PGRNseq Sample and Phenotype Processing

Format the phenotype data for downstream processing.

```
pheno.df <- data.frame(id=nomiss.pheno$SM_ID,
                       sex=as.factor(nomiss.pheno$Sex),
                       ANC=as.numeric(nomiss.pheno$ANC_nadir_1000cells.per.mcL),
                       site=as.factor(nomiss.pheno$Study.site),
                       race=as.factor(nomiss.pheno$Race),
                       dose=as.factor(nomiss.pheno$Dose..mg.))
```

Expirical Distributions of Phenotypes

```
anc.plt <- ggplot(pheno.df, aes(x=ANC)) + geom_histogram()
print(anc.plt + ggtitle('ANC'))
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

```
loganc.plt <- ggplot(pheno.df, aes(x=log10(ANC))) + geom_histogram()
print(loganc.plt + ggtitle('log10(ANC)'))
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

PCA

Prune SNP set for LD, filter on MAF, and run PCA.

```
ld.snpset <- snpgdsLDpruning(exome.geno)
```

```
## SNP pruning based on LD:
##  Sliding window: 500000 basepairs, Inf SNPs
##  |LD| threshold: 0.2
## Removing 5465 non-autosomal SNPs
## Removing 154888 SNPs (monomorphic, < MAF, or > missing rate)
## Working space: 253 samples, 82543 SNPs
## Chromosome 1: 18.98%, 4684/24674
## Chromosome 2: 20.53%, 3529/17190
## Chromosome 3: 20.07%, 2916/14528
## Chromosome 4: 22.99%, 2345/10198
## Chromosome 5: 21.66%, 2451/11314
## Chromosome 6: 17.86%, 2709/15171
```

