Web-based Supplementary Materials for "Detecting genomic aberrations using products in a multi-scale analysis"

Xuesong Yu,Timothy W. Randolph, Hua Tang, Li Hsu

July 14, 2009

Web Appendix A: Proof of consistency of change point estimator

Wang (1995) proved that with probability approaching to 1, the maximum of absolute values of wavelet coefficients at scale s occurs only within the support of wavelet function around the true change point. However the exact location of a maximum will depend on the shape of the function f and the wavelet function ψ. In the following, we show that for a step function f and Haar wavelet family, the location of the maximum of absolute values of the MODWT coefficients converges to the location of the true change point as the number of markers goes to infinity.

Without loss of generality, we consider a step function, f, with exactly one change point located at $\tau \in (0,1)$. Specifically, assume

$$f(t) = \begin{cases} \mu_1, & t \leq \tau, \\ \mu_2, & t > \tau \end{cases}$$

and $d \equiv \mu_2 - \mu_1 \neq 0$. Also assume that f is discretely sampled at n uniformly-spaced index points, $\{i/n\}_{i=1}^{n}$.

The level-j (Haar) MODWT coeffiencts $W_{j,i}$ $(i = 1, ..., n)$ give rise to the two-scale product $U_{j,i} = W_{j,i}W_{j+1,i}$. At each location i, M_i denotes the maximum, across levels, of the two-scale product, $M_i = \max_j\{U_{j,i}\}$. Here we can think of M_i as $U_{j^*,i}$ which calculated at "optimal" level j^*. In the following proof, consider the test statistics M_i as $U_{j^*,i}$. An estimated change point location, $\hat{c} = \hat{c}_{n,\alpha}$, is defined as a location of a local maximum in $\{U_{j^*,i}\}_{i=1}^{n}$ (viewed as a function of i/n), at which the adjusted p-value $< \alpha$

<u>*Claim*</u>*:* $\lim_{n\to\infty} \hat{c} = \tau$.

Proof. We introduce some notation following the convention in Percival and Walden (2000). Let $h = \{h_l : l = -L/2, \ldots, -1, 0, \ldots, L/2-1\}$ be a wavelet filter, where L is the width of the filter. For the Haar wavelet, L=2. Define $h_l = 0$ for $l < -L/2$ and $l \geq L/2$. Then the empirical MODWT wavelet coefficient $W_{j,i}$ at level j for the ith marker locus can be written as

$$W_{j,i} = \sum_{l=-L_j/2}^{L_j/2-1} h_{j,l} Y_{i-l},$$

where $\{h_{j,l}\}$ is the corresponding wavelet filter at level j and L_j is the width of $\{h_{j,l}\}$. For the Haar wavelet, $L_j = 2^j$. For wavelet coefficients at the two boundaries, we extend the data Y beyond its boundaries Y_1 and Y_n in a symmetric manner: $Y_{1-k} = Y_{1+k}$ and $Y_{n+k} = Y_{n-k}$, where $k = 0, 1, \ldots, n$.

If for sufficiently high level j and $I = \{i : \lfloor n\tau \rfloor - 2^{j-1} < i \leq \lfloor n\tau \rfloor + 2^{j-1}\}$, the local maximum of a 2-scale product $\{U_{j,i}\}$ within I converges to τ in probability. then the location of the local maximum of $\{M_i\}$ within I converges to τ. To see this, we first show that the location of local maximum of $\{|W_{j,i}|\}$ converges to τ as $n \to \infty$. Let

$$\begin{aligned} W_{j,i} &= Wf(x_i) + W\epsilon_i \\ &= \sum_{l=-2^{j-1}}^{2^{j-1}-1} h_{j,l} f(i-l) + \sum_{l=-2^{j-1}}^{2^{j-1}-1} h_{j,l}\epsilon_{i-l} \\ &= \left(\frac{d}{2} - \frac{|i - \lfloor n\tau \rfloor| d}{2^j}\right) 1_{|i-\lfloor n\tau \rfloor| \leq 2^{j-1}}(i) + \frac{\sum_{l=0}^{2^{j-1}-1} \epsilon_{i+l} - \sum_{l=1}^{2^{j-1}} \epsilon_{i-l}}{2^j} \end{aligned}$$

