COPAR Manual

Binhua Tang (bh.tang@outlook.com)

July 28, 2016

About COPAR

COPAR (ChIP-seq Optimal Peak AnalyzeR) is designed in R, and used for optimal estimation of peak alignment, and genomic feature extraction for ChIP-seq data. It contains five main functions, i.e.

1. **copar.Aligner**: the function is to detect the optimal peak candiate from the raw ChIP-seq input data (in BED format mapped using Bowtie or other tools);
2. **copar.Miner**: the function is to identify the optimal peak candiate and feature extraction analysis for the raw ChIP-seq input data (in BED format);
3. **copar.optiPN**: the function is to find the optimal peak count candiate and corresponding statistically meaningful FDR (e.g. <= 0.05);
4. **copar.SigPattern**: this function will analyze the signal pattern of the detected peak count candidate list. Together it will generate a heatmap plot (TIFF format) for the randomized sequence of the detected peak count candiate list;
5. **sigProcess**: COPAR singal process function.

ChIP-seq datasets used as study cases in this manual:

1. SRR399019, Reference [1];
2. ERR022052, Reference [2];
3. SRR015350, Reference [3].

References:

1. Tang, B., et al., Hierarchical modularity in ERa transcriptional network is associated with distinct functions and implicates clinical outcomes. NPG Scientific Reports, 2012. 2.
2. Hurtado, A., et al., FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nat Genet, 2011. 43(1): p. 27-33.
3. Welboren, W.J., et al., ChIP-Seq of ERa and RNA polymerase II defines genes differentially responding to ligands. The EMBO Journal, 2009. 28(10): p. 1418-1428.

1. copar.Aligner

The function to detect the optimal peak candiate from the raw ChIP-seq input data (in BED format);

```
library(COPAR)
```

```
## Loading required package: signal
```

```
##
## Attaching package: 'signal'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, poly
```

```
#?copar.Aligner

# copar.Aligner
# Usage
# Main steps to run COPAR.Aligner:
# The function can only run in Linux environment:
# (1) The engine package "BELT1.0.2_linux64" and input data should be located on # the reachable direct
# (2) Open R (in Linux), and run:
# copar.Aligner(hg = "hg18",
#               ws = seq(from = 100, to = 500, by = 50),
#               pv = seq(from = 0.951, to = 0.999, by = 0.003),
#               bf = "SRR399019_Sorted.bed",
#               otf = "SRR399019_copar")
```

```
## Aligned result
data(SRR399019_copar)
head(BS,5)
```

```
##   BinSize  Perc PeakNumber Threshol SignalReads NoiseLevel      TPR
## 1     100 0.951      16349  10.0170     2229888   96.7166% 99.0764%
## 2     100 0.954      16414  10.2696     2227957   96.7194%  99.141%
## 3     100 0.957      13784  10.5215     2149540   96.8349% 99.9057%
## 4     100 0.960      14171  10.8672     2139155   96.8502%   99.88%
## 5     100 0.963      13554  11.2679     2046238    96.987% 99.9336%
##        FDR
## 1 11.2891%
## 2 11.2439%
## 3 13.0895%
## 4 12.8752%
## 5  13.561%
```

2. copar.Miner

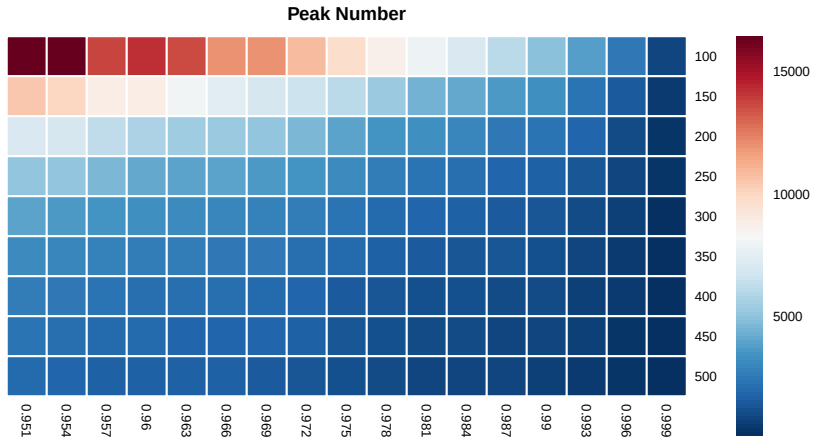
The function is to identify the optimal peak candiate from the raw ChIP-seq input data (in BED format);

```
library(COPAR)
# ?copar.Miner
# copar.Miner(input = "SRR399019_copar.rda",
#             plotPN = TRUE,
#             plotFDR = TRUE,
#             maxPN = TRUE)
```

```
## Aligned Peak Number Candidates
data(SRR399019_copar.PN)
head(BS.PN,5)
```

```
##     0.951 0.954 0.957  0.96 0.963 0.966 0.969 0.972 0.975 0.978 0.981
```

