BIOINFORMATICS

Vol. 30 ECCB 2014, pages i393–i400
doi:10.1093/bioinformatics/btu444

Broad-Enrich: functional interpretation of large sets of broad genomic regions

Raymond G. Cavalcante[1], Chee Lee[1], Ryan P. Welch[1,2], Snehal Patil[3], Terry Weymouth[3], Laura J. Scott[2] and Maureen A. Sartor[1,2,3,*]

[1]Department of Computational Medicine and Bioinformatics, [2]Department of Biostatistics and [3]Center of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

ABSTRACT

Motivation: Functional enrichment testing facilitates the interpretation of Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) data in terms of pathways and other biological contexts. Previous methods developed and used to test for key gene sets affected in ChIP-seq experiments treat peaks as points, and are based on the number of peaks associated with a gene or a binary score for each gene. These approaches work well for transcription factors, but histone modifications often occur over broad domains, and across multiple genes.

Results: To incorporate the unique properties of broad domains into functional enrichment testing, we developed Broad-Enrich, a method that uses the proportion of each gene's locus covered by a peak. We show that our method has a well-calibrated false-positive rate, performing well with ChIP-seq data having broad domains compared with alternative approaches. We illustrate Broad-Enrich with 55 ENCODE ChIP-seq datasets using different methods to define gene loci. Broad-Enrich can also be applied to other datasets consisting of broad genomic domains such as copy number variations.

Availability and implementation: http://broad-enrich.med.umich.edu for Web version and R package.

Contact: sartorma@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) identifies transcription factor (TF) binding sites and the locations of histone modifications (HMs) across the genome (Barski *et al.,* 2007), and is a step toward better understanding the gene regulatory programs of living organisms. Numerous algorithms, termed peak callers, have been developed to detect the genomic regions of significant signal (peaks) within the millions of aligned reads resulting from ChIP-seq experiments (Bailey *et al.*, 2013; Valouev *et al.*, 2008; Zhang *et al.*, 2008). Some of these peak callers are geared specifically to HMs, which are known to exhibit broader enriched domains on average compared with TFs (Zang *et al.*, 2009). HMs are numerous and varied, and like TFs, often drive the regulation of a specific biological program, such as cellular differentiation (Sen *et al.*, 2008) or growth (Bernstein *et al.*, 2006). Specific signatures often occur at HM intersections, such as the bivalent domains observed for H3K4me3 and H3K27me3, which mark genes expected to be activated on cellular differentiation (Bernstein *et al.*, 2006; Pan *et al.*, 2007). Other histone changes occur in disease progression (Chi *et al.*, 2010) or in response to environmental signals (Kaelin and McKnight, 2013). Such signatures are likely often cell-type and context specific, and therefore, assessing the biological commonalities among the targeted genes is a question of intense interest.

Gene set enrichment (GSE) is a common approach to infer biological function given a set of experimentally derived genes (Draghici *et al.*, 2003). GSE was originally developed to biologically interpret lists of differentially expressed genes derived from microarray studies (Curtis *et al.*, 2005) in terms of particular biological functions, processes or pathways [e.g. Gene Ontology (GO) (Ashburner *et al.*, 2000) or KEGG Pathways (Kanehisa and Goto, 2000)]. An early enrichment tool is DAVID (Huang *et al.*, 2008), which uses a slightly modified Fisher's exact test (FET) to determine whether experimentally derived genes significantly overlap a gene set representing a biological concept, relative to the remaining genes. Under the null hypothesis of no more overlap than expected by chance, FET assumes that each gene has the same probability of being detected as significant. In the context of GSE with ChIP-seq data, FET assumes that each gene has an equal probability of being associated with a peak. Although FET has been used with ChIP-seq data (Blow *et al.*, 2010; Han *et al.*, 2013), it is typically used only with peaks within or near gene promoters. When all peaks are used, the presence of a peak in a gene locus is often correlated with the length of the locus (Ovcharenko *et al.*, 2005), thereby violating the FET assumption. We refer to this correlation as the locus length bias. Given that some gene sets contain genes that have, overall, significantly longer (e.g. nervous system, development and transcription related) or shorter locus length (e.g. metabolic processes and stimulus responses) than the average locus length, the possibility of confounding exists when no correction is made for locus length (Taher and Ovcharenko, 2009). Using FET with only peaks near gene promoters removes nearly all of the length bias, but also ignores a large portion of the data.

Recent GSE tools for ChIP-seq experiments have attempted to correct for this length bias. One such tool, Genomic Regions Enrichment of Annotations Tool (GREAT), uses a binomial-based test to test whether the total number of peaks within the loci in a gene set is greater than expected relative to the total number of peaks, the total locus length of the gene set and the non-gapped length of the genome (McLean *et al.*, 2010). In contrast to FET, the binomial test of GREAT assumes that the number of peaks in a locus and the locus length are proportional. Thus, FET and the binomial test have opposing assumptions regarding the relationship between the presence of a peak in a