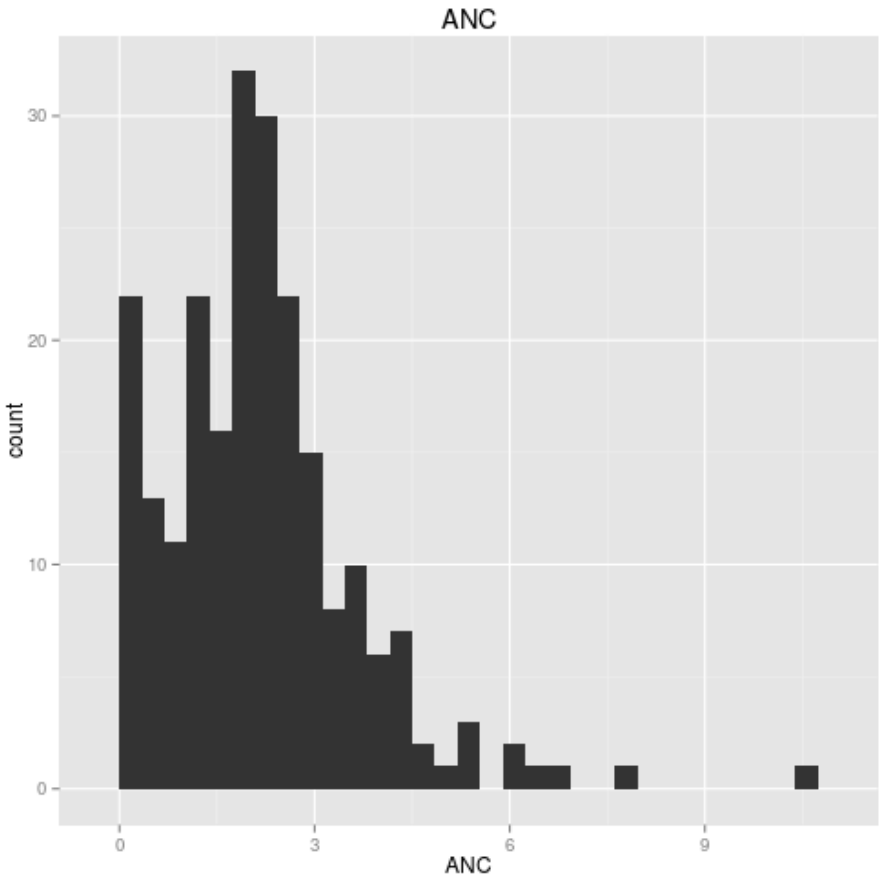


Figure 1: plot of chunk anc_plot

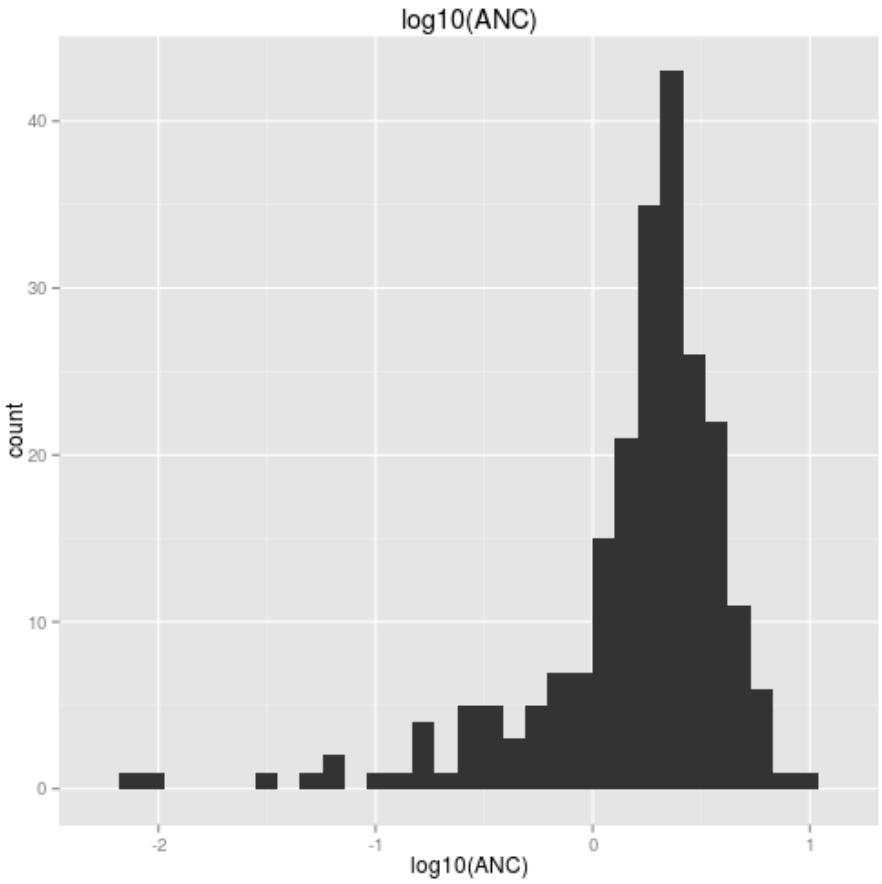


Figure 2: plot of chunk loganc_plt

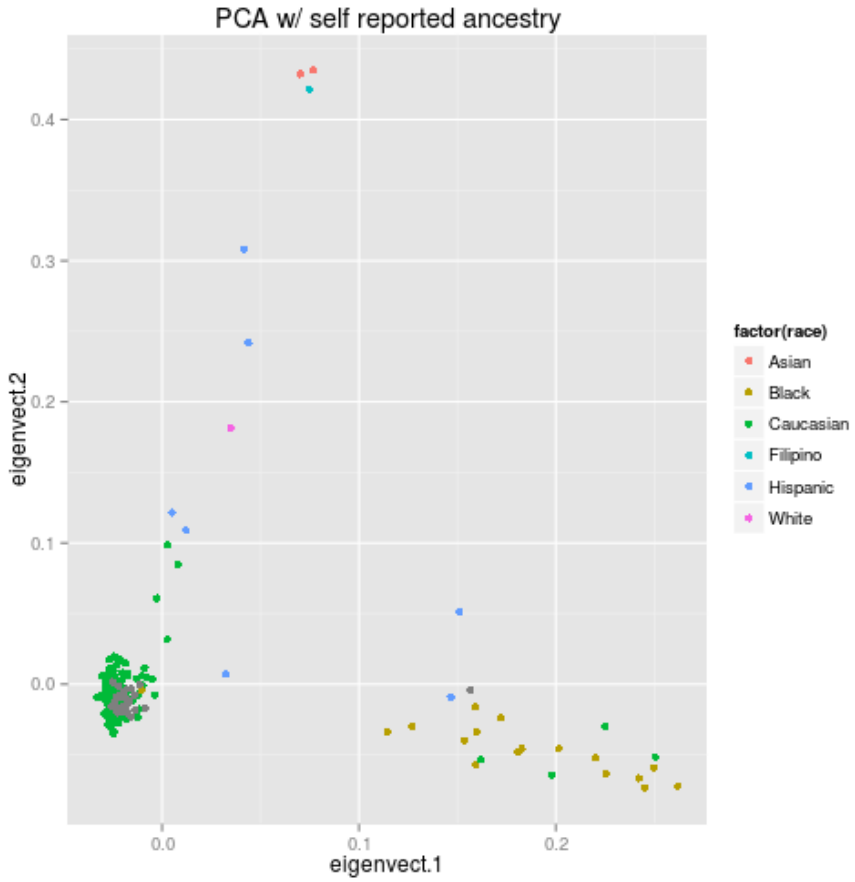
```
## Chromosome 7: 21.35%, 2368/11092
## Chromosome 8: 21.63%, 1930/8921
## Chromosome 9: 20.29%, 2070/10201
## Chromosome 10: 21.79%, 2072/9511
## Chromosome 11: 18.47%, 2868/15528
## Chromosome 12: 19.89%, 2450/12318
## Chromosome 13: 24.91%, 1091/4380
## Chromosome 14: 20.14%, 1555/7721
## Chromosome 15: 19.69%, 1629/8273
## Chromosome 16: 17.80%, 1848/10384
## Chromosome 17: 16.84%, 2183/12963
## Chromosome 18: 24.22%, 908/3749
## Chromosome 19: 15.47%, 2314/14955
## Chromosome 20: 19.38%, 1246/6428
## Chromosome 21: 21.08%, 595/2823
## Chromosome 22: 18.52%, 946/5109
## 46707 SNPs are selected in total.
```

```
pca <- snpgdsPCA(exome.geno, snp.id = unlist(ld.snpset), maf = 0.05)
```

```
## Principal Component Analysis (PCA) on SNP genotypes:
## Removing 35400 SNPs (monomorphic, < MAF, or > missing rate)
## Working space: 253 samples, 11307 SNPs
##  Using 1 CPU core.
## PCA: the sum of all working genotypes (0, 1 and 2) = 1641066
## PCA: Mon Aug 11 14:18:31 2014    0%
## PCA: Mon Aug 11 14:18:32 2014    100%
## PCA: Mon Aug 11 14:18:32 2014    Begin (eigenvalues and eigenvectors)
## PCA: Mon Aug 11 14:18:32 2014    End (eigenvalues and eigenvectors)
```

```
pca.df <- data.frame(sample.id=pca$sample.id, eigenvect=pca$eigenvect)
pheno.df <- merge(pca.df, pheno.df, by.x = "sample.id", by.y = "id")
```