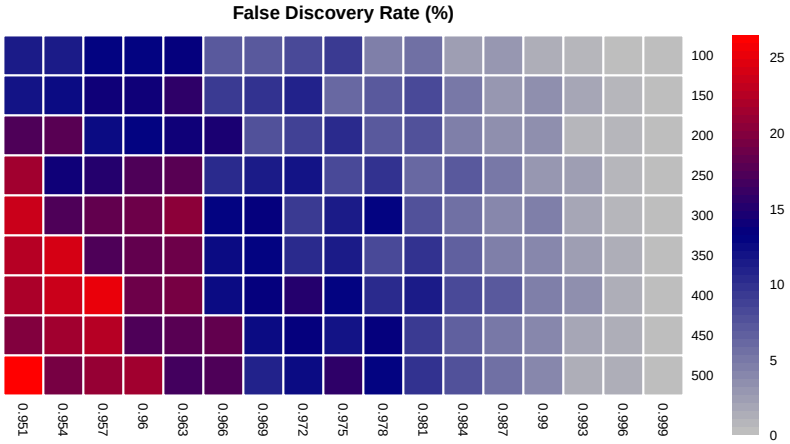
```
## 100 16349 16414 13784 14171 13554 11997 11940 10812  9693  8676  7803
## 150 10504  9982  8892  8864  7955  7383  6952  6547  6036  5224  4472
## 200  7032  6912  6165  5735  5443  5324  5057  4560  3971  3522  3271
## 250  5106  5083  4657  4046  3928  3922  3653  3406  3069  2617  2327
## 300  3979  3621  3421  3381  3172  2980  2894  2715  2304  2027  1859
##     0.984 0.987 0.99 0.993 0.996 0.999
## 100  7059  6058 4880  3865  2421   788
## 150  4158  3564 3232  2395  1573   523
## 200  2912  2548 2279  1797  1102   387
## 250  2183  1908 1697  1345   865   313
## 300  1664  1531 1360  1021   644   245
```




```
## Corresponding FDR Candidates
data(SRR399019_copar.FDR)
head(BS.FDR,5)
```

```
##       0.951   0.954   0.957    0.96   0.963   0.966    0.969    0.972
## 100 11.2891 11.2439 13.0895 12.8752 13.5610  7.1887  7.34816  8.12979
## 150 11.6954 12.3778 13.7796 13.9992 15.4876  9.2827  9.97016 10.63680
## 200 17.2967 17.8289 12.3059 13.2164 14.1304 14.6374  7.88121  8.79765
## 250 21.2034 13.8879 15.0993 17.1513 17.8471 10.2826 11.12520 12.05730
## 300 23.3512 17.4514 18.4314 18.8679 20.1678 12.7410 13.42490  9.48185
##       0.975    0.978   0.981   0.984   0.987    0.99    0.993    0.996
## 100  9.13841  4.76356 5.38186 2.50843 2.94568 1.51612 0.821468 0.526547
## 150  5.99415  7.01623 8.20404 4.93798 2.91537 3.39623 1.994060 0.598007
```

```
## 200 10.12960  6.93260 7.72167 4.59340 3.22324 3.65317 0.922722 0.657277
## 250  8.35655  9.97131 6.29691 6.97051 5.06460 2.96560 2.222220 0.954654
## 300 11.25300 12.72810 7.84615 5.78171 3.96877 4.58849 1.998000 0.801282
##        0.999
## 100 0.000000
## 150 0.000000
## 200 0.000000
## 250 0.331126
## 300 0.420168
```



3. copar.optiPN

The function is to find the optimal peak count candiate and corresponding statistically meaningful FDR (e.g. <= 0.05);

```
library(COPAR)
#?copar.optiPN
#copar.optiPN(PN=BS.PN, FDR=BS.FDR, thres_fdr=5)
```

```
## $MaxPeak
## [1] 8676
##
## $SatisFDR
```

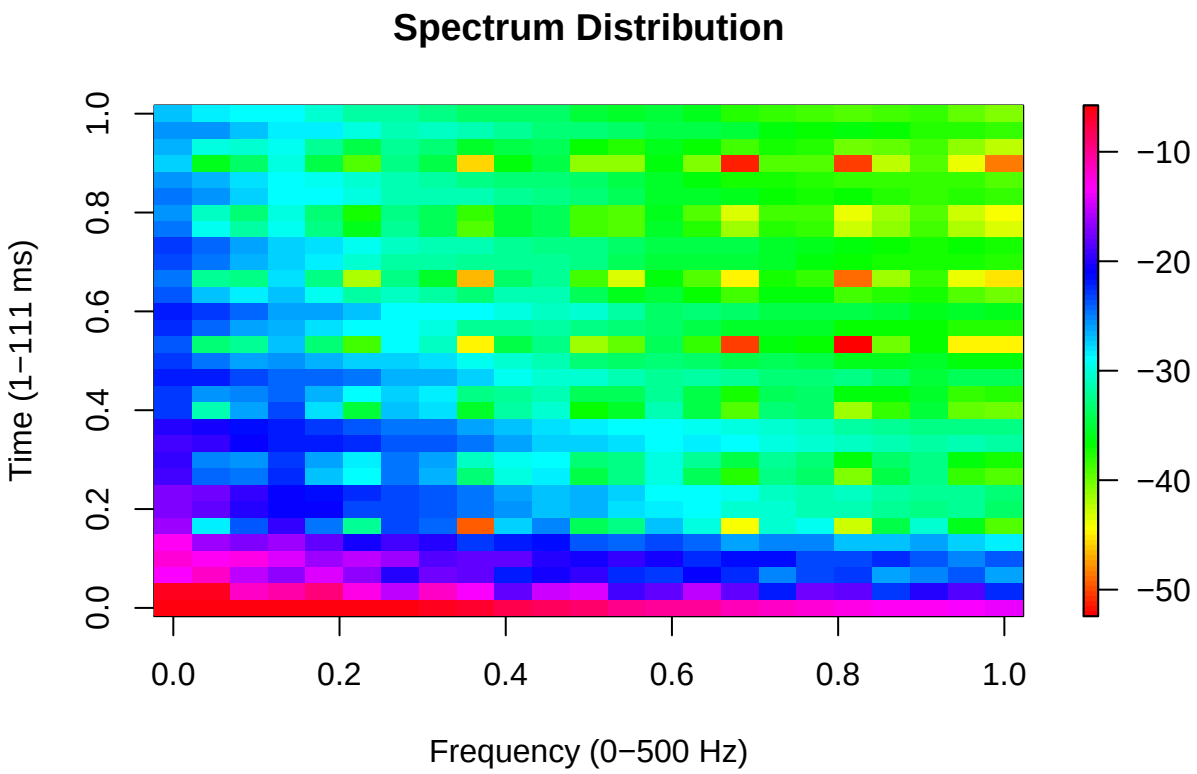
```
## [1] 4.76356
##
## $Index
##     row col
## 100   1  10
```

