Application of Gene Set Enrichment Method to ChIP-chip Data Analysis

Yidong Chen, Fan Yang, Paul S. Meltzer
Genetics Branch, Center for Cancer Research
National Cancer Institute, Bethesda, Maryland 20892
yidong@mail.nih.gov

ABSTRACT
To elucidate biological functions from gene expression profiles, gene set enrichment analysis (GSEA) is widely applied against sets of predefined genes that may yield crucial clues to their functional themes or regulatory information. However, gene list derived from array-based chromatin-immunoprecipitation (ChIP-chip) experiments, where all genes with one or more binding sites of a given protein, possesses different characteristics than that from gene expression profiles: genes are not rank-ordered by their differential expression, but rather associated with a genomic distance to its nearest binding site from the transcription start site (TSS). In this study, we proposed a unique binding-site enrichment analysis method that enabled enrichment analysis of gene list derived from whole-genome ChIP-chip experiment to gene expression data set, such as a panel of normal tissue gene expression profiles or some cancer-related expression profiles, in order to identify the essential regulatory role of the transcription factor under study.

1. INTRODUCTION
Currently, high-throughput gene expression platforms generate large sets of genes with different biological sample specimen, under different conditions and treatments. Manual evaluation of these gene sets for their biological functions or underlying processes is impractical. To alleviate this burden, Gene Set Enrichment Analysis (GSEA) was a computational method developed for statistical evaluation of sorted gene lists [3,5,9-10]. Given a rank-ordered gene list and another set of genes from, e.g., an arbitrary gene ontology category, GSEA evaluates whether or not that genes of specific functions were distributed randomly (not enriched) or amassed specifically at top (enriched) of the sorted genes list, assessed with a statistically significance score, p-value.

Different from gene expression profiling, whole-genome ChIP-chip probes entire genome for binding locations of a specific transcription factor [1]. Genes identified via ChIP-chip assay are those potentially regulated by a nearby regulatory element targeted by a transcription factor. Further different from expression profiles where genes are ranked based on their association to the phenotype of interest, e.g., t-statistics, gene lists obtained from ChIP-chip studies often do not possess rank-order and the set is generally quite large (in thousands). Taking into consideration of this property, we reversed the common GSEA application where genes derived from expression data were test for enrichment of certain biochemical pathways or gene ontology categories; instead, we proposed to test ChIP-chip gene list against various gene expression data obtained from NCBI Gene Expression Omnibus (GEO) for enrichment of possible regulation implied in these expression data. In following sections, detailed methods were presented, and examples of applications to Estrogen Receptor (ER) ChIP-chip data against a panel of ER-regulated gene expression studies, biologically irrelevant Glioma data set and normal tissue profiles were presented.

2. METHODS
Let \( G = \{g_1, g_2, \ldots, g_n\} \) denotes ChIP-chip derived gene set with \( n \) genes. Sorted gene list, \( L_k \), obtained from \( k \)th gene expression study, is obtained from a statistical test for its association with a given phenotype (e.g., ER -/+ status for ER regulated genes). A proposed enrichment statistic modified from [10] is,

\[
s_k = \frac{1}{n} \sum_{i=1}^{n} (-\log_{10} p_{i,k})
\]

where \( p_{i,k} \) is the p-value of test statistic for phenotypic association of \( i \)th gene in \( G \) that matches the same gene in \( L_k \) for \( k \)th expression study. The statistical significance of \( s_k \) can be assessed by random permutation of sample labeling (phenotype, treatment, etc) in \( k \)th expression study. However, it may be impractical when large panel of gene expression studies are involved. In this study, we proposed a simple random draw from all genes in gene list \( L_k \), such that,

\[
s_k^* = \frac{1}{n} \sum_{i=1}^{n} (-\log_{10} p_{*,k})
\]

where * indicates that the random draws were performed for the value of \( p \) in Eq. (2). The random
draw proposed in Eq. (2) is equivalent to random sample labeling when all genes are independent. With whole genome gene expression studies that commonly contain more than 40,000 genes, and more than 100 samples in one study, this approximation provides a good significance assessment. Therefore, significance of \( s_k \), or \( p_s \), is,

\[
p_s = \frac{\sum_{j=1}^{K} I(s_k > s_{k,j})}{K} \tag{3}
\]

where \( I(\bullet) \) is the indicator function giving 1 if the argument is true and 0 otherwise, \( K \) is the number of random draw. Note that \( s_k > 0 \) as defined by Eq. (1).

For multiple test correction when dealing with more than one expression studies, Eq. (3) can be easily extended to FDR controlled \( p \)-value [5, 10].

