CAPE Use Cases

The following document describes several use cases illustrating how the Coupled Polymerase Binding and Expression Tool (CAPE) can be used to relate matched RNA polymerase II (RNAPII) ChIP-Seq and RNA-seq experiments, identify features with unusual levels of RNAPII binding vs. mRNA abundance, and compare these transcripts across samples or organisms. Current use cases may always be found at http://cape.gersteinlab.org.

Use Case 1: Comparing Transcription between Worm, Fly, and Human Embryos

Our first use case uses publicly available data from the ENCODE and modENCODE consortia to show how CAPE can be used to compare orthologs between different organisms. In this case, we are comparing embryos from worm, fly, and human. Also, no additional expression files are needed as RPKM values are contained in the gtf files.

Sample Data

The following sample data will be used:

Worm RNAPII ChIP-Seq Signal Data (early embryo):
http://archive.gersteinlab.org/proj/CAPE/usecases/data/2435_Snyder_N2_POLII_eemb_combined.bw

(Note: converted to bigWig format from the public modMine file available at http://submit.modencode.org/submit/public/get_file/2435/extracted/Snyder_N2_POLII_eemb_combined.wig)

Fly RNAPII ChIP-Seq Signal Data (embryo):
http://archive.gersteinlab.org/proj/CAPE/usecases/data/3251_ON_PolII.bw

(Note: converted to bigWig format from the public modMine file available at http://submit.modencode.org/submit/public/get_file/3251/extracted/ON_PolII.wig)

Human RNAPII ChIP-Seq Signal Data (H1 embryonic stem cells):
Available from the ENCODE public data repository at
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/wgEncodeHaibTfbsH1hescPol2V0416102RawRep1.bigWig

Worm Annotation and Expression Data:
http://archive.gersteinlab.org/proj/CAPE/usecases/data/wormEmbryo.gtf

(Note: Adapted from public data available as part of the modENCODE project. This file only includes entries for known orthologous genes. Expression values are also given in this file).

Excerpt:

```
II      modENCODE_TX    gene    8651057 8658766 .       +       .       RPKM "109.609215";
gene_id "pyr-1"
II      modENCODE_TX    gene    5399522 5405988 .       +       .       RPKM "94.070177";
gene_id "mog-5"
II      modENCODE_TX    gene    13670567        13694711        .       -       .       RPKM
"61.110324"; gene_id "Y48E1A.1"
```

```
II      modENCODE_TX   gene    11661516       11675863       .       -       .       RPKM
"94.316146"; gene_id "trr-1"
```

Fly Annotation and Expression Data:
http://archive.gersteinlab.org/proj/CAPE/usecases/data/flyEmbryo.gtf

(Note: Adapted from public data available as part of the modENCODE project. This file only includes entries for known orthologous genes. Expression values are also given in this file).

Human Annotation and Expression Data:
http://archive.gersteinlab.org/proj/CAPE/usecases/data/humanEmbryo.gtf

(Note: Adapted from public data available as part of the ENCODE/GENCODE projects. This file only includes entries for known orthologous genes. Expression values are also given in this file).

Ortholog File:
http://archive.gersteinlab.org/proj/CAPE/usecases/data/wfhOrthologList.txt

(adapted from the pairwise MIT-Broad Ortholog Project files at http://compbio.mit.edu/modencode/orthologs/modencode-orths-2012-01-30/ensembl-v65/modencode.merged.orth.txt.gz)

Excerpt:

```
dhc-1   FBgn0261797    ENSG00000197102
prp-8   FBgn0033688    ENSG00000174231
ama-1   FBgn0003277    ENSG00000181222
sma-1   FBgn0004167    ENSG00000137877
pyr-1   FBgn0003189    ENSG00000084774
F33H2.5 FBgn0020756    ENSG00000177084
T08A11.2       FBgn0031266    ENSG00000115524
rme-8   FBgn0015477    ENSG00000138246
```

Generating CAPE-analyze reports for each organism

The first step is to generate individual reports for each organism using CAPE-analyze in transcript mode. We will now show the text of an example session at the commend line to generate these reports in addition to the output that was produced at each step. Please note that the directory names in the script below is tailored to our test system and should be changed if trying to reproduce these results. Also for this example, CAPE-analyze was run using Java's default heap size on Mac OSX.

