R_homework_Basic

dongxu

2019/4/10

```
## Q1 工作目录
getwd()
```

```
## [1] "C:/Users/Isaac/Desktop/Bioinformatics_selfstudy/项目实战/learning_R"
```

```
## Q2 新建六个向量 重点是字符串，数值，逻辑值
#character
chr <- c("Hello", "World","Bioinformatics")
#numeric
int <- c(1,2,3,4,5,6)
#布尔值
logic <- int <= 3
#浮点数
float <- int/5.5
#factor
factor <- factor(c("control", "treated"), levels = c("treated", "control"))
#tibble
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------- tidyverse 1.2.1 --
```

```
## √ ggplot2 3.1.0       √ purrr   0.3.2
## √ tibble  2.1.1       √ dplyr   0.8.0.1
## √ tidyr   0.8.3       √ stringr 1.4.0
## √ readr   1.3.1       √ forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
tb <- tibble(int, logic)
tb
```

```
## # A tibble: 6 x 2
##     int logic
##   <dbl> <lgl>
## 1     1 TRUE
## 2     2 TRUE
## 3     3 TRUE
## 4     4 FALSE
## 5     5 FALSE
## 6     6 FALSE
```

```
## Q3 getwd()返回值
##"C:/Users/Isaac/Desktop/Bioinformatics_selfstudy/项目实战/learning_R"
```

```
## Q4 新建数据结构（矩阵，数组，数据框，列表）
#matrix
mat <- matrix(c(1,2,3,4,5,6,7,8,9), ncol = 3, byrow = T)
mat
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]    7    8    9
```

```
#dataframe
name <- c("Steve", "Nash", "Curry", "Kid")
points <- c(20, 26, 29, 15)
assist <- c(13.5, 10.8, 6.5, 9)
df <- data.frame(name, points, assist)
df
```

```
##    name points assist
## 1 Steve     20   13.5
## 2  Nash     26   10.8
## 3 Curry     29    6.5
## 4   Kid     15    9.0
```

```
#array
arr <- array(data = int, dim = c(3,3))
arr
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    1
## [2,]    2    5    2
## [3,]    3    6    3
```

```
#list
list <- list(int = int, chr = chr, logic = logic)
list
```

```
## $int
## [1] 1 2 3 4 5 6
##
## $chr
## [1] "Hello"          "World"          "Bioinformatics"
##
## $logic
## [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE
```

```
## Q5 切片操作
# for df
df[1,]
```

```
##    name points assist
## 1 Steve     20   13.5
```

```
df[-1,]
```

```
##    name points assist
## 2  Nash     26   10.8
## 3 Curry     29    6.5
## 4   Kid     15    9.0
```

```
df[,c(1,2)]
```

```
##    name points
## 1 Steve     20
## 2  Nash     26
## 3 Curry     29
## 4   Kid     15
```

```
a <- df[,-c(2,3)]##要注意切片后可能数据类型会发生改变，因此可以使用subset()
class(a)
```

```
## [1] "factor"
```

```
a_1 <- subset(df, select = c("name"))
a_1
```

```
##    name
## 1 Steve
## 2  Nash
## 3 Curry
## 4   Kid
```

```
class(a_1)
```

```
## [1] "data.frame"
```

```
nash <- subset(df, name == "Nash")
nash
```

```
##   name points assist
## 2 Nash     26   10.8
```

```
# for matrix
mat[1:2,]
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
```

```
mat[,-3]
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    4    5
## [3,]    7    8
```

```
mat[1,3]
```

```
## [1] 3
```

```
# for array
arr[1,]
```

```
## [1] 1 4 1
```

```
# for list
list[[1]]
```

```
## [1] 1 2 3 4 5 6
```

```
list[["int"]]
```

```
## [1] 1 2 3 4 5 6
```

```
list[["int"]][2]
```

```
## [1] 2
```

```
list$chr
```

```
## [1] "Hello"         "World"         "Bioinformatics"
```

```
list$chr[2]
```

```
## [1] "World"
```

```
## Q6 用data()来加载内置数据集rivers
data("rivers")
head(rivers)
```

```
## [1] 735 320 325 392 524 450
```

```
tail(rivers)
```

```
## [1]  500  720  270  430  671 1770
```

```
max(rivers)
```

```
## [1] 3710
```

```
min(rivers)
```

```
## [1] 135
```

```
str(rivers)
```

```
##  num [1:141] 735 320 325 392 524 ...
```

```
class(rivers)
```

```
## [1] "numeric"
```

```
## Q7 读取RunInfo Table，了解数据框的基本信息
runinfo <- read.table("SraRunTable.txt", header = T, sep = "\t", quote = "", stringsAsFactors = F)
str(runinfo)
```

```
## 'data.frame':    768 obs. of  31 variables:
##  $ BioSample         : chr  "SAMN08619909" "SAMN08619908" "SAMN08619919" "SAMN08619918" ...
##  $ Experiment        : chr  "SRX3749905" "SRX3749906" "SRX3749907" "SRX3749908" ...
##  $ MBases            : int  8 11 7 18 5 11 15 14 14 14 ...
##  $ MBytes            : int  4 5 4 9 3 6 8 7 7 7 ...
##  $ Run               : chr  "SRR6790714" "SRR6790715" "SRR6790716" "SRR6790717" ...
##  $ SRA_Sample        : chr  "SRS3006136" "SRS3006149" "SRS3006140" "SRS3006150" ...
##  $ Sample_Name       : chr  "GSM3025848" "GSM3025849" "GSM3025850" "GSM3025851" ...
##  $ Assay_Type        : chr  "RNA-Seq" "RNA-Seq" "RNA-Seq" "RNA-Seq" ...
##  $ AssemblyName      : chr  "GCF_000001635.20" "GCF_000001635.20" "GCF_000001635.20" "GCF_00000163
5.20" ...
##  $ AvgSpotLen        : int  43 43 43 43 43 43 43 43 43 43 ...
##  $ BioProject        : chr  "PRJNA436229" "PRJNA436229" "PRJNA436229" "PRJNA436229" ...
##  $ Center_Name       : chr  "GEO" "GEO" "GEO" "GEO" ...
##  $ Consent           : chr  "public" "public" "public" "public" ...
##  $ DATASTORE_filetype: chr  "sra" "sra" "sra" "sra" ...
##  $ DATASTORE_provider: chr  "ncbi" "ncbi" "ncbi" "ncbi" ...
##  $ InsertSize        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Instrument        : chr  "Illumina HiSeq 2000" "Illumina HiSeq 2000" "Illumina HiSeq 2000" "Illu
mina HiSeq 2000" ...
##  $ LibraryLayout     : chr  "SINGLE" "SINGLE" "SINGLE" "SINGLE" ...
##  $ LibrarySelection  : chr  "cDNA" "cDNA" "cDNA" "cDNA" ...
##  $ LibrarySource     : chr  "TRANSCRIPTOMIC" "TRANSCRIPTOMIC" "TRANSCRIPTOMIC" "TRANSCRIPTOMIC" ...
##  $ LoadDate          : chr  "2018-03-01" "2018-03-01" "2018-03-01" "2018-03-01" ...
##  $ Organism          : chr  "Mus musculus" "Mus musculus" "Mus musculus" "Mus musculus" ...
##  $ Platform          : chr  "ILLUMINA" "ILLUMINA" "ILLUMINA" "ILLUMINA" ...
##  $ ReleaseDate       : chr  "2018-11-23" "2018-11-23" "2018-11-23" "2018-11-23" ...
##  $ SRA_Study         : chr  "SRP133642" "SRP133642" "SRP133642" "SRP133642" ...
##  $ age               : chr  "14 weeks" "14 weeks" "14 weeks" "14 weeks" ...
##  $ cell_type         : chr  "cancer-associated fibroblasts (CAFs)" "cancer-associated fibroblasts
(CAFs)" "cancer-associated fibroblasts (CAFs)" "cancer-associated fibroblasts (CAFs)" ...
##  $ marker_genes      : chr  "EpCAM-, CD45-, CD31-, NG2-" "EpCAM-, CD45-, CD31-, NG2-" "EpCAM-, CD45
-, CD31-, NG2-" "EpCAM-, CD45-, CD31-, NG2-" ...
##  $ source_name       : chr  "Mammary tumor fibroblast" "Mammary tumor fibroblast" "Mammary tumor fi
broblast" "Mammary tumor fibroblast" ...
##  $ strain            : chr  "FVB/N-Tg(MMTVPyVT)634Mul/J" "FVB/N-Tg(MMTVPyVT)634Mul/J" "FVB/N-Tg(MMT
VPyVT)634Mul/J" "FVB/N-Tg(MMTVPyVT)634Mul/J" ...
##  $ tissue            : chr  "Mammary tumor fibroblast" "Mammary tumor fibroblast" "Mammary tumor fi
broblast" "Mammary tumor fibroblast" ...
```

```
dim(runinfo)
```

```
## [1] 768  31
```

```
head(runinfo)
```

```
##      BioSample Experiment MBases MBytes        Run SRA_Sample Sample_Name
## 1 SAMN08619909 SRX3749905      8      4 SRR6790714 SRS3006136  GSM3025848
## 2 SAMN08619908 SRX3749906     11      5 SRR6790715 SRS3006149  GSM3025849
## 3 SAMN08619919 SRX3749907      7      4 SRR6790716 SRS3006140  GSM3025850
## 4 SAMN08619918 SRX3749908     18      9 SRR6790717 SRS3006150  GSM3025851
## 5 SAMN08619921 SRX3749909      5      3 SRR6790718 SRS3006142  GSM3025852
## 6 SAMN08619920 SRX3749910     11      6 SRR6790719 SRS3006141  GSM3025853
##   Assay_Type     AssemblyName AvgSpotLen  BioProject Center_Name Consent
## 1    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 2    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 3    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 4    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 5    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 6    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
##   DATASTORE_filetype DATASTORE_provider InsertSize          Instrument
## 1                sra               ncbi          0 Illumina HiSeq 2000
## 2                sra               ncbi          0 Illumina HiSeq 2000
## 3                sra               ncbi          0 Illumina HiSeq 2000
## 4                sra               ncbi          0 Illumina HiSeq 2000
## 5                sra               ncbi          0 Illumina HiSeq 2000
## 6                sra               ncbi          0 Illumina HiSeq 2000
##   LibraryLayout LibrarySelection  LlibrarySource   LoadDate     Organism
## 1        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 2        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 3        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 4        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 5        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 6        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
##   Platform ReleaseDate SRA_Study      age
## 1 ILLUMINA  2018-11-23 SRP133642 14 weeks
## 2 ILLUMINA  2018-11-23 SRP133642 14 weeks
## 3 ILLUMINA  2018-11-23 SRP133642 14 weeks
## 4 ILLUMINA  2018-11-23 SRP133642 14 weeks
## 5 ILLUMINA  2018-11-23 SRP133642 14 weeks
## 6 ILLUMINA  2018-11-23 SRP133642 14 weeks
##                             cell_type               marker_genes
## 1 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 2 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 3 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 4 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 5 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 6 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
##                source_name                    strain
## 1 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 2 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 3 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 4 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 5 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 6 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
##                    tissue
## 1 Mammary tumor fibroblast
## 2 Mammary tumor fibroblast
## 3 Mammary tumor fibroblast
## 4 Mammary tumor fibroblast
## 5 Mammary tumor fibroblast
## 6 Mammary tumor fibroblast
```

