

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 August 03, 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 RStudio 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 primarily use functions from the package *minfi*. This package has the largest set of DNA methylation functions of any on Bioconductor.

```
library(minfi)
library(magrittr)
library(knitr)
library(here)
```

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 summarized in an accessible and convenient way.

These data are a subset of samples from Liu et al 2013 (<https://pubmed.ncbi.nlm.nih.gov/23334450/>). The full dataset can be downloaded from the Genome Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42861>)

```
here() # See where the R project is located, and the default location where data will be imported from
```

```
## [1] "E:/GESS/Epigenomics_for_Social_Scientists"
sample.sheet <- read.csv(here("Data","samplesheet.csv"))
dim(sample.sheet)
```

```
## [1] 17 10
head(sample.sheet)
```

```
##      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(here("Data", "idats"), targets = sample.sheet, verbose = TRUE, extended = T)
dim(RGset)
```

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

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

```
## Formal class 'IlluminaMethylationManifest' [package "minfi"] with 2 slots
## ..@ data      :<environment: 0x0000000046c684a0>
## ..@ annotation: chr "IlluminaHumanMethylation450k"
```

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

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

```
head(annotation)
```

```
## DataFrame with 6 rows and 33 columns
```

```
##      chr      pos      strand      Name      AddressA
##      <character> <integer> <character> <character> <character>
## cg00050873      chrY      9363356      -      cg00050873      32735311
## cg00212031      chrY      21239348      -      cg00212031      29674443
## cg00213748      chrY      8148233      -      cg00213748      30703409
## cg00214611      chrY      15815688      -      cg00214611      69792329
## cg00455876      chrY      9385539      -      cg00455876      27653438
## cg01707559      chrY      6778695      +      cg01707559      45652402
##      AddressB      ProbeSeqA      ProbeSeqB
##      <character>      <character>      <character>
## cg00050873      31717405      ACAAAAAACAACACACAAC..      ACGAAAAACAACGCACAAC..
## cg00212031      38703326      CCAATTAACCACAAAAACT..      CCAATTAACCGCAAAAACT..
## cg00213748      36767301      TTTAACACCTAACACCATT..      TTTAACGCCTAACACCGTT..
## cg00214611      46723459      CTAACCTCCAAACCACACTT..      CTAACCTCCGAACCGCGCTT..
## cg00455876      69732350      AACTCTAAACTACCCAACAC..      AACTCTAAACTACCCGACAC..
## cg01707559      64689504      ACAAATTAATAAACTAAAA..      GCGAATTAATAAACTAAAA..
##      Type      NextBase      Color      Probe_rs      Probe_maf
##      <character> <character> <character> <character> <numeric>
## cg00050873      I      A      Red      NA      NA
## cg00212031      I      T      Red      NA      NA
## cg00213748      I      A      Red      NA      NA
## cg00214611      I      A      Red      NA      NA
```

##	cg00455876	I	A	Red	NA	NA
##	cg01707559	I	A	Red	NA	NA
##		CpG_rs	CpG_maf	SBE_rs	SBE_maf	Islands_Name
##		<character>	<numeric>	<character>	<numeric>	<character>
##	cg00050873	NA	NA	NA	NA	chrY:9363680-9363943
##	cg00212031	NA	NA	NA	NA	chrY:21238448-21240005
##	cg00213748	NA	NA	NA	NA	chrY:8147877-8148210
##	cg00214611	NA	NA	NA	NA	chrY:15815488-15815779
##	cg00455876	NA	NA	NA	NA	chrY:9385471-9385777
##	cg01707559	NA	NA	NA	NA	chrY:6778574-6780028
##		Relation_to_Island		Forward_Sequence		SourceSeq
##		<character>		<character>		<character>
##	cg00050873	N_Shore	TATCTCTGTCTGGCGAGGAG..	CGGGGTCCACCCACTCCAAA..		
##	cg00212031	Island	CCATTGGCCCCGCCAGTTG..	CGCACGTCTTCCGACCGCA..		
##	cg00213748	S_Shore	TCTGTGGGACCATTTTAACG..	CGCCCCCTCCTGCAGAACCT..		
##	cg00214611	Island	GCGCCGGCAGGACTAGCTTC..	CGCCCGGCCACACTGCAGC..		
##	cg00455876	Island	CGCGTGTGCCTGGACTCTGA..	GACTCTGAGCTACCGGCAC..		
##	cg01707559	Island	AGCGGCCGCTCCAGTGGTG..	CGCCCTCTGTCGCTGCAGCC..		
##		Random_Loci	Methyl27_Loci	UCSC_RefGene_Name	UCSC_RefGene_Accession	
##		<character>	<character>	<character>	<character>	
##	cg00050873			TSPY4;FAM197Y2	NM_001164471;NR_001553	
##	cg00212031			TTY14	NR_001543	
##	cg00213748					
##	cg00214611			TMSB4Y;TMSB4Y	NM_004202;NM_004202	
##	cg00455876					
##	cg01707559			TBL1Y;TBL1Y;TBL1Y	NM_134259;NM_033284;..	
##		UCSC_RefGene_Group	Phantom	DMR	Enhancer	
##		<character>	<character>	<character>	<character>	
##	cg00050873	Body;TSS1500				
##	cg00212031	TSS200				
##	cg00213748					
##	cg00214611	1stExon;5'UTR				
##	cg00455876					
##	cg01707559	TSS200;TSS200;TSS200				
##		HMM_Island	Regulatory_Feature_Name	Regulatory_Feature_Group		
##		<character>	<character>	<character>		
##	cg00050873	Y:9973136-9976273				
##	cg00212031	Y:19697854-19699393				
##	cg00213748	Y:8207555-8208234				
##	cg00214611	Y:14324883-14325218	Y:15815422-15815706	Promoter_Associated_..		
##	cg00455876	Y:9993394-9995882				
##	cg01707559	Y:6838022-6839951				
##		DHS				
##		<character>				
##	cg00050873					
##	cg00212031					
##	cg00213748					
##	cg00214611					
##	cg00455876					
##	cg01707559					

