Read in the data, look at basic statistics, generate Table 1

March 1, 2017

1 Load all libraries and functions

```
library(ggplot2)

## Warning:   package 'ggplot2' was built under R version 3.3.2

library(reshape)

## Warning:   package 'reshape' was built under R version 3.3.2

library(genefilter)
library(RColorBrewer)
library(dplyr)

##
## Attaching package:  'dplyr'
## The following object is masked from 'package:reshape':
##
##     rename
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(xtable)

source("functions.R")
```

2 Load data, add age categories

Read in data from csv file:

```
MasterFrame <- read.csv("S1_Table.csv")
```

Take a look at the dataset:

```
dim(MasterFrame)

## [1]   73 2207

head(colnames(MasterFrame))

## [1] "Status"   "Age"      "Site"     "M132T37p" "M312T36p" "M357T30p"

tail(colnames(MasterFrame))

## [1] "M325T556_2n" "M113T35_2n"  "M378T543n"   "M1013T35_1n" "M338T41n"
## [6] "CKM"

MasterFrame[1:5,1:5]

##   Status Age      Site  M132T37p  M312T36p
## 1    DMD 8.0 Australia -5.900795 -6.692130
## 2    DMD 4.0   Calgary -7.681591 -6.363260
## 3    DMD 4.3   Calgary -5.763479 -6.497395
## 4    DMD 4.3   Calgary -5.231625 -6.175428
## 5    DMD 5.0   Calgary -5.649194 -7.098781
```

Create vector of metabolites and get number of metabolites:

```
##metabolite IDs all have the format (from XCMS) of M.mass.T.retentiontime.n/p
##(for negative or positive mode)
mets <- colnames(MasterFrame)[grep("M[[:digit:]]+T[[:digit:]]+",colnames(MasterFrame))]
head(mets)

## [1] "M132T37p"    "M312T36p"    "M357T30p"    "M114T37p"    "M604T163_1p"
## [6] "M176T36p"

tail(mets)

## [1] "M861T34n"    "M325T556_2n" "M113T35_2n"  "M378T543n"   "M1013T35_1n"
## [6] "M338T41n"

length(mets)

## [1] 2203
```

Create age categories:

```r
range(MasterFrame$Age)

## [1]  4.0 28.7

MasterFrame$Category <- ""

MasterFrame$Category[MasterFrame$Age >= 4 & MasterFrame$Age < 7] <-
  "4-7 years"
MasterFrame$Category[MasterFrame$Age >= 7 & MasterFrame$Age < 11] <-
  ">7-11 years"
MasterFrame$Category[MasterFrame$Age >= 11 & MasterFrame$Age < 18] <-
  ">11-18 years"
MasterFrame$Category[MasterFrame$Age >= 18 & MasterFrame$Age < 29] <-
  ">18-29 years"

MasterFrame$Category <- factor(MasterFrame$Category,
                        levels = c("4-7 years",
                                   ">7-11 years",
                                   ">11-18 years",
                                   ">18-29 years"))

table(MasterFrame$Category)

##
##    4-7 years  >7-11 years >11-18 years >18-29 years
##           17           11           34           11
```

3 Get basic statistics

Get some basic statistics, like number of DMD cases and controls by site, age distribution for each group:

```r
##number of DMD cases and controls:
table(MasterFrame$Status)

##
## Control     DMD
##      22      51

##number of people at each site:
table(MasterFrame$Site)

##
##      Australia          Calgary              CNMC U of Pittsburgh
##              1               35                 4                7
```

```
##       UC Davis
##            26
```

```
##number of DMD cases and controls by site:
table(MasterFrame$Status, MasterFrame$Site)
```

```
##
##          Australia Calgary CNMC U of Pittsburgh UC Davis
##    Control        0      16    4               2        0
##    DMD            1      19    0               5       26
```

```
##age distribution in each group:
quantile(MasterFrame$Age[MasterFrame$Status == "DMD"])
```

```
##   0%  25%  50%  75% 100%
##  4.0  6.0 11.4 17.0 28.7
```

```
quantile(MasterFrame$Age[MasterFrame$Status == "Control"])
```

```
##      0%     25%     50%     75%    100%
##  6.0000 11.4125 13.6500 15.6750 17.8000
```

```
table(MasterFrame$Category[MasterFrame$Status == "DMD"])
```

```
##
##    4-7 years  >7-11 years >11-18 years >18-29 years
##          15            8           17           11
```

```
table(MasterFrame$Category[MasterFrame$Status == "Control"])
```

```
##
##    4-7 years  >7-11 years >11-18 years >18-29 years
##           2            3           17            0
```