Worm Embryo Analysis

```
bash-3.2$ java -jar CAPE-analyze.jar --
signalfile=/emb_study/usecase/data/2435_Snyder_N2_POLII_eemb_combined.bw --
transcriptfile=/emb_study/usecase/data/wormEmbryo.gtf --
outputfile=/emb_study/usecase/wormReport.txt --GFFkey=gene --aggoverride=500
Running in transcript mode...
Signal File: /emb_study/usecase/data/2435_Snyder_N2_POLII_eemb_combined.bw
Transcript File: /emb_study/usecase/data/wormEmbryo.gtf


Loading transcript file...
Performing aggregation with user-defined window size +/- 500 bp...
Processing 1020 features...
Processed 1000 features...
Processed 1020 features...
```

```
Determining binding and expression cutoffs...
Updating expression states for all transcripts...
Determining binding and expression cutoffs...
Updating RNAPII promoter binding states for all transcripts...
Writing output to /emb_study/usecase/wormReport.txt...
Program completed in 4 seconds
```

Fly Embryo Analysis

```
bash-3.2$ java -jar CAPE-analyze.jar --
signalfile=/emb_study/usecase/data/3251_ON_PolII.bw --
transcriptfile=/emb_study/usecase/data/flyEmbryo.gtf --
outputfile=/emb_study/usecase/flyReport.txt --GFFkey=gene
Running in transcript mode...
Signal File: /emb_study/usecase/data/3251_ON_PolII.bw
Transcript File: /emb_study/usecase/data/flyEmbryo.gtf

Loading transcript file...
Determining ideal window size with starting value +/- 1000 bp...
Processing 1015 features...
Processed 1000 features...
Processed 1015 features...

Performing aggregation with ideal peak window size +/- 250 bp...
Processing 1015 features...
Processed 1000 features...
Processed 1015 features...

Determining binding and expression cutoffs...
Updating expression states for all transcripts...
Determining binding and expression cutoffs...
Updating RNAPII promoter binding states for all transcripts...
Writing output to /emb_study/usecase/flyReport.txt...
Program completed in 4 seconds
```

Human H1 Embryonic Stem Cells Analysis

```
bash-3.2$ java -jar CAPE-analyze.jar --
signalfile=/emb_study/usecase/data/wgEncodeHaibTfbsH1hescPol2V0416102RawRep1.bigWig --
transcriptfile=/emb_study/usecase/data/humanEmbryo.gtf --
outputfile=/emb_study/usecase/humanReport.txt --GFFkey=gene
Running in transcript mode...
Signal File: /emb_study/usecase/data/wgEncodeHaibTfbsH1hescPol2V0416102RawRep1.bigWig
Transcript File: /emb_study/usecase/data/humanEmbryo.gtf

Loading transcript file...
Determining ideal window size with starting value +/- 1000 bp...
Processing 999 features...
Processed 999 features...

Performing aggregation with ideal peak window size +/- 500 bp...
Processing 999 features...
Processed 999 features...

Determining binding and expression cutoffs...
Updating expression states for all transcripts...
Determining binding and expression cutoffs...
Updating RNAPII promoter binding states for all transcripts...
Writing output to /emb_study/usecase/humanReport.txt...
Program completed in 8 seconds
```

```
bash-3.2$ java -jar CAPE-analyze.jar --
signalfile=/emb_study/usecase/data/wgEncodeHaibTfbsH1hescPol2V0416102RawRep1.bigWig --
transcriptfile=/emb_study/usecase/data/humanEmbryo.gtf --
outputfile=/emb_study/usecase/humanReport2.txt --GFFkey=gene --expressionoverride=.1,.3
Running in transcript mode...
```

```
Using expression signal thresholds of 0.1 and 0.3...
Signal File: /emb_study/usecase/data/wgEncodeHaibTfbsH1hescPol2V0416102RawRep1.bigWig
Transcript File: /emb_study/usecase/data/humanEmbryo.gtf

Loading transcript file...
Determining ideal window size with starting value +/- 1000 bp...
Processing 999 features...
Processed 999 features...

Performing aggregation with ideal peak window size +/- 500 bp...
Processing 999 features...
Processed 999 features...

Determining binding and expression cutoffs...
Updating expression states for all transcripts...
Determining binding and expression cutoffs...
Updating RNAPII promoter binding states for all transcripts...
Writing output to /emb_study/usecase/humanReport2.txt...
Program completed in 8 seconds
```