```
colnames(runinfo)
```

```
##  [1] "BioSample"          "Experiment"         "MBases"
##  [4] "MBytes"             "Run"                "SRA_Sample"
##  [7] "Sample_Name"        "Assay_Type"         "AssemblyName"
## [10] "AvgSpotLen"         "BioProject"         "Center_Name"
## [13] "Consent"            "DATASTORE_filetype" "DATASTORE_provider"
## [16] "InsertSize"         "Instrument"         "LibraryLayout"
## [19] "LibrarySelection"   "LibrarySource"      "LoadDate"
## [22] "Organism"           "Platform"           "ReleaseDate"
## [25] "SRA_Study"          "age"                "cell_type"
## [28] "marker_genes"       "source_name"        "strain"
## [31] "tissue"
```

```
class(runinfo$BioSample)
```

```
## [1] "character"
```

```
attri <- list()
findattri <- function(df){
  for ( i in 1:dim(df)[2]){
    names <- colnames(df)
    attri[names[i]] <- class(df[,i])
  }
  return(attri)
}
findattri(runinfo)
```

```
## $BioSample
## [1] "character"
##
## $Experiment
## [1] "character"
##
## $MBases
## [1] "integer"
##
## $MBytes
## [1] "integer"
##
## $Run
## [1] "character"
##
## $SRA_Sample
## [1] "character"
##
## $Sample_Name
## [1] "character"
##
## $Assay_Type
## [1] "character"
##
## $AssemblyName
## [1] "character"
##
## $AvgSpotLen
## [1] "integer"
##
## $BioProject
## [1] "character"
##
## $Center_Name
## [1] "character"
##
## $Consent
## [1] "character"
##
## $DATASTORE_filetype
## [1] "character"
##
## $DATASTORE_provider
## [1] "character"
##
## $InsertSize
## [1] "integer"
##
## $Instrument
## [1] "character"
##
## $LibraryLayout
## [1] "character"
##
## $LibrarySelection
## [1] "character"
##
```

```
## $LibrarySource
## [1] "character"
##
## $LoadDate
## [1] "character"
##
## $Organism
## [1] "character"
##
## $Platform
## [1] "character"
##
## $ReleaseDate
## [1] "character"
##
## $SRA_Study
## [1] "character"
##
## $age
## [1] "character"
##
## $cell_type
## [1] "character"
##
## $marker_genes
## [1] "character"
##
## $source_name
## [1] "character"
##
## $strain
## [1] "character"
##
## $tissue
## [1] "character"
```

```
## Q8 下载样品信息文件，读取数据，得到属性
sample <- read.csv("sample.csv", header = T, sep = ",", stringsAsFactors = F)
head(sample)
```

```
##    Accession          Title Sample.Type     Taxonomy Channels Platform
## 1 GSM3025845 SS2_15_0048_A1         SRA Mus musculus        1 GPL13112
## 2 GSM3025846 SS2_15_0048_A2         SRA Mus musculus        1 GPL13112
## 3 GSM3025847 SS2_15_0048_A3         SRA Mus musculus        1 GPL13112
## 4 GSM3025848 SS2_15_0048_A4         SRA Mus musculus        1 GPL13112
## 5 GSM3025849 SS2_15_0048_A5         SRA Mus musculus        1 GPL13112
## 6 GSM3025850 SS2_15_0048_A6         SRA Mus musculus        1 GPL13112
##      Series Supplementary.Types
## 1 GSE111229    SRA Run Selector
## 2 GSE111229    SRA Run Selector
## 3 GSE111229    SRA Run Selector
## 4 GSE111229    SRA Run Selector
## 5 GSE111229    SRA Run Selector
## 6 GSE111229    SRA Run Selector
##                                      Supplementary.Links SRA.Accession
## 1 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749902    SRX3749902
## 2 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749903    SRX3749903
## 3 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749904    SRX3749904
## 4 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749905    SRX3749905
## 5 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749906    SRX3749906
## 6 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749907    SRX3749907
##           Contact Release.Date
## 1 Kristian Pietras Nov 23, 2018
## 2 Kristian Pietras Nov 23, 2018
## 3 Kristian Pietras Nov 23, 2018
## 4 Kristian Pietras Nov 23, 2018
## 5 Kristian Pietras Nov 23, 2018
## 6 Kristian Pietras Nov 23, 2018
```

```
findattri(sample)
```

```
## $Accession
## [1] "character"
##
## $Title
## [1] "character"
##
## $Sample.Type
## [1] "character"
##
## $Taxonomy
## [1] "character"
##
## $Channels
## [1] "integer"
##
## $Platform
## [1] "character"
##
## $Series
## [1] "character"
##
## $Supplementary.Types
## [1] "character"
##
## $Supplementary.Links
## [1] "character"
##
## $SRA.Accession
## [1] "character"
##
## $Contact
## [1] "character"
##
## $Release.Date
## [1] "character"
```

```
## Q9 把runinfo与sample两个表格通过merge()函数进行关联
head(runinfo)
```

```
##      BioSample Experiment MBases MBytes        Run SRA_Sample Sample_Name
## 1 SAMN08619909 SRX3749905      8      4 SRR6790714 SRS3006136  GSM3025848
## 2 SAMN08619908 SRX3749906     11      5 SRR6790715 SRS3006149  GSM3025849
## 3 SAMN08619919 SRX3749907      7      4 SRR6790716 SRS3006140  GSM3025850
## 4 SAMN08619918 SRX3749908     18      9 SRR6790717 SRS3006150  GSM3025851
## 5 SAMN08619921 SRX3749909      5      3 SRR6790718 SRS3006142  GSM3025852
## 6 SAMN08619920 SRX3749910     11      6 SRR6790719 SRS3006141  GSM3025853
##   Assay_Type     AssemblyName AvgSpotLen  BioProject Center_Name Consent
## 1    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 2    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 3    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 4    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 5    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 6    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
##   DATASTORE_filetype DATASTORE_provider InsertSize          Instrument
## 1                sra               ncbi          0 Illumina HiSeq 2000
## 2                sra               ncbi          0 Illumina HiSeq 2000
## 3                sra               ncbi          0 Illumina HiSeq 2000
## 4                sra               ncbi          0 Illumina HiSeq 2000
## 5                sra               ncbi          0 Illumina HiSeq 2000
## 6                sra               ncbi          0 Illumina HiSeq 2000
##   LibraryLayout LibrarySelection  LibrarySource   LoadDate     Organism
## 1        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 2        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 3        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 4        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 5        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 6        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
##   Platform ReleaseDate SRA_Study      age
## 1 ILLUMINA  2018-11-23 SRP133642 14 weeks
## 2 ILLUMINA  2018-11-23 SRP133642 14 weeks
## 3 ILLUMINA  2018-11-23 SRP133642 14 weeks
## 4 ILLUMINA  2018-11-23 SRP133642 14 weeks
## 5 ILLUMINA  2018-11-23 SRP133642 14 weeks
## 6 ILLUMINA  2018-11-23 SRP133642 14 weeks
##                             cell_type               marker_genes
## 1 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 2 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 3 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 4 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 5 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 6 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
##                source_name                   strain
## 1 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 2 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 3 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 4 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 5 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 6 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
##                    tissue
## 1 Mammary tumor fibroblast
## 2 Mammary tumor fibroblast
## 3 Mammary tumor fibroblast
## 4 Mammary tumor fibroblast
## 5 Mammary tumor fibroblast
## 6 Mammary tumor fibroblast
```

```
head(sample)
```

```
##    Accession          Title Sample.Type     Taxonomy Channels Platform
## 1 GSM3025845 SS2_15_0048_A1         SRA Mus musculus        1 GPL13112
## 2 GSM3025846 SS2_15_0048_A2         SRA Mus musculus        1 GPL13112
## 3 GSM3025847 SS2_15_0048_A3         SRA Mus musculus        1 GPL13112
## 4 GSM3025848 SS2_15_0048_A4         SRA Mus musculus        1 GPL13112
## 5 GSM3025849 SS2_15_0048_A5         SRA Mus musculus        1 GPL13112
## 6 GSM3025850 SS2_15_0048_A6         SRA Mus musculus        1 GPL13112
##      Series Supplementary.Types
## 1 GSE111229    SRA Run Selector
## 2 GSE111229    SRA Run Selector
## 3 GSE111229    SRA Run Selector
## 4 GSE111229    SRA Run Selector
## 5 GSE111229    SRA Run Selector
## 6 GSE111229    SRA Run Selector
##                                      Supplementary.Links SRA.Accession
## 1 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749902    SRX3749902
## 2 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749903    SRX3749903
## 3 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749904    SRX3749904
## 4 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749905    SRX3749905
## 5 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749906    SRX3749906
## 6 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749907    SRX3749907
##            Contact Release.Date
## 1 Kristian Pietras Nov 23, 2018
## 2 Kristian Pietras Nov 23, 2018
## 3 Kristian Pietras Nov 23, 2018
## 4 Kristian Pietras Nov 23, 2018
## 5 Kristian Pietras Nov 23, 2018
## 6 Kristian Pietras Nov 23, 2018
```