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

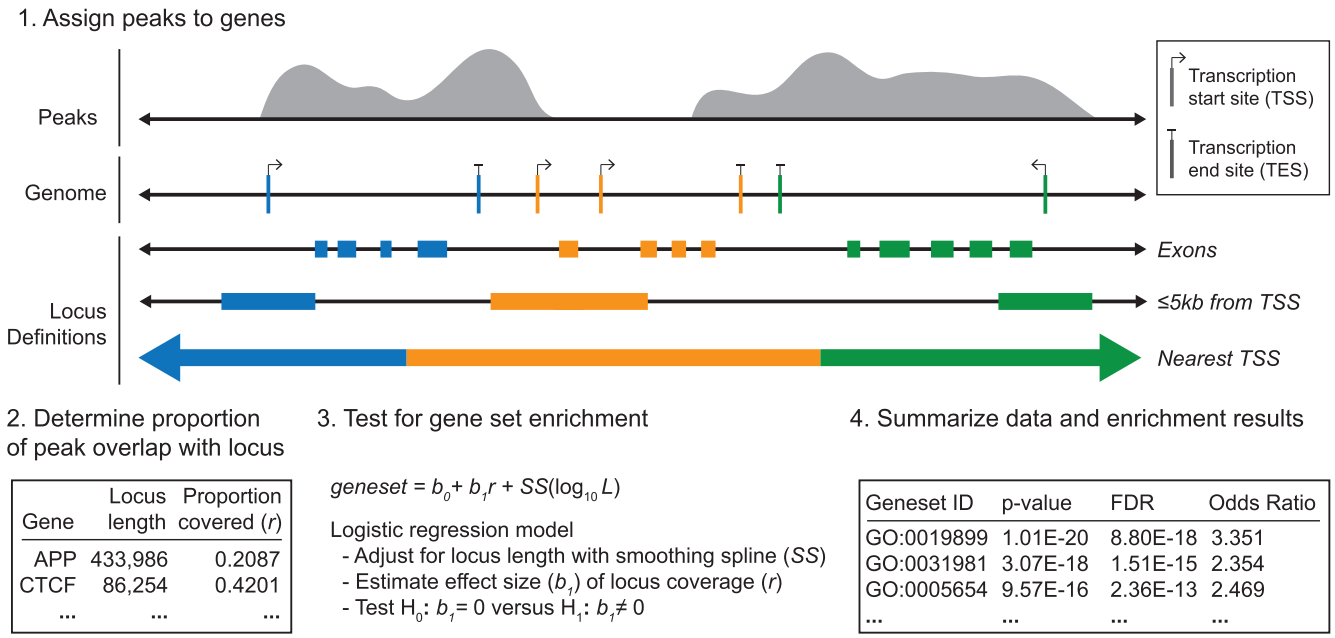


Fig. 1. Broad-Enrich functions in four steps. (1) The user selects a gene locus definition (*exons*, *≤5 kb* and *nearest TSS* are shown). (2) The proportion of each gene locus covered by ChIP-seq peaks from a given HM, or otherwise derived genomic regions, is determined. (3) For each gene set to be tested, logistic regression is performed using the model shown, where *geneset* refers to the binary vector of gene set membership, *r* refers to the vector of proportions of the gene loci covered by all peaks overlapping the respective loci, *SS* is a binomial cubic smoothing spline that corrects for any locus length bias and *L* is a vector of gene locus lengths. (4) *P*-values for enrichment or depletion are adjusted for multiple testing, and users are provided summarized functional enrichment results, peak to gene loci assignments and diagnostic plots

genomic region and the length of that region. Although FET is typically used after classifying each gene as either (i) having at least one associated peak or (ii) having no peak, the binomial test uses the total number of peaks. Both methods typically use a single nucleotide point, the midpoint or mode of the peak, to represent the entire peak region.

We examined 100 TF and 55 HM ChIP-seq experiments from ENCODE (ENCODE Project Consortium *et al.*, 2012a) for differences between peak sets from transcription factor- and histone-based ChIP-seq experiments. HM peak sets have been observed to have broader peak regions than TFs, with individual peaks often spanning multiple genes (Zang *et al.*, 2009). We hypothesized that an enrichment method using such relevant regulatory information rather than simply the midpoint of each peak, as both FET and the binomial test do, would improve performance for HMs and other experiments resulting in broad domains.

To incorporate the properties of broad-domain peak sets into functional enrichment testing, we developed Broad-Enrich to functionally interpret large sets of broad genomic regions. A unique feature of our method is that we score gene loci according to the proportion of the locus covered by all peaks overlapping the locus, which we will refer to as the coverage proportion. Broad-Enrich then uses a logistic regression model that empirically adjusts for any bias in gene locus coverage relative to locus length, avoiding the pitfalls of either FET or binomial-based tests. We show that Broad-Enrich exhibits the correct type I error rate across 55 permuted ENCODE ChIP-seq datasets. We then illustrate the benefits of Broad-Enrich across the same set of 55 datasets, concentrating on H3K4me1,-2 and -3, H3K9me3, H3K27me3 and H3K79me2 in the GM12878 cell line.

2 MATERIALS AND METHODS

2.1 Gene locus definitions

We define a gene as the region between the furthest upstream transcription start site (TSS) and furthest downstream transcription end site (TES) for that gene. The UCSC knownGene table (human genome build hg19) was used to define TSS and TES sites. We removed small nuclear RNAs, as they are likely to have different regulatory mechanisms than other genes and often reside within the boundaries of other genes. For functional enrichment testing, we use three primary definitions of a gene locus (Fig. 1.1). (i) *Nearest TSS:* the region between the upstream and downstream midpoints of a gene's TSS and the adjacent gene's TSS, equivalent to assigning each peak to the gene with the nearest TSS. (ii) *≤5 kb*: the region within 5 kb of all TSSs in a gene. If TSSs from the adjacent gene(s) are <10 kb away, we use the midpoint between the two TSSs as the boundary of the locus for each gene. (iii) *Exons*: the exons of each gene. When exons from multiple transcripts of the same gene overlap, the exons are consolidated into one continuous region. In the R package and on the Web site, we include two additional definitions. (i) *Nearest gene*: the region from the midpoint between the TSS and the adjacent gene's TSS or TES (whichever is closest) to the midpoint between the TES and the adjacent gene's TSS or TES (whichever is closest). This is equivalent to assigning peaks to the nearest gene; (ii) *≤1 kb:* same as *≤5 kb*, but within 1 kb of all TSSs in a gene.

