Supplementary materials to "CMARRT: A tool for the analysis of tiling array data by incorporating the correlation structure"

Pei Fen Kuan[1], Hyonho Chun[1] and Sündüz Keleş[1,2*]
[1]Department of Statistics,
[2]Department of Biostatistics and Medical Informatics,
1300 University Avenue, University of Wisconsin, Madison, WI 53706.
Tel: (608) 263-453. Fax: (608) 262-0032.

July 16, 2007

1 Identification of spurious peaks in Landick data

Figures 1 and 2 show examples of spurious peaks identified when the correlation structure is ignored using both the FDR and Bonferroni correction, which is known to be a very conservative approach.

2 Diagnostic plots for AR(7) model

Figure 3 shows the diagnostics plots for autoregressive model of order 7 fitted on the first 5000 probes in Krig et al. (2007).

*Corresponding author: keles@stat.wisc.edu

3 Examples of simulated data

Figure 4 shows examples of simulated data from the autoregressive and duration HMM which resembles data from a typical chromatin immunoprecipitation experiment. Figure 5 illustrate the problem of ignoring the correlation structure in the standardized moving average statistics. The histogram of p-values for the non binding probes under the independence assumption deviates from the expected uniform distribution. Similarly, the normal QQ plot under the independence assumption deviates from the standard Gaussian distribution for the non binding probes.

4 Normal quantile-quantile plots in ZNF217 ChIP-chip data

Figure 6 shows the normal QQ plots of replicates 1 and 3. The plots show improvement when the correlation structure is taken into account.

References

Krig, S., Jin, V., Bieda, M., O'Geen, H., Yaswen, P., Green, R. and Farnham, P. (2007). Identification of genes directly regulated by the oncogene znf217 using chip-chip assays, *J. Biol. Chem* **282**(13): 9703–9712.
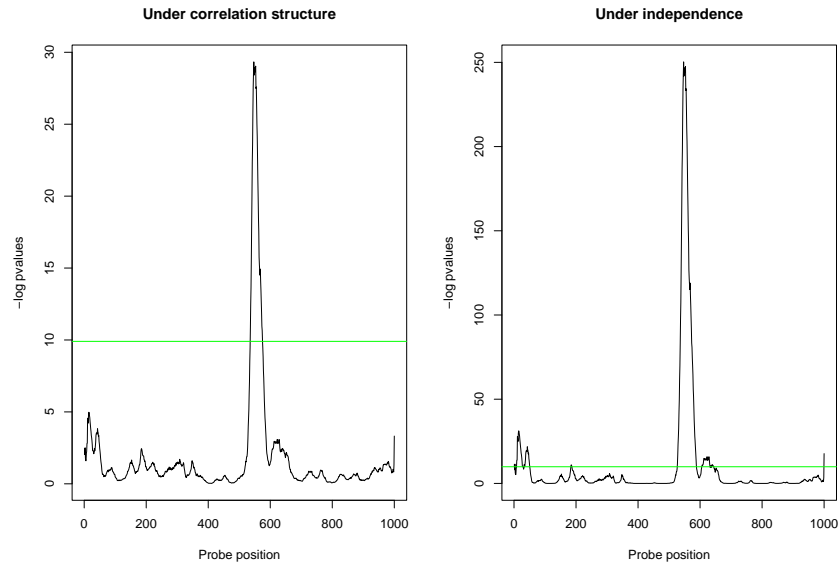
Figure 1: *p-value threshold using Bonferroni adjustment at* $\alpha = 0.05$ *in Landick.* The green line is the -log p-value threshold. Probes with -log p-value above the green line will be declared as enriched probes.