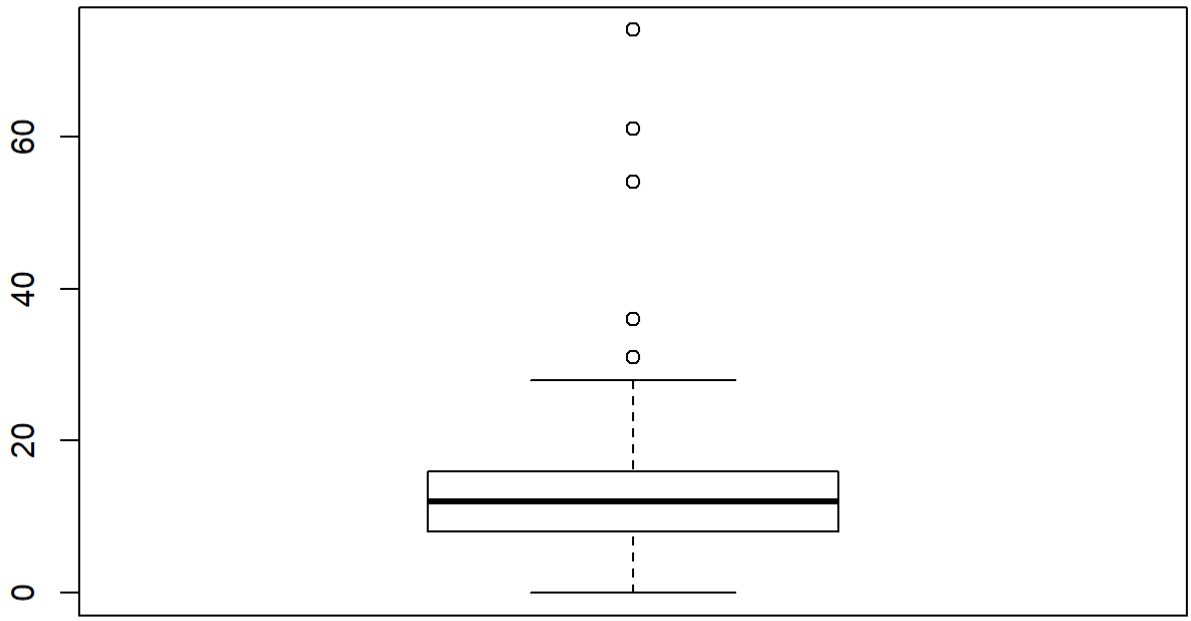
```
sample_info <- merge(runinfo, sample, by.x = "Sample_Name", by.y = "Accession")
head(sample_info)
```

```
##   Sample_Name    BioSample Experiment MBases MBytes        Run SRA_Sample
## 1  GSM3025845 SAMN08619912 SRX3749902     16      8 SRR6790711 SRS3006138
## 2  GSM3025846 SAMN08619911 SRX3749903     16      8 SRR6790712 SRS3006148
## 3  GSM3025847 SAMN08619910 SRX3749904      8      4 SRR6790713 SRS3006137
## 4  GSM3025848 SAMN08619909 SRX3749905      8      4 SRR6790714 SRS3006136
## 5  GSM3025849 SAMN08619908 SRX3749906     11      5 SRR6790715 SRS3006149
## 6  GSM3025850 SAMN08619919 SRX3749907      7      4 SRR6790716 SRS3006140
##   Assay_Type     AssemblyName AvgSpotLen  BioProject Center_Name Consent
## 1    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 2    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 3    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 4    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 5    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
## 6    RNA-Seq GCF_000001635.20         43 PRJNA436229         GEO  public
##   DATASTORE_filetype DATASTORE_provider InsertSize          Instrument
## 1                sra               ncbi          0 Illumina HiSeq 2000
## 2                sra               ncbi          0 Illumina HiSeq 2000
## 3                sra               ncbi          0 Illumina HiSeq 2000
## 4                sra               ncbi          0 Illumina HiSeq 2000
## 5                sra               ncbi          0 Illumina HiSeq 2000
## 6                sra               ncbi          0 Illumina HiSeq 2000
##   LibraryLayout LibrarySelection  LibrarySource   LoadDate     Organism
## 1        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 2        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 3        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 4        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 5        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
## 6        SINGLE             cDNA TRANSCRIPTOMIC 2018-03-01 Mus musculus
##   Platform.x ReleaseDate SRA_Study      age
## 1   ILLUMINA  2018-11-23 SRP133642 14 weeks
## 2   ILLUMINA  2018-11-23 SRP133642 14 weeks
## 3   ILLUMINA  2018-11-23 SRP133642 14 weeks
## 4   ILLUMINA  2018-11-23 SRP133642 14 weeks
## 5   ILLUMINA  2018-11-23 SRP133642 14 weeks
## 6   ILLUMINA  2018-11-23 SRP133642 14 weeks
##                            cell_type               marker_genes
## 1 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 2 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 3 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 4 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 5 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
## 6 cancer-associated fibroblasts (CAFs) EpCAM-, CD45-, CD31-, NG2-
##                source_name                   strain
## 1 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 2 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 3 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 4 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 5 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
## 6 Mammary tumor fibroblast FVB/N-Tg(MMTVPyVT)634Mul/J
##                    tissue          Title Sample.Type     Taxonomy
## 1 Mammary tumor fibroblast SS2_15_0048_A1         SRA Mus musculus
## 2 Mammary tumor fibroblast SS2_15_0048_A2         SRA Mus musculus
## 3 Mammary tumor fibroblast SS2_15_0048_A3         SRA Mus musculus
## 4 Mammary tumor fibroblast SS2_15_0048_A4         SRA Mus musculus
## 5 Mammary tumor fibroblast SS2_15_0048_A5         SRA Mus musculus
## 6 Mammary tumor fibroblast SS2_15_0048_A6         SRA Mus musculus
##   Channels Platform.y    Series Supplementary.Types
```

```
## 1        1   GPL13112 GSE111229    SRA Run Selector
## 2        1   GPL13112 GSE111229    SRA Run Selector
## 3        1   GPL13112 GSE111229    SRA Run Selector
## 4        1   GPL13112 GSE111229    SRA Run Selector
## 5        1   GPL13112 GSE111229    SRA Run Selector
## 6        1   GPL13112 GSE111229    SRA Run Selector
##                                       Supplementary.Links SRA.Accession
## 1 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749902    SRX3749902
## 2 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749903    SRX3749903
## 3 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749904    SRX3749904
## 4 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749905    SRX3749905
## 5 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749906    SRX3749906
## 6 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX3749907    SRX3749907
##            Contact Release.Date
## 1 Kristian Pietras Nov 23, 2018
## 2 Kristian Pietras Nov 23, 2018
## 3 Kristian Pietras Nov 23, 2018
## 4 Kristian Pietras Nov 23, 2018
## 5 Kristian Pietras Nov 23, 2018
## 6 Kristian Pietras Nov 23, 2018
```