2.2 Proportional assignment of peaks to genes

A unique feature of Broad-Enrich is how peaks are assigned to gene loci. For a particular gene locus definition, each locus is scored according to the proportion covered by the union of all peaks overlapping the locus (Fig. 1.1). Our approach accounts for the extent to which a locus is covered by a peak and allows coverage by multiple peaks.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

2.3 Annotation databases

Functional enrichment results presented here are performed on gene sets constructed from the GO database and the KEGG Pathways database. We construct GO terms from GO biological processes, GO cellular components and GO molecular functions using the *org.Hs.eg.db* and *GO.db* R packages. All analyses in the article were performed using R version 3.0.1. KEGG pathways are inherited from LRpath (Kim *et al.*, 2012). Eleven additional annotation databases are offered in the R package, including cytoband regions, Biocarta (Nishimura, 2001) and Panther pathways (Mi *et al.*, 2012), pFAM (Punta *et al.*, 2011) and gene sets derived from literature-based Medical Subject Heading terms (Kim *et al.*, 2012; Sartor *et al.*, 2010). Before enrichment testing, all gene sets are filtered through the user-selected gene locus definition so that only genes with a locus definition are included in the tests. By default, only gene sets containing between 10 and 2000 genes are tested. A minimum of 10 genes allows better convergence of the logistic regression model used for enrichment (Peduzzi *et al.*, 1996) and the maximum of 2000 genes avoids general, less-informative gene sets. Annotation databases were built for human (hg19), mouse (mm9 and mm10) and rat (rn4).

2.4 Broad-Enrich method for functional enrichment testing

We use a logistic regression framework to test for functional enrichment, similar to LRpath (Sartor *et al.*, 2009), an enrichment testing method developed for microarray data. The independent variable r for Broad-Enrich is the vector of proportions of each gene's locus that is covered by the union of all peaks (Fig. 1 visually represents these proportions). The dependent variable is a binary vector indicating gene set membership (1 if the gene belongs to the gene set and 0 otherwise). Let π be the proportion of genes in the gene set at a specified r value and locus length L. Then, the ratio $\pi/(1-\pi)$ is the odds that a gene with peak coverage proportion r and locus length L is a member of a given gene set. If the log odds increase as r increases, then we conclude the gene set is positively associated with the coverage proportion, and thus enriched with the experimental set of broad genomic regions. We use the model:

$$\log\frac{\pi}{1-\pi} = b_0 + b_1 r + SS(\log_{10} L)$$

where b_0 is the intercept, b_1 is the coefficient of interest for the coverage proportion, the function SS is a binomial cubic smoothing spline that adjusts for the potentially confounding effect of locus length and the $\log_{10}$ transformation is used to improve the model fit (data not shown).

The smoothing spline function is fitted using generalized cross-validation to estimate the smoothing penalty, λ, and 10 knots with the cubic spline basis as an approximation to a true cubic smoothing spline (Wood, 2006; 2010). The overall model is fitted using a penalized likelihood maximization approach with the *gam* function in the *mgcv* R package (Wood, 2010). A Wald test is used to test the null hypothesis H_0: $b_1 = 0$ versus the alternative H_1: $b_1 \neq 0$ and to calculate the P-value for the significance of the coverage proportion coefficient, b_1 (Fig. 1.3). Gene sets with $b_1 > 0$ are enriched, whereas those with $b_1 < 0$ are depleted. P-values are corrected for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) adjustment (Benjamini and Hochberg, 1995). For presented analyses, gene sets with FDR < 0.05 are considered to be significant.

2.5 Experimental ChIP-seq peak datasets

We used 155 ENCODE ChIP-seq datasets from 31 DNA binding proteins: 11 HMs and 20 TFs across five cell lines (GM12878, H1-hESC, HeLa-S3, HepG2 and K562), representing the largest complete matrix of experiments of HMs and TFs among tier 1 and tier 2 cell lines. Peaks for the 55 HM datasets were called by the ENCODE Consortium using Scripture (ENCODE Project Consortium *et al.*, 2012b), and used as is. The 100 TF datasets were originally called using a variety of peak callers according to the lab of origin. We implemented a standard peak-calling pipeline for the TF datasets (Supplementary Methods and Supplementary Table S1).

2.6 Permutations to test type I error rate

Two permutation scenarios were performed to assess the type I error rate of the enrichment tests under the null hypothesis of no true biological enrichment with gene sets from GO. In both scenarios, gene labels are permuted so that each gene is given the GO term assignments of a randomly chosen gene. Preserved in both scenarios is the number of genes in a gene set and the correlations among the gene sets inherited from their parent/child relationships.

In the first scenario (referred to as Permuted), we randomly permute gene labels relative to locus length and peak coverage proportion. The resulting permutations remove true biological association and the locus length bias inherent in the GO terms. In the second scenario (referred to as Permuted in Bins), gene labels are randomly permuted within bins of 100 genes sorted by locus length. This has the effect of preserving the relationship between locus length and peak coverage proportion in the dataset. The resulting permutations remove true biological association in the gene sets while maintaining any locus length bias. Tests exhibiting inflated type I error under this scenario in excess of the first scenario can be considered as not appropriately accounting for locus length. Each type I error estimate was based on 5404 tests.