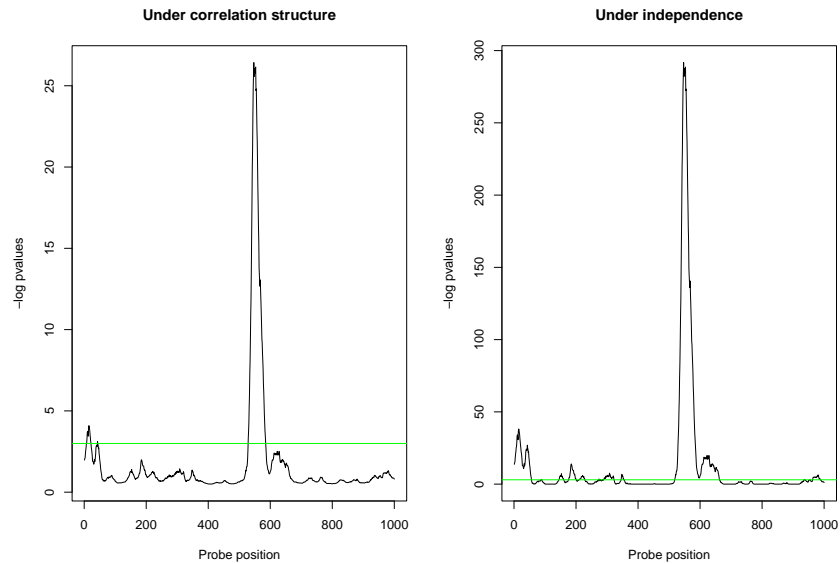Figure 2: *q-value threshold at FDR=0.05 in Landick.* The green line is the -log q-value threshold. Probes with -log q-value above the green line will be declared as enriched probes.
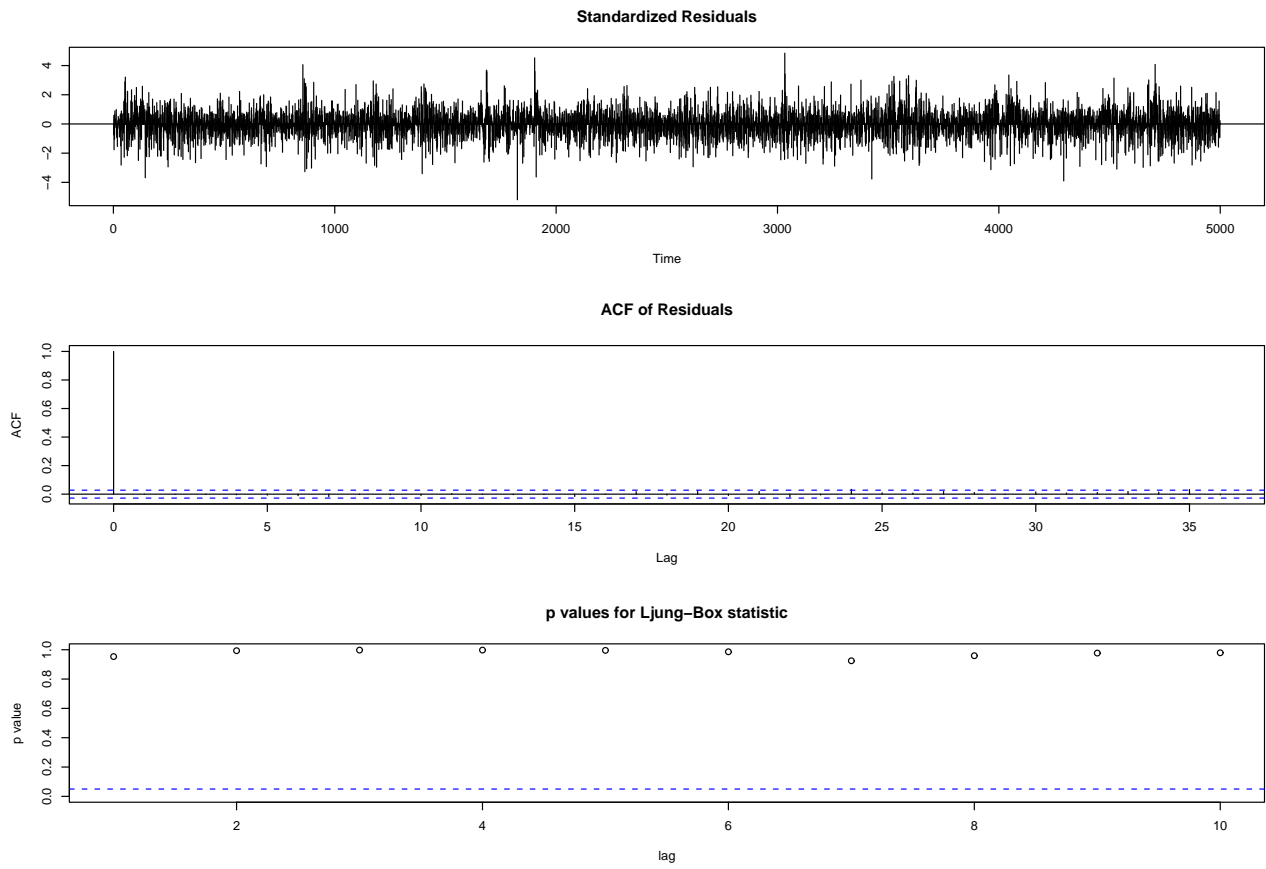
Figure 3: *Diagnostic plots for AR(7) model fitted to the data in Krig et al. (2007).*
