Data Pre-processing

Rachel Edgar

2015-11-22

Load Libraries

```
library(ggplot2)
library(reshape)
```

Read in the data

```
load("~/Documents/Presentations/graspods_workshop/graspods_redgar_GSE59685_minimal_data.RData")
```

Common Genomics Data Format

```
Betas[1:5,1:5]
```

```
##            GSM1443648 GSM1443565 GSM1443413 GSM1443723 GSM1443621
## cg05460370 0.05387941 0.04514289 0.05233722 0.04979636 0.06733567
## cg22010617 0.80553016 0.83280196 0.85400563 0.83308010 0.82661549
## cg12827601 0.72220884 0.67504912 0.68495113 0.68869547 0.68696993
## cg06931464 0.82932928 0.84416073 0.82663952 0.83383143 0.83831489
## cg02273617 0.05973617 0.06397710 0.07811459 0.06465995 0.05626568
```

```
head(Meta)
```

```
##       series_id        gsm Subject           barcode ad.disease.status
## 15129  GSE59685 GSM1443648      92 6042316066_R02C02           Exclude
## 15049  GSE59685 GSM1443565      75 6042316069_R01C01                AD
## 14909  GSE59685 GSM1443413      34 6042316069_R06C02                AD
## 15204  GSE59685 GSM1443723     109 7786923107_R05C01           Exclude
## 15102  GSE59685 GSM1443621      86 7796806029_R06C01                AD
## 14926  GSE59685 GSM1443432      38 6969568082_R06C02                AD
##       braak.stage    Sex age.blood age.brain            Tissue
## 15129           5 FEMALE        87        92    frontal cortex
## 15049           6 FEMALE        76        79    frontal cortex
## 14909           6   MALE        70        71    frontal cortex
## 15204           3 FEMALE        80        90    frontal cortex
## 15102           5   MALE        75        77    frontal cortex
## 14926           6   MALE        87        88 entorhinal cortex
```

Quality Control (QC)

```
## NA Count in samples
na_count_sample <-sapply(Betas, function(y) sum(length(which(is.na(y)))))
na_count_sample
```

```
## GSM1443648 GSM1443565 GSM1443413 GSM1443723 GSM1443621 GSM1443432
##          0          1          0          0          0          0
## GSM1443318 GSM1443273 GSM1443366 GSM1443694 GSM1443293 GSM1443307
##          1          1          1          0          0          1
## GSM1443431 GSM1443738 GSM1443653 GSM1443719 GSM1443665 GSM1443295
##          0          1          1          0          1          0
## GSM1443560 GSM1443710 GSM1443415 GSM1443696 GSM1443606 GSM1443613
##          1          0          0          0          0          1
## GSM1443559
##          1
```

```
## NA Count in probes
na_count_probe <-sapply(1:nrow(Betas), function(y) length(which(is.na(Betas[y,]))))
Betas[na_count_probe>0,]
```

```
##            GSM1443648 GSM1443565 GSM1443413 GSM1443723 GSM1443621
## cg05973337   0.308197         NA  0.2680379  0.3507813  0.2296172
##            GSM1443432 GSM1443318 GSM1443273 GSM1443366 GSM1443694
## cg05973337  0.2434264         NA         NA         NA  0.2483111
##            GSM1443293 GSM1443307 GSM1443431 GSM1443738 GSM1443653
## cg05973337   0.271292         NA  0.2312583         NA         NA
##            GSM1443719 GSM1443665 GSM1443295 GSM1443560 GSM1443710
## cg05973337  0.2740651         NA  0.1330823         NA  0.1261528
##            GSM1443415 GSM1443696 GSM1443606 GSM1443613 GSM1443559
## cg05973337  0.2159948  0.2171282  0.2651066         NA         NA
```