基于下午的统计可视化

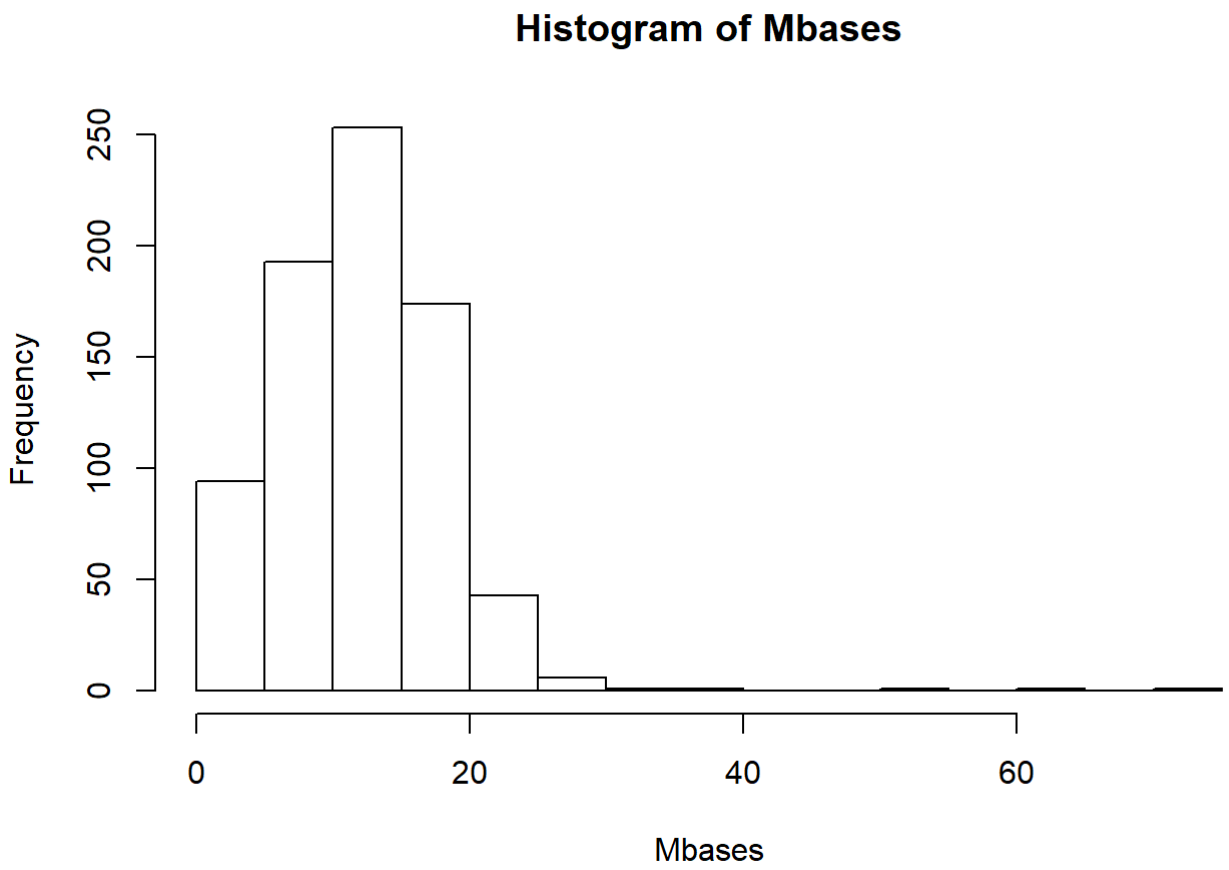
```
## Q1 Mbases的箱线图，五分位数，频数图，密度图
boxplot(runinfo$MBases)
```



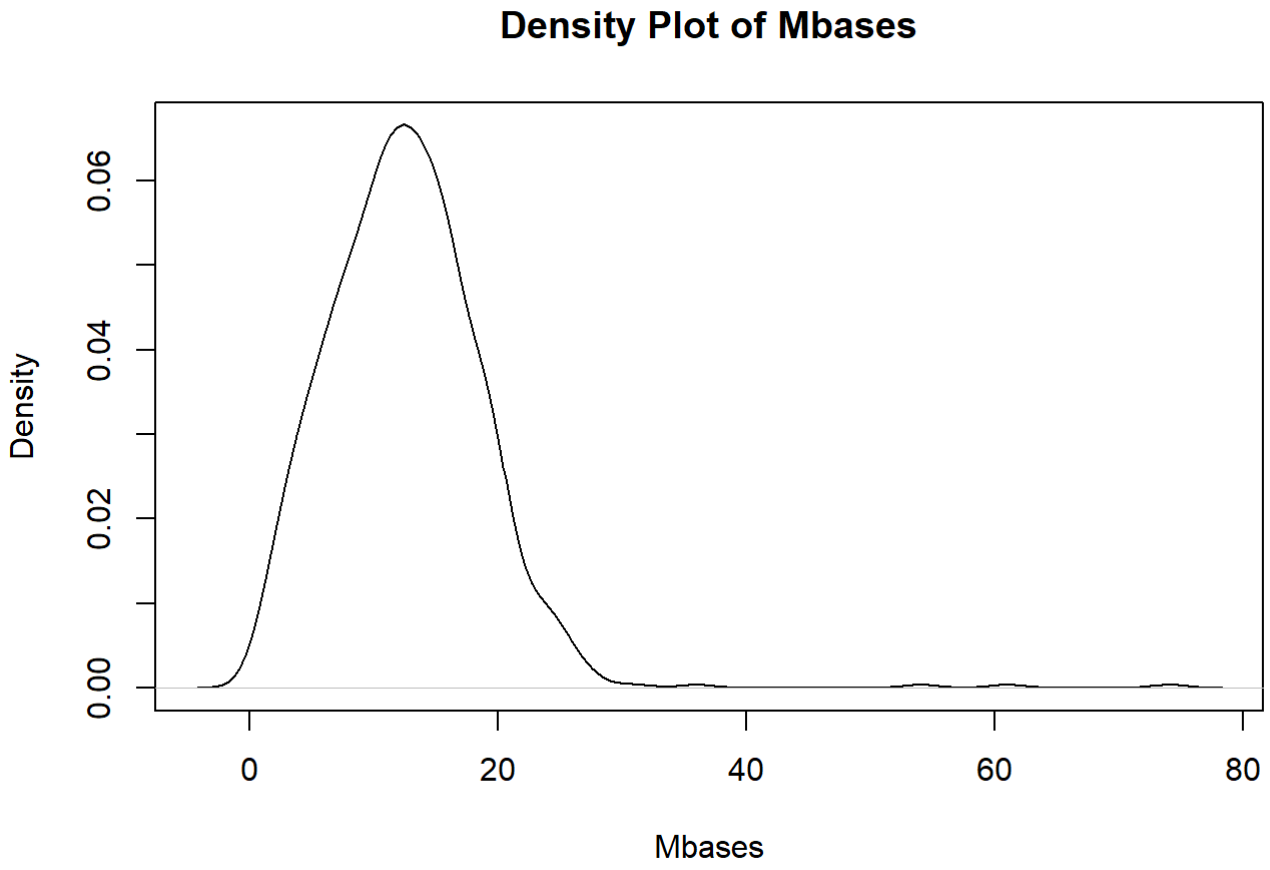

```
fivenum(runinfo$MBases)
```

```
## [1]  0  8 12 16 74
```

```
hist(runinfo$MBases, main = "Histogram of Mbases", xlab = "Mbases")
```