2.7 Alternative functional enrichment testing methods

We compared the functional enrichments for the 55 HM experiments (11 HMs across 5 cell lines) found with Broad-Enrich with those found by FET and our implementation of the binomial test of GREAT (McLean *et al.*, 2010). Additionally, we determined the type I error rate for a simplified version of the Broad-Enrich model excluding the smoothing spline [simple logistic regression (LR) model] to assess its necessity. Genes that were annotated in GO or KEGG and had a defined locus were included in the analyses. We used a two-sided FET to test for association of peak presence ($\geq$1 peak midpoint within a gene locus) and gene set membership. We used a binomial test similar to the one described in GREAT; we calculate the probability of seeing greater than or equal to the number of peaks we observe for a gene set, π, with the formula:

$$\sum_{i=k_\pi}^{n} \binom{n}{i} p_\pi^i (1-p_\pi)^{n-i}$$

where n is the total number of peaks within gene loci in any gene set, and k_π is the number of peaks annotated to gene set π. The term p_π is defined as the expected proportion of peaks in gene set π. In other words, p_π is the total non-gapped gene loci length in the gene set, divided by the total non-gapped length of loci with at least one gene set annotation. P-values are calculated as the probability of observing k_π or more peaks in the gene set.

We also used GREAT (http://bejerano.stanford.edu/great/) with hg19, the non-gapped genome as the background region, and the single nearest gene within 9999 kb association rule excluding curated regulatory domains.

3 RESULTS

3.1 Differences between histone- and TF-based ChIP-seq data

We examined peaks from 155 ENCODE ChIP-seq experiments including 20 TFs and 11 HMs in five cell lines. We find that, relative to TF-based experiments, ChIP-seq experiments

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

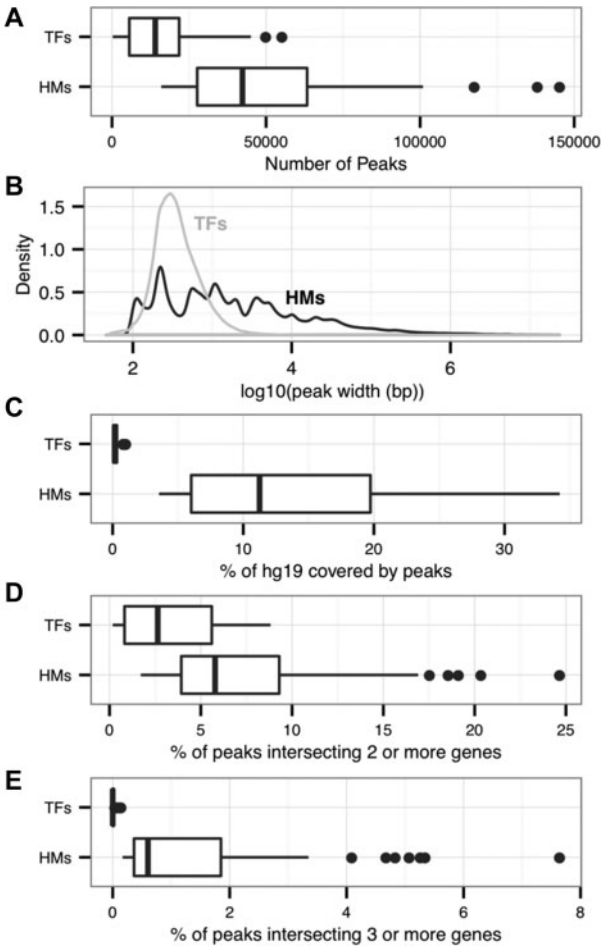


Fig. 2. HM- and TF-based peak sets exhibit several different properties, observed with 100 TF and 55 HM ENCODE ChIP-seq datasets. (**A**) There tends to be more peaks in HM experiments (median = 42 330) compared with TF experiments (median = 14 040). (**B**) The peak width distributions are significantly different. HM peaks (black) tend to be broad and highly variable (median = 1255 bp, SD = 483 279 bp), whereas TF peaks (gray) tend to be narrow and less variable (median = 330 bp, SD = 560 bp). (**C**) HM peaks consistently cover a greater percentage of hg19 (median = 11.25%) than TF peaks (median = 0.16%). (**D**) The percentage of peaks covering two or more gene loci also tends to be higher for HMs (median = 5.78%) than for TFs (median = 2.64%). (**E**) The same is true of peaks covering three or more gene loci (median = 0.6 and 0%, respectively). Both (D) and (E) use the *nearest TSS* definition

detecting HMs tend to have more peaks (Fig. 2A), broader peaks (Fig. 2B) and more variable peaks widths (Fig. 2B). We also find histone-based peaks tend to cover a much larger percentage of the hg19 genome (Fig. 2C).

In addition to more and broader peaks in the HM datasets, we observed that the HM datasets also tend to have a higher proportion of peaks intersecting two or more gene loci compared with TF datasets. With the *nearest TSS* locus definition, we find the percentage of peaks covering two or more gene loci tends to be higher for HMs (median = 5.78%, range = 1.71–24.66%) than for TFs (median = 2.64%, range = 0.17–8.82%) (Fig. 2D). Similarly, the percentage of peaks covering three or more loci is higher for HMs (median = 0.60%, range = 0.17–7.64%) than for TFs (median = 0%, range = 0.00–0.14%) (Fig. 2E). The properties observed in HM peak sets indicate current methods may be ill-suited for detecting functional enrichment in HM ChIP-seq data.

3.2 Broad-Enrich method

Based on the differences observed between TFs and HMs in ChIP-seq data, we aimed to develop an enrichment testing method that accounts for the extent to which each HM is associated with each gene. Using the number of peaks associated with a gene, as GREAT does, would yield stronger association to a gene with two narrow peaks than to a gene with one broad region covering the entire gene. Using a binary indicator of whether a gene has at least one peak associated with it, as is done with FET, would not account for any differences in the proportion of the gene locus covered. Both approaches ignore instances where a peak covers a significant portion of the loci of two or more genes.

