CMARRT: A tool for the analysis of ChIP-chip data from tiling arrays by incorporating the correlation structure

Pei Fen Kuan[1], Hyonho Chun[1] and Sündüz Keleş[1,2*]

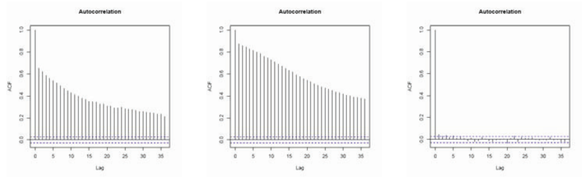
[1]Department of Statistics, [2]Department of Biostatistics and Medical Informatics, University of Wisconsin,Madison

*E-mail: keles@stat.wisc.edu

Tiling array designs

• Used in ChIP-chip experiments for studying protein-DNA interactions.

• Overlapping probe design and fragmentation of DNA sample hybridized result in correlation among probes mapping to consecutive genomic locations.



• ChIP-chip experiments produce massive amount of data.

• Require rapid and robust analysis methods.

• Commonly used methods (moving average statistics) do not address these issues.

Moving average statistics

• Can be defined over a fixed number of probes or fixed genomic distance.

• Let $Y_1,\ldots,Y_N$ denote measurements on the N probes.

• Let w_i be a window size of $2w_i+1$ (w_i probes to the left and right of i-th probe).

• Moving average statistics for probe i: $T_i = \frac{1}{2w_i+1}\sum_{j=i-w_i}^{i+w_i} Y_i$.

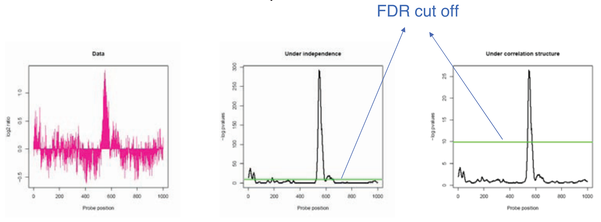
• $\mathrm{var}(T_i) = \frac{1}{(2w_i+1)^2}\left((2w_i+1)\sigma^2 + \sum_{j=i-w_i}^{i+w_i}\sum_{k\neq j}\mathrm{cov}(Y_j,Y_k)\right)$.

• Standardized moving average statistics $S_i = \frac{T_i}{\sqrt{\mathrm{var}(T_i)}}$.

• Standard practice ignores the covariance term in $\mathrm{var}(T_i)$ and obtains null distribution under hypothesis of no binding at probe i by permutation or Gaussian approximation (assuming Y_i's iid distributed as normal random variables).

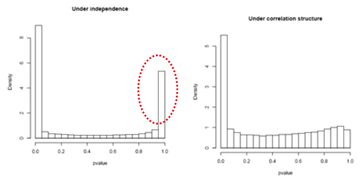
Problem with ignoring the correlation structure

• Increase in the number of false positives.

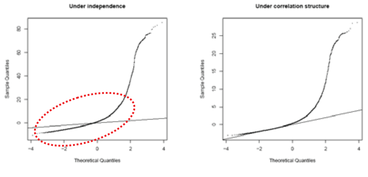


• Misspecification of distribution for S_i .

➢ Histogram of the p-values:
Expect a mixture of uniform distribution between 0 and 1 and a non-uniform distribution concentrated near 0.



➢ Normal QQ plots:
Expect quantiles of unbound probes to fall along a 45° reference line.



Estimating the correlation structure

• CMARRT (Correlation, Moving Average, Robust and Rapid method for Tiling array) :A fast empirical method using sample autocorrelation function $\hat{\rho}(k)$.

$$\hat{\rho}(k) = \frac{\sum_{t=1}^{T-k}(Y_t-\bar{Y})(Y_{t+k}-\bar{Y})}{\sum_{t=1}^{T}(Y_t-\bar{Y})^2}, \qquad \hat{\mathrm{cov}}(Y_j,Y_{j+k}) = \hat{\rho}(k)\hat{\sigma}^2$$

•Algorithm

➢Remove top M % of outlying/bound probes.

➢ Identify segments of at least N consecutive probes. Regions flanking large gaps or repeat masked regions are treated as two separate segments.

➢ For each segment j, compute $\hat{\rho}_j(k)$.

➢ For any lag k, let $\hat{\rho}(k)$ be average of $\hat{\rho}_j(k)$ over j.

Simulation studies

• Simulation 1 (Autoregressive (AR) model): Generate unbound data from an autoregressive (AR) model under different order and correlations.

• Simulation 2 (Duration Hidden Markov Model (HMM)) : Generate data from Hidden Markov Model (HMM) with explicit duration distribution to introduce direct dependencies at probe level observations.

• Compare the performances of:

➢ CMARRT: our proposed moving average method that estimates the correlation structure.

➢ Indep: moving average method that ignores the correlation structure as implemented in ChIPOTle.