The above runs also show several of the options available to customize analyzes in CAPE. For example, when analyzing the worm data we used the "—aggoverride=500" command line option to tell CAPE to use a window size of +/- 500 bp around the start position given in the annotation file. Worm transcripts have what are called splice leaders that can be included in the annotation files, resulting in a shift of the polymerase-binding site from the annotated start position. CAPE will detect the maximum regardless of its position inside the initial aggregation window, but this "play" in the annotation can produce an overly broad aggregation profile and hence, CAPE-analyze would choose a larger ideal window size. We chose to set a manual window size in this instance. We also performed two different analyses on human. The first uses CAPE's default boundaries for low and high cutoffs for binding and expression (the 25th and 75th percentiles). The second run overrides the expression cutoffs with defined RPKM values using the –expressionoverride option. This was done as an exercise to show that one can refine CAPE-analyze cutoffs using either percentile or raw data values.

CAPE-analyze Output

CAPE-analyze produces four output files from the above runs. For convenience, these files can be obtained at the following links and a snippet of output is provided.

Fly CAPE-analyze Report:
http://archive.gersteinlab.org/proj/CAPE/usecases/reports/flyReport.txt

```
#Transcript File: /emb_study/usecase/data/flyEmbryo.gtf
#Signal File: /emb_study/usecase/data/3251_ON_PolII.bw
#Mode: transcript
#Promoter Binding Low Percentile Threshold: 25.0 (Binding <= 1.972
#Promoter Binding High Percentile Threshold: 75.0 (Binding >= 15.744
#Expression Low Percentile Threshold: 25.0 (Expression <= 25.209
#Expression High Percentile Threshold: 75.0 (Expression >= 62.315

#A. Transcripts with low promoter binding and high expression: 25
#B. Transcripts with high promoter binding and low expression: 26
#C. Transcripts with no promoter binding and no expression: 0
#D. Transcripts with no promoter binding: 0
#E. Transcripts with no expression: 0
#F. "Normal" transcripts: 964

#Category      Transcript ID  Chromosome     Start   End     Strand  PromoterSignal
       BodySignal     Ratio  (Stalling Index) Expression Value
A      FBgn0004603    2R      1868785 1900039 +       -5.832  3.463   -1.684  65.285
```

```
A       FBgn0033062     2R      1968333 1973125 -       -0.213  9.449   -0.023  81.819
A       FBgn0016697     2R      13300269        13301274        +       1.262   8.521   0.148
        111.608
A       FBgn0035046     2R      20551921        20552962        +       1.683   10.89   0.155
        65.167
```

Worm CAPE-analyze Report:
http://archive.gersteinlab.org/proj/CAPE/usecases/reports/wormReport.txt

Human CAPE-analyze Report:
http://archive.gersteinlab.org/proj/CAPE/usecases/reports/humanReport.txt

Human CAPE-analyze Report (using manual expression cutoffs):
http://archive.gersteinlab.org/proj/CAPE/usecases/reports/humanReport-expOverride.txt

Combining and Visualizing Results Using CAPE-compare

CAPE-compare is used to combine and visualize CAPE-analyze reports into a summary format containing information about all organisms. Since we are exploring different organisms in this use case, all with different gene/transcript ID nomenclatures, an ortholog file must be used for the comparison. If we were comparing organisms mapped against the same annotation set (such as diseased human cells vs. healthy human cells), an ortholog file would not be necessary.

A sample run of CAPE-compare follows, using the CAPE-analyze reports generated above. Note that for human H1 embryonic stem cells, we are using the report generated using CAPE's default options:

```
bash-3.2$ java -jar CAPE-compare.jar -i
/emb_study/usecase/wormReport.txt,/emb_study/usecase/flyReport.txt,/emb_study/usecase/hum
anReport.txt -l Worm,Fly,Human -p /emb_study/usecase/output/wfhEmbryoComparison -o
/emb_study/usecase/data/wfhOrthologList.txt

Running CAPE-compare on the following label/file pairs:
Worm    /emb_study/usecase/wormReport.txt
Fly     /emb_study/usecase/flyReport.txt
Human   /emb_study/usecase/humanReport.txt

Initializing variables...
Parsing reports...
Comparing lists and writing results...
Using ortholog file /emb_study/usecase/data/wfhOrthologList.txt...
Writing raw comparison data to /emb_study/usecase/output/wfhEmbryoComparison.txt...
Writing summary tables to /emb_study/usecase/output/wfhEmbryoComparison.html...
Writing R script to generate Venn Diagrams to
/emb_study/usecase/output/wfhEmbryoComparison.r...
Complete!
```