We first define the gene locus definitions, which capture the main trends of where HMs tend to occur relative to exons and TSSs. In this article, we use (i) the region(s) within 5 kb of every TSS of a gene ($\leq 5\,kb$), (ii) the combined exon regions for a given gene (*exons*) and (iii) the region between the upstream and downstream midpoints between a gene's TSS and the adjacent gene's TSS (*nearest TSS*) (Fig. 1). These locus definitions represent binding in the greater promoter regions, throughout gene bodies and anywhere in the surrounding genomic region including enhancers (assigned to the gene with the nearest TSS), respectively.

Given a locus definition, the proportion of each gene locus covered by all peaks overlapping the locus is determined. To test for significant enrichment, we use a logistic regression approach with gene set membership as the outcome and the proportion of a locus covered as the predictor. Because of the known confounding effect of locus length relative to the presence of ≥ 1 peak (Taher and Ovcharenko, 2009), we examined and observed a similar relationship between locus length and peak coverage proportion (Supplementary Fig. S1). We correct for $\log_{10}$ locus length empirically using a binomial cubic smoothing spline (see Section 2 for more details). *P*-values are then calculated for enrichment and adjusted for multiple testing.

Broad-Enrich outputs three tab-delimited text files: (i) peak-to-gene locus assignments from the input peak set with lengths of peaks, loci and overlap; (ii) the gene locus coverage information after aggregating over all peaks overlapping a locus; (iii) the enrichment results, with significance values and summary information for tested gene sets. QC plots showing the relationship between $\log_{10}$ locus length and the proportion of the locus covered by a peak are also output (Supplementary Fig. S1).

3.3 Investigation of type I error

Under the null hypothesis of no true GSE, the type I error rate, or proportion of false-positive results, for a dataset at a given threshold α is the proportion of gene sets with *P*-value less than α. A method with type I error rate higher than the expected α level will result in an overabundance of false-positive results. We investigated the type I error rates for Broad-Enrich, the simple LR model, the binomial-based test and FET, for 55 HM datasets

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

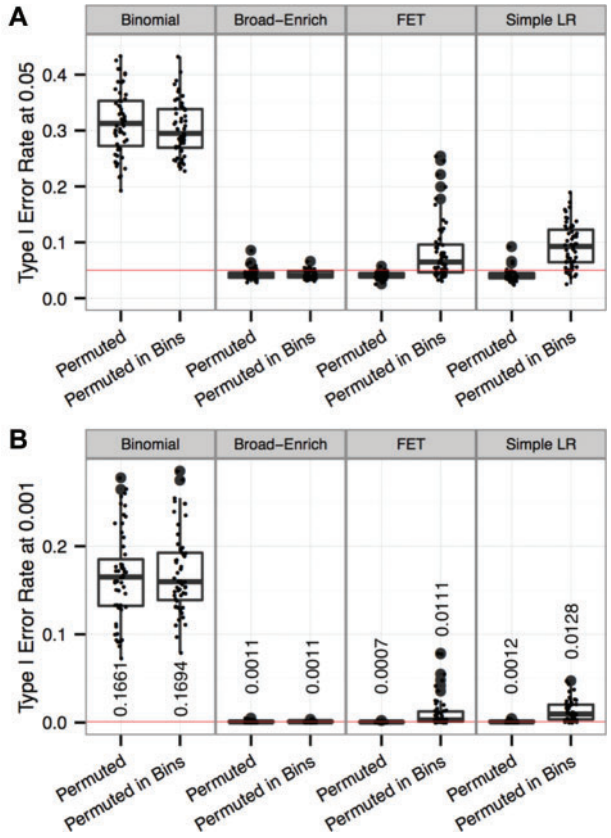


Fig. 3. Type I error rates of the binomial-based test, Broad-Enrich, the simple LR model and FET under the two permutation scenarios with the *nearest TSS* locus definition. Each point represents 1 of the 55 HM datasets (Supplementary Table S2). (**A**) At $\alpha = 0.05$ (red line), we find inflated type I error for the binomial test under both permutation scenarios, the correct error rate for Broad-Enrich and the correct error rate for permutations eliminating length bias but often inflated error for permutations preserving length bias for both the simple LR model and FET. (**B**) At $\alpha = 0.001$ (red line), we observe results similar to $\alpha = 0.05$. Mean error rates are given inset

under two permutation scenarios using the *nearest TSS* locus definition. Both permutations remove any true biological association between gene sets and the genes they contain. The first scenario (Permuted) assesses type I error of the enrichment test under no locus length bias. The second scenario (Permuted in Bins) has the effect of preserving the locus length properties of the gene sets and illustrates the extent to which the type I error rate is affected by locus length.

We find that Broad-Enrich exhibits the correct type I error rates in both permutation scenarios and at different α levels. The binomial test exhibits severely inflated type I error in both scenarios, and both the simple LR model and FET exhibit the correct type I error rate in the 'Permuted' scenario, but have inflated error for the 'Permuted in Bins' scenario [Fig. 3A ($\alpha = 0.05$), Fig. 3B ($\alpha = 0.001$) and Supplementary Table S2]. Comparing Broad-Enrich with the simple LR model, we conclude that the smoothing spline is essential for Broad-Enrich's well-calibrated type I error. None of the 55 datasets tested exhibited correct type I error for the binomial-based test. Welch *et al*. identified significant extra variability (beyond that expected by the binomial test) in the number of peaks assigned to genes in ENCODE ChIP-seq data; they show this, together with the incorrect assumption of the binomial test with respect to locus length, accounts for the inflated type I error (Welch *et al*., 2014). In contrast, FET resulted in correct type I error for 16 of 55 datasets under both permutation scenarios (Fig. 3A and Supplementary Table S2). The inflated type I error of the remaining 39 datasets results from FET being unable to account for the locus length bias present in these datasets (Welch *et al*., 2014; Taher and Ovcharenko, 2009). We compare the enrichment results for these 16 datasets with those of Broad-Enrich in Section 3.5.

