

CMARRT Package Example Version 1.0

Pei Fen Kuan¹, Adam Hinz², Hyonho Chun¹, Sündüz Keleş^{1,3}

¹Department of Statistics, University of Wisconsin,
Madison, WI 53706.

²Department of Computer Science, University of Wisconsin,
Madison, WI 53706.

³Department of Biostatistics and Medical Informatics, University of Wisconsin,
Madison, WI 53706.

December 3, 2008

1 Overview

CMARRT package extends the standard moving average approach (Keles et al.; 2006; Buck et al.; 2005) commonly used in the analysis of ChIP-chip data by incorporating the correlation structure in identifying bound regions as proposed in Kuan et al. (2008) for data from tiling arrays. The package can be loaded with command:

```
> library(CMARRT)
```

The package requires a data frame which consists of chromosome ID, start coordinate, stop coordinate and measurement for each probe. The measurement could be an average log base 2 ratio of the two channels across replicates or (regularized) paired t-statistic for arrays with two channels (e.g., Nimblegen) and a (regularized) two sample t-statistic for single channel arrays (Affymetrix). We assume that the data have been properly normalized. The example data (`data.cmarrt`) used for illustration of the package functionality consists of 20000 probes. An example of input data:

```
> data(data.cmarrt)
> colnames(data.cmarrt)
[1] "chr" "start" "stop" "logR"
> dim(data.cmarrt)
[1] 20000 4
> data.cmarrt[1:10, ] # First 10 rows of the data
```

	chr	start	stop	logR
1	1	103451	103500	-0.21095485
2	3	408226	408275	0.09955185
3	3	333226	333275	-0.27800406
4	2	206676	206725	0.50951252
5	2	268651	268700	-0.17240536
6	3	499401	499450	-0.11449531
7	2	128526	128575	0.26870394
8	2	276426	276475	0.69391971
9	3	421526	421575	-0.04824883
10	3	442401	442450	2.10444395

Probes are declared as bound using adjusted p-values for multiple comparisons under the Gaussian approximation. The main function in **CMARRT** is `cmarrt.ma` which computes the p-values for each probe by taking into account the correlation structure.

CMARRT is developed using the Gaussian approximation approach and thus it is important to check if this assumption is violated. The function `plot.cmarrt` produces the diagnostic plots (histogram of p-values and normal QQ plots) for comparing the distribution of standardized MA statistics under correlation and independence. An example of the diagnostic plot is given in Figure 1. If the distribution of the standardized moving average statistics S_i^* is correctly specified, the quantiles of S_i^* for unbound probes fall along a 45° reference line against the quantiles from the standard Gaussian distribution. In addition, the p-values obtained should be a mixture of uniform distribution between 0 and 1 and a non-uniform distribution concentrated near 0. The list of bound regions is obtained using the function `cmarrt.peak` for a given error rate control which adjusts the p-values for multiple comparisons.

This is an example of using the **CMARRT** package in R. The details of these functions are given in the next section.

```
> library(CMARRT)
> data(data.cmarrt)
> # Run CMARRT
> out <- cmarrt.ma(data.cmarrt, M = 0.02,
+ frag.length=500,window.opt = "fixed.probe")

> str(out)
List of 6
 $ data.sort:'data.frame':      20000 obs. of  5 variables:
  ..$ regID: num [1:20000] 1 1 1 1 1 1 1 1 1 1 ...
```

```

..$ chr : num [1:20000] 1 1 1 1 1 1 1 1 1 1 ...
..$ start: num [1:20000] 1 26 51 76 101 126 151 176 201 226 ...
..$ stop : num [1:20000] 50 75 100 125 150 175 200 225 250 275 ...
..$ logR : num [1:20000] -0.2414 0.0268 -0.1130 0.0227 -0.8147 ...
$ ma : num [1:20000] -0.2414 -0.1073 -0.1092 -0.0762 -0.2239 ...
$ z.cmarrrt : num [1:20000] -1.681 -0.775 -0.788 -0.565 -1.563 ...
$ z.indep : num [1:20000] -3.49 -1.61 -1.64 -1.18 -3.25 ...
$ pv.cmarrrt: num [1:20000] 0.954 0.781 0.785 0.714 0.941 ...
$ pv.indep : num [1:20000] 1.000 0.946 0.949 0.880 0.999 ...

> # Plotting Figure 1
> plot.cmarrrt(out)

> # Identifying peak regions
> bdd.reg <- cmarrrt.peak(out, alpha = 0.05, method = "BH", minrun = 4)

> str(bdd.reg)
List of 2
 $ cmarrrt.bound:List of 6
  ..$ Chr : num [1:23] 1 1 1 1 1 1 3 3 3 3 ...
  ..$ Start : num [1:23] 32851 35651 50676 97951 109726 ...
  ..$ Stop : num [1:23] 33125 36400 51525 98750 110500 ...
  ..$ n.probe: int [1:23] 10 29 33 31 30 41 32 5 30 30 ...
  ..$ min.pv : num [1:23] 1.43e-04 2.46e-62 1.77e-60 4.65e-68 4.97e-60 ...
  ..$ ave.pv : num [1:23] 5.25e-04 3.01e-05 1.49e-05 1.67e-05 8.34e-06 ...
 $ indep.bound :List of 6
  ..$ Chr : num [1:184] 1 1 1 1 1 1 1 1 1 1 ...
  ..$ Start : num [1:184] 1751 14151 15226 16826 19601 ...
  ..$ Stop : num [1:184] 2525 14575 15350 17225 20000 ...
  ..$ n.probe: int [1:184] 30 16 4 15 15 17 11 20 15 32 ...
  ..$ min.pv : num [1:184] 5.24e-06 6.74e-06 1.93e-03 8.15e-05 1.08e-05 ...
  ..$ ave.pv : num [1:184] 0.00099 0.00113 0.00361 0.00240 0.00101 ...