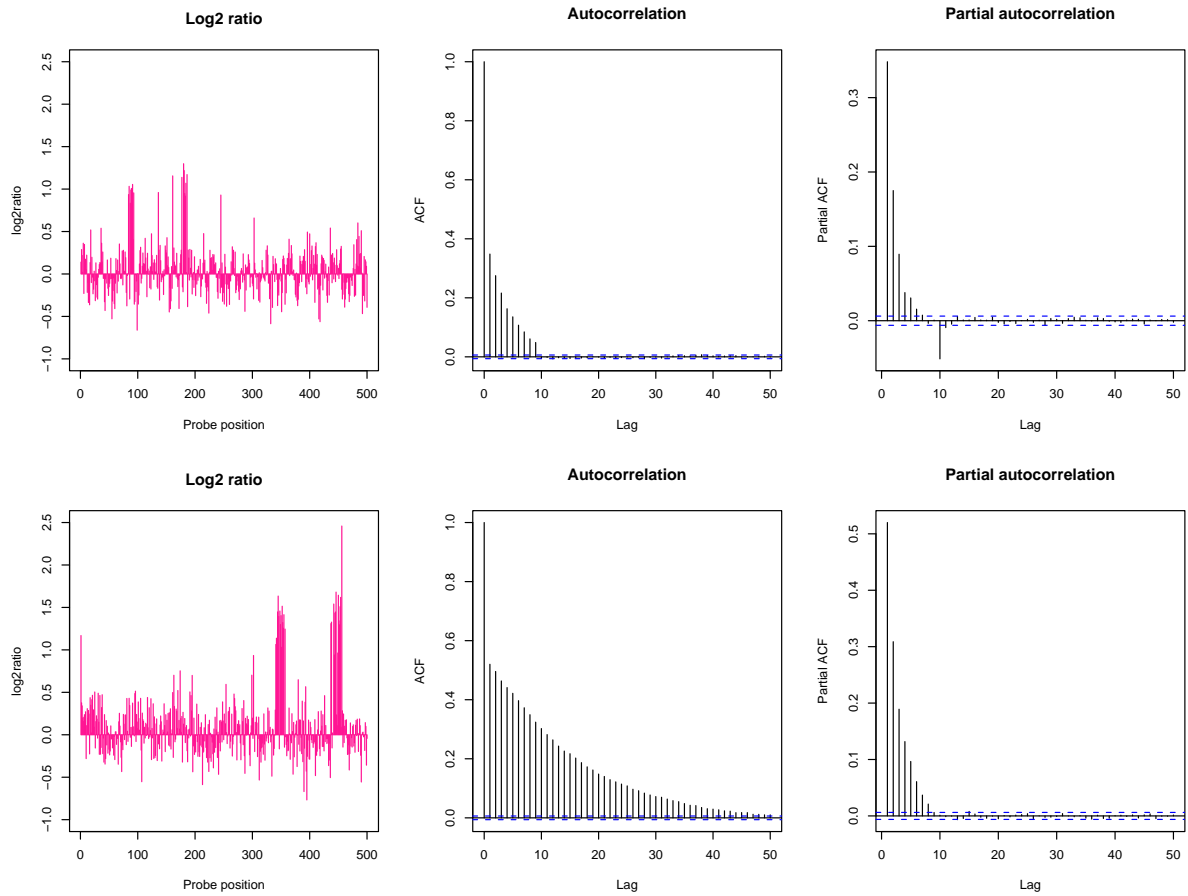
Figure 4: *Log base 2 ratio, autocorrelation and partial autocorrelation plots for simulated data.* The top panels are from the autoregressive model and the bottom panels are from the duration HMM. The simulated data resemble data from a typical chromatin immunoprecipitation experiment.
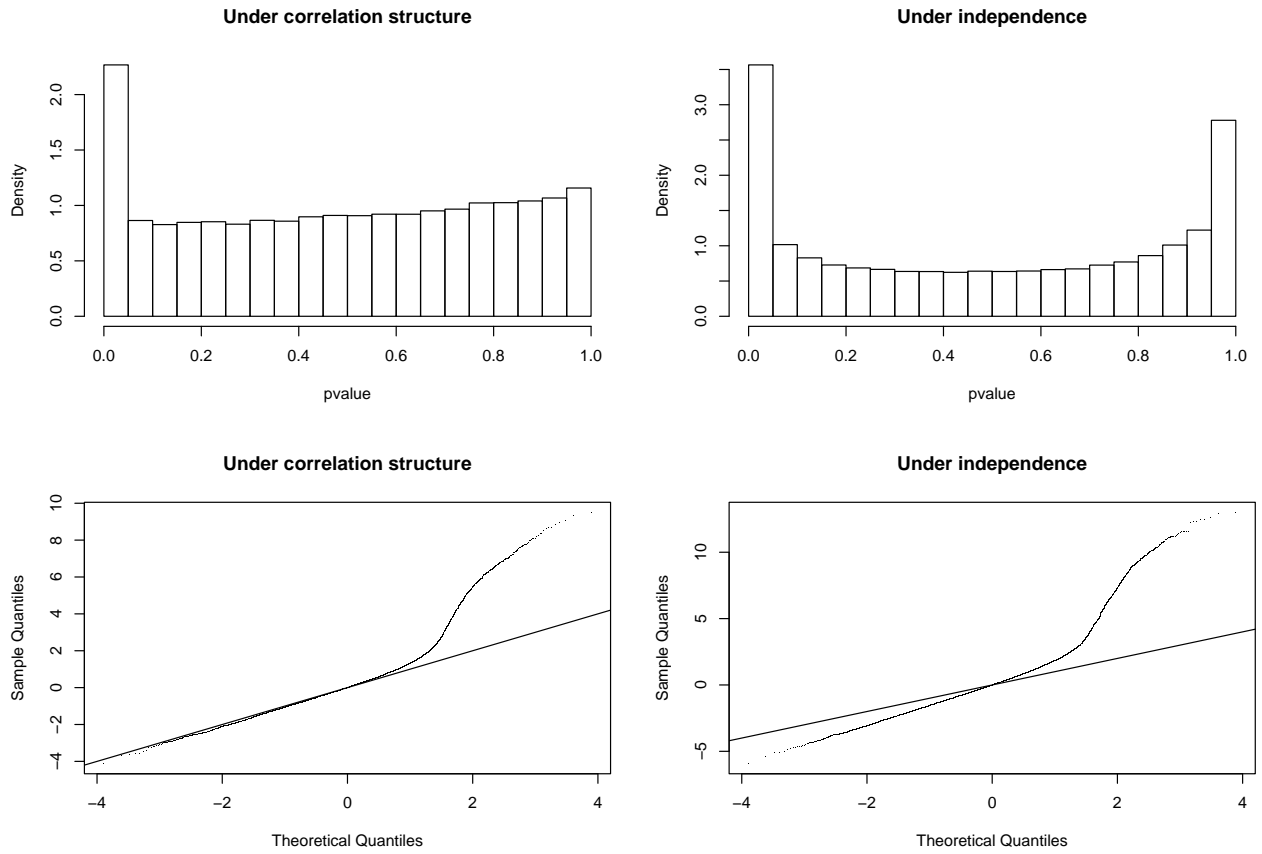
