# How To Use LncDM

Hui Zhi

July 22, 2016

## Contents

## 1 Overview

The Illumina Infinium HumanMethylation450 BeadChip (450k) is a cost-effective array and cover over 450,000 CpG sites within human genome. LncDM is a package that based on reannotation method to use Illumina HumanMethylation450 BeadChip and Gencode's transcript information to get lncRNAs methylation value. This vignette demonstrates how to easily use the LncDM package. The package can get the methylation matrix and identify the differential methylation sites, elements or the regions who are in the elements of lncRNA etc. The annotation file of Illumina Infinium HumanMethylation450 BeadChip is from GEO and the annotation file of human transcripts is from GENCODE. LncDM gets a matrix of methylation values of the samples (see the section 2.1) and the phenotype of samples (see the section 2.2). Then it can calculate the methylation values of elements (see the section 3). This package uses linear and t test to identify the differential methylation or the sites, elements and regions that related with the phenotype (see the section 4).

## 2 Load and preprocessed raw data

This section introduces how to prepare DNA methylation data and preprocessed the methylation signal intensity files.

## 2.1 preparing methylation sample files

In order to get the different methylation pattern, the user need to prepare methylation signal or beta value files under one directory. Methylation signal files include 4 columns, orderly, there are CpGID, Methylated Intensity, Unmethylated Intensity and Detection P value. Beta value files have 2 columns including CpGID and beta value.

```
> ##the directory of 450k methylation data
> Dir <- system.file("extdata/localdata/Level_2",package="LncDM")
> setwd(Dir)
> files <- list.files(Dir)
> files

[1] "TCGA-3C-AAAU-01A-11D-A41Q-05.txt" "TCGA-3C-AALI-01A-11D-A41Q-05.txt"
[3] "TCGA-3C-AALJ-01A-31D-A41Q-05.txt" "TCGA-BH-A1F0-11B-23D-A138-05.txt"
[5] "TCGA-E2-A15I-11A-32D-A138-05.txt"

> this.sample <- read.table(files[1],sep="\t",header=F);
> head(this.sample)

                     V1                           V2
1     Hybridization REF TCGA-3C-AAAU-01A-11D-A41Q-05
2 Composite Element REF         Methylated_Intensity
3            cg00000029             306.682834933964
4            cg00000108             7303.07609187233
5            cg00000109             4046.71928665482
6            cg00000165             312.809873078168
                            V3                           V4
1 TCGA-3C-AAAU-01A-11D-A41Q-05 TCGA-3C-AAAU-01A-11D-A41Q-05
2       Unmethylated_Intensity            Detection_P_value
3             2652.92460702799                            0
4             354.663183794135                            0
5             772.365266311358                            0
6             2900.05083827659                            0
```

## 2.2 preparing sample phenotype file

User should provide each sample's phenotype.

```
> ##the directory of phenotype file
> Dir <- system.file("extdata/localdata",package="LncDM")
> setwd(Dir)
> ##phenotype file
> files <- list.files(Dir)
> files

[1] "BRCA_pheno.txt" "Level_2"        "loadData.Rdata"

> pheno <- read.table(files[1],sep="\t",header=T);
> head(pheno)

                        sample    type
1 TCGA-3C-AAAU-01A-11D-A41Q-05    case
```

```
2 TCGA-3C-AALI-01A-11D-A41Q-05    case
3 TCGA-3C-AALJ-01A-31D-A41Q-05    case
4 TCGA-E2-A15I-11A-32D-A138-05 control
5 TCGA-BH-A1F0-11B-23D-A138-05 control
```