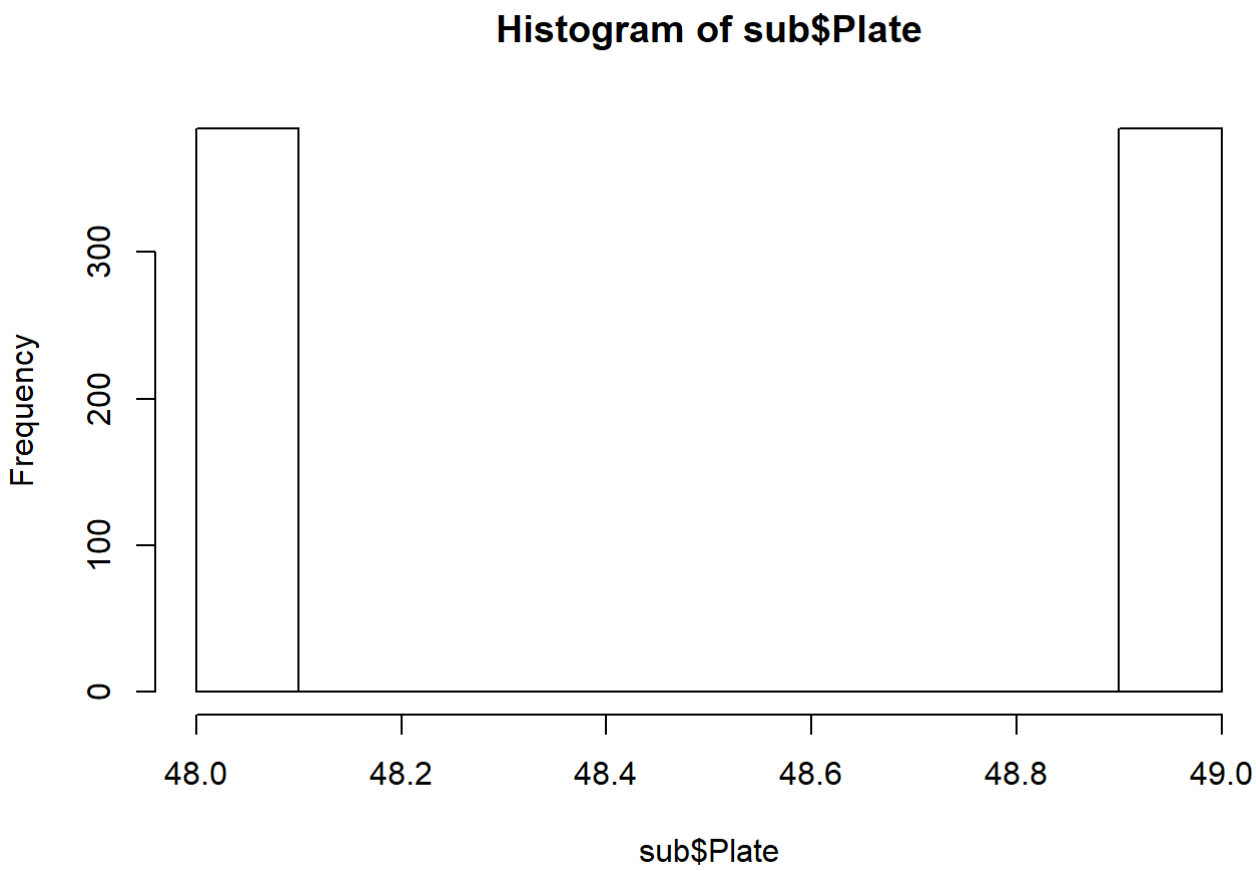

```
plot(density(runinfo$MBases), main = "Density Plot of Mbases", xlab = "Mbases")
```




```
## Q2 sample列表中根据下划线分割看第三列元素的统计情况
sub <- subset(sample_info, select = c("Title", "MBases"))
stringsplit <- function(string){
  strsplit(string, "_")[[1]][3]
}
sub$Plate <- unlist(lapply(sub$Title, stringsplit))
sub$Plate <- as.numeric(sub$Plate)
class(sub$Plate)
```

```
## [1] "numeric"
```

```
hist(sub$Plate)
```

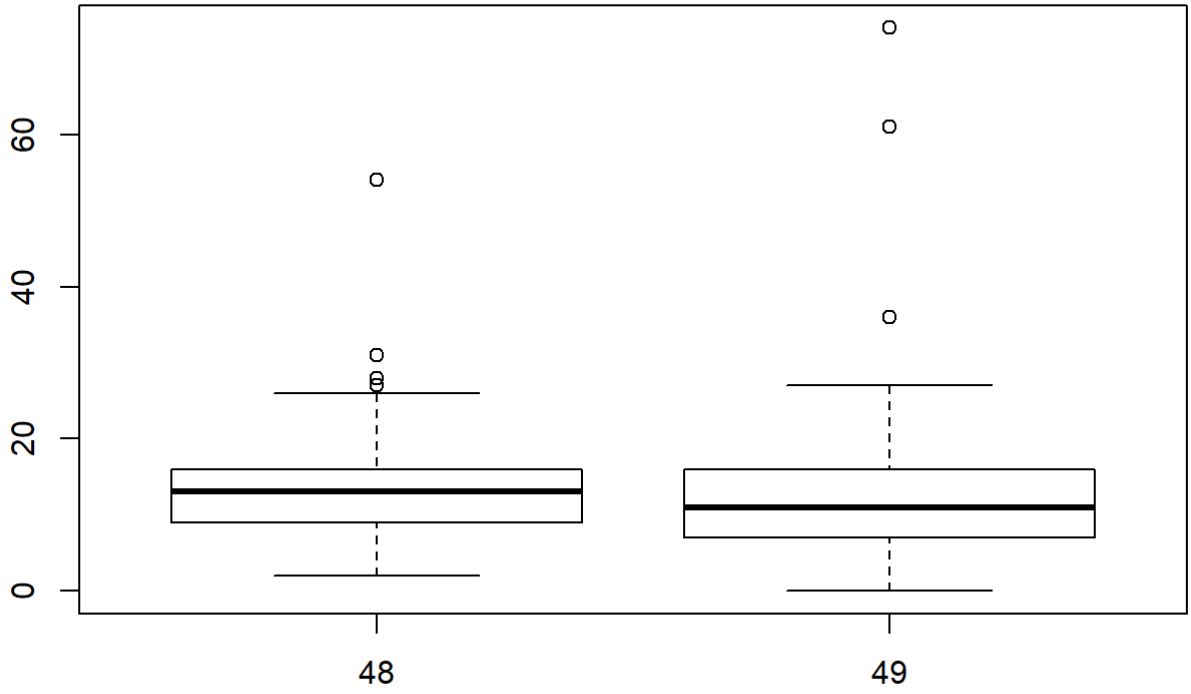



```
## Q3 plate与MBases进行关联，检验是否有统计学差异
t.test(sub$MBases ~ sub$Plate)
```

```
##
##  Welch Two Sample t-test
##
## data:  sub$MBases by sub$Plate
## t = 2.3019, df = 728.18, p-value = 0.02162
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1574805 1.9831445
## sample estimates:
## mean in group 48 mean in group 49
##         13.08854         12.01823
```