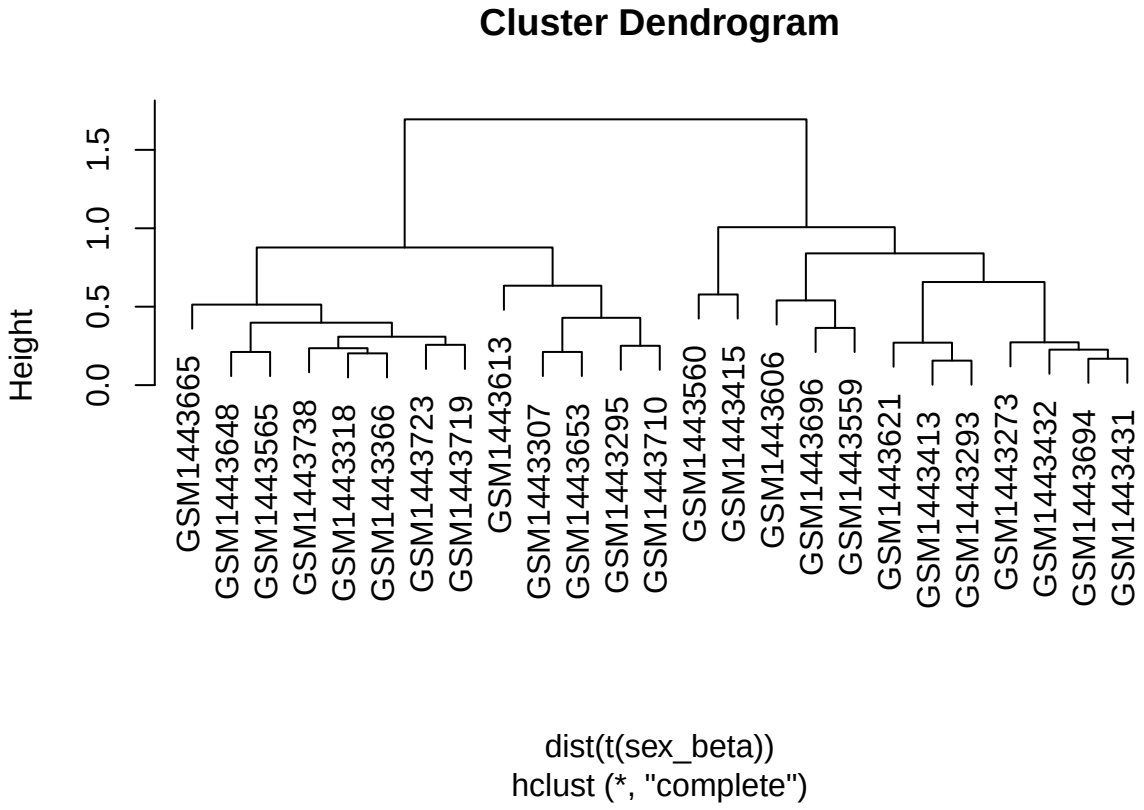
```
## filter to probes with no NAs (stringent threshold should be more relaxed in a full dataset)
Betas_clean<-Betas[na_count_probe==0,]
```