```

The list of bound regions obtained under independence (ignoring the correlation structure), `bdd.reg$indep.bound` is for comparison. It is not recommended to use this list for downstream analysis

2 Function descriptions

2.1 `cmarrrt.ma`

This function extends the moving average approach by incorporating the correlation structure. It also outputs the p-values of the standardized moving average statistics under the

Gaussian approximation.

Usage

```
cmarrt.ma(datafile, M = NULL, frag.length, window.opt = "fixed.probe")
```

Arguments

- **datafile**: `data.frame` with col 1 (chromosome), col 2 (start coordinate), col 3 (stop coordinate), col 4 (probe measurement).
- **M**: rough estimate of the percentage of bound probes. If unknown, leave it `NULL`.
- **frag.length**: average fragment length from sonication. This parameter is used to calculate the width of the moving average window. For most ChIP-chip experiments, this is around 500 bps.
- **window.opt**: option for sliding window, either `"fixed.probe"` or `"fixed.gen.dist"`. Default is `"fixed.probe"`.

Details

The `datafile` contains at least four columns with column 1 containing the chromosome ID, column 2 is the start coordinate, column 3 is the stop coordinate and column 4 is the probe measurement (usually log ratio of intensities). Computation using `window.opt = "fixed.probe"` calculates the moving average statistics within a fixed number of probes and is computationally more efficient. Use this option if the tiling array is regular with approximately constant resolution. `window.opt="fixed.gen.dist"` computes the moving average statistics over a fixed genomic distance.

Value

- **data.sort**: data file sorted by genomic position.
- **ma**: unstandardized moving average (MA) statistics.
- **z.cmarrt**: standardized MA statistics under correlation structure.
- **z.indep**: standardized MA statistics under independence (ignoring correlation structure).
- **pv.cmarrt**: p-values of probes under correlation.
- **pv.indep**: p-values of probes under independence (ignoring correlation structure).

Note

The p-values are obtained under the Gaussian approximation. Therefore, it is important to check the normal quantile-quantile plot if the Gaussian approximation is valid. The function also outputs the computation under independence (ignoring the correlation structure) for comparisons.

2.2 `plot.cmarrt`

Plot the histograms of p-values and normal QQ plots under correlation structure and independence.

Usage

```
plot.cmarrt(cmarrt.ma)
```

Arguments

- `cmarrt.ma`: output object from `cmarrt.ma`.

Details

Diagnostic plots for comparing the distribution of standardized MA statistics under correlation and independence.

Value

Produces histogram of p-values and normal QQ plots under correlation structure and independence.

2.3 `cmarrt.peak`

Obtain bound regions under a given error rate control using correction method from `p.adjust` which adjusts the p-values for multiple comparisons. Type `?p.adjust` for further details.

Usage

```
cmarrt.peak(cmarrt.ma, alpha, method, minrun)
```

Arguments

- `cmarrt.ma`: output object from `cmarrt.ma`.
- `alpha`: error rate control for declaring bound region.

- `method`: correction method inherited from `p.adjust`. ("holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none")
- `minrun`: minimum number of consecutive probes to be called a bound/peak region.