Equations (1) to (3) provide a simplified GSEA version that in fact applicable to expression data as well. To compare the ChIP derived gene list along with the expression derived gene lists, a modified significance assessment procedure was proposed as follows: for one reference gene list \( L_{ref} \), we first randomly selected \( d \) genes \( v \) times (e.g., \( v = 10,000 \)) to form an empirical distribution \( f_d(s) \), as the distribution under the null hypothesis (no enrichment), and its 99% upper limit, \( c_{0.99} \), was then determined. For any gene list to be tested (from top \( M \) significant differential expressed genes or randomly from all ChIP derived genes), randomly selected \( d \) genes \( v \) times to form \( f_i(s) \), such that,

\[
q = \int_{c_{0.99}}^{s} f_1(x)dx \tag{4}
\]

where \( q \) is the probability that enrichment score \( s \) falls within 99% upper limit. \( q = 0 \) when gene list to be tested is highly enriched in \( L_{ref} \), while \( q = 1 \) when gene list is not enriched. The advantage of applying Eq. (4), rather than Eq. (3) is that we standardize \( d \) genes with random selection, regardless of the gene set sizes from various expression studies, enabling meaningful comparison cross data sets. Moreover, Eq. (3) controls false positive rate, while Eq. (4) treats enrichment score as a distribution and thus controls both Type I and Type II error.

3. RESULTS

Estrogen receptor ChIP-chip data were obtained from [1] that contained 3,665 binding sites. As a comparison, 3,629 Pol-II binding sites were also provided from [1]. We have prepared three ER +/- expression profiling data sets: 1) van de Vijver study [11] with 295 primary breast tumors; 2) Sotiriou study with 189 primary operable invasive breast tumors [7]; 3) Neve study [6] with 51 breast cancel cell lines. All data sets were analyzed for differential expression due to ER +/- status via R/Limma package, and genes were then ordered via the adjusted \( p \)-value (Benjamini-Hochberg multiple test correction). For comparison purpose, we have also prepared a set of Glioma expression profiles [2] and genes were rank ordered by their differential expression comparing the grade III to IV. The Glioma data set was selected as a negative control irrelevant to ER data set biologically.

Figure 1 illustrated the results from the proposed enrichment test: solid blue bar-chart was derived from null hypothesis (random selection of 200 genes) by using van de Vijver data set as the reference data set, and the left-most line anchored the perfect result should test sample be van de Vijver as well. The red-line showed enrichment of ER binding genes to ER regulated genes (via van de Vijver expression data). Two other ER +/- expression studies showed significant enrichment (light blue and green lines), which were expected due to the similarity of studies, while Glioma data set showed much weak enrichment (\( p \approx 0.05 \)). For ER binding genes within 100kbp (total of 1,335), the enrichment score was about 3.54 (\( p < 0.01 \)), which was barely different from Glioma (negative control). By selecting genes within 10kbp (solid red-line), the enrichment score was significantly improved (\( p \approx 5 \times 10^{-13} \)) as shown in Figure 1.

![Figure 1](image-url)

Figure 1. Enrichment test. Black bar chart was derived from Eq. 2 with randomly selected 200 genes from van de Vijver data set (worst case) and the solid blue-line was obtained from randomly selected 200 genes out of top 1000 genes (best case), also from van de Vijver data set (reference data set). Enrichment scores derived from two other ER regulated gene expression studies (light blue and green) were also provided with their enrichment \( p \)-values. The red line with ‘+’ markers was derived from 1,335 genes with ER binding sites less than 100kbp away from TSS (200 genes were randomly selected to show the variation). Significant enrichment was observed with ER binding genes within 10kbp (red-line). Densities were estimated via kernel smoothing method.

To visualize the specificity of gene sets relatively to the percent of genes selected for certain criteria, we selected top \( m\% \) genes according to ER regulation significance for expression studies, and \( m\% \) of genes according to the distance to binding sites (ER or Pol-II). Figure 2 illustrated the results for gene set

\[
f_1(x) = \sum_{j=1}^{K} I(s_k > s_{k,j}) 
\]
specification, in which 50% genes from van de Vijver data set were significant (expected since the reference data set was van de Vijfer), other expression studies shown specification when greater than 20% genes were enriched in ER regulation, except for that from Glioma data where no significance at any selection level. For ChIP-chip results, ER binding genes were significant if we constrained by their distance between binding site to TSS (about 5%). On the other hand, the Pol-II ChIP data yielded no significance, indicating no specificity to ER regulation. Specificity test of ER binding genes were carried out with panel of normal tissues (Novartis set [8] with 46 tissue types profiled with Affymetrix U95A arrays). Only tissues from Placenta and Trachea showed some enrichment in ER regulated genes.

![Figure 2](image)

**Figure 2.** Specificity of the enrichment test. Percent of genes selected from each data set that overlapped with gene symbols between two array platforms (van de Vijver data set as reference data), rather than the exact number of genes, were obtained and their enrichment significance *p*-value were evaluated by Eq. (4). Null hypothesis distribution was generated from 10,000 selections (with replacement) of 200 random genes from van de Vijver data set. When the number of selected genes under various criteria was less than 200, bootstrapping method was applied. The number of genes selected was determined by the adjusted *p*-value per expression study, while ChIP genes were selected by their distance to the nearest binding site.

4. CONCLUSIONS
Considering the uniqueness of ChIP-chip experiment derived gene set, we proposed a gene set enrichment method that explored the gene set enrichment in various publicly available gene expression studies, rather than common GSEA applications that provided enrichment to functional categories or biological pathways. By doing so, the connection of ChIP derived gene set to specifically designed expression studies can be established, and thus further functional study may be carried out for their implied GO categories and biochemical pathways, as well as additional hypotheses may be formulated.

5. REFERENCES