It is clear that, $\arg\max_{i\in I}(|W_{j,i}|)/n \to \tau$, for large j and $|i - \lfloor n\tau \rfloor| \leq 2^{j-1}$, as $n \to \infty$. For the 2-scale product $U_{j,i}$, we have

$$\begin{aligned} U_{j,i} &= W_{j,i} W_{j+1,i} \\ &= \left\{\left(\frac{d}{2} - \frac{|i - \lfloor n\tau \rfloor| d}{2^j}\right) 1_{|i-\lfloor n\tau \rfloor| \leq 2^{j-1}}(i) + o_p(1)\right\} \left\{\left(\frac{d}{2} - \frac{|i - \lfloor n\tau \rfloor| d}{2^{j+1}}\right) 1_{|i-\lfloor n\tau \rfloor| \leq 2^{j}}(i) + o_p(1)\right\} \\ &= d^2 \left(\frac{1}{2} - \frac{|i - \lfloor n\tau \rfloor|}{2^j}\right)\left(\frac{1}{2} - \frac{|i - \lfloor n\tau \rfloor|}{2^{j+1}}\right) 1_{|i-\lfloor n\tau \rfloor| \leq 2^{j-1}}(i) + o_p(1) \end{aligned}$$

Hence, $\arg\max_{i\in I}(U_{j,i})/n \to \tau$, as $n \to \infty$.

This proves the location of the local maxima of $\{M_i\}_{i=1}^n$ converges to τ. By the definition of $\hat{c}$, under the null, type I error rate is α. So as $\alpha \to 0$, $\lim_{n\to\infty} \hat{c} = \tau$.