Details

The function returns two objects, `cmarrt.bound` and `indep.bound`. Each object is a list of bound regions which can be accessed by `Chr` (chromosome), `Start` (start coordinate of each bound region), `Stop` (stop coordinate of each bound region), `n.probe` (number of probes within each bound region), `min.pv` (minimum p-values of each bound region), `ave.pv` (average p-values of each bound region). `min.pv` or `ave.pv` can be used to rank the peaks.

Value

- `cmarrt.bound`: list of bound regions obtained under correlation structure. This is in the bed file format and can be used with UCSC genome browser.
- `indep.bound`: list of bound regions obtained under independence (ignoring correlation).

Note

The list of bound regions obtained under independence (ignoring the correlation structure) is for comparison. It is not recommended to use this list for downstream analysis.

2.4 Warning messages

These are some warning messages from CMARRT.

- `Warning:Contiguous regions too short for estimating the correlation structure reliably`: If the largest segment of contiguous probes contains fewer than 200 probes, the correlation structure is not estimated and analysis is performed by assuming independence.
- `Warning:Window size=0. Cannot compute moving average. Check fragment size and array resolution--Terminated`: If the estimated window size is zero or 1, analysis is terminated.

References

Buck, M., Nobel, A. and Lieb, J. (2005). Chipotle: a user-friendly tool for the analysis of chip-chip data, *Genome Biol* **6**(11).

Keles, S., van der Laan, M. J., Dudoit, S. and Cawley, S. (2006). Multiple testing methods for chip-chip high density oligonucleotide array data, *Journal of Computational Biology* **13**(3): 579–613.

Kuan, P., Chun, H. and Keles, S. (2008). Cmarrrt: A tool for the analysis of chip-chip data from tiling arrays by incorporating the correlation structure, *Proceedings of the Pacific Symposium of Biocomputing* **13**: 515–526.

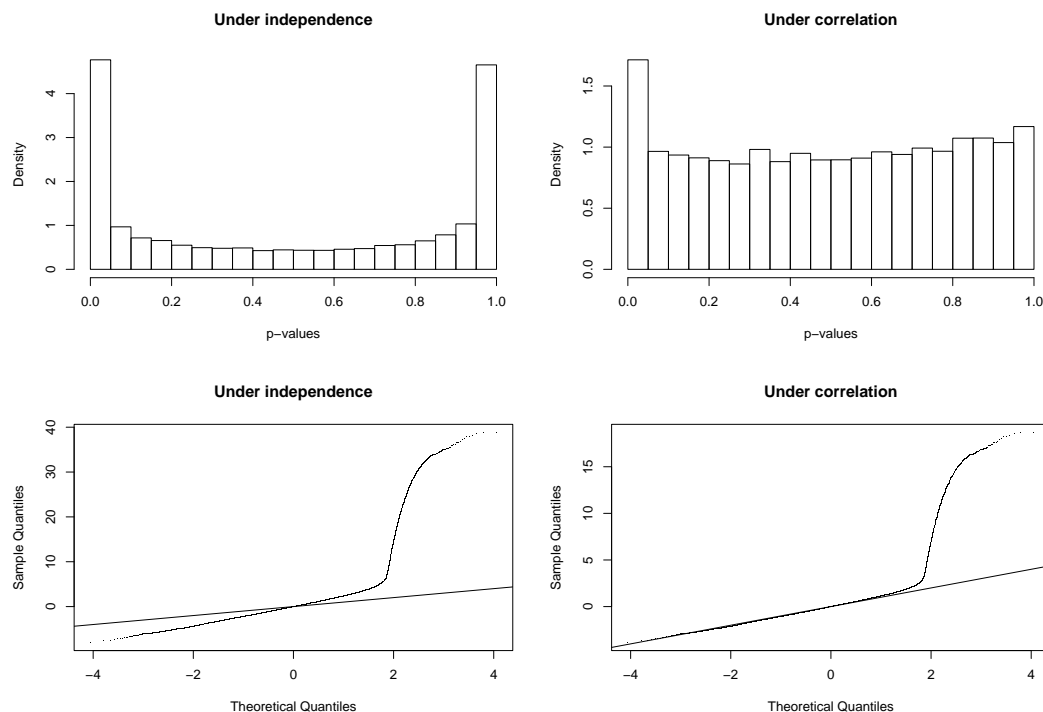


Figure 1: *Histograms of p-values and normal quantile-quantile plots (qqplot)*. The left panels show the distribution of p-values and qqplot of S_i^* when the correlation structure is ignored. The top left panel shows that if the correlation structure is ignored, the distribution of p-values for unbound probes deviates from the uniform distribution for larger p-values. The bottom left panel shows that if the correlation structure is ignored, the distribution of S_i^* s for unbound probes deviates from the standard Gaussian distribution.