3.4 Summary of ENCODE HM enrichment results

We tested for GSE using Broad-Enrich in the same 55 HM ChIP-seq datasets from the ENCODE Consortium. We find that significantly enriched gene sets outnumber significantly depleted gene sets by ~3:1 over all the datasets (Supplementary Table S3). The number of enriched gene sets varies greatly among experiments, with as few as 8 for H3K9me3 in K562 and as many as 1058 for H3K4me2 in H1-hESC (median number of enriched gene sets = 664) of 5591 total gene sets tested from GO and KEGG, and using the *nearest TSS* locus definition. For a fixed histone, the number of enriched gene sets can vary greatly across the five cell lines (e.g. H2az range = 74–767 and H3K9me3 range = 8–253), suggesting different biological activity for such HMs across GM12878, H1-hESC, HeLa-S3, HepG2 and K562.

For each HM, we determined the extent of overlap among significantly enriched gene sets across the five cell lines with the *nearest TSS* locus definition (Supplementary Table S4). GM12878 and H1-hESC tend to have the highest percentage of unique enrichments across all HMs. This could be an indication of more specific regulation via HMs in these cell lines compared with the others. H3K36me3 and H3K79me2 exhibit the highest percentage of enriched gene sets common to all cell lines (39% each). Both modifications tend to occur within the gene body, and the observation of many mutually enriched gene sets could be a result of their necessary functions in constitutively expressed gene groups required by cells, such as transcription and RNA processing (ENCODE Project Consortium *et al*., 2012a). H2az had the smallest percent (0.1%) of mutually enriched gene sets among all five cell lines, with the most uniquely occurring in the embryonic stem cell line.

3.5 Comparison of Broad-Enrich to FET and GREAT

FET has an acceptable type I error rate (≤ 0.05 at $\alpha = 0.05$ level) in only 16 of 55 datasets (Fig. 3A and Supplementary Table S2). These datasets tend to have fewer peaks overall, and more peaks located within 5 kb of the TSS compared with the 39 HM datasets with type I error rate > 0.05. For each of these 16 datasets, we compared the average peak coverage proportion of gene loci in the gene sets uniquely enriched by Broad-Enrich with those uniquely enriched by FET. The gene sets uniquely enriched by Broad-Enrich have a consistently higher proportion of the gene locus covered (Supplementary Table S5). We also examined the percentage of significant enrichments that were stronger in one

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

method versus the other by comparing the FDR values of gene sets enriched in either method. Broad-Enrich resulted in stronger enrichment signal in 12 of 16 datasets (Supplementary Table S5). Finally, we compared the power of Broad-Enrich with FET in the 16 datasets by varying the proportion of genes with a peak, and the proportion of each gene locus covered by a peak (Supplementary Methods). We find that Broad-Enrich has higher power than FET in nearly all cases (Supplementary Table S6).

For comparison with GREAT (v1.8.2), we selected six histone datasets (H3K4me1,-2,-3, H3K9me3, H3K27me3 and H3K79me2 in the cell line GM12878) representing a mixture of activators/repressors and binding close/distal to TSSs. We tested all GO terms using the 'single nearest gene' within 9999 kb gene regulatory domain definition provided in GREAT because it is most similar to the *nearest TSS* definition in Broad-Enrich. We compared relative ranks of enrichments, as the binomial-based test implemented in GREAT has overly significant *P*-values (inflated type I error rate). Comparing the top 20 ranked GO terms for each enrichment test, we find that compared with GREAT, Broad-Enrich consistently finds gene sets with higher coverage in terms of the proportion of each gene locus having the HM (Supplementary Table S7).

The GM12878 cell line is a lymphoblastoid cell line. Lymphoblasts are naïve lymphocytes, which is the term used for any of the three types of white blood cell (leukocytes) in the vertebrate immune system. H3K4me1 is a known general transcriptional activator. The top 20 ranked GO terms for H3K4me1 in Broad-Enrich include leukocyte activation, lymphocyte activation, regulation of lymphocyte activity, positive regulation of immune response and regulation of leukocyte activation (Table 1 and Supplementary Table S8). None of the above (and only one immune-related term) is in the top 20 ranked GO terms according to GREAT. In contrast, the top terms ranked by GREAT included mitochondrion- and ribonucleotide binding-related gene sets, which are not as strongly related to the known properties of GM12878 (Table 2 and Supplementary Table S8).

H3K27me3 is a known repressor of differentiation and developmental genes. Within the top 20 ranked GO terms from Broad-Enrich, we find tissue development, organ morphogenesis, epithelium cell differentiation and regionalization. According to GREAT, none of the above or related GO terms is ranked in the top 20, and only one is in the top 100 (Supplementary Table S9). Moreover, the top terms ranked by GREAT included metabolic processes and energy/transport-related gene sets, which are not commonly associated with the regulatory targets of H3K27me3.

In both instances, we find that the binomial test not only finds an overabundance of significant (FDR < 0.05) terms, as indicated by its inflated type I error rate, but also that Broad-Enrich ranks biologically relevant terms better than GREAT.

