Epigenomics for Social Scientists

01 Installing and loading packages, reading in datasets

Kelly Bakulski, Shan Andrews, John Dou, Jonah Fisher, Erin Ware

Last compiled on July 12, 2021

Setup

Install required packages (Do not run the following code chunk if on rstudio cloud account for EGESS)

There are many useful packages for DNA methylation pre-processing and analysis. The following packages are largely downloaded from **Bioconductor** which holds myriad useful packages for bioinformatics; the remaining packages come from CRAN which is a standard repository for general R packages. Note that your rsudio cloud environment will already have these installed so there is no need to run the following (it will take a long time). We include this code in the document so you can have a resource for package installation in future analyses you may do.

Load relevant packages

This should be done whenever you start a new r session. For this script we only use functions from the package *minfi*. This package has essentially the largest set of functions of any on Bioconductor.

```
library(minfi)
library(data.table)
library(magrittr)
```

Setting file paths for data

Here we just set up paths for loading of data

```
# Setting file paths for data

# The rest of this script assumes that your data are in a folder called "project" on the Cloud.
# As you work on your own computer, you will need to specify the folder locations.

# Folder location of the data files
data_dir <- "E:/GESS/2060001/Data/"# for rstudio cloud switch to: "/cloud/project/Data/"
data_dir
```

```
## [1] "E:/GESS/2060001/Data/"
```

Read in the data

Here we read in our phenotype data and our RGChannelSet. The RGset is a single large object that is an amalgamation of the .idat files where the data are organized and summarised in an accessible and convenient way.

```
pheno <- data.table::fread(file.path(data_dir, "samplesheet.csv")) #Base R used read.csv() to read in c
dim(pheno)
```

```
## [1] 17 10
```

```
head(pheno)
```

```
##         GEOID celltype casestatus age gender    smoking  Array      Slide
## 1: GSM1051870      PBL         RA  60      F      never R03C02 7766130158
## 2: GSM1052024      PBL         RA  29      F      never R01C02 5730053010
## 3: GSM1052035      PBL    Control  48      M occasional R04C02 5730053011
## 4: GSM1051874      PBL    Control  51      F    current R02C01 7766130166
## 5: GSM1051871      PBL         RA  57      F      never R05C02 7766130158
## 6: GSM1051872      PBL         RA  64      F    current R06C02 7766130158
##                         Basename Batch
## 1: GSM1051870_7766130158_R03C02     2
## 2: GSM1052024_5730053010_R01C02     1
## 3: GSM1052035_5730053011_R04C02     1
## 4: GSM1051874_7766130166_R02C01     2
## 5: GSM1051871_7766130158_R05C02     2
## 6: GSM1051872_7766130158_R06C02     2
```

```
RGset <- read.metharray.exp(file.path(data_dir, "idats"), targets = pheno, verbose = TRUE)
dim(RGset)
```

```
## [1] 622399     17
```

```
manifest <- getManifest(RGset)
str(manifest)
```

```
## Formal class 'IlluminaMethylationManifest' [package "minfi"] with 2 slots
##   ..@ data      :<environment: 0x000000004ac4f560>
##   ..@ annotation: chr "IlluminaHumanMethylation450k"
```

```
annotation <- getAnnotation(RGset)
dim(annotation)
```

```
## [1] 485512     33
```

```
annotation[1:2, ]
```

```
## DataFrame with 2 rows and 33 columns
##                    chr       pos      strand        Name    AddressA
##            <character> <integer> <character> <character> <character>
## cg00050873        chrY   9363356           -  cg00050873    32735311
## cg00212031        chrY  21239348           -  cg00212031    29674443
##               AddressB                                        ProbeSeqA
##            <character>                                      <character>
## cg00050873    31717405 ACAAAAAAACAACACACAACTATAATAATTTTTAAAATAAATAAACCCCA
## cg00212031    38703326 CCCAATTAACCACAAAAACTAAACAAATTATACAATCAAAAAAACATACA
##                                                ProbeSeqB        Type
##                                              <character> <character>
## cg00050873 ACGAAAAAACAACGCACAACTATAATAATTTTTAAAATAAATAAACCCCG           I
## cg00212031 CCCAATTAACCGCAAAAACTAAACAAATTATACGATCGAAAAAACGTACG           I
##               NextBase       Color    Probe_rs Probe_maf      CpG_rs   CpG_maf
##            <character> <character> <character> <numeric> <character> <numeric>
## cg00050873           A         Red          NA        NA          NA        NA
## cg00212031           T         Red          NA        NA          NA        NA
```

```
##                 SBE_rs   SBE_maf           Islands_Name Relation_to_Island
##            <character> <numeric>            <character>        <character>
## cg00050873          NA        NA   chrY:9363680-9363943            N_Shore
## cg00212031          NA        NA chrY:21238448-21240005             Island
##
##
## cg00050873 TATCTCTGTCTGGCGAGGAGGCAACGCACAACTGTGGTGGTTTTTGGAGTGGGTGGACCC[CG]GCCAAGACGGCCTGGGCTGACCAGAC
## cg00212031 CCATTGGCCCGCCCCAGTTGGCCGCAGGGACTGAGCAAGTTATGCGGTCGGGAAGACGTG[CG]TTAAAGGGCTGAAGGGGAGGGACGGA
##                                                    SourceSeq Random_Loci
##                                                  <character> <character>
## cg00050873 CGGGGTCCACCCACTCCAAAAACCACCACAGTTGTGCGTTGCCTCCTCGC
## cg00212031 CGCACGTCTTCCCGACCGCATAACTTGCTCAGTCCCTGCGGCCAACTGGG
##            Methyl27_Loci UCSC_RefGene_Name UCSC_RefGene_Accession
##              <character>       <character>            <character>
## cg00050873                 TSPY4;FAM197Y2 NM_001164471;NR_001553
## cg00212031                         TTTY14              NR_001543
##            UCSC_RefGene_Group     Phantom         DMR    Enhancer
##                   <character> <character> <character> <character>
## cg00050873       Body;TSS1500
## cg00212031             TSS200
##                     HMM_Island Regulatory_Feature_Name Regulatory_Feature_Group
##                    <character>             <character>              <character>
## cg00050873   Y:9973136-9976273
## cg00212031 Y:19697854-19699393
##                        DHS
##            <character>
## cg00050873
## cg00212031
```