## 2.3 preprocessed raw data

The function `loaddata` will return beta value of CpG sites and the annotation information for lincRNA, processed transcript, protein coding gene and pseudogene. In this section users can control results by some arguments: `XYchrom`, `sitefilter`, `sitefilterperc`, `snpfilter` is designed to filter CpG sites; `samplefilter`, `samplefilterperc` is designed to filter unqualified samples; `normalization` decide whether to normalize the different chips using quantile normalization; `transfm` decide whether to transform beta value; `imputation` is designed to fill the NA. This function also provide genomewide methylation value's visualization(see figure 1,2,3,4,5 and 6).

```
> ##the directory of phenotype and 450k methylation's sample data
> Dir <- system.file("extdata/localdata",package="LncDM")
> setwd(Dir)
> ##name of phenotype file
> groupfile <- "BRCA_pheno.txt"

>loadData <- loaddata(fileDir="Level_2",is_beta=FALSE,beta_method="M/(M+U)",
groupfile=groupfile,samplefilter = TRUE,contin="OFF",samplefilterperc = 0.75,
XYchrom = c(FALSE, "X","Y"),sitefilter = TRUE, sitefilterperc = 0.75,
filterDecetP=0.05,normalization  = FALSE,transfm = FALSE,snpfilter=c(FALSE,"prob_snp"),
gcase="case",gcontrol="control",skip=2,imputation="knn",knn.k=10)
>save(loadData,file="loadData.Rdata")
```