Figure 5: *Histogram of p-values and normal QQ plots for simulated data.* Both the top and bottom right panels show the problem if the correlation structure is ignored.
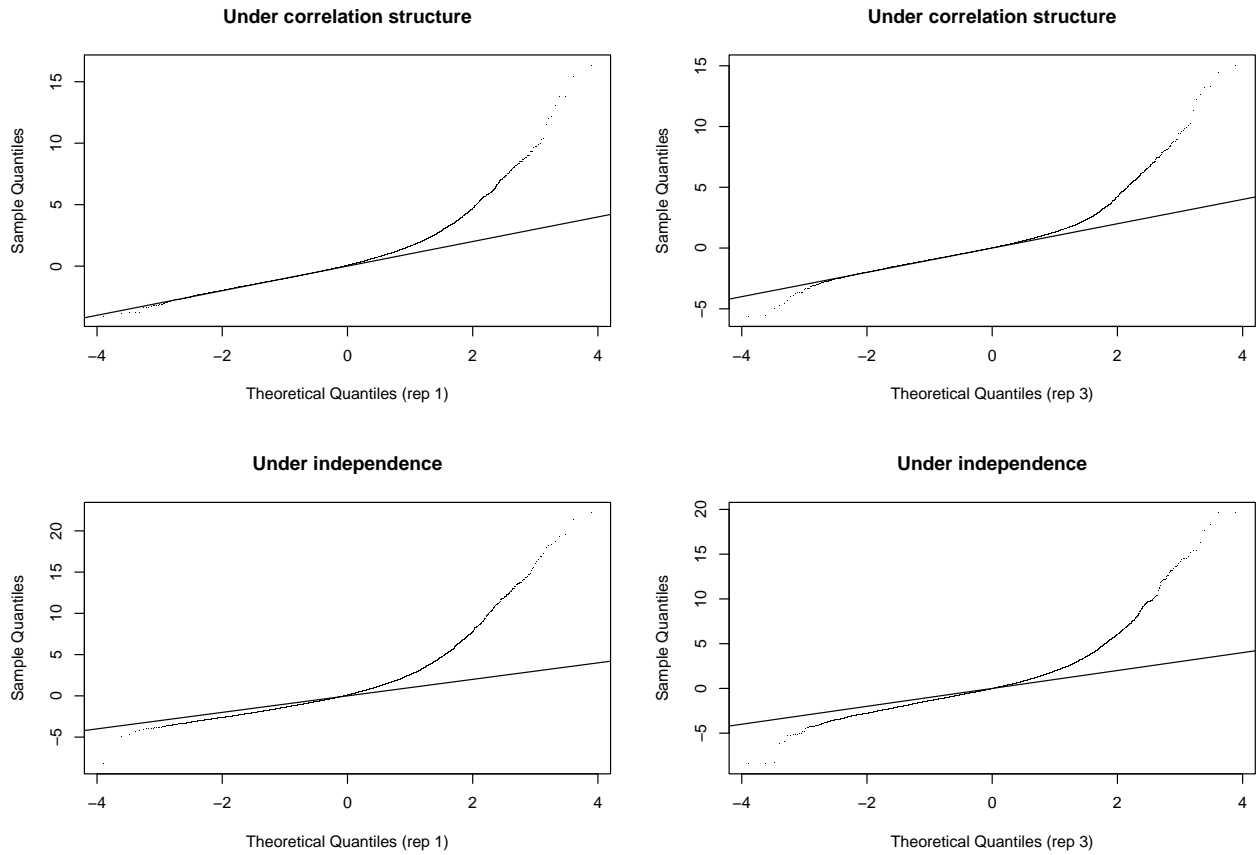
Figure 6: *Normal QQ plots for replicates 1 and 3.* The plots show improvement when the correlation structure is taken into account.

