are the sensitivities (vertical axis) with the standard error bars for the BDSAcgh (in white), cghMCR with recurrence rate of 0.5 (in gark gray) and cghMCR with recurrence rate of 0.2 (in light gray) as function of increasing segment length (horizontal axis). The top panel is for $\tau = 0.1$ (low noise), middle panel is for $\tau = 0.2$ (medium noise), and bottom panel is for $\tau = 0.3$ (high noise).

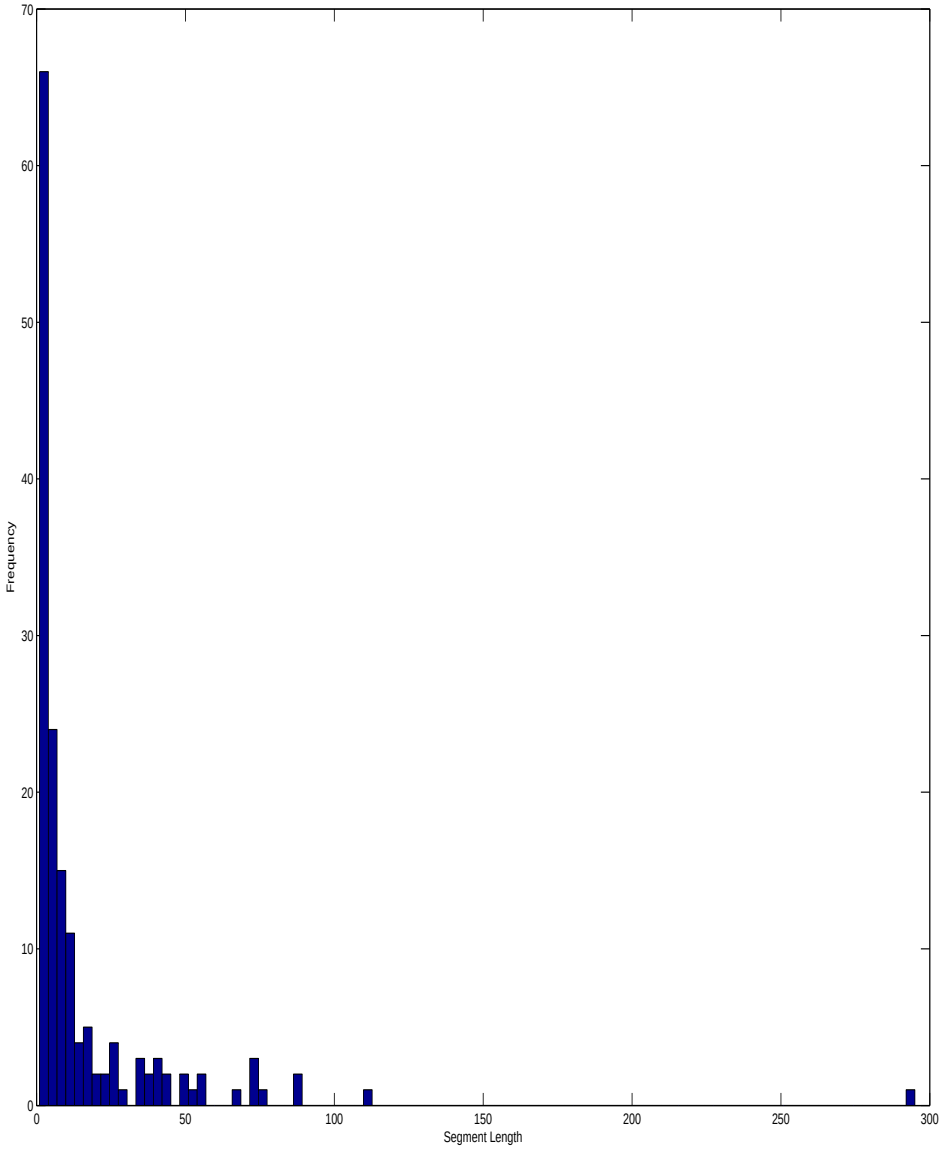

Figure 4: Histogram of segment lengths found by BDSAcgh for chromosome 1 for the SC group.