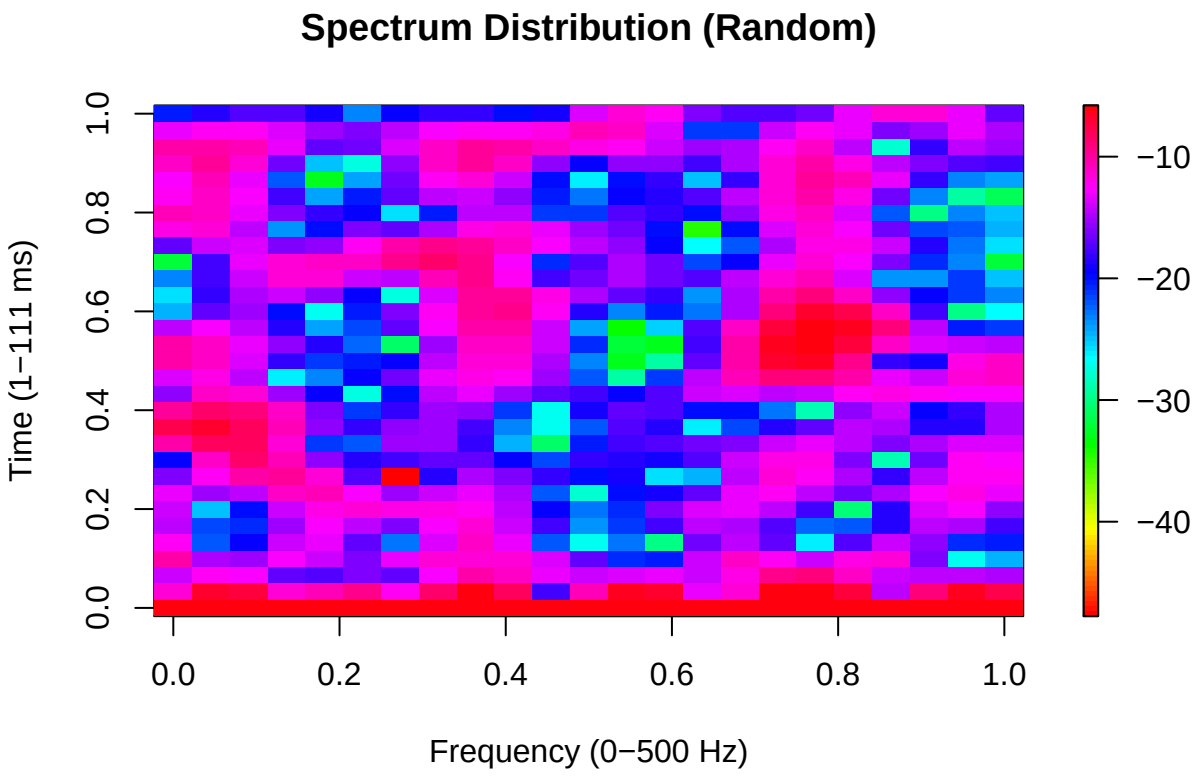
4. copar.SigPattern

This function will analyze the signal pattern of the detected peak count candidate list. Together it will generate a heatmap plot (TIFF format) for the randomized sequence of the detected peak count candiate list;

```
library(COPAR)
# ?copar.SigPattern
# copar.SigPattern(inpFile="SRR399019_copar.rda",plotSigPattern=T)
```

```
## Loading required package: spam

## Loading required package: grid

## Spam version 1.3-0 (2015-10-24) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.

##
## Attaching package: 'spam'

## The following objects are masked from 'package:base':
##
##     backsolve, forwardsolve

## Loading required package: maps

##
##  # maps v3.1: updated 'world': all lakes moved to separate new #
##  # 'lakes' database. Type '?world' or 'news(package="maps")'.  #
```


