Principal component analysis, generate Figure 1 and S2 Table

March 1, 2017

```
opts_chunk$set(echo=TRUE, fig.path='figures/', cache=FALSE)
```

1 Load all libraries, functions, and data

Load everything, check structure of data objects:

```
library(ggplot2)

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

library(reshape2)

## Warning:  package 'reshape2' was built under R version 3.3.2

library(genefilter)
library(RColorBrewer)
library(xtable)

source("functions.R")

load("MasterFrame.RData")
```

2 Principal components analysis

First perform PCA decomposition:

```
pca <- prcomp(MasterFrame[,mets],
              center=TRUE, scale=TRUE)$x
dim(pca)

## [1] 73 73
```

We also want to save the percent of the total variance that is explained by each component.

```
##get variance of each component
varComp <- diag(cov(pca))
##get fraction of variance explained by each component
varComp[1]/sum(varComp)

##       PC1
## 0.2331287

varComp[2]/sum(varComp)

##        PC2
## 0.05979062

##check this with the output from the R object:
summary(prcomp(MasterFrame[,mets],
               center=TRUE, scale=TRUE))$importance[,1:5]

##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     22.66236 11.47688 9.668406 9.443508 7.626386
## Proportion of Variance  0.23313  0.05979 0.042430 0.040480 0.026400
## Cumulative Proportion   0.23313  0.29292 0.335350 0.375830 0.402230

##add them to MasterFrame object, so it is easier to use aesthetics
MasterFrame <- cbind(MasterFrame,
                     pca[,1:5])
```

2.1 Figure 1

Make PCA plots for age categories and sites side-by-side to generate Figure 1 from paper:

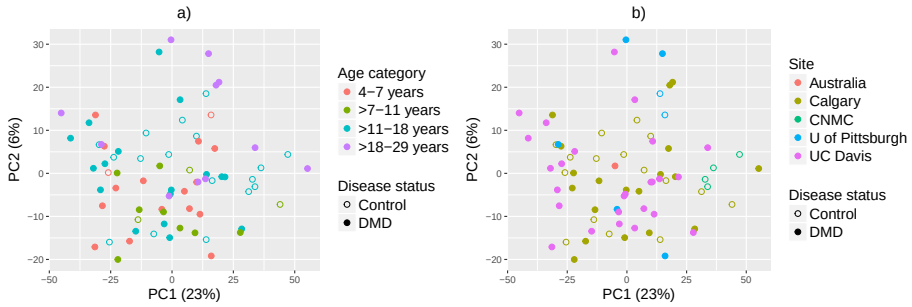
```
ggPlot1 <- ggplot(MasterFrame, aes(x=PC1, y=PC2)) +
  geom_point(size=2.2, aes(shape=Status, color=Category)) +
  scale_color_discrete(name="Age category") +
  scale_shape_manual(values=c(1,19), name="Disease status") +
  xlab(paste("PC1 (", round(varComp[1]/sum(varComp)*100), "%)", sep="")) +
  ylab(paste("PC2 (", round(varComp[2]/sum(varComp)*100), "%)", sep="")) +
  theme(plot.title = element_text(size = 15, hjust = 0.5, vjust=1.5),
        legend.title = element_text(size = 14),
        legend.text = element_text(size=14),
        axis.title = element_text(size=14)) +
  guides(color = guide_legend(order=1),
         shape = guide_legend(order=2)) +
```

```
  ggtitle("a)")

ggPlot2 <- ggplot(MasterFrame, aes(x=PC1, y=PC2)) +
  geom_point(size=2.2, aes(color=Site, shape=Status)) +
  scale_shape_manual(values=c(1,19), name="Disease status") +
  xlab(paste("PC1 (", round(varComp[1]/sum(varComp)*100), "%)", sep="")) +
  ylab(paste("PC2 (", round(varComp[2]/sum(varComp)*100), "%)", sep="")) +
  theme(plot.title = element_text(size = 15, hjust = 0.5, vjust=1.5),
        legend.title = element_text(size = 14),
        legend.text = element_text(size=14),
        axis.title = element_text(size=14)) +
  guides(color = guide_legend(order=1),
         shape = guide_legend(order=2)) +
  ggtitle("b)")

multiplot(ggPlot1, ggPlot2, cols=2)

## Loading required package:  grid
```