3.6 Effect of locus definition on enrichment

It is known that some histone marks preferentially occur in particular locations relative to gene features. To investigate the effect of locus definition on enrichment signal, we ran Broad-Enrich for each of the 55 HM ChIP-seq datasets with the *nearest TSS, exons* and *≤5 kb* locus definitions. We hypothesized that using a

Table 1. A subset of the top 20 gene sets, as ranked by Broad-Enrich, for H3K4me1 in the GM12878 cell line using the *nearest TSS* definition

GO ID	Description	Broad-enrich rank	GREAT rank	% GS average coverage
GO:0002684	Positive regulation of immune system process	3	165	32
GO:0002764	Immune response-regulating signaling pathway	4	647	37
GO:0045321	Leukocyte activation	5	74	31
GO:0046649	Lymphocyte activation	7	80	32
GO:0051249	Regulation of lymphocyte activation	10	182	34
GO:0035556	Intracellular signal transduction	11	63	25
GO:0050778	Positive regulation of immune response	13	426	33
GO:0012501	Programmed cell death	14	26	26
GO:0031347	Regulation of defense response	15	452	33
GO:0002694	Regulation of leukocyte activation	16	148	32

Table 2. A subset of the top 20 gene sets, as ranked by GREAT (v1.8.2), for H3K4me1 in the GM12878 cell line using the 'single nearest gene' within 9999 kb gene regulatory definition

GO ID	Description	Broad-enrich rank	GREAT	% GS average coverage
GO:0031981	Nuclear lumen	27	1	26
GO:0046907	Intracellular transport	47	3	28
GO:0002376	Immune system process	1	5	28
GO:0005524	ATP binding	366	7	22
GO:0043687	Post-translational protein modification	2009	8	22
GO:0032553	Ribonucleotide binding	338	10	22
GO:0006917	Induction of apoptosis	30	11	31
GO:0017076	Purine nucleotide binding	308	12	22
GO:0033554	Cellular response to stress	112	13	26
GO:0005739	Mitochondrion	281	16	25

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

locus definition better conforming to the known genomic location of the histone mark would result in stronger enrichment signal.

H3K4me2, known to occur in promoters (Pekowska *et al*,. 2010), tends to have strongest enrichment signal with the $\leq 5\,kb$ locus definition across the five cell lines (Supplementary Fig. S2). H3K4me3, also known to occur in promoters (Bernstein *et al*., 2006), shows results similar to H3K4me2 (not shown). H3K79me2 binds near the 5′ end of gene bodies, and overall we see the strongest enrichment signal when using the $\leq 5\,kb$ definition (Supplementary Fig. S3). In contrast, H3K36me3 binds near the 3′ end of the gene body, and we see a somewhat stronger enrichment when using the *exons* definition compared with the $\leq 5\,kb$ definition (Supplementary Fig. S4) (Barth and Imhof, 2010; ENCODE Project Consortium *et al*., 2012a). Histone acetylation, such as H3K9ac, tends to occur near TSSs (Barth and Imhof, 2010), and we observe stronger enrichment signal for the $\leq 5\,kb$ locus definition across the five cell lines (Supplementary Fig. S5). H3K27me3 gives stronger enrichment signal with the *exons* definition for all cell lines except H1-hESC, which performs best with the $\leq 5\,kb$ locus definition (Supplementary Fig. S6). This may be indicative of a different regulatory regime for H3K27me3 in embryonic stem cells versus the other cell lines, consistent with current literature (Xie *et al*., 2013). H3K4me1 is considered a distal activating mark (Dong *et al*., 2012), and exhibits stronger enrichment signal with the *nearest TSS* locus definition in GM12878 and HepG2 but stronger signal with $\leq 5\,kb$ in H1-hESC, HeLa-S3 and K562 (Supplementary Fig. S7). Broad-Enrich results from the additional tier 2 ENCODE cell lines A549, Huvec and Monocytes-CD14+, and using the same three locus definitions resulted in the same overall conclusions for the 11 HMs above (not shown). Overall, we observed that the locus definition closest to the known locations of an HM provided the strongest enrichment results. These results should be interpreted in light of the fact that *nearest TSS* is the only locus definition to include all peak regions; thus, important information about individual genes within enriched gene sets may be lost for the $\leq 5\,kb$ or *exons* definitions.

4 DISCUSSION

Functional enrichment testing leverages our collective biological knowledge together with high-throughput genomic technologies in a statistical framework to functionally interpret new biological data. Unique properties observed in ChIP-seq data for HMs have led to the use of specialized peak-calling algorithms. These properties, combined with the bias observed in gene loci coverage relative to locus length, present challenges to existing functional enrichment methods. We have developed Broad-Enrich to address these issues in functionally interpreting large sets of broad genomic regions. Our approach uses the proportion of a gene locus covered by all peaks overlapping the locus, and a correction accounting for the locus length in a logistic regression model with gene set membership as the outcome.

Inflated type I error rates result in an overabundance of false-positive results, while well-calibrated type I error rates result in accurately reported FDRs. We demonstrate that Broad-Enrich has a well-calibrated type I error rate across 55 HM ChIP-seq datasets representing a wide variety of technical and biological characteristics. In contrast, the binomial-based test consistently exhibits inflated type I error, while FET has the correct type I error for only 16 of the 55 datasets. These 16 HMs represent transcriptional activators, or HMs occurring in actively transcribed genes. Even for these 16 HMs, Broad-Enrich tends to provide stronger enrichment signal than FET. Compared with GREAT, Broad-Enrich finds more biologically relevant terms in the top ranked gene sets, as illustrated with immune function-related terms for H3K4me1 and H3K27me3 in the context of lymphoblastoid cell line GM12878. While rank comparisons are not ideal, in the absence of a gold standard, we rely on known biological roles for the HMs combined with known characteristics in cellular context.

