ADSP's WES samples: phenotype and more

Xulong Wang (xulong.wang@jax.org)

April 6, 2016

```r
rm(list = ls())
setwd("~/Dropbox/GitHub/wes")

options(stringsAsFactors = F)
mdata <- read.delim("./docs/ADSP_wes_samplesheet_Jan2016_Updated_01262016.txt")
names(mdata)
```

```
##  [1] "X"                      "SMID"
##  [3] "WellADSPQC"             "ADSPID"
##  [5] "ADSP_SM_ID"             "freeze_no"
##  [7] "SampleSelection"        "StudyName"
##  [9] "STID"                   "CID"
## [11] "PlatformType"           "Gender"
## [13] "PrevAD"                 "IncAD"
## [15] "status"                 "Age"
## [17] "APOE"                   "Autopsy"
## [19] "Braak"                  "Race"
## [21] "Ethnicity"              "FamID"
## [23] "AD"                     "Consent"
## [25] "dbGapID"                "lssc"
## [27] "numlanes"               "totbases"
## [29] "totmapbases"            "totunimapbases"
## [31] "percent_aligned"        "percent_unique"
## [33] "error_rate"             "total_reads"
## [35] "targeted_insert_length" "SampleSource"
## [37] "avgcov"                 "tgtbases20x"
## [39] "ontgtbases"             "avgtgtdepth"
## [41] "sra_sample_id"          "file_updated"
## [43] "SRR"
```

In ADSP's WES study, 3 sequencing centers - WashU, Baylor, Broad - sequenced 9949 samples. 10 of the 9949 samples were sequenced repeatedly by the 3 centers. This led to 20 redundant sequencing results. Noteworthy, depending on the specific sample, sequencing results by any of the 3 centers could be flagged as "SeqControl", which we removed from downstream analysis.

```r
table(mdata$SampleSelection)
```

```
##
## Case Control     Enriched   SeqControl
##         9133          816           20
```

```r
seqcontrol = mdata[mdata$SampleSelection == "SeqControl", ]
table(mdata$ADSPID %in% seqcontrol$ADSPID)
```

```
##
## FALSE  TRUE
##  9939    30
```

```r
seqcontrol_all = mdata[mdata$ADSPID %in% seqcontrol$ADSPID, ]
seqcontrol_all[, c("ADSPID", "SampleSelection", "Gender", "status", "lssc")]
```

```
##              ADSPID SampleSelection Gender  status   lssc
## 1152 A-ACT-AC002970    Case Control      1 control  WashU
## 1153 A-ACT-AC002970      SeqControl      1 control Baylor
## 1154 A-ACT-AC002970      SeqControl      1 control  Broad
## 1155 A-ACT-AC002972    Case Control      0 control  WashU
## 1156 A-ACT-AC002972      SeqControl      0 control Baylor
## 1157 A-ACT-AC002972      SeqControl      0 control  Broad
## 1265 A-ACT-AC003403    Case Control      1    case  WashU
## 1266 A-ACT-AC003403      SeqControl      1    case Baylor
## 1267 A-ACT-AC003403      SeqControl      1    case  Broad
## 1274 A-ACT-AC003410    Case Control      1    case  WashU
## 1275 A-ACT-AC003410      SeqControl      1    case Baylor
## 1276 A-ACT-AC003410      SeqControl      1    case  Broad
## 1402 A-ADC-AD000263    Case Control      1 control  Broad
## 1403 A-ADC-AD000263      SeqControl      1 control Baylor
## 1404 A-ADC-AD000263      SeqControl      1 control  WashU
## 2647 A-ADC-AD003250    Case Control      0    case  Broad
## 2648 A-ADC-AD003250      SeqControl      0    case Baylor
## 2649 A-ADC-AD003250      SeqControl      0    case  WashU
## 2676 A-ADC-AD003299    Case Control      1    case  Broad
## 2677 A-ADC-AD003299      SeqControl      1    case Baylor
## 2678 A-ADC-AD003299      SeqControl      1    case  WashU
## 7648    C-CHS-30738    Case Control      1    case Baylor
## 7649    C-CHS-30738      SeqControl      1    case  Broad
## 7650    C-CHS-30738      SeqControl      1    case  WashU
## 7762    C-CHS-50100    Case Control      1 control Baylor
## 7763    C-CHS-50100      SeqControl      1 control  Broad
## 7764    C-CHS-50100      SeqControl      1 control  WashU
## 8017    C-CHS-51381    Case Control      0 control Baylor
## 8018    C-CHS-51381      SeqControl      0 control  Broad
## 8019    C-CHS-51381      SeqControl      0 control  WashU
```

