Genome-wide Kozak Sequence Over-represented Motif Analysis

Mariana Rius, Joshua Rest

Abstract

Bioinformatic approach to identifying over-represented motifs in the region framing the start codon (25 bp up and downstream) for genes annotated in the three sequenced Labyrinthulomycete genomes (*Aurantiochytrium limacinum, Schizochytrium aggregatum, and Aplanochytrium kergulense*).

Citation: Mariana Rius, Joshua Rest Genome-wide Kozak Sequence Over-represented Motif Analysis. **protocols.io** dx.doi.org/10.17504/protocols.io.hikb4cw

Published: 31 Mar 2017

Protocol

Download gene annotation (gff) file and fasta file for species of interest

Step 1.

Schizochytrium aggregatum

Schag1_GeneCatalog_genes_20121220.gff

Schag1_AssemblyScaffolds.fasta from

http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Schag1

Aurantiochytrium limacinum

Aurli1_GeneCatalog_genes_20120618.gff

Aurli1_AssemblyScaffolds.fasta from

http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Aurli1

Aplanchytrium kergulense

Aplke1_GeneCatalog_genes_20121220.gff

Aplke1_AssemblyScaffolds.fasta from

http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Aplke cmd COMMAND (R - 3.3.2)

```
ShGeneCat <-
 read.delim("Schag1_GeneCatalog_genes_20121220.gff", header=FALSE, stringsAsFactors=FALSE)
```

Create working gene catalog for organism of interest. Schizochytrium aggregatum (Schag1) code provided herein as an example.

NOTES

Mariana Rius 31 Mar 2017

Using R version 3.3.2 and the following packages:

doBy (doBy_4.5-15)

data.table (data.table_1.10.0)

seqinr (seqinr_3.3-3)

OPTIONAL: Create .rda file to facilitate access to annotations

Step 2.

Create subset of annotation file.

cmd COMMAND (R - 3.3.2)

```
colnames(ShGeneCat) <- c("contig","V2","type","start","stop","V6","strand","num","V9")
getPID <- function(dx){
  a <- regmatches(dx,gregexpr("proteinId (\\d+)",dx,perl=T))[[1]]
  ifelse(is.na(a[1]),NA,as.numeric(unlist(regmatches(a,gregexpr("\\d+",a,perl=T)))))
}
PID <- vapply(ShGeneCat$V9,FUN=getPID,double(1))
ShGeneCat <- cbind(ShGeneCat,PID)

getExonNum <- function(dx){
  a <- regmatches(dx,gregexpr("exonNumber (\\d+)",dx,perl=T))[[1]]
  ifelse(is.na(a[1]),NA,as.numeric(unlist(regmatches(a,gregexpr("\\d+",a,perl=T)))))
}
ExonN <- vapply(ShGeneCat$V9,FUN=getExonNum,double(1))
ShGeneCat <- cbind(ShGeneCat,ExonN)
save(ShGeneCat,file=paste(species,"GeneCat.rda",sep=""))
```

Example of ShGeneCat.

Identify the coordinates of 25 base pairs up and downstream of all annotated coding start sites

Step 3.

Retain only genes with a protein ID

cmd COMMAND (R - 3.3.2)

```
ShGeneCat <- ShGeneCat[!(is.na(ShGeneCat$PID)),]
```

Identify the coordinates of 25 base pairs up and downstream of all annotated coding start sites

Step 4.

Identify species and term

cmd COMMAND

```
term <- "wg" #whole genome
species <- "Sh"  #Ap, Sh, or Au
```

Identify the coordinates of 25 base pairs up and downstream of all annotated coding start sites

Step 5.

Create new destination for identified coordinates

cmd COMMAND (R - 3.3.2)

```
ShGeneWg <- ShGeneCat[]
```

Identify the coordinates of 25 base pairs up and downstream of all annotated coding start sites

Step 6.

Write table with coordinates of region of interest for each gene. Here 25 bases up and downstream were isolated as region of interest.

cmd COMMAND (R - 3.3.2)

```
promC <- do.call("rbind",lapplyBy(~PID,data=ShGeneWg,function(dx){
 if(dx$strand[1]=="+"){
 return(c(dx$contig[1],dx[dx$ExonN==1,"start"]-26,dx[dx$ExonN==1,"start"]+27,dx$PID[1],"1",
"+")) #returns the first codon
  }
  if(dx$strand[1]=="-"){
    return(c(dx$contig[1],dx[dx$ExonN==1,"stop"]-28,dx[dx$ExonN==1,"stop"]+25,dx$PID[1],"1"
,"-")) #returns the first codon
  }
}))
colnames(promC) <- c("chr","start","stop","name","frame","strand")
save(promC,file=paste(species,term,"promC","rda",sep="."))
```

Identify the coordinates of 25 base pairs up and downstream of all annotated coding start sites

Step 7.

Change any negative start sites to 1

cmd COMMAND (R - 3.3.2)

```
promC[promC[,'start'] < 1, 'start'] <- 1

write.table(promC,file=paste(species,term,"promC","gff",sep="."),quote=FALSE,row.names=FALS
E,col.names=FALSE,sep="\t")
```

Create FASTA file containing region of interest

Step 8.

Using FASTA files previously downloaded:

Schag1_AssemblyScaffolds.fasta

Aurli1_AssemblyScaffolds.fasta

Aplke1_AssemblyScaffolds.fasta

Run bedtools command to retrieve sequence data.

cmd COMMAND

```
bedtools getfasta -s -fi Schag1_AssemblyScaffolds.fasta -bed Sh.wg.promC.gff -fo Sh.wg.promC.fasta -name
```

bedtools 2.15.0

Create FASTA file containing region of interest

Step 9.

Use bioawk to discard any sequences not containing an 'ATG' as the start codon.

cmd COMMAND

```
bioawk -c fastx 'substr($seq,26,3) ~ /ATG/ { print ">"$name"\n"$seq; }' Sh.wg.promC.fasta >Sh.wg.promC.ATG26.fasta
```

bioawk version 20110810

Use RSATprotist to identify over-represented motifs in sequences

Step 10.

Use RSATprotist online in the web interface

http://rsat01.biologie.ens.fr/rsa-tools/

Input FASTA file:

Sh.wg.promC.ATG26.fasta

1 - Choose your type of data to analyse

ChIP-seq

List of gene names

Sequences

Matrices (PSSM)

Coordinates (BED)

List of variants

2 - Choose your biological question/ analysis to perform

Are there over-represented motifs in these sequences?

I want to scan these sequences with a motif

3 - Relevant RSAT programs

oligo-analysis (words)

dyad-analysis (spaced pairs)

🔗 LINK:
http://rsat01.biologie.ens.fr/rsa-tools/

ANNOTATIONS

Mariana Rius 31 Mar 2017

Example output can be viewed at

https://you.stonybrook.edu/labyrinthulomycetes/regulatory-element-discovery-in-labyrinthulomycete-genomes/