CAPE-compare Output

Up to three files will be generated for each CAPE-compare run. In all cases, a tab-delimited text file containing the category breakdown for each transcript in an ortholog set will be produced. In cases where two, three, or four CAPE-analyze reports are being compared, two additional files will also be generated. The first is a summary breakdown of transcripts shared between organisms within each CAPE-analyze category. Clicking the numbers in the table will take you the corresponding list of feature IDs. The third file is a ready-to-run R script that will generate Venn diagrams for each CAPE-analyze category showing the breakdown by organism. Note that the free VennDiagram R package must be installed to use the R script (see Users Guide) and that Venn diagrams will only be produced for categories with at least one data point. Links to the files generated by this run appear below, as well as the Venn diagrams produced by R.

CAPE-compare tab-delimited report (raw data):
http://archive.gersteinlab.org/proj/CAPE/usecases/reports/wfhEmbryoComparison.txt

Excerpt:

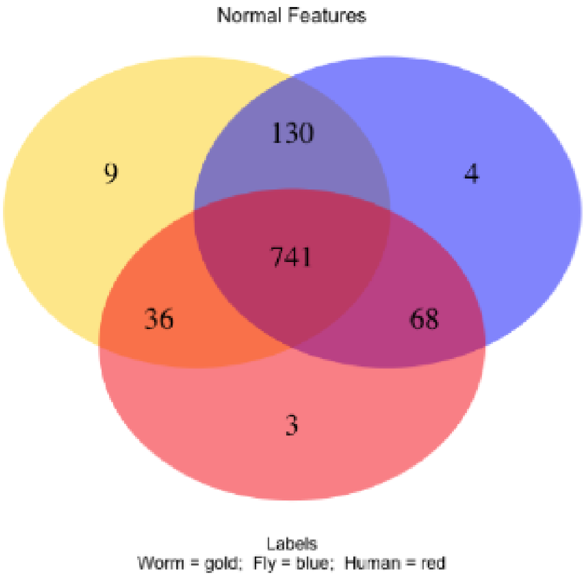
```
#A. Transcripts with low promoter binding and high expression
#B. Transcripts with high promoter binding and low expression
#C. Transcripts with no promoter binding and no expression
#D. Transcripts with no promoter binding
#E. Transcripts with no expression
#F. "Normal" transcripts
#NA. No data available in CAPE-analyze report file

#Worm Feature  Fly Feature    Human Feature  Worm State     Fly State      Human State
dhc-1   FBgn0261797    ENSG00000197102        F       F       F
prp-8   FBgn0033688    ENSG00000174231        F       F       F
ama-1   FBgn0003277    ENSG00000181222        F       F       F
sma-1   FBgn0004167    ENSG00000137877        F       F       D
pyr-1   FBgn0003189    ENSG00000084774        F       F       F
F33H2.5 FBgn0020756    ENSG00000177084        F       F       A
…
…
…
```

CAPE-compare HTML Summary Report:
http://archive.gersteinlab.org/proj/CAPE/usecases/reports/wfhEmbryoComparison.html

CAPE-compare R Script (Produces the Venn Diagrams):
http://archive.gersteinlab.org/proj/CAPE/usecases/reports/wfhEmbryoComparison.r

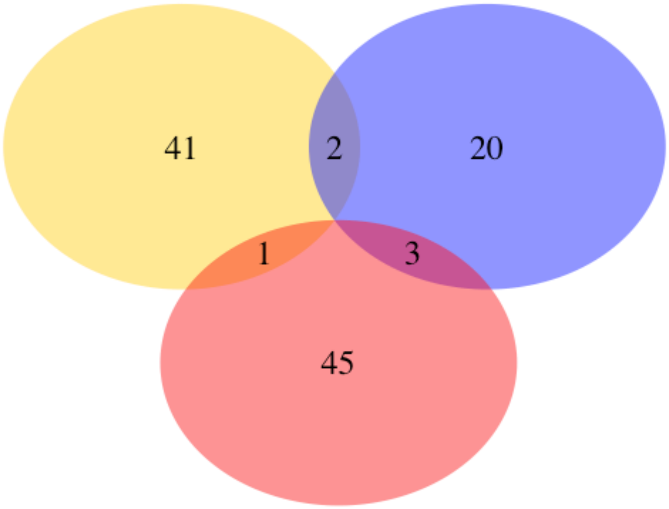
When comparing 1,010 possible orthologous genes across worm early embryo, fly embryo, and human H1 embryonic stem cells, we see that most transcripts fall in the "normal" category for all three organisms. That is, 741 orthologs do not show an extreme difference between the degrees of mRNA abundance and RNAPII binding. 250 orthologs show an extreme case in at least one organism. 29 orthologs were not included in the comparison due to missing data in at least one organism (in most cases, this was due to the human ortholog not meeting an iIDR quality cutoff of > 1). In cases where data is missing for at least one member of an ortholog set, the entire set will be ignored by CAPE-compare.



