SOURCES FOR GENE SETS

	Database	Ver	Species
Gene Ontology	GO Biological Processes	3.4.2	hs,mm,rn, dm,dr
	GO Cellular Component		
	GO Molecular Function		
Other Functional Annotations	BioCarta Pathway	6.0	hs,mm,rn
	KEGG Pathway	3.2.3	hs,mm,rn
	Panther Pathway	3.5	hs,mm,rn
	pFAM	31.0	hs,mm,rn
	Reactome	61	hs,mm,rn, dm,dr
Literature	MeSH (gene2MeSH)	~2013	hs,mm,rn
MSigDB Derived	Hallmark	6.0	hs
	Immunologic		
	Oncogenic		
Targets	Comparative Toxicogenomics Database (CTD)	Jul 06 2017	hs
	Drug Bank	5.0.7	hs,mm,rn
	MicroRNA (MSigDB)	6.0	hs,mm,rn
	Transcription Factors (MSigDB)	6.0	hs,mm,rn
Interaction	Protein Interaction BioGRID	3.4.151	hs,mm,rn
Other	Metabolite (NCIBI)		hs,mm,rn
	Cytoband (NCBI)		hs

All data are stored in the *chipenrich.data* package.

ENRICHMENT METHODS

ChIP-Enrich Statistical Method

ChIP-Enrich uses a logistic regression approach to simultaneously 1) adjust for the gene locus length and mappability, and 2) test for gene set enrichment. Our model is:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 GS_i + f(\log_{10}(LL_i * m_i + 1))$$

The probabilities π_i are defined as the probability of gene *i* being assigned a peak given the data. Our dependent variable is then a binary vector with 1 if the gene has a peak assigned to it, and 0 otherwise. The parametric term geneset is also a binary vector, where 0 denotes that the gene does not belong in the set of genes being tested and 1 otherwise. The function $f(\log_{10}(LL_i * m_i + 1))$ is a cubic smoothing spline term that takes into account both the locus length (LL) and the average mappability (m) of each gene's locus. More detailed methods are provided in the ChIP-Enrich publication[1].

Poly-Enrich Statistical Method

Poly-Enrich uses a negative binomial generalized linear model to model the number of peaks assigned to each gene. The model is:

$$\log(\mu_i) = \beta_0 + \beta_1 GS_i + f(\log_{10}(LL_i * m_i + 1))$$

Where for each gene i, μ_i is the mean of the negative binomial distribution and the overdispersion parameter θ is estimated so that $Var(Y|GS) = \mu + \theta\mu^2$, where Y is the number of peaks for each gene. The function f is a negative binomial cubic smoothing spline that adjusts for the gene's locus length (LL) and optional mappability (m). More detailed methods are provided in the Poly-Enrich publication[2].

Hybrid

The hybrid test runs both ChIP-Enrich and Poly-Enrich methods, and then combines the p-values from both by:

$$p_{hybrid} = 2 * \min(p_{CE}, p_{PE})$$

Proofs for why this works are described in the Poly-Enrich publication[2].

Fisher's exact test

This is the plain two-sided fisher's exact test testing for association between if a gene has a peak and gene set membership. As we showed that this test heavily inflates Type 1 error[1], especially due to locus length differences, we only recommend using this test for the <1kb and <5kb locus definitions.

Weighting Peaks (Poly-Enrich only)

Poly-Enrich also has the capability to adjust for peaks' signal values by giving each peak a weight proportional to the their log signal value and normalizing so that the mean of all peak weights is equal to 1. For every peak assigned to a gene, we can then add up all the weights and use that as our "count" data. We can then use the same model, except assuming a quasi-negative binomial family to accommodate for the non-whole number data. The calculations end up being identical.

LOCUS DEFINTIONS

All supplied locus definitions are predefined. The <1kb, <5kb, <10kb locus definitions are equivalent to assigning peaks within 1kb, 5kb, or 10kb of the transcription start site (TSS) respectively. The >10kb upstream definition is assigning peaks farther away from 10kb of any TSS to its nearest TSS. The Exon and Intron locus definitions are equivalent to using only peaks that occur within an annotated exon or intron, respectively. The Nearest Gene and the Nearest TSS locus definitions are equivalent to assigning peaks to the nearest gene or TSS respectively.

Our recommendations are:

<1kb	for proximal promoter regions
<5kb	for greater promoter regions
<10kb	for promoters and nearby enhancers
>10kb upstream	for possible enhancers
Exon	for exon regions only
Intron	for intron regions only
Nearest Gene	for regulation from anywhere when peaks often occur in middle or near end of genes
Nearest TSS	default method for regulation from anywhere

We also provide the user to define their own locus definition, where the format should be have the following columns: entrez gene ID, chromosome, start of locus, end of locus.

OTHER OPTIONS

Filtering gene sets

One may also limit the gene sets by giving a maximum number of genes in the gene set to be tested with a default of 2000. Decreasing the maximum number of genes required for a gene set to be tested may result in faster completion of test.

Mappability

Adjusting for mappability is optional. We have pre-calculated mappability values for each gene in each pre-defined locus definition. Details on how mappability is defined can be found in the ChIP-Enrich publication[1].
In the case where the locus definition is 'User Defined,' user defined mappability values should also be provided.
If locus length is not adjusted for mappability, this is equivalent to setting all mappability values to equal 1, i.e. as if all loci are equally mappable.

Peak-Threshold number (ChIP-Enrich only):

ChIP-Enrich uses "if a gene as assigned at least one peak" as its outcome, but one can use any other whole number for its threshold. For example, one may think one peak may not be sufficient for regulation or if a single peak may likely be a false positive.

REFERENCES

1. R.P. Welch, C. Lee, R.A. Smith,S. Patil, T. Weymouth, P. Imbriano, L.J. Scott, M.A. Sartor. "ChIP-Enrich: Gene set enrichment testing for ChIP-seq data." NAR. 2014.

2. Lee CT, Cavalcante RC, Lee C, Qin T, Patil S, Wang S, Boyle AP, Sartor MA. “Poly-Enrich: Count-based Methods for Gene Set Enrichment Testing with Genomic Regions.” bioRxiv 488734; doi: https://doi.org/10.1101/488734. Preprint.