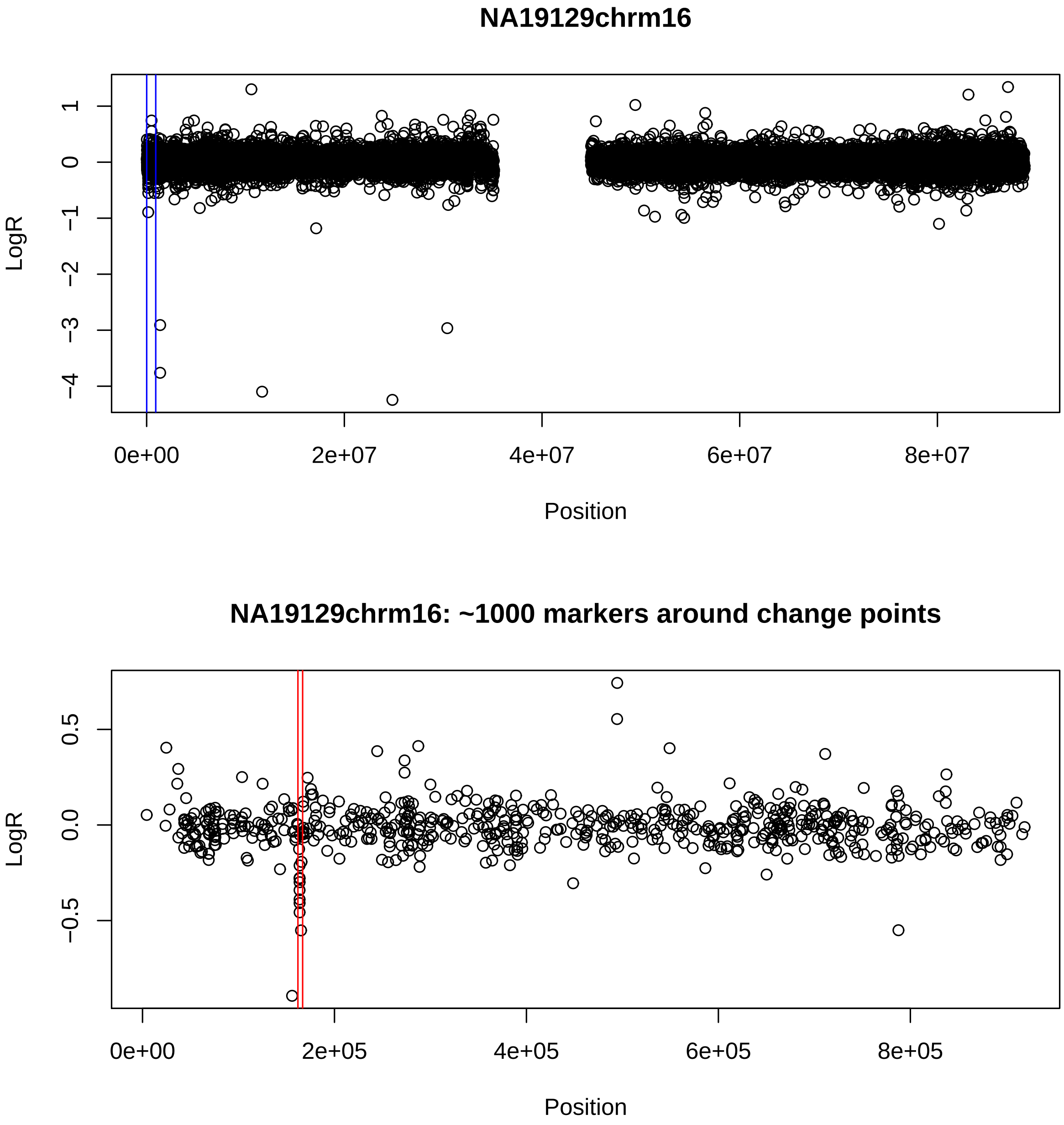
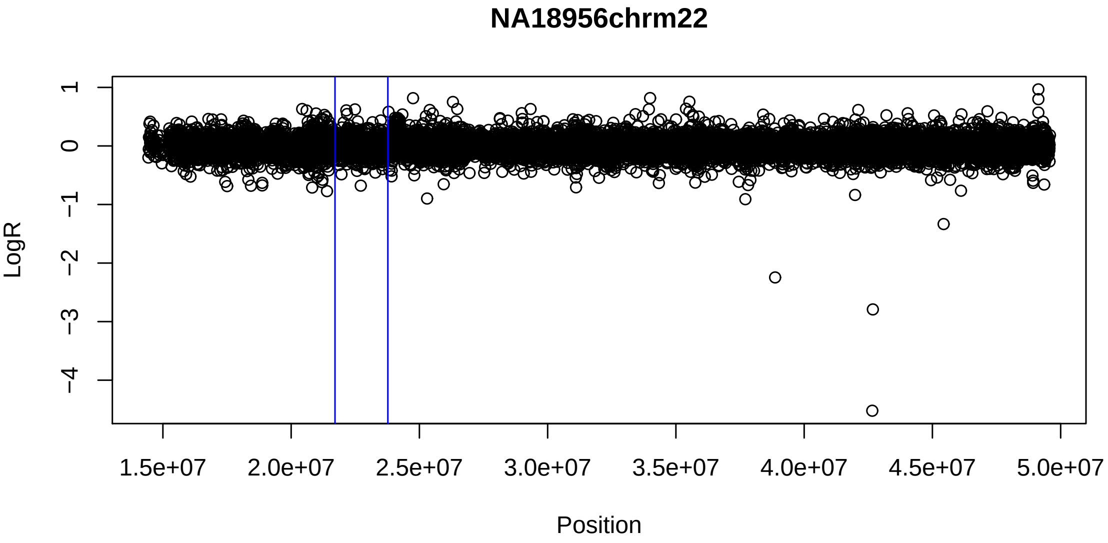


Figure 1: Top: Scatter plot for NA19129 chromosome 16 (SNR =-3.13). The vertical lines indicate region that surrounds the validated change points. The blank spot is the centromere. Bottom: Zoomed-in scatter plot of the region that surrounds the change points (about 100 SNPs on left side and 500 SNPs on right side of change points). The vertical lines indicate the validated change points.



