Autism Project EDA

Carlos Ramírez

August 4, 2017

Exploratory Data Analysis on Autism SNP data

Illumina raw results are provided as input files for Plink, Partek and GenomeStudio. Additionally, three important files can be observed named: i) "*_FinalReport.txt", ii)"*.opa", and iii)"SampleSheet.*.csv". FinalReport.txt file contains info about the nucleotide composition for each SNP in every sample. The".opa" file summarizes SNP genomic information such as chromosome location and position. Finally, "sampleSheet.csv" seems to contain patient sample ID.
This report shows data processing regarding these three files.

```
## loading data from local repositories
## local directory
pathReport <- "~/dataBases/GoldenGate_U15_025 AT/FinalReport/U15_025_ID_FinalReport.txt"
pathOpa <- "~/dataBases/GoldenGate_U15_025 AT/GenomeStudio/GS0014416-OPA.opa"
pathSampleSheet <- "~/dataBases/GoldenGate_U15_025 AT/GenomeStudio/SampleSheet_ID.csv"

## loading into R
freport <- read.csv(file = pathReport,
            sep = "\t",
            header = TRUE,
            skip = 9,
            stringsAsFactors = FALSE,
            colClasses = "character")

opa <- read.csv(file = pathOpa,
                skip = 12,
                header = TRUE)

sample <- read.csv(file = pathSampleSheet,
         skip = 8,
         header = TRUE)

## display files composition
lapply(list(freport[1:10],opa,sample), names)
```

```
## [[1]]
##  [1] "X"         "AU_118_1" "AU_103_1" "X1004"    "AU_126_1" "X40_1026"
##  [7] "AU_106_1" "X40_1029" "X1005"    "AU_113_1"
##
## [[2]]
##  [1] "Ilmn.ID"         "Name"            "oligo1"
##  [4] "oligo2"          "oligo3"          "IllumiCode.name"
##  [7] "Illumicode.Seq"  "Ilmn.Strand"     "SNP"
## [10] "CHR"             "Ploidy"          "Species"
## [13] "MapInfo"         "TopGenomicSeq"   "CustomerStrand"
##
## [[3]]
## [1] "Sample_ID"         "SentrixBarcode_A"  "SentrixPosition_A"
```

```
## [4] "Path"              "Aux"
```

As can be inferred from inspecting the files and from the given output FinalReport.txt file contains in its first field (column) called "X" the illumina SNP ID. The rest of the fields (AU_118_1, AU_103_1, ...) corresponds to sample IDs also present in the sampleSheet.csv file. The entries of the matrix given in this file contains observed nucleotides.

In the rest of the report frequency of nucleotide markers are obtained from finalReport.txt file. Using the information on SNP location from .opa file a bar graph of frequencies for chromosome 22 is constructed.

The next code shows data filtering process.

```
## keeping only SNP data
for (i in 2:ncol(freport)) {
        freport[,i] <- gsub("\\|.*","",freport[i,2:ncol(freport)])
}

## recording existing SNPs
bag <- c()
for (j in 2:ncol(freport)) {
        for (i in 1:nrow(freport)) {
                if ( ! ( freport[i,j] %in% bag ) ) {
                        bag <- c(bag,freport[i,j])
                }
        }
}
bag <- sort(bag)

## matrix of frequencies definition
frequencies <- matrix(data = 0,
                      nrow = nrow(freport),
                      ncol = length(bag))
colnames(frequencies) <- bag


## getting frequencies from haplotypes
for (i in 1:nrow(freport)) {
        res <- table(gsub("\\|.*","",freport[i,2:ncol(freport)]))
        frequencies[i,names(res)] <- res
}

frequencies.df <- as.data.frame(frequencies)
frequencies.df$"X" <- freport[,"X"]
head(frequencies.df)
```

```
##   -- AA AC AG AT CC CG GC GG TC TG TT         X
## 1  0 11  0 51  0  0  0  0 34  0  0  0 rs7053244
## 2  0  1  0 22  0  0  0  0 73  0  0  0 rs4804833
## 3  0  0  0  2  0 36  0  0  1 44  0 13 rs4274855
## 4  0 83  0  1  9  0  0  0  2  0  0  1 rs2236379
## 5  0  2  0  1  0 51  0  0  1 13  0 28    rs6940
## 6  0  5  0 24  0  1  0  0 65  1  0  0 rs7209406
```

The next code relates so created frequencies.df dataframe to their chromosome location located in the .opa file.

```
chrPosition <- opa[,c("Name","CHR")]
names(chrPosition) <- c("X","CHR")
frequencies.df <- merge(frequencies.df,chrPosition,by="X",all = FALSE)
head(frequencies.df)
```

```
##            X -- AA AC AG AT CC CG GC GG TC TG TT CHR
## 1 rs10013071  0 26  0 21  0  4  1  0 43  1  0  0   4
## 2 rs10013254  0  6  1  3  0 22  0  0 21 21  1 21   4
## 3 rs10029775  0  8  1  6  0 46  1  0 15  8  2  9   4
## 4  rs1003576  1 14  1  9  0 17  0  0 25  8  2 19  20
## 5  rs1004008  0  2  0  1  0 51  0  0  1 13  0 28   3
## 6  rs1004212  0 20 36  1  1 27  1  0  7  0  0  3  14
```