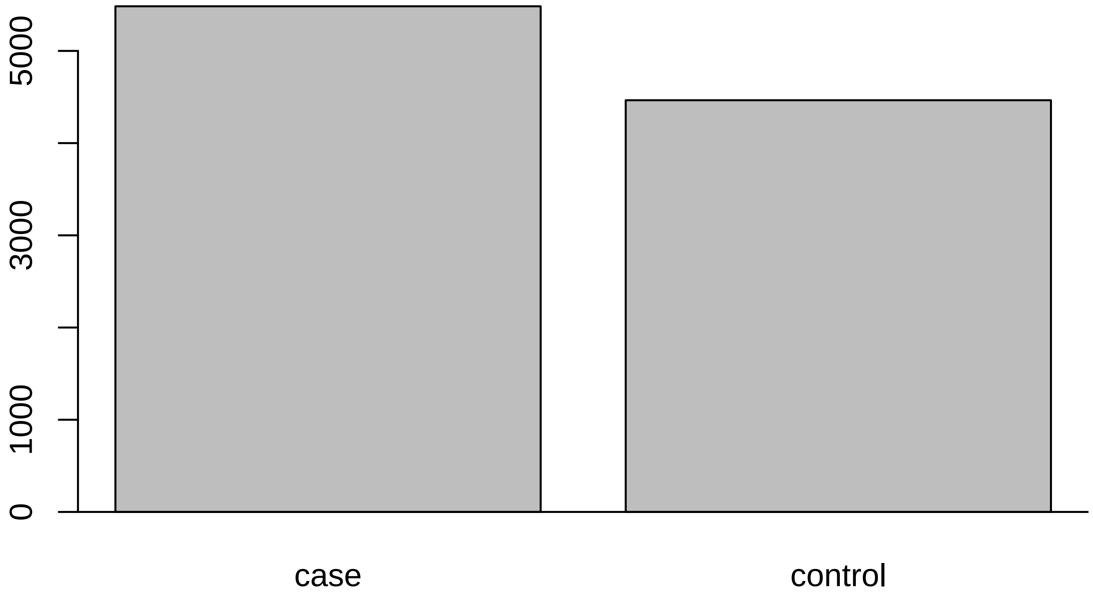
```r
mdata = mdata[mdata$SampleSelection != "SeqControl", ]
```

This led to 9949 sequencing results of 9949 people. Each of the 9949 people was diagnosed as control or case in Alzheimer's disease status.

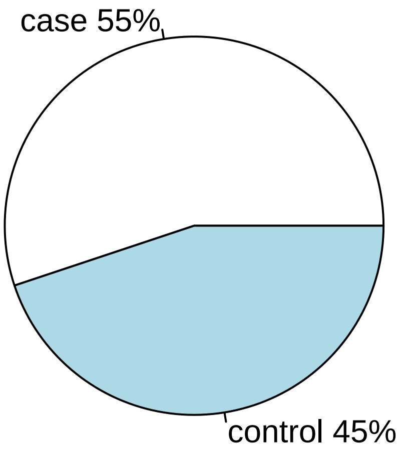
```r
(status = table(mdata$status))
```

```
##
##    case control
##    5484    4465
```

```r
barplot(status); abline(h = 0)
```



```
pct <- round(status/sum(status)*100)
lbs <- paste(paste(names(status), pct), "%", sep = "")

# pdf("./pdf/ad_pie.pdf", family = "Helvetica")

# par(cex = 1.7, col = "grey30")
pie(status, labels = lbs)
```



```
# dev.off()
```