7

FDR=0.01

	CMARRT	Indep	TileMap
Common peaks	0.829(679/819)	0.819(1423/1736)	0.718(799/1113)
% of peaks within ±2kb	0.373	0.278	0.136
% of peaks within ±10kb	0.638	0.565	0.442
% of peaks within ±100kb	0.916	0.903	0.824

FDR=0.05

	CMARRT	Indep	TileMap
Common peaks	0.834(972/1165)	0.790(1796/2272)	0.714(978/1370)
% of peaks within ±2kb	0.326	0.267	0.134
% of peaks within ±10kb	0.609	0.565	0.431
% of peaks within ±100kb	0.904	0.900	0.826

FDR=0.10

	CMARRT	Indep	TileMap
Common peaks	0.836(1128/1350)	0.779(2096/2689)	0.703(1071/1524)
% of peaks within ±2kb	0.316	0.265	0.135
% of peaks within ±10kb	0.590	0.561	0.428
% of peaks within ±100kb	0.903	0.894	0.821

FDR=0.15

	CMARRT	Indep	TileMap
Common peaks	0.828(1261/1523)	0.763(2301/3051)	0.701(1171/1671)
% of peaks within ±2kb	0.300	0.259	0.136
% of peaks within ±10kb	0.578	0.552	0.434
% of peaks within ±100kb	0.902	0.890	0.827

FDR=0.20

	CMARRT	Indep	TileMap
Common peaks	0.82(1394/1695)	0.759(2506/3300)	0.693(1245/1797)
% of peaks within ±2kb	0.293	0.246	0.132
% of peaks within ±10kb	0.569	0.540	0.429
% of peaks within ±100kb	0.895	0.887	0.824

Table 1: *Distance of ZNF217-binding sites relative to TSS.* Each subpanel shows the results at different FDR threshold. The first row shows the percentage of overlaps in the binding sites between the two replicates. The next 3 rows show the number of binding sites located within the prespecified range of TSS.