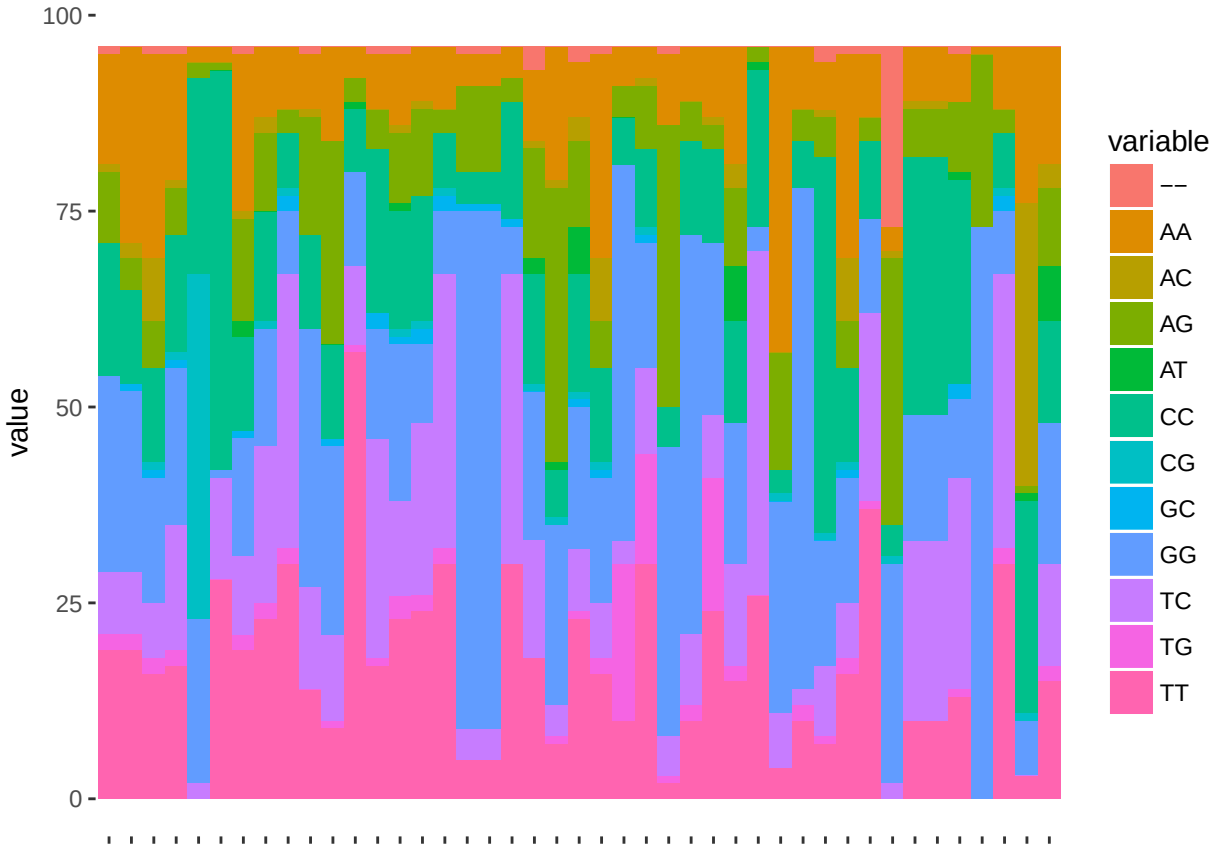
Finally, as an example the next plot shows the frequencies of each marker located in the chromosome 20 in the sample population.

```
library(ggplot2)
library(reshape2)

chrNumber <- 20
n <- as.character(chrNumber)
frequencies.m <- melt(frequencies.df[frequencies.df$CHR==n,])
```

```
## Using X, CHR as id variables
```

```
g <- ggplot(data = frequencies.m, aes(x = X,y = value,fill=variable))
g <- g + scale_color_manual(values=variable)
g <- g + geom_bar(stat="identity",width = 1)
g <- g + theme(axis.line=element_blank(),
               axis.text.x=element_blank(),
               axis.title.x=element_blank(),
               panel.background=element_blank(),
               panel.border=element_blank(),
               panel.grid.major=element_blank(),
               panel.grid.minor=element_blank(),
               plot.background=element_blank())
plot(g)
```



Conclusiones

El análisis exploratorio preliminar de los datos muestra que la información puede ser obtenida a partir de varios archivos proveídos por Illumina como los archivos de entrada de Plink, Partek y genomeStudio. Adicionalmente, también son proveídos resúmenes con los marcadores observados para cada SNP en cada muestra en el archivo FinalReport.txt en cual se basó este reporte para obtener las frequencias de marcadores en la población muestreada.

Como perspectiva inmediata se tiene un análisis exploratorio de los datos partiendo de los archivos de entrada a Plink, Partek y GenomeStudio. Después de esto se procederá al análisis descriptivo de los datos mediante por ejemplo la observación de marcadores genómicos sobrerrepresentados en los pacientes cuando se tengan los datos sobre los controles. Otra perspectiva es analizar las bases de datos disponibles publicamente.