Spectrum Distribution
Time (1−111 ms)
0.0
0.2
0.4
0.6
0.8
1.0
Frequency (0−500 Hz)
−10
−20
−30
−40
−50



5. sigProcess

COPAR singal process function.

```
library(COPAR)
# sigProcess(x, inpName="SRR015350_COPA")
```

6. Study Case 2: ERR022052

Reference: Hurtado, A., et al., FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nat Genet, 2011. 43(1): p. 27-33.

```
## Aligned result
data(ERR022052_copar)
head(BS,5)
```

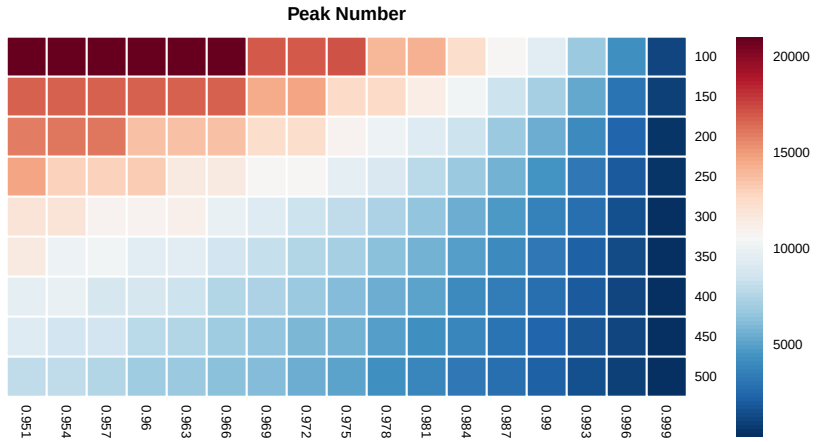
```
##   BinSize  Perc PeakNumber Threshol SignalReads NoiseLevel     TPR
## 1     100 0.951      20853  3.65152      538500    93.6163 96.8302
## 2     100 0.954      20976  3.69101      535384    93.6532 76.8640
## 3     100 0.957      20976  3.73039      535361    93.6535 76.7830
## 4     100 0.960      21008  3.76966      534855    93.6595 76.7517
## 5     100 0.963      21025  3.83836      534663    93.6617 76.8561
```

```
##        FDR
## 1 13.72420
## 2  2.09087
## 3  2.09202
## 4  2.09023
## 5  2.08665
```

```
library(COPAR)
# ?copar.Miner
#copar.Miner(input = "ERR022052_copar.rda",
#            plotPN = TRUE,
#            plotFDR = TRUE,
#            maxPN = TRUE)
```

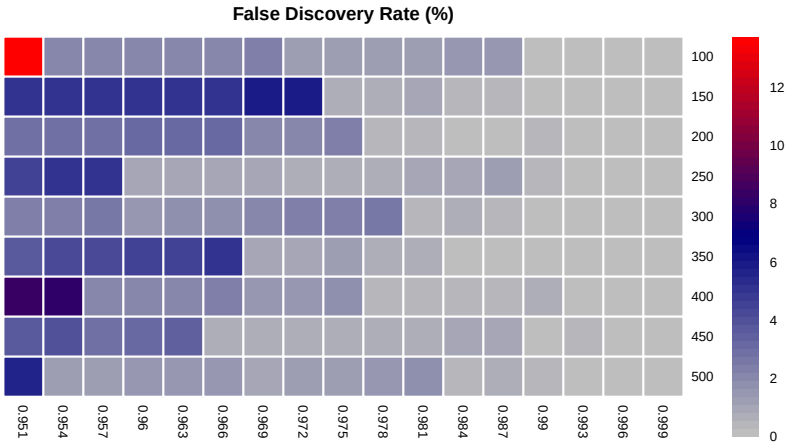
```
## Aligned Peak Number Candidates
data(ERR022052_copar.PN)
head(BS.PN,5)
```

```
##     0.951 0.954 0.957  0.96 0.963 0.966 0.969 0.972 0.975 0.978 0.981
## 100 20853 20976 20976 21008 21025 21029 17038 17061 17173 14116 14184
## 150 16657 16658 16668 16766 16809 16817 14558 14634 12605 12689 11436
## 200 16007 16098 16103 13565 13647 13672 12350 12451 10955 10146  9173
## 250 14739 13037 13054 13117 11459 11535 10630 10692  9700  9014  7738
## 300 11958 12029 11001 11017 11082  9924  9298  8495  8039  7425  6579
##     0.984 0.987 0.99 0.993 0.996 0.999
## 100 12339 10631 9525  6761  4352  1140
## 150 10197  8338 7091  5225  3101   832
## 200  8510  6724 5573  4065  2479   605
## 250  6752  5609 4475  3240  1921   481
## 300  5528  4689 3688  2711  1605   388
```




```
## Corresponding FDR Candidates
data(ERR022052_copar.FDR)
head(BS.FDR,5)
```