Cluster by sex chromosomes

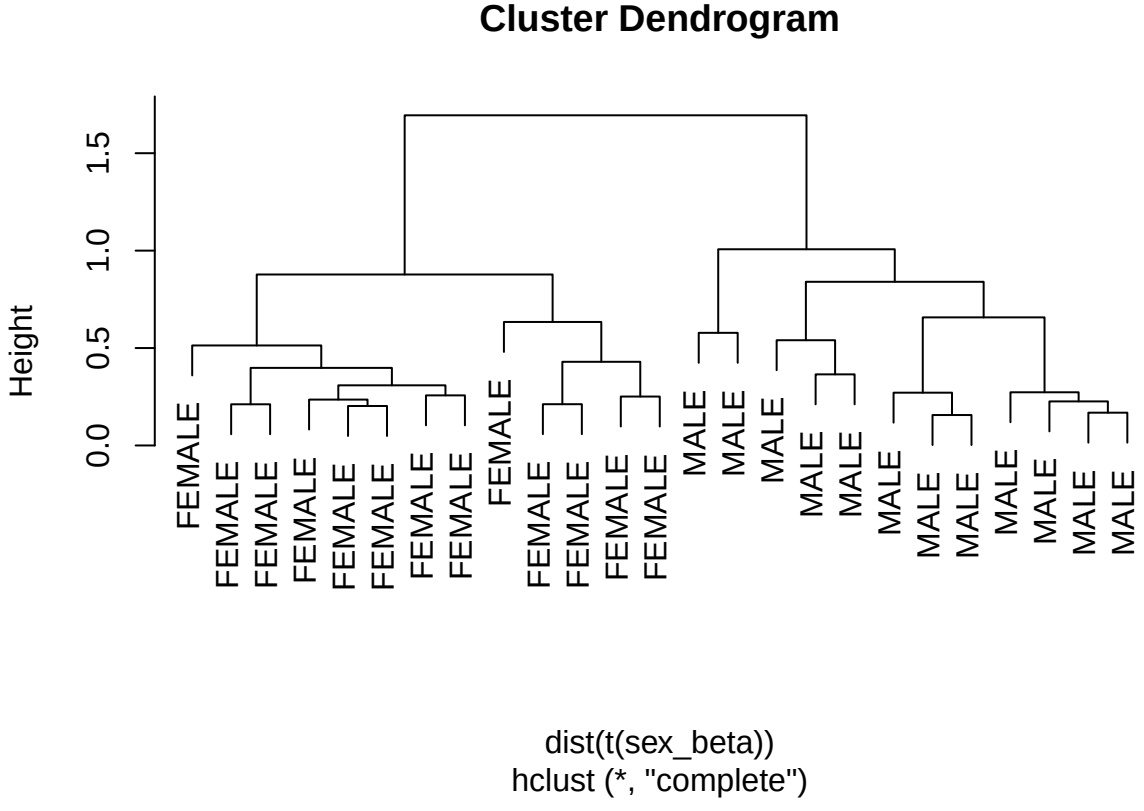
```
# filter to probes on X and Y
# note we are using Betas_clean here

sex<-Annotation[which(Annotation$CHR%in%c("Y","X")),]
sex_beta<-Betas_clean[which(rownames(Betas_clean)%in%sex$ILMNID),]
sex_beta<-sex_beta[complete.cases(sex_beta),]

#cluster and plot
sex_clust <- hclust(dist(t(sex_beta)))
plot(sex_clust)
```




```
## Add fancy labels
plot(sex_clust, labels=Meta$Sex)
```