Explore the dataset

```
class(annotation)
```

```
## [1] "DFrame"  
## attr("package")  
## [1] "S4Vectors"
```

```
class(RGset)
```

```
## [1] "RGChannelSetExtended"  
## attr("package")  
## [1] "minfi"
```

```
getClass(RGset)
```

```
## class: RGChannelSetExtended  
## dim: 622399 17  
## metadata(0):  
## assays(5): Green Red GreenSD RedSD NBeads  
## rownames(622399): 10600313 10600322 ... 74810490 74810492  
## rowData names(0):  
## colnames(17): GSM1051870_7766130158_R03C02 GSM1052024_5730053010_R01C02  
## ... GSM1052032_5730053011_R06C01 GSM1052037_5730053011_R06C02  
## 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> <DNAStrngSet>  
## 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  
##      <DNAStrngSet> <DNAStrngSet> <integer>  
## 1 ACAAAAAAAC...ATAAACCCCA ACGAAAAAAC...ATAAACCCCG 2  
## 2 CCCAATTAAC...AAAACATACA CCCAATTAAC...AAAACGTACG 4  
## 3 TTTTAACACC...AAAAAAAACA TTTTAACGCC...AAAAAAAACG 3  
## 4 CTAACCTCCA...AACACAAACA CTAACCTCCG...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 # Pull out the participant/demographics/phenotype information
pd$sex <- pd$gender

dim(pd)

## NULL

head(pd)

## $GEOID
## [1] "GSM1051870" "GSM1052024" "GSM1052035" "GSM1051874" "GSM1051871"
## [6] "GSM1051872" "GSM1051866" "GSM1051863" "GSM1052025" "GSM1052021"
## [11] "GSM1052029" "GSM1051879" "GSM1051878" "GSM1051883" "GSM1051877"
## [16] "GSM1052032" "GSM1052037"
##
## $celltype
## [1] "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL"
## [13] "PBL" "PBL" "PBL" "PBL" "PBL"
##
## $casestatus
## [1] "RA"      "RA"      "Control" "Control" "RA"      "RA"      "RA"
## [8] "Control" "RA"      "Control" "RA"      "RA"      "RA"      "Control"
## [15] "RA"      "RA"      "Control"
##
## $age
## [1] 60 29 48 51 57 64 44 43 55 53 46 47 37 52 53 65 59
##
## $gender
## [1] "F" "F" "M" "F" "F" "F" "M" "F" "F" "M" "F" "M" "M" "F" "M" "F" "M"
##
## $smoking
## [1] "never"      "never"      "occasional" "current"     "never"
## [6] "current"    "never"      "current"     "occasional" "ex"
## [11] "ex"         "ex"         "ex"          "current"     "ex"
## [16] "current"    "never"

table(pd$casestatus)

##
## Control      RA
##          6      11

table(pd$sex)

##

```

```
## F M
## 10 7

table(pd$sex, pd$casestatus)

##
##      Control RA
## F      3  7
## M      3  4

table(pd$Batch)

##
## 1 2
## 7 10

table(pd$Batch, pd$casestatus)

##
##      Control RA
## 1      3  4
## 2      3  7

summary(pd$age)

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

summary(pd$age[pd$sex == "M"])

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

summary(pd$age[pd$sex == "M"])

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

head(pd)

## $GEOID
## [1] "GSM1051870" "GSM1052024" "GSM1052035" "GSM1051874" "GSM1051871"
## [6] "GSM1051872" "GSM1051866" "GSM1051863" "GSM1052025" "GSM1052021"
## [11] "GSM1052029" "GSM1051879" "GSM1051878" "GSM1051883" "GSM1051877"
## [16] "GSM1052032" "GSM1052037"
##
## $celltype
## [1] "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL" "PBL"
## [13] "PBL" "PBL" "PBL" "PBL" "PBL"
##
## $casestatus
## [1] "RA"      "RA"      "Control" "Control" "RA"      "RA"      "RA"
## [8] "Control" "RA"      "Control" "RA"      "RA"      "RA"      "Control"
## [15] "RA"      "RA"      "Control"
##
## $age
## [1] 60 29 48 51 57 64 44 43 55 53 46 47 37 52 53 65 59
##
## $gender
```

```
## [1] "F" "F" "M" "F" "F" "F" "M" "F" "F" "M" "F" "M" "M" "F" "M" "F" "M"
##
## $smoking
## [1] "never"      "never"      "occasional" "current"     "never"
## [6] "current"    "never"      "current"    "occasional" "ex"
## [11] "ex"         "ex"         "ex"         "current"     "ex"
## [16] "current"    "never"
```

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", "RGset.rda"))
```