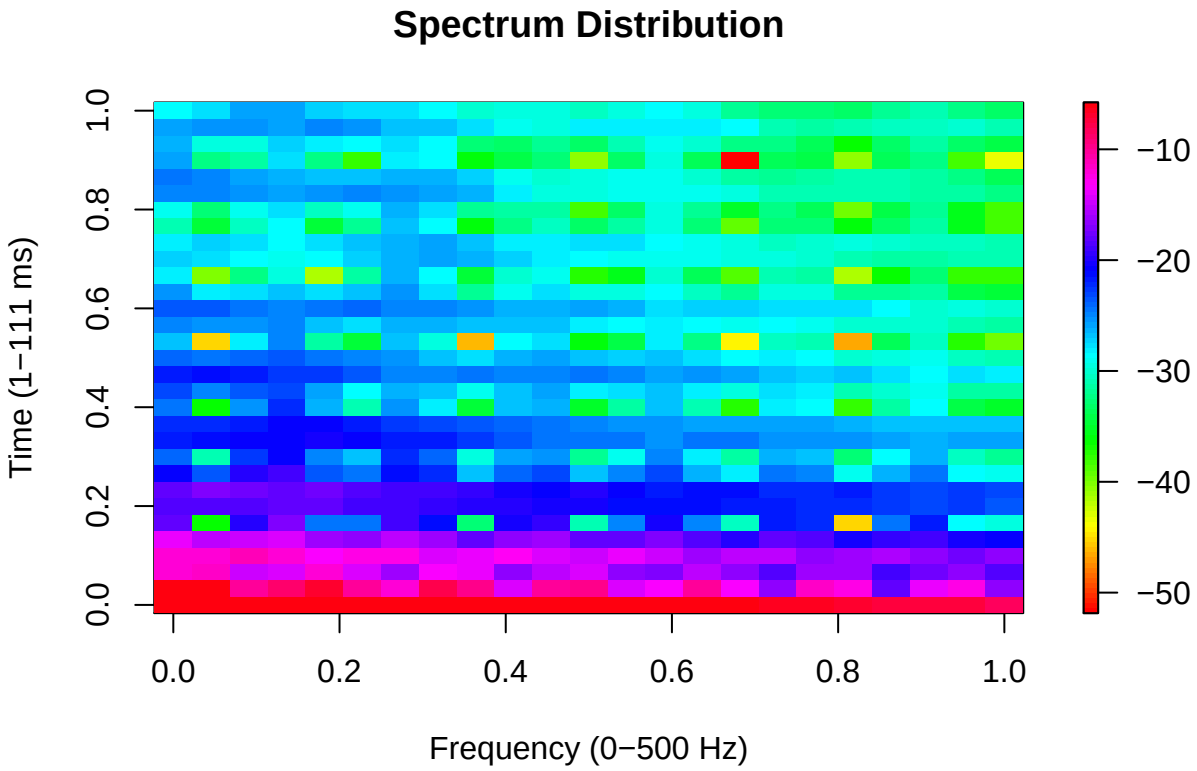
```
##        0.951   0.954   0.957     0.96    0.963    0.966   0.969    0.972
## 100 13.72420 2.09087 2.09202 2.090230 2.086650 2.084610 2.29649 1.115650
## 150  5.04842 5.04900 5.04755 5.029110 5.019370 5.015660 5.85398 5.842210
## 200  2.77428 2.77944 2.77815 3.050090 3.047860 3.041550 1.95138 1.941750
## 250  4.50868 5.20810 5.21237 0.846638 0.921659 0.925846 1.01370 0.620592
## 300  2.34542 2.33584 2.55999 1.635660 1.657710 1.876430 2.01171 2.197420
##        0.975    0.978    0.981    0.984    0.987      0.99     0.993
## 100 1.140360 1.204290 1.209820 1.405090 1.613500 0.1695450 0.0743494
## 150 0.798545 0.794837 0.880404 0.384426 0.480885 0.1135400 0.1350570
## 200 2.209450 0.346569 0.382681 0.201613 0.269461 0.3256740 0.0000000
## 250 0.682876 0.746186 0.867988 1.037190 1.299390 0.3830550 0.1553280
## 300 2.363660 2.581500 0.489596 0.581501 0.343053 0.0273898 0.0372439
##         0.996 0.999
## 100 0.0232883     0
## 150 0.2278650     0
## 200 0.0406339     0
## 250 0.0000000     0
## 300 0.0630517     0
```

