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

HU BO

Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation,
9500 Euclid Ave, Cleveland, OH 44195, USA
Email: 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

JIE NA

Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation,
9500 Euclid Ave, Cleveland, OH 44195, USA
Email: na.jie@mayo.edu

BYRON H. LEE

Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic Foundation,
9500 Euclid Ave, Cleveland, OH 44195, USA
Email: leeb@ccf.org

YAOMIN XU †

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

† Corresponding author

Deconvolution using Bayesian mixture model

We use the case of DNA methylation for illustration. For each subject i, let Y_{ij} be the tag count for window $j, j = 1, \dots, n,$ assuming that Y_{ij} follows a three-component Poisson mixture model. To simplify the notation, we ignore the subscript i in this section. The mixture model can be written as

$$g(y_j) = p_1 f(y_j; \lambda_1) + p_2 f(y_j; \lambda_2) + p_3 f(y_j; \lambda_{j3}),$$

where p_1, p_2 and p_3 are the mixing weights, and $f(y_j; \lambda)$ represents a Poisson density function with rate parameter λ. In this step, our inference is based on Bayesian method for each subject, where we consider the vector of p_1, p_2, p_3 follows a Dirichlet prior. As stated in the main text, $f_1 = f(y_j; \lambda_1)$ represents lack of signals captured during enrichment or generated by sequencing. $f_2 = f(y_j; \lambda_2)$ represents genomic regions containing few CpG sites with low methylation enrichment. Therefore, we assume that $\lambda_1 = 0$ and $\lambda_2 (> 0)$ follows a Gamma prior. $f_3 = f(y_j; \lambda_{j3})$ represents regions with very strong methylation enrichment, and we assume $\lambda_{j3} > \lambda_2$. In addition, the window-specific λ_{j3} follows a common Gamma prior.

To fit the three-component mixture model, we introduce a multinomial variable $Z_j = (z_{j1}, z_{j2}, z_{j3})^T$ indicating the component membership. Then, for each subject, the likelihood across n windows given the observed tag counts and missing membership indicators is

$$\prod_{j=1}^{n} f(y_j;\ \lambda_1)^{z_{j1}} f(y_j;\ \lambda_2)^{z_{j2}} f(y_j;\ \lambda_{j3})^{z_{j3}}\ w_{j1}^{z_{j1}} w_{j2}^{z_{j2}} w_{j3}^{z_{j3}},$$

where $w_{jl} = P(z_{jl} = 1)$ is the probability that the tag count y_j comes from component l for $l = 1,2,3$.

To accommodate additional uncertainty and flexibility, we consider a hierarchical prior structure. Markov chain Monte Carlo with Gibbs sampler[2] is used to obtain the posterior estimation of the unknown parameters. The three-component model and the Poisson distribution of f 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 π_{ij} denote the event probability for a given window j ($j = 1, \dots, n$) on subject i ($i = 1, \dots, m$),

$$\pi_{ij} = \begin{cases} w_{ij}, \text{if window } j \text{ on subject } i \text{ is methylated} \\ 1 - w_{ij}, \text{if window } j \text{ on subject } i \text{ is not methylated,} \end{cases}$$

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

$$w_{ij} = w_{ij3} = P\left(z_{ij3} = 1\right).$$

To represent any given pattern at window j on subject i, we defined an event score, e_{ij}, as

$$e_{ij} = \alpha_i \frac{\pi_{ij}}{1 - \pi_{ij}},$$

where α_i is a normalization constant defined as $\alpha_i = {(1 - \bar{\pi}_i)}/{\bar{\pi}_i}$, with $\bar{\pi}_i$ as the chromosomal mean of π_{ij}. The composite event score, e_j^c, is constructed across samples as

$$e_j^c = h\left(e_{1j,} \dots, e_{ij}, \dots, e_{mj}\right),$$

where h 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 w_{ij} 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 $\{e_j^c > -1, j = 1, \dots, n\}$, 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 statistics[1] to identify these sites. The discrete binary scan statistic $S_{n,r}$ in a sequence of n binary trials, $U_j = I[e_j^c > -1], j = 1,2, \dots, n,$ is defined as the maximum number of events within any r consecutive windows, which can be formulated as,

$$S_r^j = \sum_{i=j}^{j+r-1} U_i \, , j = 1, \dots, n - r + 1,$$

$$S_{n,r} = \max_{1 \le j \le n-r+1} S_r^j \, .$$

Using $P(S_{n,r} < s)$, the probability that $S_{n,r}$ is less than any given positive number s, 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 Glaz[1] 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 $P(S_{n,r} < s)$, and clusters with probabilities less than 0.001 are selected as event sites for downstream analysis. r and s are chosen to ensure the approximation and bounds theory are valid when calculating $P(S_{n,r} < s)$. 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

$$\boldsymbol{Y} = \boldsymbol{G}\beta + \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{v} + \boldsymbol{\epsilon},$$

$\boldsymbol{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_i, \dots, \boldsymbol{y}_m)^T$ is the vector of methylation signal, where $\boldsymbol{y}_i = (y_{i1}, \dots, y_{ij}, \dots, y_{in})^T$ takes the natural logarithm of the posterior estimate of the mean signals obtained during deconvolution for n windows within the same site for subject i. $\boldsymbol{G}$ 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 $\boldsymbol{G} = (0,0,0,0,0,0,1,1,1,1,1,1)^{\boldsymbol{T}}$. β is the regression coefficient, and its estimate provides the differential ratios between the contrasted groups. $\boldsymbol{X}$ represents experimental co-factors

or conditions with $\boldsymbol{\gamma}$ as the regression coefficient. $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_m)^T$ is the vector of the random effects of the multiple windows within the same site, letting $\boldsymbol{v} \sim N(0, \boldsymbol{V})$. The structure of variance-covariance matrix $\boldsymbol{V}$ is determined by the experimental design. For instance, if all subjects are independent, $\boldsymbol{V}$ is a block diagonal matrix with the diagonal block V_i being an n by n matrix. When handling complex sampling schemes, such as paired samples and hierarchically nested designs, more complicated variance-covariance structure is needed. $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_i, \ldots, \boldsymbol{\epsilon}_m)^T$ represents the random error and is independently distributed with $\boldsymbol{v}$ and $\boldsymbol{\epsilon}_i \sim N(0, D_i)$ for subject i. The variance components of V_i and D_i are estimated using the restricted maximum likelihood method[3].

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).