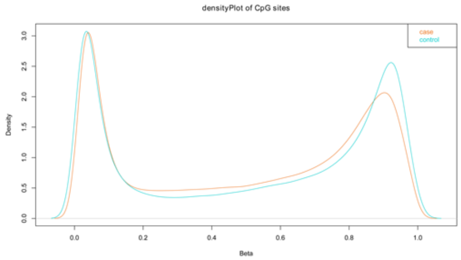

Figure 1: density plot of case vs control after preprocessing

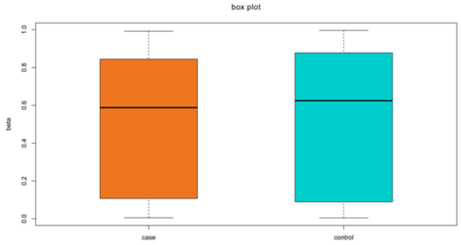

Figure 2: box plot of case vs control after preprocessing

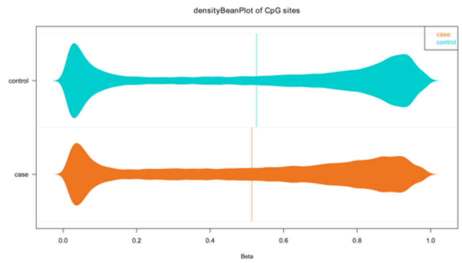

Figure 3: densityBean plot of case vs control after preprocessing

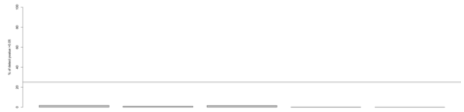

Figure 4: samples detect pvalue's distribution

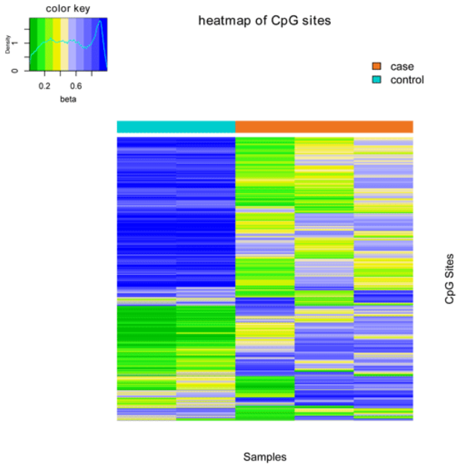

Figure 5: heat map of all CpG sites

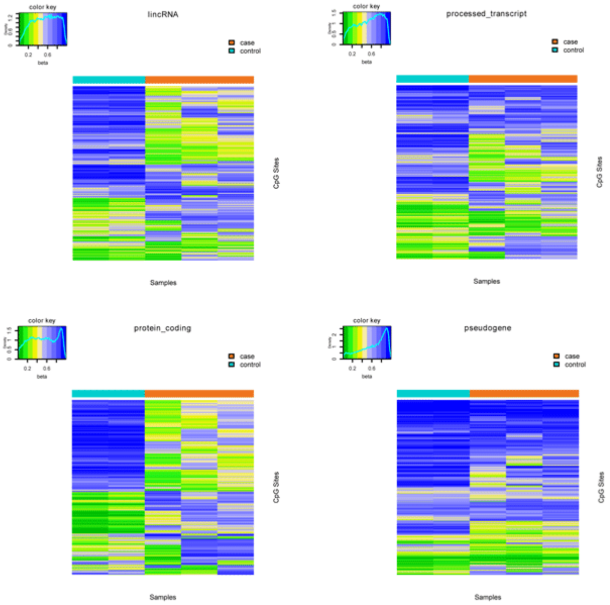

Figure 6: heat map of all kinds of transcript (lincRNA, processed transcript, protein coding, pseudogene) CpG sites

## 3 Methylation value of gene function element

The function `regionLevel` will return beta value of gene, lincRNA, processed-transcript and pseudogenes' elements. Elements include TSS1500, TSS200, intron, genebody, 1 exon, 5'UTR and 3'UTR ( non coding transcript do not have 5'UTR and 3'UTR ). If the element has only one CpG site, it's beta value is the element's methylation value; otherwise user can chose mean or median method to caculate methylation value. This section also provide transcript element's visualization that consist statistic for transcripts and CpG sites and beta value distribution ( see figure 7 and 8 ).

```
> dir <- system.file("extdata/localdata",package="LncDM")
> ##load the result of loaddata()
> load(paste(dir,"/loadData.Rdata",sep=""))
> Region <- regionLevel(data=loadData,indexmethod = "mean",classes="lincRNA")
```

```
There some information about sub_regions of  lincRNA :
intron  region contains: 6948 GENCODE  lincRNA  transcript
TSS1500  region contains: 3076 GENCODE  lincRNA  transcript
1_exon  region contains: 1572 GENCODE  lincRNA  transcript
TSS200  region contains: 1829 GENCODE  lincRNA  transcript
genebody  region contains: 1359 GENCODE  lincRNA  transcript

A RegionMethy450 class is created and the slotNames are:
 groupinfo annotation transAnno transBeta
```

## 4 Different methylation pattern

LncDM can identify different methylation site (dms), different methylation region (dmr) and different methylation element (dme) in specific disease. The dmr is the longest region which methylation state

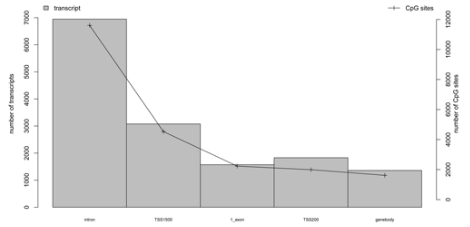

Figure 7: transctipt and CpG site statistic for each element

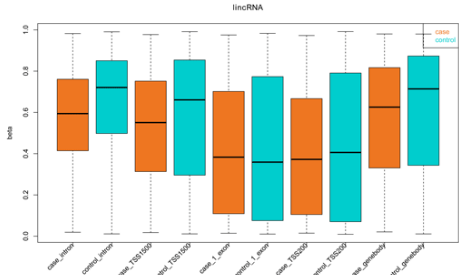

Figure 8: case vs control's beta value distribution for each element

is all up-regulation or down-regulation. For each specific site or region, wilcox, limma, satterthwaite or t test is used for uncontinuous phenotype, if the phenotype is continuous linear regression will be used.