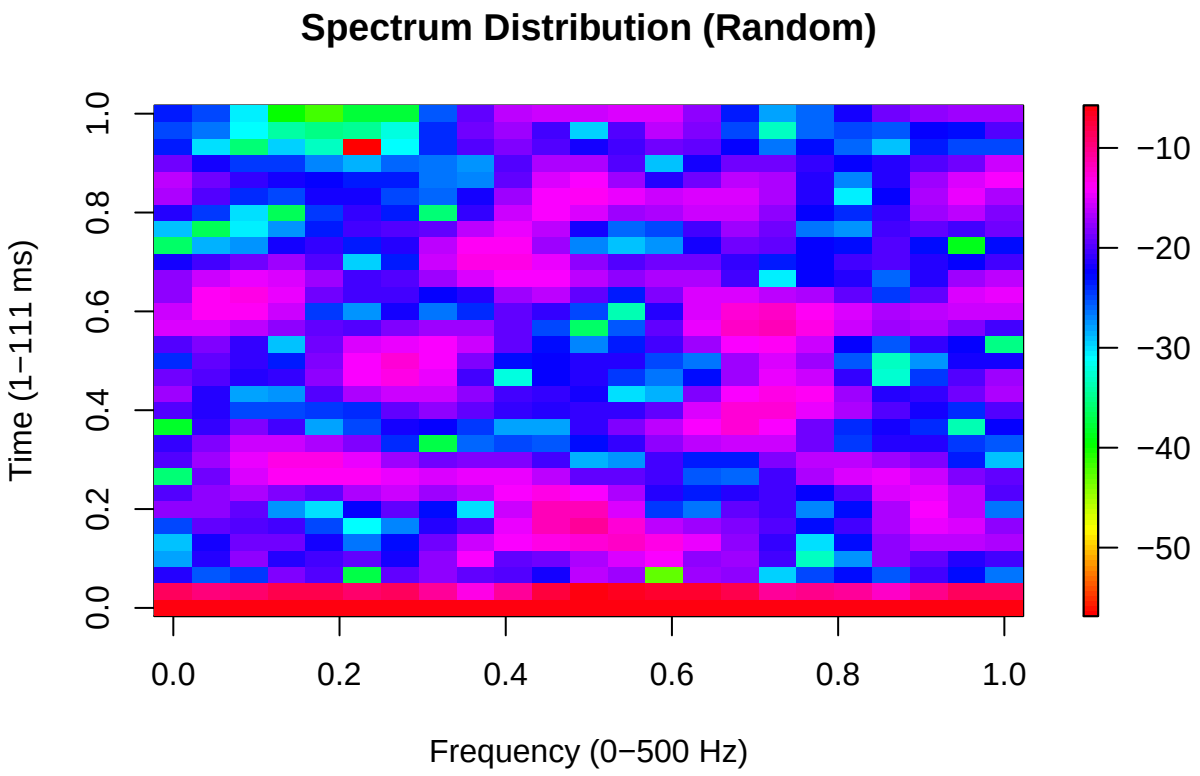



```
library(COPAR)
#?copar.optiPN
#copar.optiPN(PN=BS.PN, FDR=BS.FDR, thres_fdr=5)
```

```
## $MaxPeak
## [1] 21029
##
## $SatisFDR
## [1] 2.08461
##
## $Index
##     row col
## 100   1   6
```

```
library(COPAR)
# ?copar.SigPattern
# copar.SigPattern(inpFile="SRR015352_copar.rda", plotSigPattern=T)
```


Spectrum Distribution
Time (1−111 ms)
0.0
0.2
0.4
0.6
0.8
1.0
Frequency (0−500 Hz)
−10
−20
−30
−40
−50



7. Study Case 3: SRR015350

Reference: Welboren, W.J., et al., ChIP-Seq of ERa and RNA polymerase II defines genes differentially responding to ligands. The EMBO Journal, 2009. 28(10): p. 1418-1428.

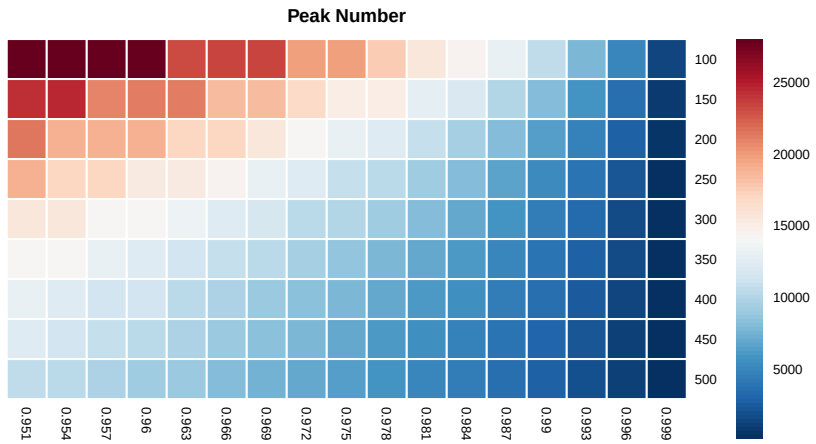
```
## Aligned result
data(SRR015350_copar)
head(BS,5)
```

```
##   BinSize  Perc PeakNumber Threshol SignalReads NoiseLevel      TPR
## 1     100 0.951      27862  4.63084     1104394   90.4973% 99.6519%
## 2     100 0.954      27961  4.69271     1100792   90.5283% 99.5851%
## 3     100 0.957      28000  4.79269     1100204   90.5333% 99.5786%
## 4     100 0.960      28016  4.95279     1099956   90.5355% 99.6074%
## 5     100 0.963      23252  5.11243     1034998   91.0944% 99.7678%
##         FDR
## 1 2.76762%
## 2 2.77006%
## 3 2.75995%
## 4 2.75802%
## 5 3.38514%
```

```
library(COPAR)
# ?copar.Miner
# copar.Miner(input = "SRR015350_copar.rda",
#             plotPN = TRUE,
#             plotFDR = TRUE,
#             maxPN = TRUE)
```