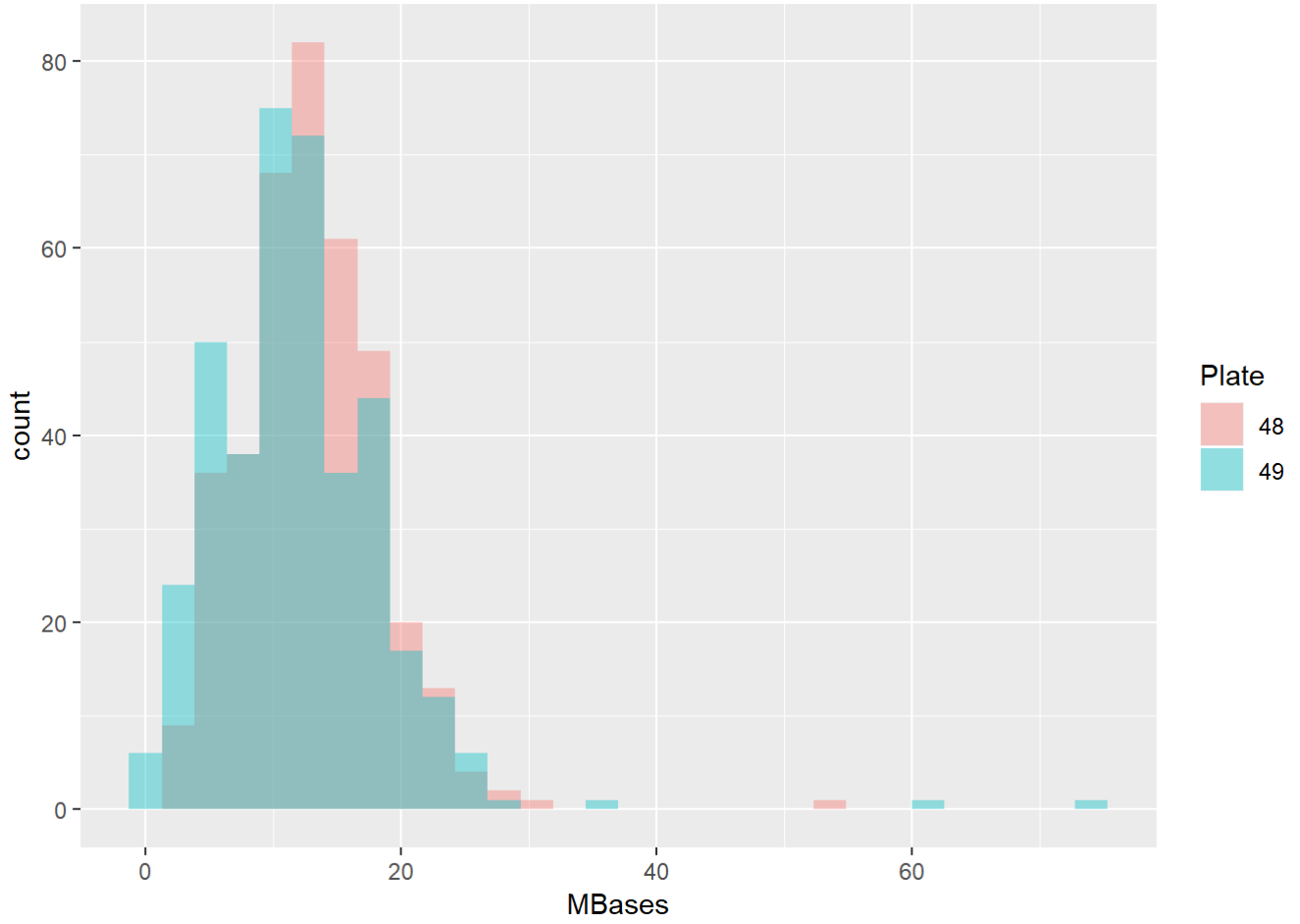
```
## p-value = 0.02162 < 0.05
```

```
## Q4-5 分组绘制箱线图，频数图，密度图
library(ggplot2)
boxplot(sub$MBases~sub$Plate)
```

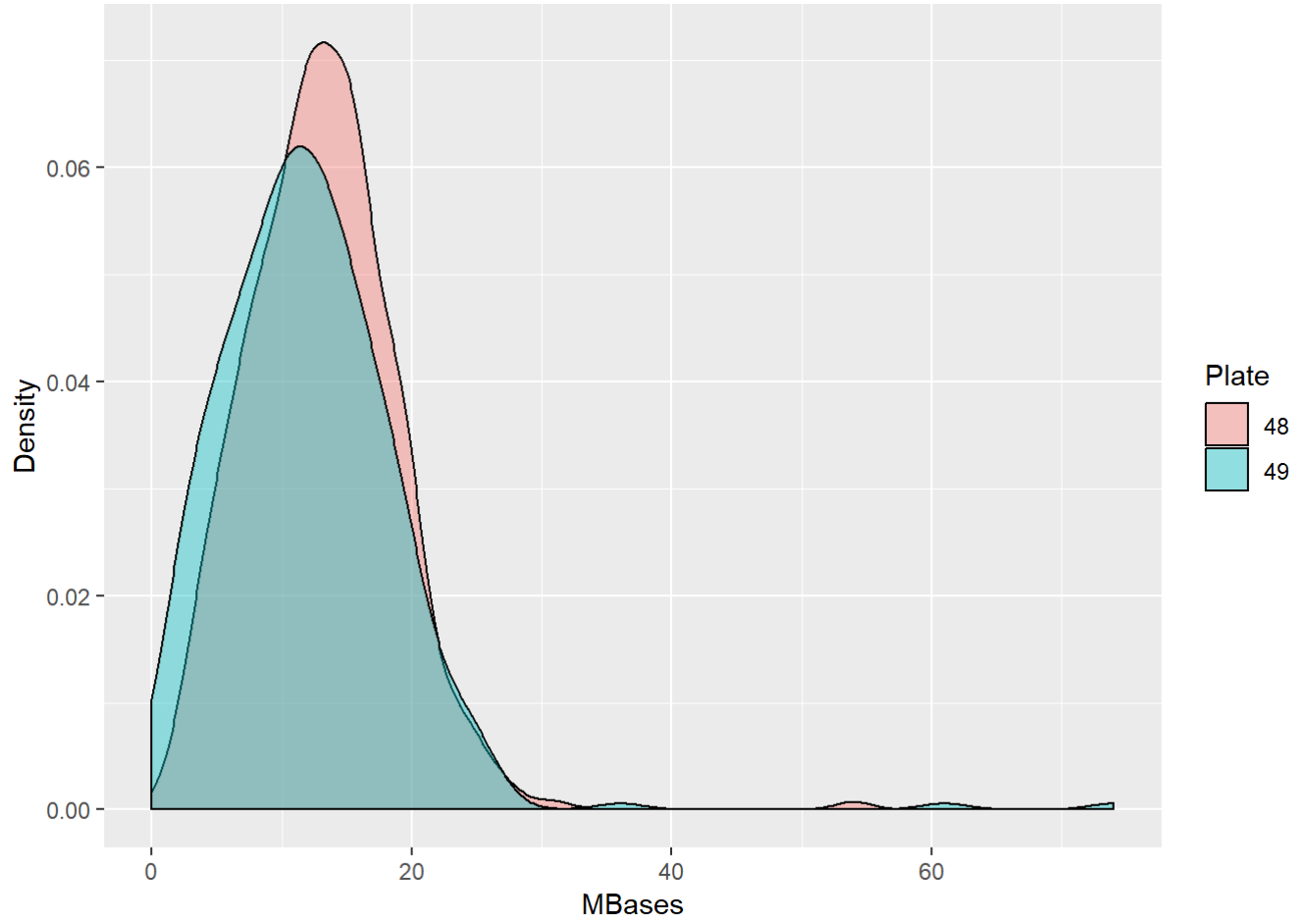



```
sub$Plate <- factor(sub$Plate)
ggplot(sub, aes(x = MBases, fill = Plate))+
  geom_histogram(position = "identity", alpha = .4)
```