Explore the dataset

```
typeof(annotation)
```

```
## [1] "S4"
```

```
typeof(RGset)
```

```
## [1] "S4"
```

```
getClass(RGset)
```

```
## class: RGChannelSet
## dim: 622399 17
## metadata(0):
## assays(2): Green Red
## rownames(622399): 10600313 10600322 ... 74810490 74810492
## rowData names(0):
## colnames: NULL
## colData names(11): GEOID celltype ... Batch filenames
## Annotation
##   array: IlluminaHumanMethylation450k
##   annotation: ilmn12.hg19
```

```
manifest
```

```
## IlluminaMethylationManifest object
```

```
## Annotation
##   array: IlluminaHumanMethylation450k
## Number of type I probes: 135476
## Number of type II probes: 350036
## Number of control probes: 850
## Number of SNP type I probes: 25
## Number of SNP type II probes: 40
```

```
head(getProbeInfo(manifest))
```

```
## DataFrame with 6 rows and 8 columns
##          Name    AddressA    AddressB       Color       NextBase
##   <character> <character> <character> <character> <DNAStringSet>
## 1  cg00050873    32735311    31717405         Red              A
## 2  cg00212031    29674443    38703326         Red              T
## 3  cg00213748    30703409    36767301         Red              A
## 4  cg00214611    69792329    46723459         Red              A
## 5  cg00455876    27653438    69732350         Red              A
## 6  cg01707559    45652402    64689504         Red              A
##                 ProbeSeqA              ProbeSeqB      nCpG
##            <DNAStringSet>          <DNAStringSet> <integer>
## 1 ACAAAAAAAC...ATAAACCCCA ACGAAAAAAC...ATAAACCCCG         2
## 2 CCCAATTAAC...AAAACATACA CCCAATTAAC...AAAACGTACG         4
## 3 TTTTAACACC...AAAAAAAACA TTTTAACGCC...AAAAAAAACG         3
## 4 CTAACTTCCA...AACACAAACA CTAACTTCCG...AACGCGAACG         5
## 5 AACTCTAAAC...AAAAAACTCA AACTCTAAAC...AAAAAACTCG         2
## 6 ACAAATTAAA...ACAAAAAACA GCGAATTAAA...ACAAAAAACG         6
```

```
dim(getProbeInfo(manifest))
```

```
## [1] 135476      8
```

```
table(getProbeInfo(manifest)$Color)
```

```
##
##   Grn   Red
## 46289 89187
```

```
pd <- RGset@colData@listData ; setDT(pd) #This setDT function lets us format our data as data.table

dim(pd)
```

```
## [1] 17 11
```

```
head(pd)
```

```
##         GEOID celltype casestatus age gender    smoking  Array      Slide
## 1: GSM1051870      PBL         RA  60      F      never R03C02 7766130158
## 2: GSM1052024      PBL         RA  29      F      never R01C02 5730053010
## 3: GSM1052035      PBL    Control  48      M occasional R04C02 5730053011
## 4: GSM1051874      PBL    Control  51      F    current R02C01 7766130166
## 5: GSM1051871      PBL         RA  57      F      never R05C02 7766130158
## 6: GSM1051872      PBL         RA  64      F    current R06C02 7766130158
##                         Basename Batch
## 1: GSM1051870_7766130158_R03C02     2
## 2: GSM1052024_5730053010_R01C02     1
## 3: GSM1052035_5730053011_R04C02     1
## 4: GSM1051874_7766130166_R02C01     2
```

```
