**CEseek**

CEseek uncovers composite transcriptional regulatory elements within a set of test sequences using a suitable background set. Simple TF motifs (PWMs or PPMs) are provided as inputs and used to scan for all possible CE combinations and configurations in a pairwise manner. The user can specify the range of spacing distances between the simple TF motifs.

**R packages required**

Biostrings

GenomicRanges

seqLogo

brew

BiocGenerics

**R packages optional**

SortableHTMLTables

**Key Functions**

CEseek has two key functions (i) **CEseek\_scanner** and (ii) **seekCEsFromScanned** whose command lines in R and descriptions are provided below.

CEscan.out <- **CEseek\_scanner** (input.seqs, tf.pwms , length.gap, half.gap, match.thres)

input.seqs : DNAStringsetObject with unique names for each sequence. Test and background DNA sequences are combined. All sequences must be uniquely named using character arrays

tf.pwms : TF motifs as a list of PWMs or PPMs (see SimpleMotifs.CEseek\_format)

length.gap : Limits of allowed bp gap between simple TF motifs as a vector of (minimum-gap, maximum-gap) e.g., c(0,11) where the total number of allowable gap positions is an even number

half.gap : Number of bp allowed in length.gap divided by 2. This is to ensure that the total number of allowable gap positions is an even number

match.thres : threshold for matching simple TF motifs with test and background sequences, required by matchPWM function from Biostrings. It should be denoted as a character string e.g., “80%”.

CEscan.out : output of CEseek\_scan that contains a nested list of sequences for each CE type and its configurations. Each sequence “hit” is extracted as an end-to-end motif pair and is used by the function described below to perform statistical analysis.

final.output <- **seekCEsFromScanned** (CEscan.out, test.names, bg.names, p.cut, out.dir, output.html = “no”)

CEscan.out : output from CEseek\_scanner

test.names : a character array containing the sequence names that are treated as test sequences.

bg.names : a character array containing the sequence names that are treated as background sequences.

p.cut : P value cut off for generating the motif logos and motif ppms. Note that final output in R data object will contain all CEs with associated statistics irrespective of p.cut value

out.dir : directory name for storage of motif logos. If output.html = “yes” then html page will be stored in directory e.g. “B\_cell\_CEs”

output.html = “no”: if set to yes require SortableHTMLTables

Components of final.output from **seekCEsFromScanned**

TestProbMatrix : Position probability matrices (PPMs) of CEs derived from hits in test sequences that are enriched on the basis of P value defined by p.cut

BackgroundProbMatrix : Position probability matrices (PPMs) of CEs derived from hits in background sequences (depleted in test sequences) on the basis of P value defined by p.cut

CEEnrichment.Stats : A data.frame object that contains the statistical analysis report of CEs with enrichment P value defined by “p.cut”

CE.All.Stats : A data.frame object containing statistical analyses of all scanned CE configurations irrespective of P-value cut off

TFcombn.Enrichment : Statistical analysis of TF motif pairs (CE types), co-occurrence of simple motif hits irrespective of orientation and within specified length.gap

**Auxiliary Function**

This function operates in the same way as CEseek\_scanner and allows user to specify a single TF motif anchor, with anchor.tf.string, and then search for CEs with all other TF motifs. anchor.tf.string is a unique character string used by conventional grep function to identify the TF motif of interest by its name in tf.pwms library.

CEscan.out <- **CEseek\_scanner**\_anchored (input.seqs, tf.pwms, length.gap, half.gap, match.thres, anchor.tf.string)