4 Table 1

Generate Table 1 from paper:

```
tabBySite <- table(MasterFrame$Site, MasterFrame$Status)
tabBySite <- data.frame(tabBySite)
tabBySite
```

```
##              Var1    Var2 Freq
## 1       Australia Control    0
## 2         Calgary Control   16
## 3            CNMC Control    4
```

```
## 4   U of Pittsburgh Control    2
## 5          UC Davis Control    0
## 6          Australia     DMD    1
## 7           Calgary     DMD   19
## 8              CNMC     DMD    0
## 9   U of Pittsburgh     DMD    5
## 10         UC Davis     DMD   26
```

```r
colnames(tabBySite) <- c("Site", "Status", "N")
reshape(tabBySite, timevar="Status", idvar="Site", direction="wide")
```

```
##               Site N.Control N.DMD
## 1         Australia         0     1
## 2           Calgary        16    19
## 3              CNMC         4     0
## 4 U of Pittsburgh         2     5
## 5          UC Davis         0    26
```

```r
tabBySite <- reshape(tabBySite, timevar="Status", idvar="Site", direction="wide")
##reorder to put DMD first
tabBySite
```

```
##               Site N.Control N.DMD
## 1         Australia         0     1
## 2           Calgary        16    19
## 3              CNMC         4     0
## 4 U of Pittsburgh         2     5
## 5          UC Davis         0    26
```

```r
tabBySite <- tabBySite[,c(1,3,2)]
##reorder to sort by total numbers
tabBySite <- tabBySite[order(rowSums(as.matrix(tabBySite[,-1])),
                             decreasing=TRUE), ]

##now have a separate table for age
tabByAge <-
  data.frame(Age = "Age",
             DMD =
               paste(round(quantile(MasterFrame$Age[MasterFrame$Status == "DMD"],
                                    probs = c(0, 0.5, 1)),1), sep="", collapse=", "),
             Controls =
               paste(round(quantile(MasterFrame$Age[MasterFrame$Status == "Control"],
                                    probs = c(0, 0.5, 1)),1), sep="", collapse=", "))

tabByAge
```

```
##   Age            DMD       Controls
## 1 Age 4, 11.4, 28.7 6, 13.6, 17.8
```

```r
##add in total by age categories
tabByAgeGroup <- table(MasterFrame$Category, MasterFrame$Status)
tabByAgeGroup <- data.frame(tabByAgeGroup)
tabByAgeGroup
```

```
##              Var1    Var2 Freq
## 1    4-7 years Control    2
## 2  >7-11 years Control    3
## 3 >11-18 years Control   17
## 4 >18-29 years Control    0
## 5    4-7 years     DMD   15
## 6  >7-11 years     DMD    8
## 7 >11-18 years     DMD   17
## 8 >18-29 years     DMD   11
```

```r
colnames(tabByAgeGroup) <- c("Age", "Status", "N")
reshape(tabByAgeGroup, timevar="Status", idvar="Age", direction="wide")
```

```
##           Age N.Control N.DMD
## 1    4-7 years         2    15
## 2  >7-11 years         3     8
## 3 >11-18 years        17    17
## 4 >18-29 years         0    11
```

```r
tabByAgeGroup <- reshape(tabByAgeGroup, timevar="Status", idvar="Age", direction="wide")
##reorder to put DMD first
tabByAgeGroup
```

```
##           Age N.Control N.DMD
## 1    4-7 years         2    15
## 2  >7-11 years         3     8
## 3 >11-18 years        17    17
## 4 >18-29 years         0    11
```

```r
tabByAgeGroup <- tabByAgeGroup[,c(1,3,2)]
colnames(tabByAge) <- c("Age", "N.DMD", "N.Control")
tabByAge <- sapply(tabByAge, as.character)
tabByAgeGroup <- sapply(tabByAgeGroup, as.character)
tabByAge <- rbind(tabByAge, tabByAgeGroup)
tabByAge
```

```
##           Age            N.DMD           N.Control
## tabByAge "Age"          "4, 11.4, 28.7" "6, 13.6, 17.8"
##          "4-7 years"    "15"            "2"
##          ">7-11 years"  "8"             "3"
##          ">11-18 years" "17"            "17"
##          ">18-29 years" "11"            "0"
```

```
##add in totals at the end
tabByAge <- rbind(tabByAge,
                c("Total", sum(MasterFrame$Status == "DMD"),
                  sum(MasterFrame$Status == "Control")))
tabByAge

##          Age           N.DMD           N.Control
## tabByAge "Age"         "4, 11.4, 28.7" "6, 13.6, 17.8"
##          "4-7 years"   "15"            "2"
##          ">7-11 years" "8"             "3"
##          ">11-18 years" "17"           "17"
##          ">18-29 years" "11"           "0"
##          "Total"       "51"            "22"

rownames(tabBySite) <- rownames(tabByAge) <- NULL
```

```
xtable(tabBySite)
```

	Site	N.DMD	N.Control
1	Calgary	19	16
2	UC Davis	26	0
3	U of Pittsburgh	5	2
4	CNMC	0	4
5	Australia	1	0

```
xtable(tabByAge)
```

	Age	N.DMD	N.Control
1	Age	4, 11.4, 28.7	6, 13.6, 17.8
2	4-7 years	15	2
3	>7-11 years	8	3
4	>11-18 years	17	17
5	>18-29 years	11	0
6	Total	51	22

5   Save objects

```
save(list=c("MasterFrame","mets"), file="MasterFrame.RData")
```

6 Get session info

Session info:

```
sessionInfo()

## R version 3.3.1 (2016-06-21)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 14393)
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] xtable_1.8-2      dplyr_0.5.0        RColorBrewer_1.1-2
## [4] genefilter_1.56.0 reshape_0.8.6      ggplot2_2.2.1
## [7] knitr_1.15.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.9         highr_0.6           plyr_1.8.4
##  [4] bitops_1.0-6        tools_3.3.1         digest_0.6.11
##  [7] annotate_1.52.1     evaluate_0.10       RSQLite_1.1-2
## [10] memoise_1.0.0       tibble_1.2          gtable_0.2.0
## [13] lattice_0.20-33     Matrix_1.2-6        DBI_0.5-1
## [16] parallel_3.3.1      stringr_1.1.0       S4Vectors_0.12.1
## [19] IRanges_2.8.1       stats4_3.3.1        grid_3.3.1
## [22] Biobase_2.34.0      R6_2.2.0            AnnotationDbi_1.36.1
## [25] XML_3.98-1.5        survival_2.40-1     magrittr_1.5
## [28] scales_0.4.1        BiocGenerics_0.20.0 splines_3.3.1
## [31] assertthat_0.1      colorspace_1.3-2    stringi_1.1.2
## [34] RCurl_1.95-4.8      lazyeval_0.2.0      munsell_0.4.3
```