Finally, we examined the effect of locus definition on the enrichment signal from Broad-Enrich. We see the strongest enrichment signal by using the locus definition closest to the known locations of the HM. For two HMs, we observe differences in the optimal locus definition. For H3K27me3, the *exons* locus definition performs best in all cell lines except for H1-hESC, where $\leq 5\,kb$ performs best. This difference could be explained by the role H3K27me3 plays in embryonic stem cells, where it is known to often occur in promoters of genes having CpG islands to regulate differentiation of ES cells (Deaton and Bird, 2011; Xie *et al*., 2013). For H3K4me1, we observe that *nearest TSS* performs best for GM12878 and HepG2, whereas $\leq 5\,kb$ performs best for the remaining cell lines. This might indicate that GM12878 and HepG2 cells rely more heavily on long-range enhancer activity for gene activation than the other three cell lines. These results emphasize that the definition with strongest enrichment signal tends to mirror the currently understood location of HM binding. Our implementation of Broad-Enrich allows users to define their own custom locus definition to fit their own experimental contexts.

In addition to functionally interpreting single HM experiments, it is also possible to examine bivalent or trivalent HM signatures together (e.g. H3K4me3 and H3K27me3) with Broad-Enrich and compare the results with the HMs individually to determine if bivalency leads to unique biological function. Broad-Enrich is also applicable to other types of broad domain experiments, such as copy number variations.

As the regulatory programs of living organisms are better understood, Broad-Enrich may be improved with distal regulatory information from Hi-C experiments, allowing for more accurate locus definitions. The significance or strength of each peak region reported by peak callers may also be incorporated in the enrichment model. Such future changes may bring functional interpretation of broad genomic regions closer to making optimal use of peak information.

Funding: This work was supported by the National Institutes of Health grants from the National Cancer Institute [R01CA158286-01A1 to M.A.S.]; the National Human Genome Research Institute [T32-HG000040 to R.C.]; National Institute of Environmental Health Sciences P30 Core Center [P30-ES017885-01A1 to M.A.S.].

Conflict of interest: none declared.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bailey,T. *et al.* (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.*, **9**, e1003326.

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Barth,T.K. and Imhof,A. (2010) Fast signals and slow marks: thedynamics of histone modifications. *Trends Biochem. Sci.*, **35**, 618–626.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.*, 289–300.

Bernstein,B.E. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.

Blow,M.J. *et al.* (2010) ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.*, **42**, 806–810.

Chi,P. *et al.* (2010) Covalent histone modifications–miswritten, misinterpreted and mis-erased in human cancers. *Nat. Rev. Cancer*, **10**, 457–469.

Curtis,R.K. *et al.* (2005) Pathways to the analysis of microarray data. *Trends Biotechnol.*, **23**, 429–435.

Deaton,A.M. and Bird,A. (2011) CpG islands and the regulation of transcription. *Gene Dev.*, **25**, 1010–1022.

Dong,X. *et al.* (2012) Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.*, **13**, R53.

Draghici,S. *et al.* (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

ENCODE Project Consortium. *et al.* (2012a) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

ENCODE Project Consortium. *et al.* (2012b) Histone Modifications by ChIP-seq from ENCODE/Broad Institute. Available from https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeBroadHistone (16 March 2014, date last accessed).

Han,J. *et al.* (2013) ER-stress-induced transcriptional regulation increases protein synthesis leading to cell death. *Nat. Cell Biol.*, **15**, 481–490.

Huang,D.W. *et al.* (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Kaelin,W.G. Jr and McKnight,S.L. (2013) Influence of metabolismon epigenetics and disease. *Cell*, **153**, 56–69.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kim,J.H. *et al.* (2012) LRpath analysis reveals common pathways dysregulated via DNA methylation across cancertypes. *BMC Genomics*, **13**, 526.

McLean,C.Y. *et al.* (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 1630–1639.

Mi,H. *et al.* (2012) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377–D386.

Nishimura,D. (2001) BioCarta. *Biotechnol. Softw. Internet Rep*, **2**, 117–120.

Ovcharenko,I. *et al.* (2005) Evolution and functional classification of vertebrate gene deserts. *Genome Res.*, **15**, 137–145.

Pan,G. *et al.* (2007) Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell*, **1**, 299–312.

Peduzzi,P. *et al.* (1996) A simulation study of the number of events per variable in logistic regression analysis. *J. Clin.Epidemiol.*, **49**, 1373–1379.

Pekowska,A. *et al.* (2010) A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res.*, **20**, 1493–1502.

Punta,M. *et al.* (2011) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Sartor,M.A. *et al.* (2009) LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, **25**, 211–217.

Sartor,M.A. *et al.* (2010) ConceptGen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics*, **26**, 456–463.

Sen,G.L. *et al.* (2008) Control of differentiation in a self-renewing mammalian tissue by the histone demethylase JMJD3. *Gene Dev.*, **22**, 1865–1870.

Taher,L. and Ovcharenko,I. (2009) Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. *Bioinformatics*, **25**, 578–584.

Valouev,A. *et al.* (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.

Welch,R.P. *et al.* (2014) ChIP-Enrich: gene set enrichment testing for ChIP-seq data. *Nucleic Acids Res.*, **42**, e105.

Wood,S.N. (2006) *Generalized Additive Models: an Introduction with R*. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Wood,S.N. (2010) *mgcv: GAMs with GCV/AIC/REML Smoothness Estimation and GAMMs by PQL*. R package version, 1.6-2.

Wood,S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Series B Stat. Methodol.*, **73**, 3–36.

Xie,W. *et al.* (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, **153**, 1134–1148.

Zang,C. *et al.* (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952–1958.

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014