2.2 Association of PC1 with different variables

Look to see if PC1 is associated with DMD, age, their interaction, or study site:

```
##make a table for the results
PC1.AIC.df <-
  data.frame(Model = c("DMD status, Age, (DMD status) x Age, Site",
                       "DMD status, Age, (DMD status) x Age",
                       "DMD status, Age, Site",
                       "DMD status, Age",
                       "Age, Site",
                       "DMD status, Site",
                       "DMD status",
                       "Age",
                       "Site"),
```

```
             AIC.PC1 = c(AIC(lm(PC1 ~ Status*Age+Site, data=MasterFrame)),
                         AIC(lm(PC1 ~ Status*Age, data=MasterFrame)),
                         AIC(lm(PC1 ~ Status+Age+Site, data=MasterFrame)),
                         AIC(lm(PC1 ~ Status+Age, data=MasterFrame)),
                         AIC(lm(PC1 ~ Age+Site, data=MasterFrame)),
                         AIC(lm(PC1 ~ Status+Site, data=MasterFrame)),
                         AIC(lm(PC1 ~ Status, data=MasterFrame)),
                         AIC(lm(PC1 ~ Age, data=MasterFrame)),
                         AIC(lm(PC1 ~ Site, data=MasterFrame))))
PC1.AIC.df
```

```
##                                       Model  AIC.PC1
## 1 DMD status, Age, (DMD status) x Age, Site 660.4427
## 2       DMD status, Age, (DMD status) x Age 664.6513
## 3                     DMD status, Age, Site 659.2032
## 4                           DMD status, Age 662.7182
## 5                                 Age, Site 657.2034
## 6                           DMD status, Site 660.9136
## 7                                 DMD status 664.5963
## 8                                       Age 663.9808
## 9                                      Site 658.9139
```

```
##get minimum AIC value
argMinAIC1 <- which.min(PC1.AIC.df$AIC.PC1)
##get information-theoretic interpretation
PC1.AIC.df$probRatio1 <- exp((-PC1.AIC.df$AIC.PC1+PC1.AIC.df$AIC.PC1[argMinAIC1])/2)

PC1.AIC.df <- PC1.AIC.df[,c("Model","AIC.PC1","probRatio1")]
PC1.AIC.df[,-1] <- sapply(PC1.AIC.df[,-1], round, 2)
PC1.AIC.df
```

```
##                                       Model AIC.PC1 probRatio1
## 1 DMD status, Age, (DMD status) x Age, Site  660.44       0.20
## 2       DMD status, Age, (DMD status) x Age  664.65       0.02
## 3                     DMD status, Age, Site  659.20       0.37
## 4                           DMD status, Age  662.72       0.06
## 5                                 Age, Site  657.20       1.00
## 6                           DMD status, Site  660.91       0.16
## 7                                 DMD status  664.60       0.02
## 8                                       Age  663.98       0.03
## 9                                      Site  658.91       0.43
```

2.2.1 S2 Table

```
xtable(PC1.AIC.df[,1:3])
```

	Model	AIC.PC1	probRatio1
1	DMD status, Age, (DMD status) x Age, Site	660.44	0.20
2	DMD status, Age, (DMD status) x Age	664.65	0.02
3	DMD status, Age, Site	659.20	0.37
4	DMD status, Age	662.72	0.06
5	Age, Site	657.20	1.00
6	DMD status, Site	660.91	0.16
7	DMD status	664.60	0.02
8	Age	663.98	0.03
9	Site	658.91	0.43

Look at the top model in more detail:

```
topLM.PC1 <- lm(PC1 ~ Age+as.factor(Site), data=MasterFrame)
summary(topLM.PC1)

##
## Call:
## lm(formula = PC1 ~ Age + as.factor(Site), data = MasterFrame)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.11 -15.73   0.00  12.38  45.63
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -10.7136    20.9116  -0.512    0.610
## Age                            0.7161     0.3831   1.869    0.066 .
## as.factor(Site)Calgary         1.7338    21.0430   0.082    0.935
## as.factor(Site)CNMC           37.8955    23.2575   1.629    0.108
## as.factor(Site)U of Pittsburgh  3.3463    22.3121   0.150    0.881
## as.factor(Site)UC Davis        -5.0700    21.1719  -0.239    0.811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.69 on 67 degrees of freedom
## Multiple R-squared:  0.2247,Adjusted R-squared:  0.1668
## F-statistic: 3.883 on 5 and 67 DF,  p-value: 0.003767

anova(lm(PC1 ~ Age+as.factor(Site), data=MasterFrame),
      lm(PC1 ~ Age, data=MasterFrame))
```

```
## Analysis of Variance Table
##
## Model 1: PC1 ~ Age + as.factor(Site)
## Model 2: PC1 ~ Age
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     67 28669
## 2     71 35102 -4   -6432.7 3.7583 0.0081 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3 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] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] xtable_1.8-2       RColorBrewer_1.1-2 genefilter_1.56.0
## [4] reshape2_1.4.2     ggplot2_2.2.1      knitr_1.15.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.9          AnnotationDbi_1.36.1 magrittr_1.5
##  [4] splines_3.3.1        IRanges_2.8.1        BiocGenerics_0.20.0
##  [7] munsell_0.4.3        lattice_0.20-33      colorspace_1.3-2
## [10] stringr_1.1.0        highr_0.6            plyr_1.8.4
## [13] tools_3.3.1          parallel_3.3.1       Biobase_2.34.0
## [16] gtable_0.2.0         DBI_0.5-1            survival_2.40-1
## [19] digest_0.6.11        lazyeval_0.2.0       assertthat_0.1
## [22] tibble_1.2           Matrix_1.2-6         S4Vectors_0.12.1
```

```
## [25] bitops_1.0-6          RCurl_1.95-4.8       memoise_1.0.0
## [28] RSQLite_1.1-2         evaluate_0.10        labeling_0.3
## [31] stringi_1.1.2         scales_0.4.1         XML_3.98-1.5
## [34] stats4_3.3.1          annotate_1.52.1
```