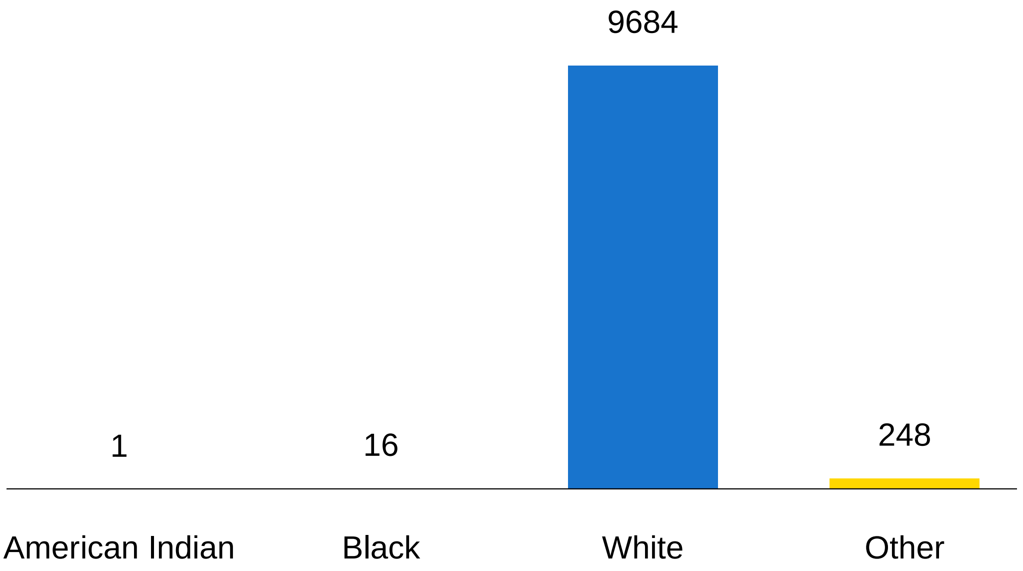
Age, Sex, Race, and APOE genotypes were also reported for each sample.

White race constituted 97.3% of the total.

```
race = table(mdata$Race)
names(race) = c("American Indian", "Black", "White", "Other")
race
```

```
## American Indian           Black           White           Other
##               1              16            9684             248
```

```
# pdf("./pdf/race.pdf", width = 6, height = 3)

# op <- par(mar = c(5, 4, 4, 3))
mycol <- c("grey70", "firebrick1", "dodgerblue3", "gold1", "chartreuse3", "darkorchid2")
bar <- barplot(race, ylim = c(0, max(race) + 2e3), axes = F, border = NA, las = 1, space = 0.75, col =
abline(h = 0, lwd = 1, col = "black")
text(x = bar, y = race + 1e3, labels = race)
```



```
# dev.off()
```

Unpleasantly, all 90 years older people were annotated as “90+”. This accounted for 1307 people, and 13.1% of the total. We coded these people as 90 years old exactly in the downstream analysis.

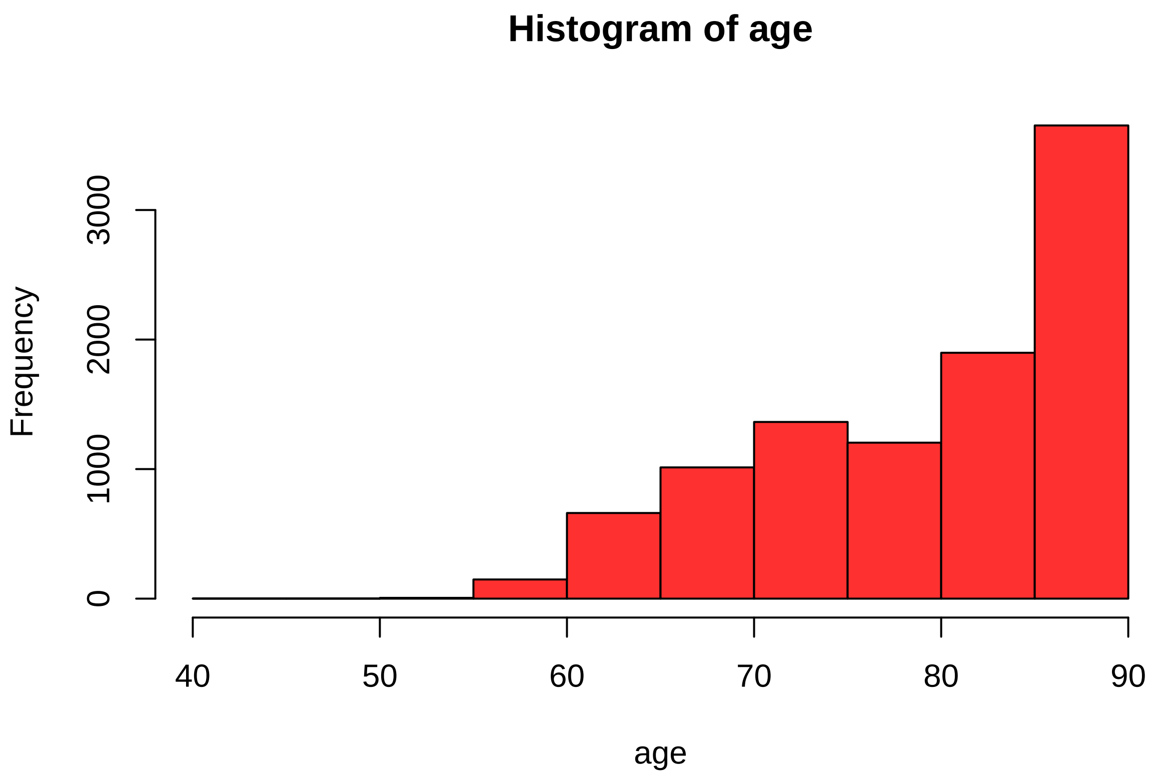
```
table(mdata$Age == "90+")
```

```
##
## FALSE  TRUE
##  8642  1307
```

```
age = mdata$Age = as.numeric(gsub("\\+", "", mdata$Age))
summary(age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    42.0    73.0    82.0    79.9    88.0    90.0
```

```
# pdf("./pdf/age.pdf", family = "Helvetica", height = 5)

# par(cex = 1.7, col = "grey30")
hist(age, col = "firebrick1")
```