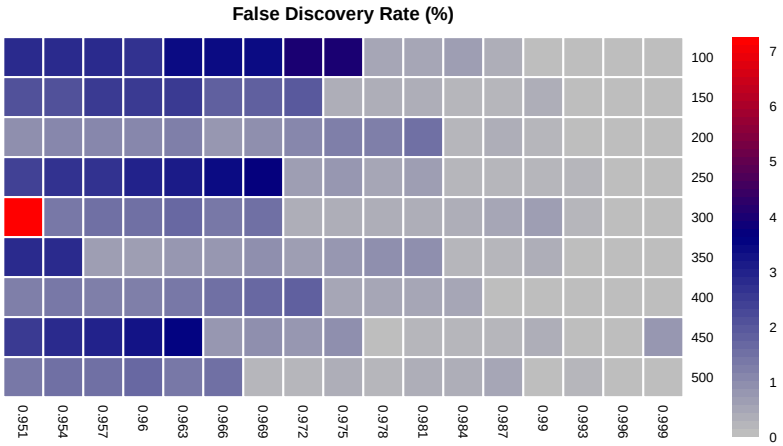
```
## Aligned Peak Number Candidates
data(SRR015350_copar.PN)
head(BS.PN,5)
```

```
##     0.951 0.954 0.957  0.96 0.963 0.966 0.969 0.972 0.975 0.978 0.981
## 100 27862 27961 28000 28016 23252 23277 23361 19746 19837 17481 15656
## 150 24366 24399 21038 21041 21133 18344 18427 16615 15023 15062 12859
## 200 21354 18939 18943 19003 16875 16922 15484 14135 13209 12317 10919
## 250 18938 17080 17126 15414 15456 14359 13154 12369 10984 10399  9103
## 300 15570 15578 14206 14246 13269 12322 11611 10400  9892  9085  8077
##     0.984 0.987  0.99 0.993 0.996 0.999
## 100 14351 13153 10644  7787  4878  1296
## 150 12085 10127  8012  5898  3567   817
## 200  9378  7950  6522  4609  2745   547
## 250  8034  6786  5297  3852  2193   391
## 300  6863  5816  4548  3266  1779   287
```



```
## Corresponding FDR Candidates
data(SRR015350_copar.FDR)
head(BS.FDR,5)
```

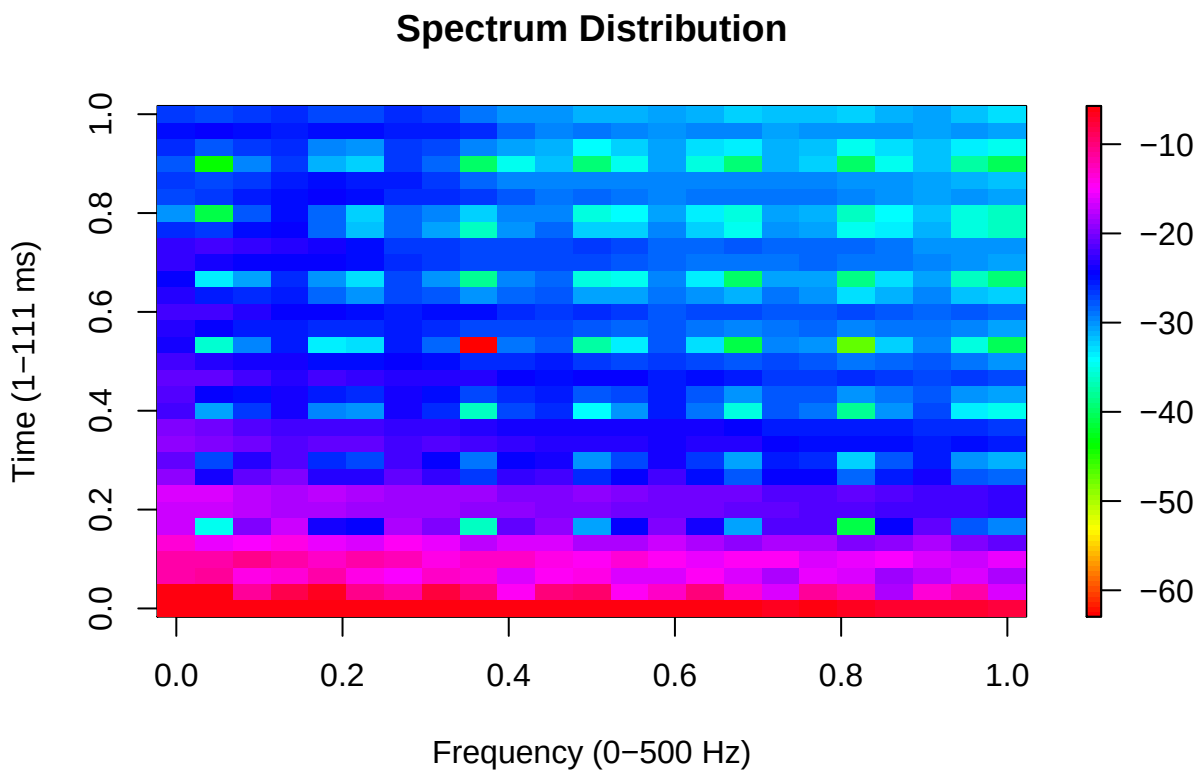
```
##        0.951   0.954   0.957    0.96   0.963    0.966    0.969    0.972
## 100 2.767620 2.77006 2.75995 2.75802 3.38514 3.385930 3.379930 4.054320
## 150 2.106160 2.10403 2.49230 2.49056 2.48418 1.775630 1.770200 2.000470
## 200 0.927955 1.08639 1.08627 1.08929 1.23935 0.835983 0.919555 1.069590
## 250 2.395770 2.71060 2.71724 3.04562 3.04943 3.354110 3.729860 0.694444
## 300 7.258890 1.42566 1.59437 1.59227 1.73531 1.424660 1.551970 0.375361
##        0.975    0.978    0.981    0.984    0.987      0.99     0.993
## 100 4.047740 0.486855 0.555130 0.627834 0.328595 0.0756716 0.0129266
## 150 0.326623 0.325928 0.397320 0.241365 0.287442 0.3880820 0.0680156
## 200 1.181430 1.281430 1.523530 0.267151 0.315736 0.2156830 0.0654022
## 250 0.799056 0.510057 0.636453 0.199875 0.236581 0.1895730 0.2602810
## 300 0.394697 0.429705 0.335362 0.394449 0.481596 0.6578950 0.1532800
##         0.996 0.999
## 100 0.0412456     0
## 150 0.0000000     0
## 200 0.1094890     0
## 250 0.0000000     0
## 300 0.0000000     0
```