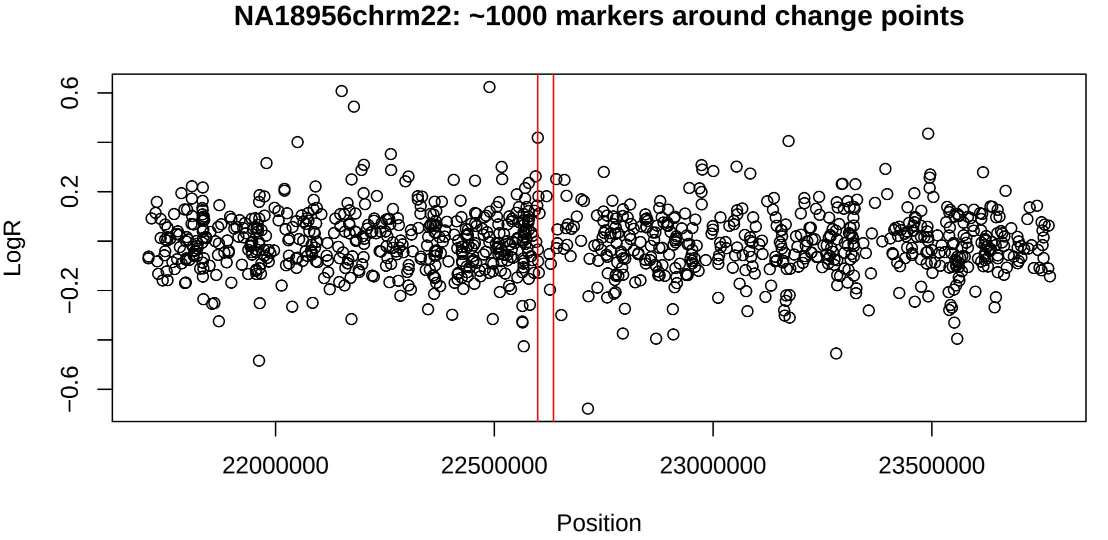


Figure 2: Top: Scatter plot for NA18956 chromosome 22 (SNR =0.29). The vertical lines indicate region that surrounds the validated change points. Bottom: Zoomed-in scatter plot of the region that surrounds the change points(about 500 SNPs on either side of change points). The vertical lines indicate the validated change points.