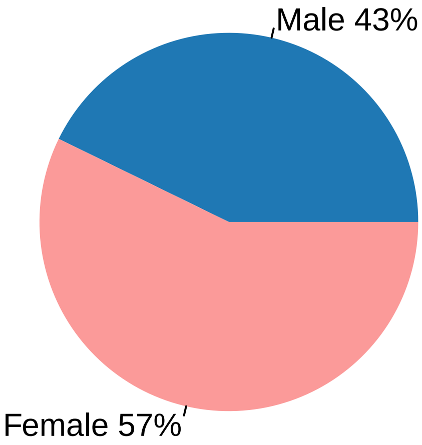
```
# dev.off()
```

There were 13% more females than males.

```
(sex = table(mdata$Gender))
```

```
##
##    0    1
## 4254 5695
```

```
pct <- round(sex/sum(sex)*100)
lbs <- paste(paste(c("Male", "Female"), pct), "%", sep = "")

# pdf("./PDF/sex.pdf", family = "Helvetica")

# par(cex = 1.7, col = "grey30")
pie(sex, label = lbs, col = c("#1f78b4", "#fb9a99"), border = F)
```



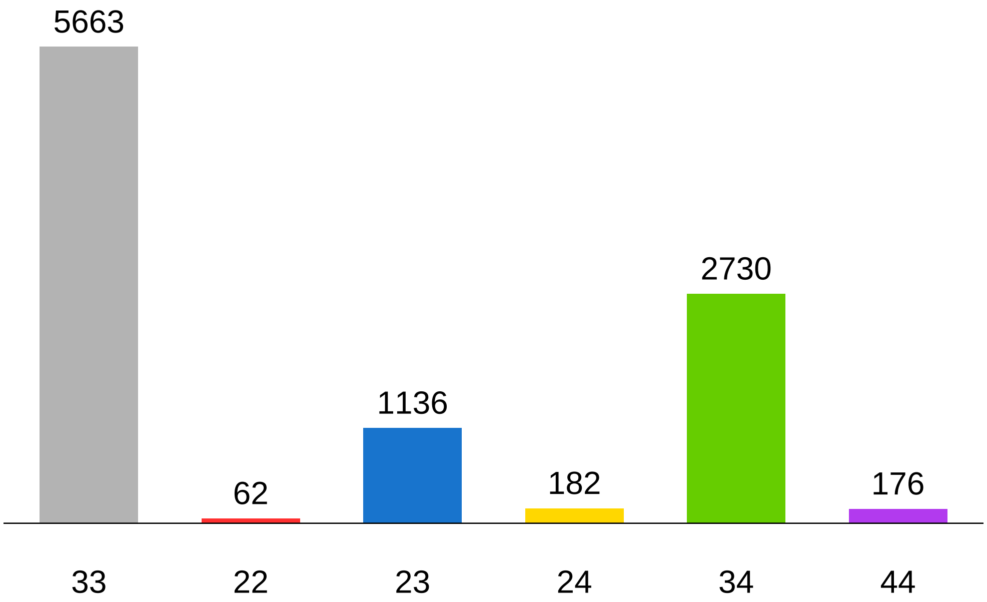
```
# dev.off()
```

We had 6 unique APOE genotypes, which filled all the possible bi-allelic combinations of APOE's 3 alleletypes, e2/e3/e4. Homozygous e3 was predominantly the majority, while homozygous e2 or e4 were rare. Further, e4/e4 carriers showed a 15-fold increased risk of developing AD comparing to the e3/e3 carriers.