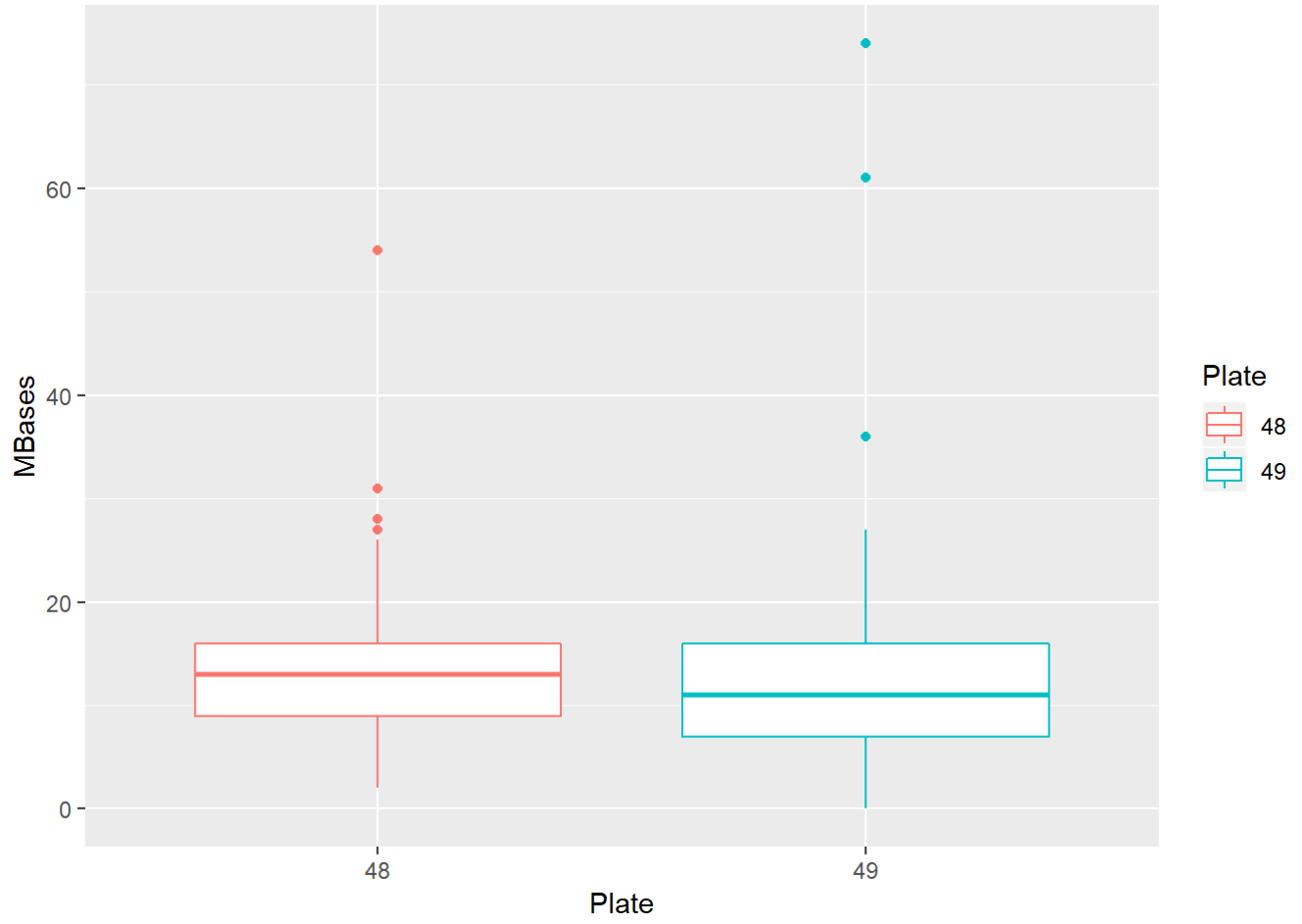
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(sub, aes(x = MBases, fill = Plate))+
  geom_density(position = "identity", alpha = .4)+
  ylab("Density")
```




```
ggplot(sub, aes(x = Plate, y = MBases, color = Plate))+
  geom_boxplot()
```



```
## Q6 使用ggpubr绘制上面三个图
library(ggpubr)
```

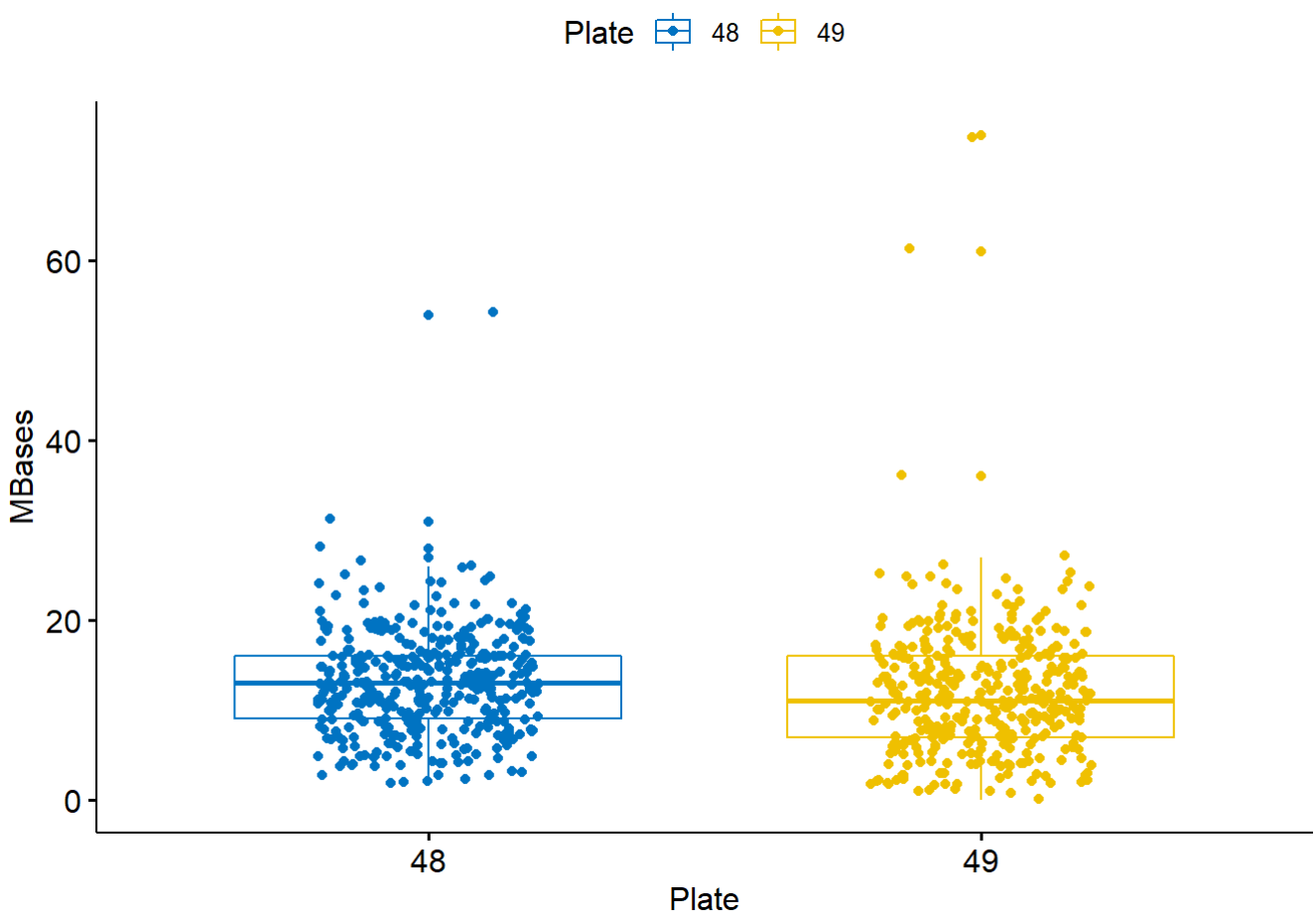
```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