Distributions of data are a useful general QC tool

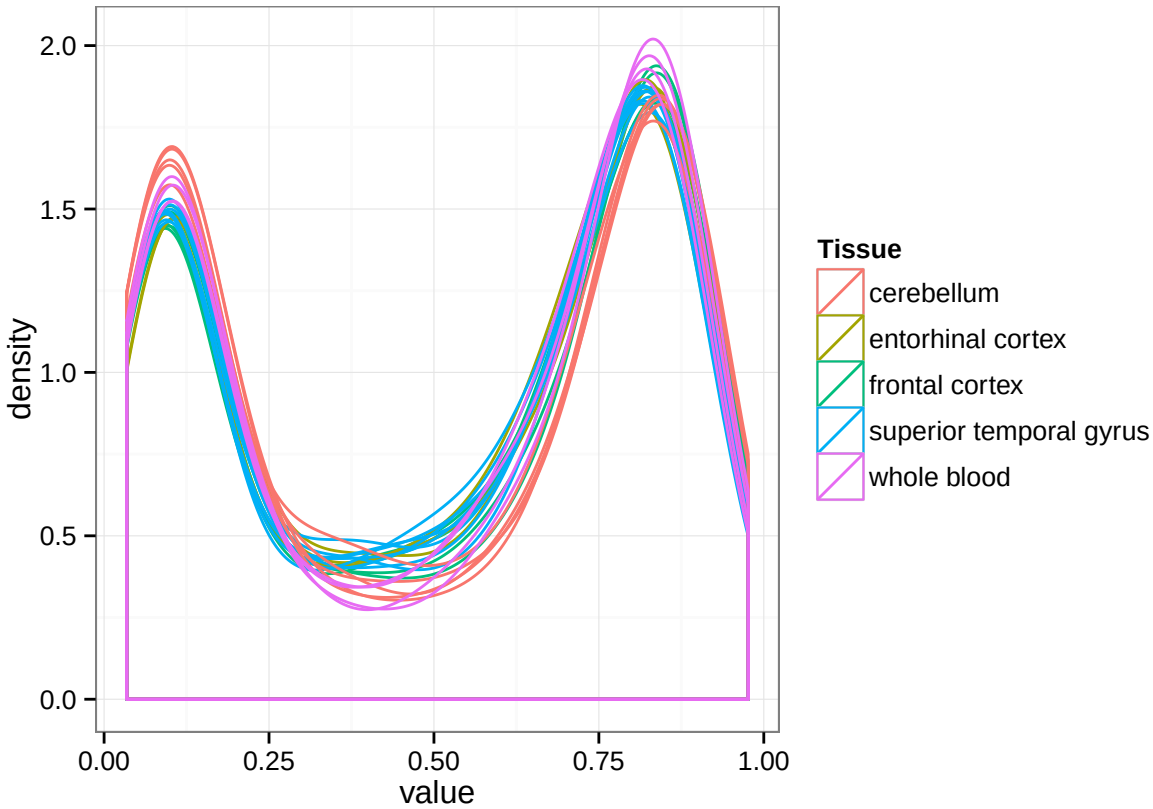
```
## melt is a crazy magic R function that reshapes data
Beta_Plot<- melt(Betas_clean)
```

```
## Using  as id variables
```

```
#add meta data
Beta_Plot<-merge(Beta_Plot,Meta, by.x="variable", by.y="gsm")
head(Beta_Plot)
```