## 5: GSM1051871_7766130158_R05C02     2
## 6: GSM1051872_7766130158_R06C02     2
##                                                    filenames
## 1: E:/GESS/2060001/Data//idats/GSM1051870_7766130158_R03C02
## 2: E:/GESS/2060001/Data//idats/GSM1052024_5730053010_R01C02
## 3: E:/GESS/2060001/Data//idats/GSM1052035_5730053011_R04C02
## 4: E:/GESS/2060001/Data//idats/GSM1051874_7766130166_R02C01
## 5: E:/GESS/2060001/Data//idats/GSM1051871_7766130158_R05C02
## 6: E:/GESS/2060001/Data//idats/GSM1051872_7766130158_R06C02
```

```
pd[, table(casestatus)]
```

```
## casestatus
## Control      RA
##       6      11
```

```
pd[, table(gender)]
```

```
## gender
##  F  M
## 10  7
```

```
pd[, table(gender, casestatus)]
```

```
##       casestatus
## gender Control RA
##      F       3  7
##      M       3  4
```

```
pd[, table(Batch)]
```

```
## Batch
##  1  2
##  7 10
```

```
pd[, table(Batch, casestatus)]
```

```
##      casestatus
## Batch Control RA
##     1       3  4
##     2       3  7
```

```
pd[, summary(age)]
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   46.00   52.00   50.76   57.00   65.00
```

```
pd[gender == "M", summary(age)]
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   37.00   45.50   48.00   48.71   53.00   59.00
```

```
pd[gender == "F", summary(age)]
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   47.25   53.50   52.20   59.25   65.00
```

```
head(pd)
```

```
##         GEOID celltype casestatus age gender    smoking  Array      Slide
```

```
## 1: GSM1051870      PBL         RA  60      F      never R03C02 7766130158
## 2: GSM1052024      PBL         RA  29      F      never R01C02 5730053010
## 3: GSM1052035      PBL    Control  48      M occasional R04C02 5730053011
## 4: GSM1051874      PBL    Control  51      F    current R02C01 7766130166
## 5: GSM1051871      PBL         RA  57      F      never R05C02 7766130158
## 6: GSM1051872      PBL         RA  64      F    current R06C02 7766130158
##                        Basename Batch
## 1: GSM1051870_7766130158_R03C02     2
## 2: GSM1052024_5730053010_R01C02     1
## 3: GSM1052035_5730053011_R04C02     1
## 4: GSM1051874_7766130166_R02C01     2
## 5: GSM1051871_7766130158_R05C02     2
## 6: GSM1051872_7766130158_R06C02     2
##                                                    filenames
## 1: E:/GESS/2060001/Data//idats/GSM1051870_7766130158_R03C02
## 2: E:/GESS/2060001/Data//idats/GSM1052024_5730053010_R01C02
## 3: E:/GESS/2060001/Data//idats/GSM1052035_5730053011_R04C02
## 4: E:/GESS/2060001/Data//idats/GSM1051874_7766130166_R02C01
## 5: E:/GESS/2060001/Data//idats/GSM1051871_7766130158_R05C02
## 6: E:/GESS/2060001/Data//idats/GSM1051872_7766130158_R06C02
```

Save RGset object

While in our 17 sample example for lab no process takes especially long, once you scale up the number of samples you will see larger and larger increases in computation time. Therefore, saving large intermediate data products such as the RGChannelSet is helpful.

```
# Save RGChannelSet object
save(RGset, file = file.path(data_dir, "RGset.rda"))
```