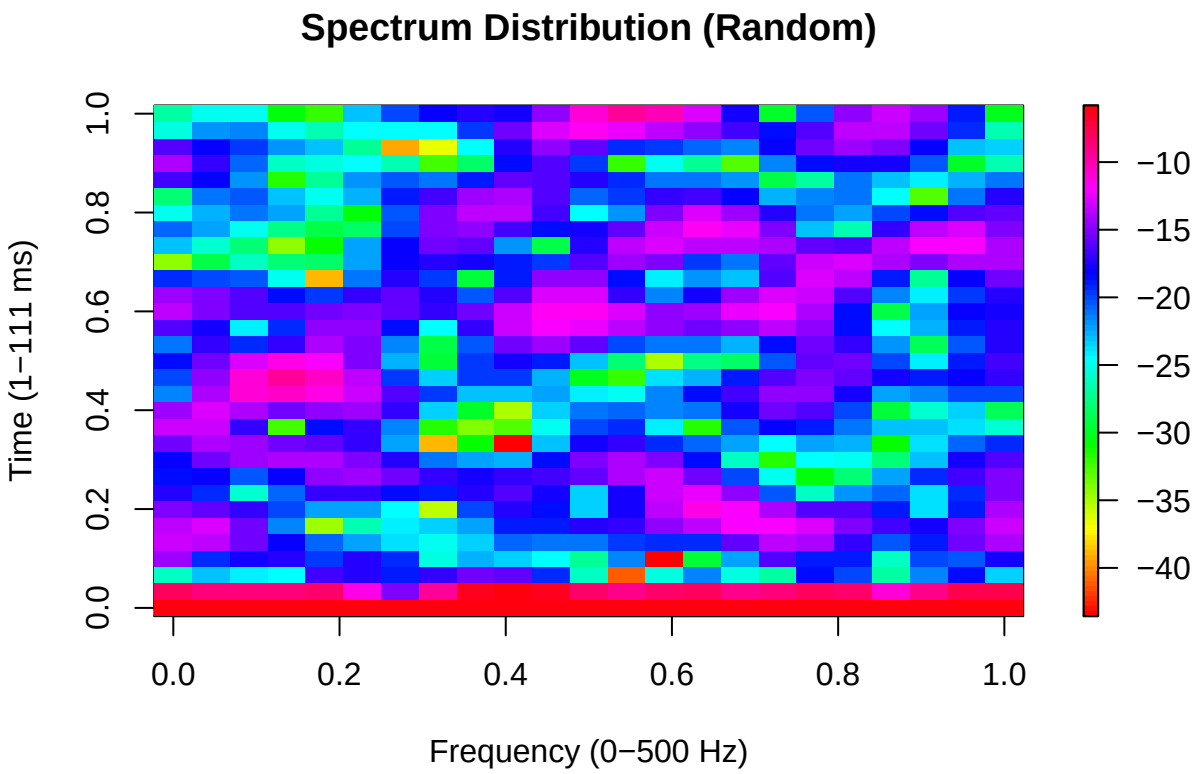


```
library(COPAR)
#?copar.optiPN
#copar.optiPN(PN=BS.PN, FDR=BS.FDR, thres_fdr=5)
```

```
## $MaxPeak
## [1] 28016
##
## $SatisFDR
## [1] 2.75802
##
## $Index
##     row col
## 100   1   4
```

```
library(COPAR)
# ?copar.SigPattern
# copar.SigPattern(inpFile="SRR015350_COPAR.rda",plotSigPattern=T)
```




Spectrum Distribution (Random)
Time (1−111 ms)
0.0
0.2
0.4
0.6
0.8
1.0
Frequency (0−500 Hz)
−10
−15
−20
−25
−30
−35
−40