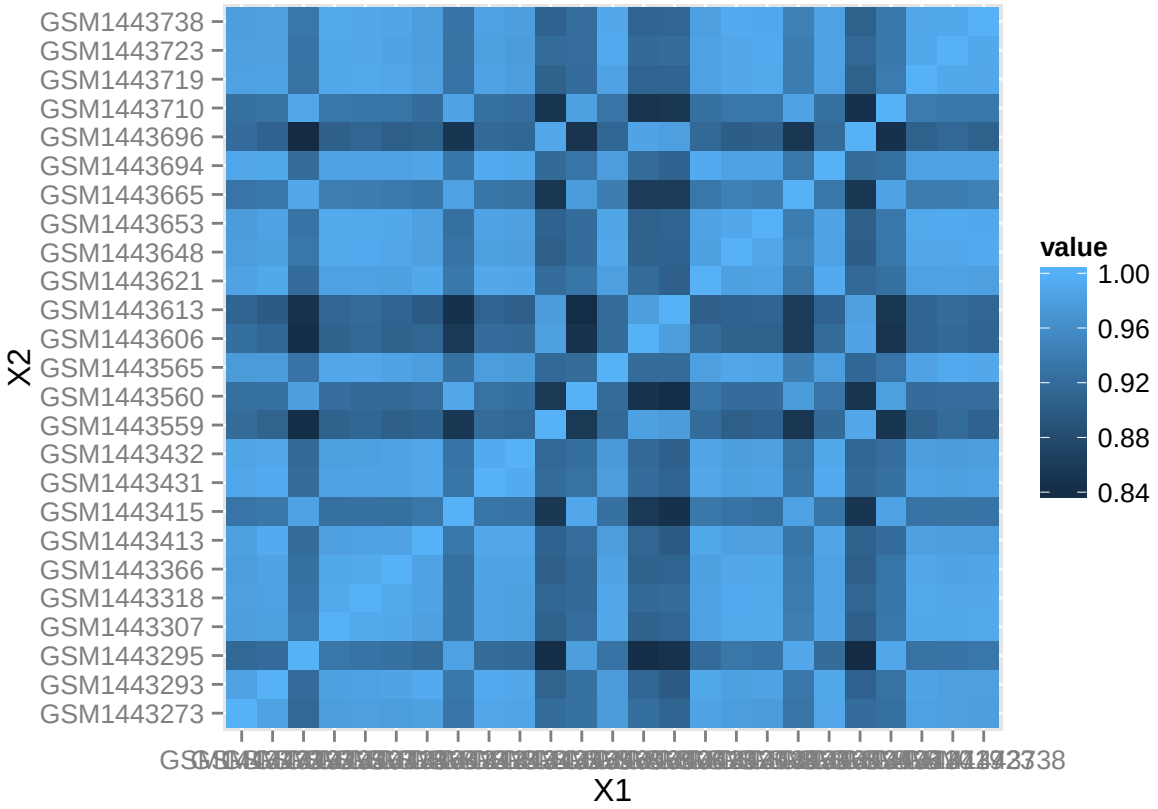
```
##     variable      value series_id Subject           barcode
## 1 GSM1443273 0.05221738  GSE59685       6 6969568084_R02C01
## 2 GSM1443273 0.76780572  GSE59685       6 6969568084_R02C01
## 3 GSM1443273 0.71074176  GSE59685       6 6969568084_R02C01
## 4 GSM1443273 0.87005948  GSE59685       6 6969568084_R02C01
## 5 GSM1443273 0.07554355  GSE59685       6 6969568084_R02C01
## 6 GSM1443273 0.06362516  GSE59685       6 6969568084_R02C01
##   ad.disease.status braak.stage  Sex age.blood age.brain            Tissue
## 1                 C           1 MALE        78        78 entorhinal cortex
## 2                 C           1 MALE        78        78 entorhinal cortex
## 3                 C           1 MALE        78        78 entorhinal cortex
## 4                 C           1 MALE        78        78 entorhinal cortex
## 5                 C           1 MALE        78        78 entorhinal cortex
## 6                 C           1 MALE        78        78 entorhinal cortex
```

```
ggplot(Beta_Plot, aes(value, group=variable, color=Tissue))+geom_density()+theme_bw()
```



Sample correlation heat maps can show outliers

```
qplot(x=X1, y=X2, data=melt(cor(Betas_clean)), fill=value, geom="tile")
```



Normalization

The types are specific and the packages are large. So we won't do it today. See the slides notes for a starting place on researching the best normalization for your data.