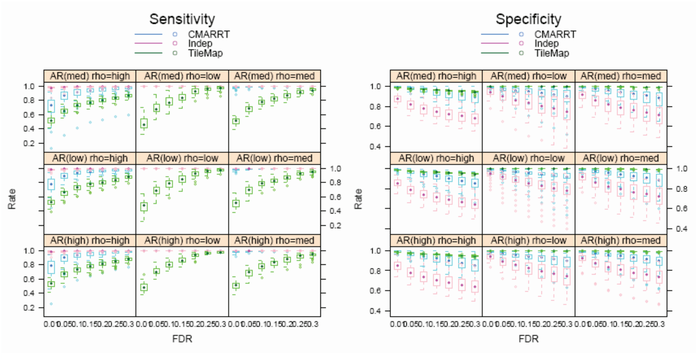
➢ TileMap (HMM option): a HMM approach for inferring bound or unbound hidden states.

• A probe is declared bound if its adjusted p-value (Y.Benjamini *et al.* (1995)) is smaller than FDR cutoff for CMARRT and Indep.

• For TileMap, use direct posterior probability approach (M.Newton *et al.* (2004)) to control FDR.

• Calculate sensitivity and specificity at various FDR controls.

Results: Simulation 1

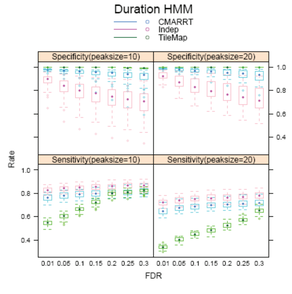


Results: Simulation 2

• CMARRT is superior in both sensitivity and specificity in all simulations.

• Indep has significantly lower specificity than CMARRT.

• TileMap more conservative than moving average approach.

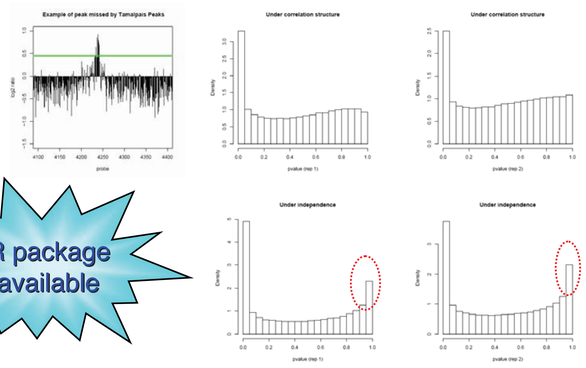


Case study: ZNF-217 ChIP-chip data

• Tiling arrays on ENCODE regions from Krig *et al.* One 50-mer every 38 bp.

• Krig *et al.* identified bound region with Tamalpais Peaks program.

• Analyze 2 replicates.



R package available

• Determine the extent of overlaps between the set of peaks in the two replicates.

• Determine the location of identified bound regions relative to transcription start sites using GENECODE genes.

FDR=0.01	CMARRT	Indep	TileMap
Common peaks	0.803	0.819	0.718
%of peaks within±2kb	0.334	0.278	0.136
%of peaks within±10kb	0.619	0.565	0.442
%of peaks within±100kb	0.911	0.903	0.824
FDR=0.05	CMARRT	Indep	TileMap
Common peaks	0.806	0.790	0.714
%of peaks within±2kb	0.321	0.267	0.134
%of peaks within±10kb	0.589	0.565	0.431
%of peaks within±100kb	0.903	0.900	0.826
FDR=0.10	CMARRT	Indep	TileMap
Common peaks	0.805	0.779	0.703
%of peaks within±2kb	0.300	0.265	0.135
%of peaks within±10kb	0.579	0.561	0.428
%of peaks within±100kb	0.904	0.894	0.821
FDR=0.15	CMARRT	Indep	TileMap
Common peaks	0.794	0.763	0.701
%of peaks within±2kb	0.284	0.259	0.136
%of peaks within±10kb	0.564	0.552	0.434
%of peaks within±100kb	0.899	0.890	0.827

References

• M.J.Buck *et al.* (2005), Genome Biol. 6(11).

• T.H.Kim *et al.* (2005), Nature 436:876-880.

• H.Ji *et al.* (2005), Bioinformatics 21(18):3629-3636.

• S.R. Krig *et al.* (2007), J.Biol.Chem. 282(13):9703-9712.

• Y.Benjamini *et al.* (1995), JRSS-B. 57:289-300.

• M.Newton *et al.* (2004), Biostatistics 5:155-176.

• Kuan *et al.* (2007),*To appear in the Proceedings of the Pacific Symposium of Biocomputing.*

Acknowledgements:

Professor Robert Landick, Dept of Bacteriology, UW-Madison. This research has been supported in part by a PhARMA Foundation Research Starter Grant (P.K. and S.K.) and NIH grants 1-R01-HG03747-01 (S.K.) and 4-R-37-GM038660-20 (H.C.).