- 741 orthologs shared between all three organisms
- 250 orthologs show variation between organisms
- 29 orthologs omitted due to missing data (iIDR threshold, no signal data, etc).

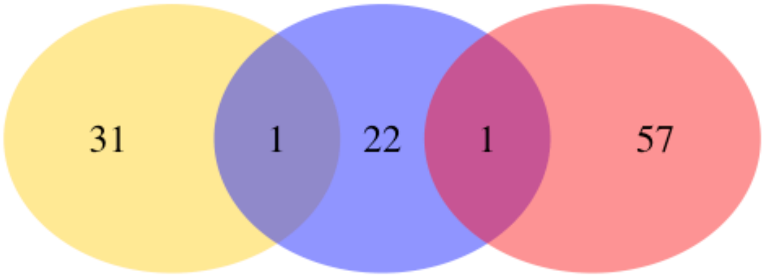
The 250 orthologs that are classified differently in at least one organism are interesting, as this difference may indicate differential regulation between worm, fly, and human embryos. Examining both genes with a stalled polymerase (high RNAPII binding, low expression) and genes that are undergoing a burst of transcription or not transcribed by RNAPII (high expression, low RNAPII binding), we find that affected genes are predominantly organism-specific. 37 orthologs did not have RNAPII binding data for human H1 embryonic stem cells from ChIP-Seq, either due to these promoters falling in unmappable regions or due to a genuine lack of RNAPII binding. No orthologs in this analysis fell into the "No expression" or the "No binding, no expression" categories. Gene IDs for each category and grouping can be found in the CAPE-compare HTML summary report linked above.

Features with high promoter binding and low expression

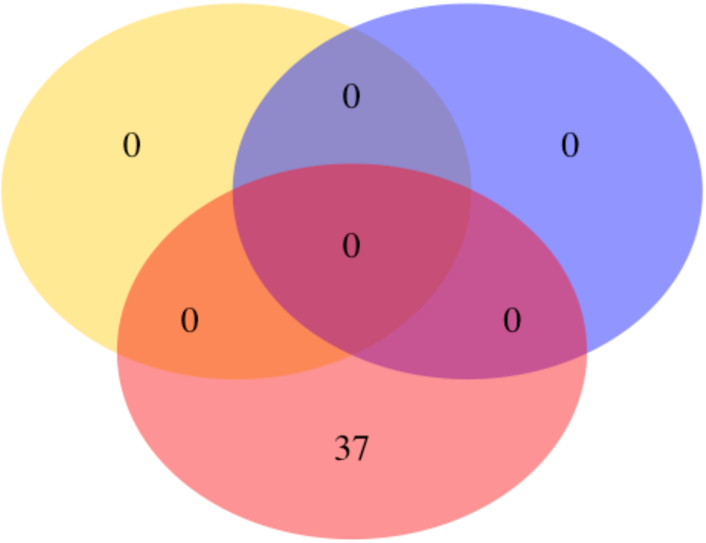


Labels
Worm = gold; Fly = blue; Human = red

Features with low promoter binding and high expression



Labels
Worm = gold; Fly = blue; Human = red



Features with no promoter binding

Labels
Worm = gold; Fly = blue; Human = red

Use Case 2: Classifying ChIP-Seq Peaks Using CAPE-analyze in Binding Mode

This use case will demonstrate CAPE's binding mode, an additional mode that can help researchers better annotate ChIP-Seq peaks. The functionality of this mode is similar to that of BedTools' closestBed program (http://code.google.com/p/bedtools/), but supports ENCODE formats such as bigwig, bigBed, and narrowPeak files. This mode combines ChIP-Seq peak annotations with transcript or gene annotations and mRNA abundance from RNA-seq. The output is a table of each peak that identifies the nearest feature within a user-defined window, and whether a peak should be classified as associated with a transcription start site (TSS), a transcription termination site (TTS), or neither. For more information about binding mode, please see the Users Guide.