## 4.1 different methylation site

The function `dms` can return both of dms and dms's beta value data. This section also provide visualization for dms ( see figure 9, 10, 11 ).

```
> dir.create(paste(Dir,"/dms",sep=""))
> setwd(paste(Dir,"/dms",sep=""))
> dms <- dms(data=loadData,contin="OFF",classes="lincRNA",testmethod = "t.test", Padj = "fdr",
+ gcase = "case", gcontrol = "control", paired = FALSE,rawpcut = 0.05, adjustpcut = 0.05,
+ betadiffcut = 0.3,XY=c(FALSE,"X","Y"),tlog=FALSE,num=1)
```

```
Performing t.test...
plot Manhattan
plot heatmap of DMS
plot boxplot of difference methylation
```

```
> head(dms)
```

```
                P-Value Adjust Pval beta-Difference  Mean_case Mean_control
cg18833705 9.583241e-05  0.03707674      -0.3223531 0.07104251    0.3933957
cg26603116 3.117195e-04  0.04316763      -0.3173426 0.61760625    0.9349489
cg22505202 1.440704e-04  0.03798099      -0.7134701 0.05774222    0.7712123
cg00556100 2.408574e-04  0.04295571       0.5970915 0.86313669    0.2660452
cg02063759 4.487971e-04  0.04586051       0.4691142 0.66099928    0.1918851
cg16714755 1.845997e-04  0.04128335       0.3998665 0.53258946    0.1327230
```

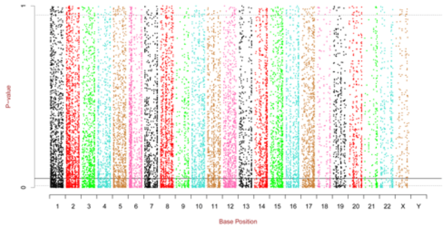

Figure 9: manhattan plot of genomewide different methylation analysis result

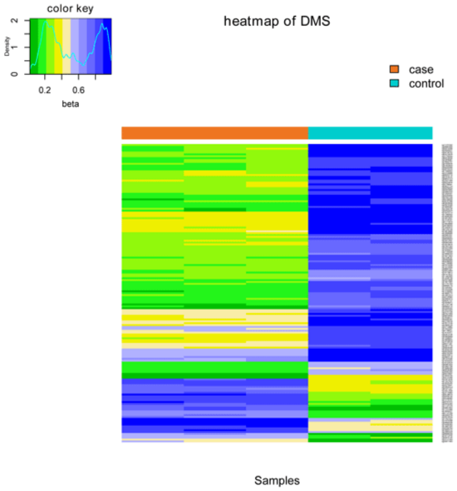

Figure 10: heat map of different methylation site

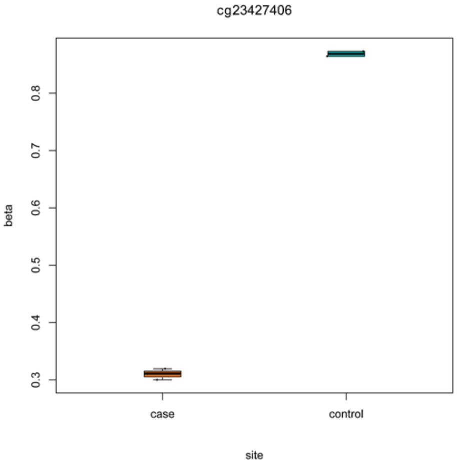

Figure 11: box plot of the most different methylation site

## 4.2 different methylation element

The function `dme` can return both of dme and dme's beta value data. This section also provide visualization for dme ( see figure 12, 13 ).

```
> dir.create(paste(Dir,"/dme",sep=""))
> setwd(paste(Dir,"/dme",sep=""))
> dme <- dme(data=Region,classes="lincRNA",contin="OFF",testmethod = "t.test", Padj = "fdr",
+ gcase = "case", gcontrol = "control", paired = FALSE,rawpcut = 0.05, adjustpcut = 0.05,
+ betadiffcut = 0.3,num=1)
```

```
calculating intron
Performing t.test...
calculating TSS1500
Performing t.test...
calculating 1_exon
Performing t.test...
calculating TSS200
Performing t.test...
calculating genebody
Performing t.test...
```