Principal Components Analysis

```
PCA_full<-princomp(Betas_clean)
Loadings<-as.data.frame(unclass(PCA_full$loadings))

Loadings[1:5,1:5]
```

```
##                Comp.1       Comp.2     Comp.3     Comp.4      Comp.5
## GSM1443648 -0.1998090  0.014288247 -0.1597080 -0.1918554 -0.02416354
## GSM1443565 -0.2022003 -0.018978660 -0.1180153 -0.2235868  0.21717052
## GSM1443413 -0.2066050 -0.008377242 -0.1750504  0.2366216 -0.25638883
## GSM1443723 -0.2027681 -0.017223556 -0.1226995 -0.1821064  0.31213952
## GSM1443621 -0.2051718 -0.019639417 -0.1349820  0.2399803  0.24838975
```

```r
# Correlate PC1 and age of the sample
cor(Loadings[,1],as.numeric(Meta$age.blood))
```

```
## [1] 0.2054007
```

```r
# ANOVA PC1 and Tissue of the sample
summary(aov(Loadings[,1]~Meta$Tissue))
```

```
##             Df    Sum Sq   Mean Sq F value   Pr(>F)
## Meta$Tissue  4 0.0003698 9.245e-05   12.76 2.55e-05 ***
## Residuals   20 0.0001449 7.240e-06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

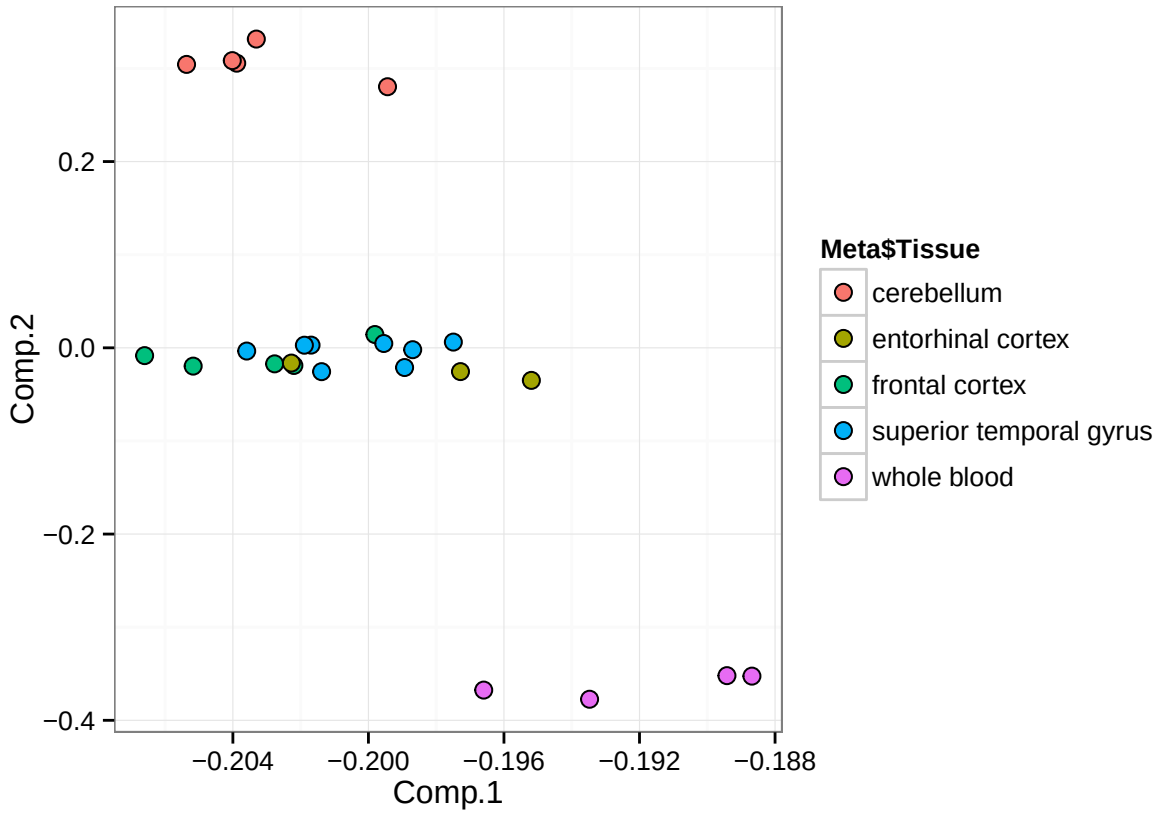
```r
# ANOVA PC2 and Tissue of the sample
summary(aov(Loadings[,2]~Meta$Tissue))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Meta$Tissue  4 0.9959 0.24898    1310 <2e-16 ***
## Residuals   20 0.0038 0.00019
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# ANOVA PC2 and sex of the sample
summary(aov(Loadings[,2]~Meta$Sex))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Meta$Sex     1 0.0552 0.05522   1.345  0.258
## Residuals   23 0.9445 0.04107
```

```r
# its weird probably because it is a tiny dataset...
ggplot(Loadings,aes(Comp.1,Comp.2, fill=Meta$Tissue))+geom_point(shape=21, color="black", size=3)+theme
```


0.2
0.0
−0.2
−0.4
Comp.2
−0.204
−0.200
−0.196
−0.192
−0.188
Comp.1
Meta$Tissue
cerebellum
entorhinal cortex
frontal cortex
superior temporal gyrus
whole blood