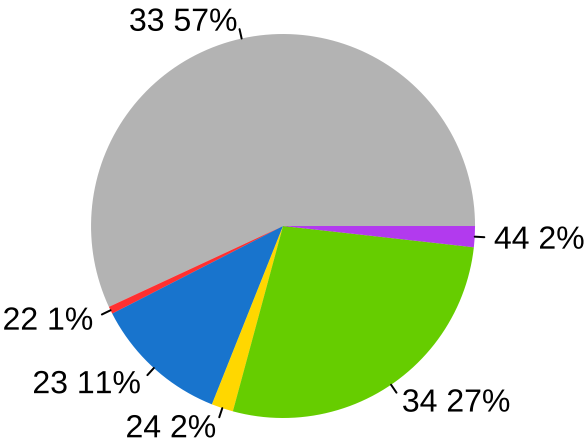
```
mdata$APOE = factor(mdata$APOE, levels = c("33", "22", "23", "24", "34", "44"))
(Apoe = table(mdata$APOE))
```

```
##
##   33   22   23   24   34   44
## 5663   62 1136  182 2730  176
```

```
pct <- round(Apoe/sum(Apoe)*100)
lbs <- paste(paste(names(Apoe), pct), "%", sep = "")

# pdf("./pdf/apoe_bar.pdf", width = 6, height = 4)

op <- par(mar = c(5, 4, 4, 3))
bar <- barplot(Apoe, ylim = c(0, max(Apoe) + 5e2), axes = F, border = NA, las = 1, space = 0.65, col = r
abline(h = 0, lwd = 1, col = "black")
text(x = bar, y = Apoe + 3e2, labels = Apoe)
```



```
# dev.off()

# pdf("./PDF/apoe.pdf", family = "Helvetica")

# par(cex = 1.7, col = "grey30")
pie(Apoe, label = lbs, col = mycol, border = F)
```



```
# dev.off()

(x = as.matrix(table(mdata$status, mdata$APOE)))
```

```
##
##            33   22   23   24   34   44
##   case    2709   20  334  115 2142  164
##   control 2954   42  802   67  588   12
```

```
(or44 = (x["case", "44"] / x["case", "33"]) / (x["control", "44"] / x["control", "33"]))
```

```
## [1] 14.90267
```

```r
fisher.test(x[, c("44", "33")])
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  x[, c("44", "33")]
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   8.277122 29.488797
## sample estimates:
## odds ratio
##   14.90362
```

Make the final phenotype data for downstream analysis

```r
tfam <- read.table("./plink/wes.tfam")
mdata = mdata[match(tfam$V1, mdata$ADSP_SM_ID), ]
all(tfam$V1 == mdata$ADSP_SM_ID)
```

```
## [1] TRUE
```

```r
mdata$APOE <- factor(mdata$APOE, levels = c("33", "22", "23", "24", "34", "44")) # ref:33

mdata$Apoe2 <- sapply(strsplit(as.character(mdata$APOE), ""), function(x) sum(x == 2))
mdata$Apoe3 <- sapply(strsplit(as.character(mdata$APOE), ""), function(x) sum(x == 3))
mdata$Apoe4 <- sapply(strsplit(as.character(mdata$APOE), ""), function(x) sum(x == 4))

mdata$Gender <- factor(mdata$Gender, levels = c("0", "1")) # ref:male
mdata$Sex = as.integer(mdata$Gender) - 1

race.map = data.frame(code = c(1, 4:6), name = c("American Indian", "Black", "White", "Other"))
mdata$Race = race.map$name[match(mdata$Race, race.map$code)]

mdata$AD = 0
mdata$AD[mdata$status == "case"] = 1

save(mdata, file = "./mdata.rdt")
```