# Package 'BiFET'

May 24, 2024

**Type** Package

**Title** Bias-free Footprint Enrichment Test

**Version** 1.25.0

**Date** 2021-11-20

**Maintainer** Ahrim Youn <Ahrim.Youn@jax.org>

**Description** BiFET identifies TFs whose footprints are over-represented in target regions compared to background regions after correcting for the bias arising from the imbalance in read counts and GC contents between the target and background regions. For a given TF k, BiFET tests the null hypothesis that the target regions have the same probability of having footprints for the TF k as the background regions while correcting for the read count and GC content bias. For this, we use the number of target regions with footprints for TF k, t_k as a test statistic and calculate the p-value as the probability of observing t_k or more target regions with footprints under the null hypothesis.

**License** GPL-3

**biocViews** ImmunoOncology, Genetics, Epigenetics, Transcription, GeneRegulation, ATACSeq, DNaseSeq, RIPSeq, Software

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 6.0.1

**Imports** stats, poibin, GenomicRanges

**Suggests** rmarkdown, testthat, knitr

**VignetteBuilder** knitr

**git_url** https://git.bioconductor.org/packages/BiFET

**git_branch** devel

**git_last_commit** 0e5eca9

**git_last_commit_date** 2024-04-30

**Repository** Bioconductor 3.20

**Date/Publication** 2024-05-24

**Author** Ahrim Youn [aut, cre],
Eladio Marquez [aut],
Nathan Lawlor [aut],
Michael Stitzel [aut],
Duygu Ucar [aut]

# Contents



---

`calculate_enrich_p` *Function to calculate p-value testing if footprints of a TF are over-represented in the target set of peaks compared to the background set of peaks correcting for the bias arising from the imbalance of GC-content and read counts between target and background set*

---

## Description

Function to calculate p-value testing if footprints of a TF are over-represented in the target set of peaks compared to the background set of peaks correcting for the bias arising from the imbalance of GC-content and read counts between target and background set

## Usage

```
calculate_enrich_p(GRpeaks, GRmotif)
```

## Arguments

| | |
|---|---|
| `GRpeaks` | ATAC-seq or DNase-seq peaks in GRanges class where each row represents the location of each peak. In addition there must be 3 metadata columns called "reads" (representing read counts in each peak), "GC" (representing the GC content), and lastly "peaktype" which designates each peak as either ("target","background","no"). |
| `GRmotif` | Footprint calls from a footprint algorithm in GRanges class where each row represents the location of each PWM occurrence. The footprint calls in the forward strand and those in the backward strand from the same PWM are not differentiated. The row names of GRmotif are the motif IDs (e.g. MA01371 STAT1). |

## Details

In this example, the file input_peak_motif.Rdata was obtained as follows: we used ATAC-seq data obtained from five human PBMC (Ucar, et al., 2017) and five human islet samples (Khetan, et al., 2017) and called peaks using MACS version 2.1.0 (Zhang, et al., 2008) with parameters "-nomodel -f BAMPE". The peak sets from all samples were merged to generate one consensus peak set (N = 57,108 peaks) by using package DiffBind_2.2.5. (Ross-Innes, et al., 2012), where only the peaks

present at least in any two samples were included in the analysis. We used the **summits** option to re-center each peak around the point of greatest enrichment and obtained consensus peaks of same width (200bp).

Out of these consensus peaks, we defined regions that are specifically accessible in PBMC samples as regions where at least 4 PBMC samples have a peak, whereas none of the islet samples have a peak (n=4106 peaks; these regions are used as target regions in this example). Similarly, we defined islet-specific peaks as those that were called as a peak in at least 4 islet samples but none in any of the PBMC samples (n=12886 peaks). The rest of the peaks excluding the PBMC/islet-specific peaks were used as the background (i.e., non-specific) peaks in our analyses (n=40116 peaks). For each peak, GC content was obtained using peak annotation program annotatePeaks.pl from the HOMER software (Heinz. et al., 2010).

In this example, TF footprints were called using PIQ algorithm (Sherwood, et al., 2014) using the pooled islet samples and pooled PBMC samples to increase the detection power for TF footprints. We used only the TF footprints that have a purity score greater than 0.9. The example file contains footprint calls for only five PWMs from the JASPAR database to reduce computing time.

## Value

Returns a list of parameter alpha_k, p values from BiFET algorithm and p values from the hypergeometric test

## Note

The function calculate_enrich_p first generates a TF binding matrix M where i_th row represents i_th PWM and j_th column represetns j_th peak with M_i, j=1 if the footprint of i_th PWM overlaps j_th peak and 0 otherwise. Finally, the function "calculate_enrich_p" returns a list of parameter alpha_k, enrichment p values from BiFET algorithm and enrichment p values from the hypergeometric test.

## Author(s)

Ahrim Youn

## Examples

```
# Load in the peak file and footprint calls from a footprint algorithm
peak_file <- system.file("extdata", "input_peak_motif.Rdata",
 package = "BiFET")
load(peak_file)
result <- calculate_enrich_p(GRpeaks,GRmotif)
```

# Index