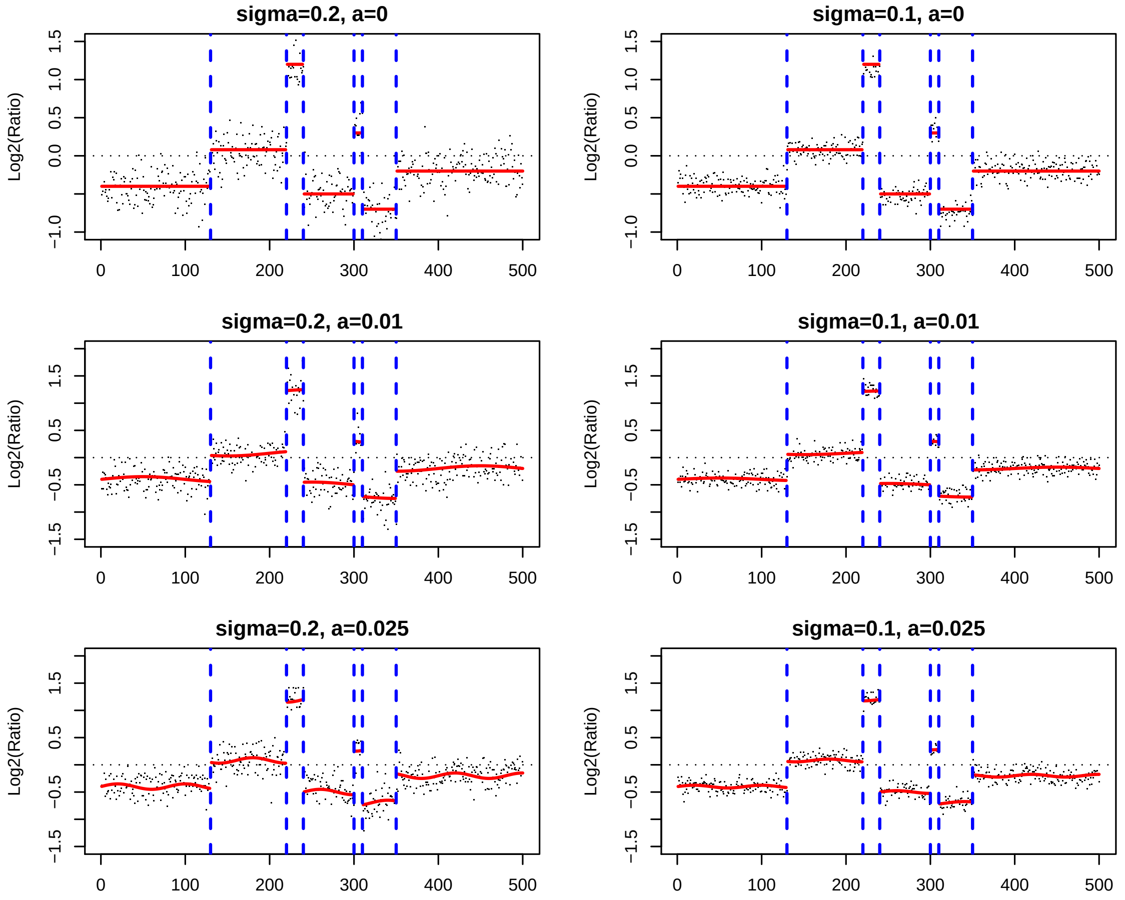


Figure 3: An example of a simulated data with $\sigma = 0.1, 0.2$ and $a = 0$ for no trend (top panel), $a = 0.01$ for long period (middle panel) and $a = 0.025$ for short period trend (bottom panel). The solid lines indicate the mean plus trend, i.e., $f(i) + 0.25\sigma\sin(a\pi i)$, the vertical dashed lines indicate the locations of change points.

Table 1: Summary of results under the complete null for when ϵ is i.i.d t distributed with df= 3, 2, and 1, respectively. A significance level 0.01 was used.

				$\hat{R}$				
df	Method	0	1	2	3	4	5+	Type I error
3	Permute W_1	463	36	1	0	0	0	0.074
	Permute $\hat{\epsilon}(0.10)$	498	2	0	0	0	0	0.004
	Permute $\hat{\epsilon}(0.05)$	498	1	1	0	0	0	0.004
	CBS	494	0	6	0	0	0	0.012
2	Permute W_1	439	56	5	0	0	0	0.122
	Permute $\hat{\epsilon}(0.10)$	496	4	0	0	0	0	0.008
	Permute $\hat{\epsilon}(0.05)$	494	5	1	0	0	0	0.012
	CBS	498	0	2	0	0	0	0.004
1	Permute W_1	306	182	11	1	0	0	0.388
	Permute $\hat{\epsilon}(0.10)$	493	6	1	0	0	0	0.014
	Permute $\hat{\epsilon}(0.05)$	497	2	1	0	0	0	0.006
	CBS	495	0	5	0	0	0	0.010

References

1. Wang, Y. (1995). Jump and sharp cusp detection by wavelets. Biometrika 82: 385-397.