# Online Methods

# iDPT: An Integrative approach for *de novo* pattern recognition of differentially expressed events in enrichment-based next-gen sequencing studies

Angela H. Ting

Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic Foundation  
9500 Euclid Ave, Cleveland, OH 44195, USA  
Email: [tinga@ccf.org](mailto:tinga@ccf.org)

Hu Bo

Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation,   
9500 Euclid Ave, Cleveland, OH 44195, USA  
Email: [hub@ccf.org](mailto:hub@ccf.org)

Li Zhang  
*Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation,   
9500 Euclid Ave, Cleveland, OH 44195, USA  
Email:* [*li.zhang@ucsf.edu*](mailto:li.zhang@ucsf.edu)

Jie Na   
*Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation,   
9500 Euclid Ave, Cleveland, OH 44195, USA  
Email:* [*najiek@yahoo.com*](mailto:najiek@yahoo.com)

Byron H. Lee

*Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic Foundation,   
9500 Euclid Ave, Cleveland, OH 44195, USA  
Email:* [*leeb@ccf.org*](mailto:leeb@ccf.org)

# Yaomin Xu [[1]](#footnote-1)† *Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, OH 44195, USA Department of Biostatistics, Vanderbilt University School of Medicine Department of Biomedical Informatics, Vanderbilt Univeristy School of Medicine 2220 Pierce Avenue, Nashville, TN 37232, USA Email:* [*yaomin.xu@vanderbilt.edu*](mailto:yaomin.xu@vanderbilt.edu)

## Deconvolution using Bayesian mixture model

We use the case of DNA methylation for illustration. For each subject , let be the tag count for window assuming that follows a three-component Poisson mixture model. To simplify the notation, we ignore the subscript in this section. The mixture model can be written as

where and are the mixing weights, and represents a Poisson density function with rate parameter . In this step, our inference is based on Bayesian method for each subject, where we consider the vector of follows a Dirichlet prior. As stated in the main text, represents lack of signals captured during enrichment or generated by sequencing. represents genomic regions containing few CpG sites with low methylation enrichment. Therefore, we assume that and follows a Gamma prior. represents regions with very strong methylation enrichment, and we assume. In addition, the window-specific follows a common Gamma prior.

To fit the three-component mixture model, we introduce a multinomial variable

indicating the component membership. Then, for each subject the likelihood across windows given the observed tag counts and missing membership indicators is

where is the probability that the tag count comes from component for .

To accommodate additional uncertainty and flexibility, we consider a hierarchical prior structure. Markov chain Monte Carlo with Gibbs sampler2 is used to obtain the posterior estimation of the unknown parameters. The three-component model and the Poisson distribution of were chosen for simplicity and computational efficiency and can be refined with a more complex model for better performance. We use the simplistic model here for a proof-of-concept.

## Pattern recognition with posterior membership probability

### *Defining enrichment events based on posterior membership probability*

An event is defined as the presence of a pre-specified differential enrichment pattern. Let denote the event probability for a given window (on subject ,

where the posterior membership probability of window on subject , is from the 3rd component in the mixture model, i.e.,

.

To represent any given pattern at window on subject , we defined an event score, , as

where is a normalization constant defined as , with as the chromosomal mean of . The composite event score, , is constructed across samples as

where is a summary function across all subjects. We used the *minimum* of 25th *quantile* within groups. This approach reduces bias due to outliers or sampling errors and identifies sites with consistent patterns in the majority (75%) of samples within each group. We calculated the composite event scores using the posterior estimation of obtained from the mixture model.

### *Selecting windows based on the composite event score*

Based on the pattern recognition framework above, a simplistic cutoff strategy for feature extraction of significant events is to select windows that satisfy {, which is analogous to an odds ratio-based decision making by claiming a window harbouring the pre-specified pattern if it is 0.5 fold higher than the chromosomal average. We used the cutoff of -1 to minimize the ambiguity of claiming multiple patterns for a given window while allowing for higher sensitivity to select event windows.

## Whole genome scan of event sites based on scan statistics

An event site is defined as a stretch of contiguous sequence on the genome that is enriched with unusually large number of events with the same enrichment pattern. We utilized discrete scan statistics1 to identify these sites. The discrete binary scan statistic in a sequence of binary trials, is defined as the maximum number of events within any consecutive windows, which can be formulated as,

Using , the probability that is less than any given positive number , we can probabilistically select the event sites with variable sizes. However, the exact evaluation of this quantity is computationally intractable. Therefore, we used an approximation and bounds developed by Glaz1 for efficient calculation. Based on this, we applied a tree-based dynamic search algorithm to systematically identify all the event sites throughout the genome. Briefly, for each whole-genome search process, we first constructed a tree using the event positions and subsequently searched through the tree to identify all event sites. The decision of unusualness is based on and clusters with probabilities less than 0.001 are selected as event sites for downstream analysis. and are chosen to ensure the approximation and bounds theory are valid when calculating . We used 0.001 as a cutoff to allow detection of not only sites containing consecutive windows but also those containing small gaps.

## Differential testing with linear mixed-effects model

For each event site identified during the whole genome pattern recognition and scan, a linear mixed-effects model was used to assess the significance and the effect size of the differential ratio, so

is the vector of methylation signal, where takes the natural logarithm of the posterior estimate of the mean signals obtained during deconvolution for windows within the same site for subject . is the design matrix of the group comparisons. In an example of two-group comparison with two samples in each group, assuming three windows in the site, we have. is the regression coefficient, and its estimate provides the differential ratios between the contrasted groups. represents experimental co-factors or conditions with as the regression coefficient. is the vector of the random effects of the multiple windows within the same site, letting The structure of variance-covariance matrix is determined by the experimental design. For instance, if all subjects are independent, is a block diagonal matrix with the diagonal block being an by matrix. When handling complex sampling schemes, such as paired samples and hierarchically nested designs, more complicated variance-covariance structure is needed. represents the random error and is independently distributed with and for subject . The variance components of and are estimated using the restricted maximum likelihood method3.

## Software

An open source software package based on R is freely available at <http://idpt.github.com/dptscan/>.

# References

1. Glaz, J., Pozdnyakov, V. & Wallenstein, S. *Scan Statistics: Methods and Applications* (Birkhauser, 2009).

2. Casella, G. & George, E. I. Explaining the Gibbs sampler. *American Statistician* 167-174 (1992)

3. Pinheiro, J. C. & Bates, D. M. Mixed-effects models in S and S-PLUS (2000) New York. *NY: Springer*  (2000).

1. † Corresponding author [↑](#footnote-ref-1)