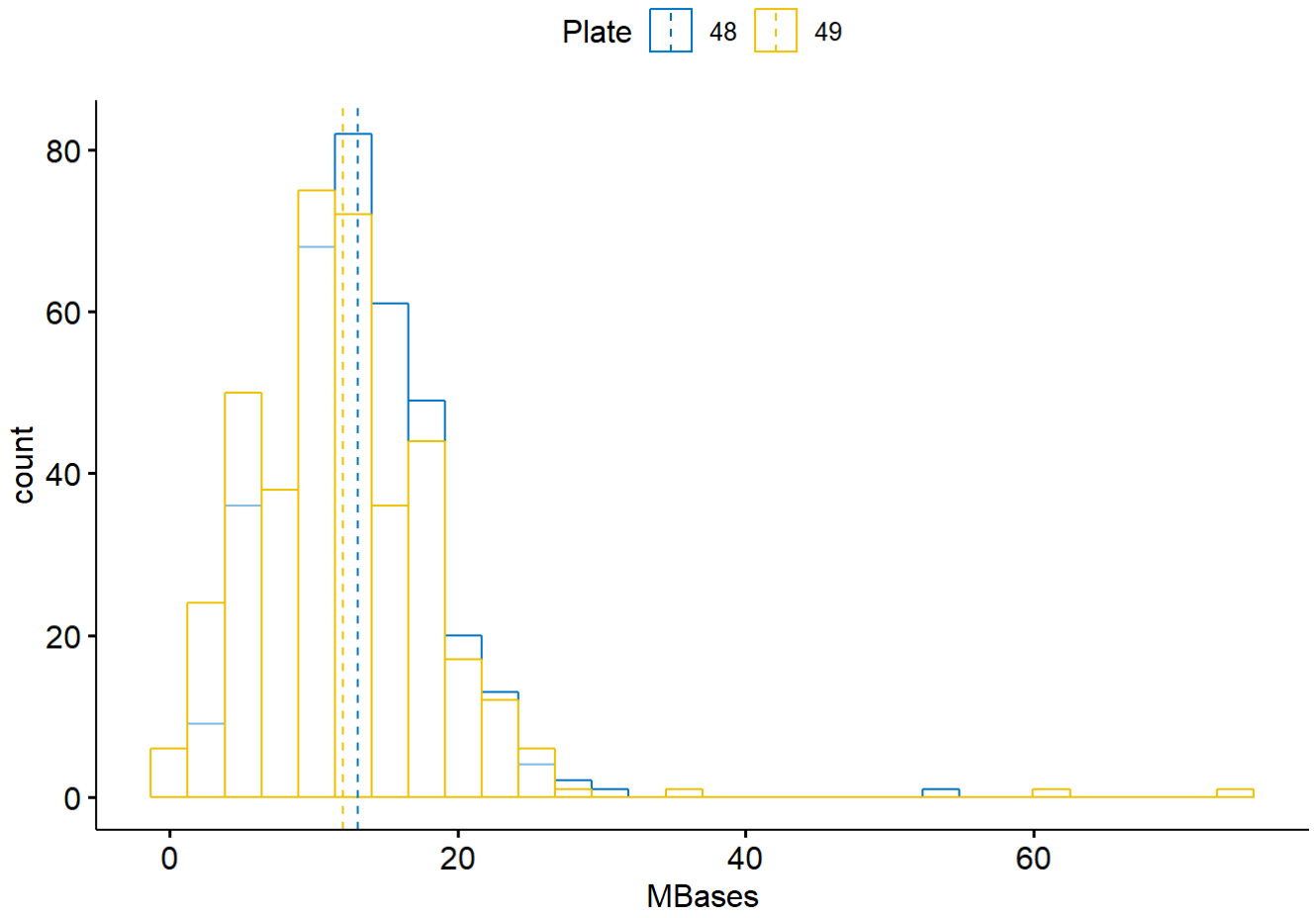
```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

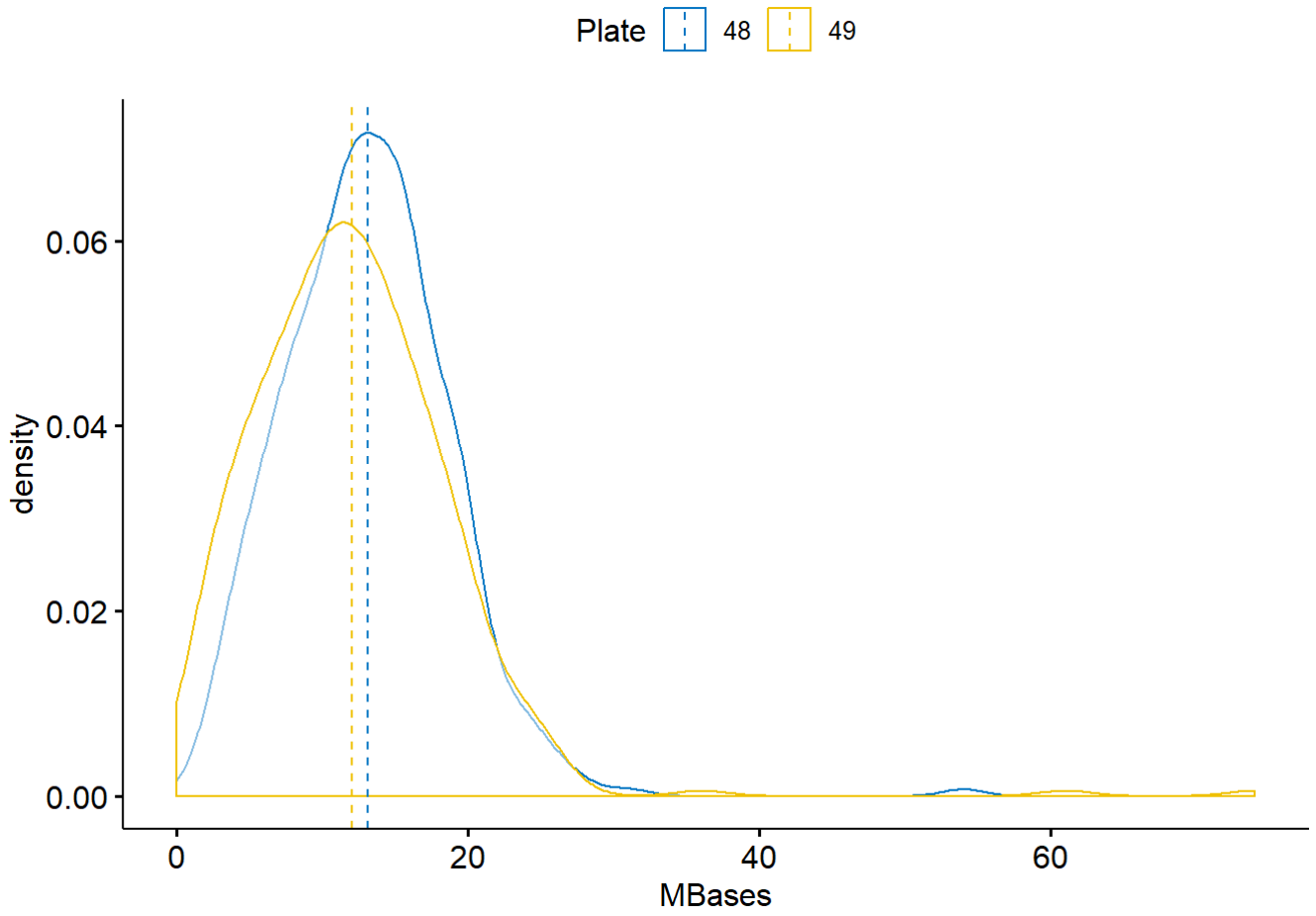
```
ggboxplot(sub, x = "Plate", y = "MBases", color = "Plate", palette = "jco", add = "jitter")
```




```
gghistogram(sub, x = "MBases", color = "Plate", palette = "jco", add = "mean")
```




```
ggdensity(sub, x= "MBases", color = "Plate", palette = "jco", add = "mean")
```



```
## Q7 随机取384个MBases的信息
sub_1 <- sample(nrow(sub), 384)
sub_2 <- sub[sub_1,][,c(3,2,1)]
dim(sub_2)
```

```
## [1] 384   3
```