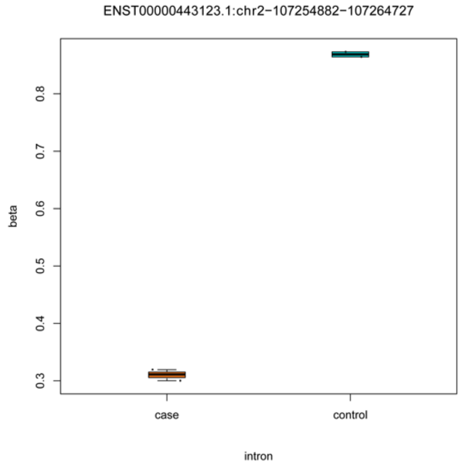

Figure 12: box plot of the most different methylation element

## 4.3 different methylation region

The function `dmr` can return both of dmr and dmr's beta value data. This section also provide visualization for dmr ( see figure 14, 15 ).

```
> dir.create(paste(Dir,"/dmr",sep=""))
> setwd(paste(Dir,"/dmr",sep=""))
> dmr <- dmr(data=loadData,contin="OFF",classes="lincRNA",testmethod = "t.test", Padj = "fdr",
+ gcase = "case", gcontrol = "control", paired = FALSE,rawpcut = 0.05, adjustpcut = 0.05,
+ betadiffcut = 0.3,num=1,sole=FALSE)
```

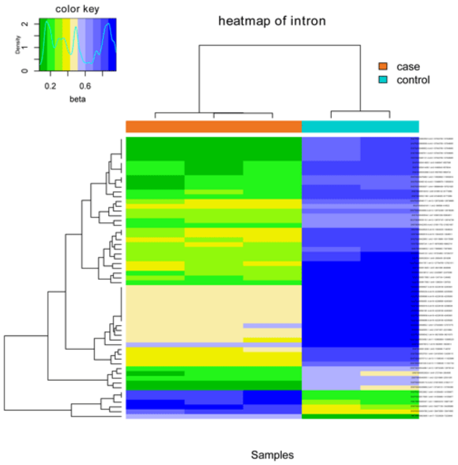

Figure 13: heat map of LincRNA intron region

```
Performing t.test...
Performing t.test...
plot boxplot of difference methylation
plot heatmap of DMR
```

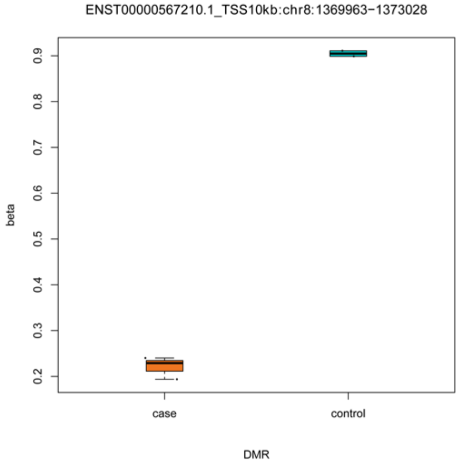

Figure 14: box plot of the most different methylation region

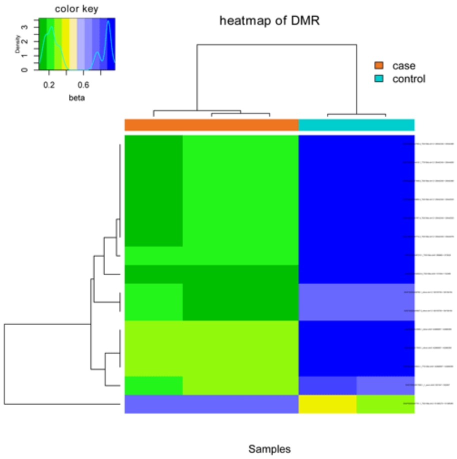

Figure 15: heat map of different methylation region

# 5 Session Info

The script runs within the following session:

```
R version 3.3.1 Patched (2016-07-20 r70946)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 8 (jessie)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] LncDM_1.0

loaded via a namespace (and not attached):
 [1] MASS_7.3-45          plyr_1.8.4           limma_3.28.16
 [4] gplots_3.0.1         tools_3.3.1          Rcpp_0.12.6
 [7] KernSmooth_2.23-15   reshape_0.8.5        preprocessCore_1.34.0
[10] gdata_2.17.0         impute_1.46.0        caTools_1.17.1
[13] bitops_1.0-6         WriteXLS_4.0.0       gtools_3.5.0
[16] beanplot_1.2
```