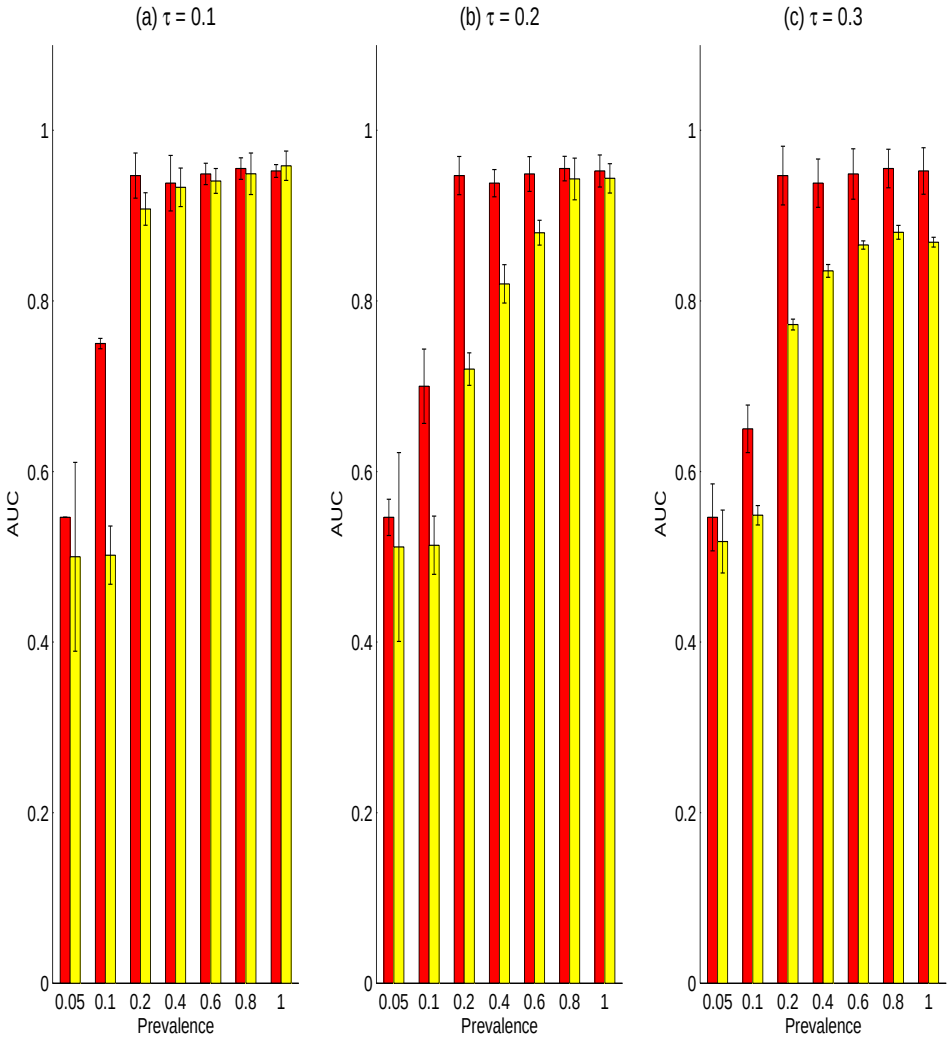

Figure 5: *Simulation study:* Shown here the bar graphs of the mean area under the curves (AUCs) with the standard error bars for the BDSAcgh (in red) and cghMCR (in yellow). The vertical axes are the mean AUC with the horizontal axes sorted by increasing prevalence. The leftmost panel is for $\tau = 0.1$ (low noise), middle panel is for $\tau = 0.2$ (medium noise), and rightmost panel is for $\tau = 0.3$ (high noise).