This use case uses publicly available ChIP-Seq signal and peak files for RNA polymerase II from the ENCODE consortium. Annotations were produced by the GENCODE Project using expression data from the ENCODE consortium. All data were generated from the K562 human cell line.

Sample Data

Signal File:

http://archive.gersteinlab.org/proj/CAPE/usecases/data/wgEncodeSydhTfbsK562Pol2StdSig.bigWig

(Note: this file is unaltered from its original version available from the ENCODE public data repository at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/wgEncodeSydhTfbsK562Pol2StdSig.bigWig)

Transcript File:

http://archive.gersteinlab.org/proj/CAPE/usecases/data/K562-HighIDRTranscripts.gtf

(Note: this file is adapted from the GENCODE project file by taking only those transcripts with iIDR [a quality metric] >= 1.0. The original file is available at http://genome.crg.es/~jlagarde//encode/pre-DCC/wgEncodeCshlLongRnaSeq//20120220_long_quantifications_gencodev10_cufflinks_cshl_NOT_SUBMITTED/LID16629-LID16630_TranscriptGencV10IAcuff.gtf

Peak File:

http://archive.gersteinlab.org/proj/CAPE/usecases/data/wgEncodeSydhTfbsK562Pol2StdPk.narrowPeak

(Note: this file is unaltered from its original version available from the ENCODE public data repository at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/wgEncodeSydhTfbsK562Pol2StdPk.narrowPeak.gz)

Generating a CAPE-analyze Report in Binding Mode

A sample run of CAPE-analyze in binding mode follows. Please note that the "—binding" flag must be specified to run the tool in binding mode, as CAPE's default behavior is to run in transcript mode. Please note that directories are specific to our test system and should be changed if trying to reproduce this result.

```
bash-3.2$ java -jar CAPE-analyze.jar --
signalfile=/ptemp/wgEncodeSydhTfbsK562Pol2StdSig.bigWig --transcriptfile=/ptemp/
gtf/K562-HighIDRTranscripts.gtf --outputfile=/ptemp/CAPE-K562BindingReport.xls --
peakfile=/ptemp/ChIP-Seq/peaks/wgEncodeSydhTfbsK562Pol2StdPk.narrowPeak --binding
```

```
Running in binding mode...
Loading ChIP-Seq peak file...
Loading transcript file...
Finding maxima and corresponding signal levels...
Finding nearest TSS and TTS for maxima...
Writing output to /ptemp/CAPE-K562BindingReport.xls...
Program completed in 38 seconds
```

Sample Output

Output from CAPE-analyze running in binding mode:
http://archive.gersteinlab.org/proj/CAPE/usecases/reports/CAPE-K562BindingReport.xls

Excerpt:

```
#Transcript File: ptemp/NewApproach/gtf/K562-HighIDRTranscripts.gtf

#Signal File: ptemp/wgEncodeSydhTfbsK562Pol2StdSig.bigWig
#Peak File: /ptemp/NewApproach/ChIP-Seq/peaks/wgEncodeSydhTfbsK562Pol2StdPk.narrowPeak
#Upstream pad = 1000 bp; Downstream pad = 1000
#Mode: binding

#Pad values used: 1000bp upstream, 1000bp downstream

#Chromosome    Start   End    PeakPosition   PeakScore      DistanceToNearestTSS
        TSSTranscriptID       TSS_RPKM       DistanceToNearestTTS   TTSTranscriptID
        TTS_RPKM       Association
chr1    713770 714492  713983 514.7   23     ENST00000428504.1      2.328  3588
        ENST00000457084.1     0.278   TSS
chr1    762552 763294  762819 163.9   83     ENST00000473798.1      0.415  1233
        ENST00000473798.1     0.415   TSS
chr1    839851 840391  840212 90.4    42228  ENST00000483767.1      0.112  39372
        ENST00000327044.6     5.482   Neither
chr1    878395 878889  878597 35.1    3843   ENST00000483767.1      0.112  987
        ENST00000327044.6     5.482   TTS
chr1    894411 894815  894624 185.8   12     ENST00000469563.1      0.902  998
        ENST00000469563.1     0.902   TSS
chr1    901209 902557  902316 112.5   3384   ENST00000481067.1      0.455  1221
        ENST00000338591.3     1.553   Neither
chr1    935246 936536  935441 136.5   111    ENST00000428771.2      1.113  1099
        ENST00000428771.2     1.113   TSS
```