Look at the first two PCs, and color the samples by self-reported ancestry.

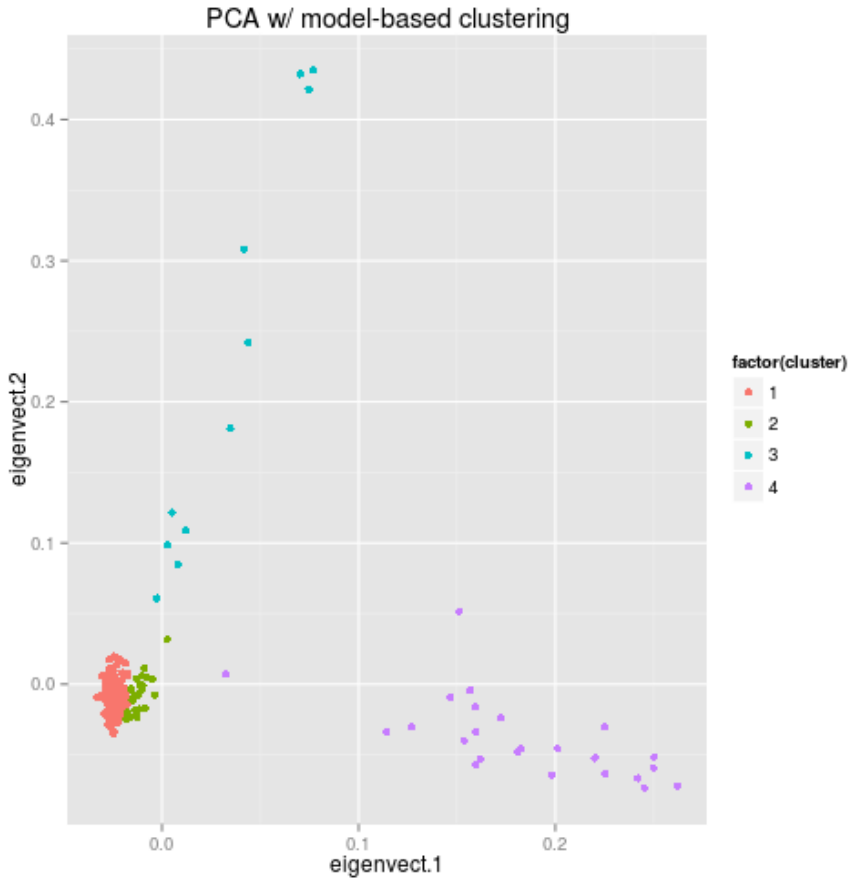


Model-based Clustering

Add principal components to the phenotype data frame, then cluster samples in the first two PCs.

```
names(pheno.df)[which(names(pheno.df) == "sample.id")] <- "id"
cluster3 <- Mclust(pheno.df[,c("eigenvect.1", "eigenvect.2")], G=3) # specify 3 mixture comp
clustern <- Mclust(pheno.df[,c("eigenvect.1", "eigenvect.2")]) # select # of components usi
pheno.df$cluster <- clustern$classification
euroclust <- which(table(pheno.df$cluster) == max(table(pheno.df$cluster))) # euro clust is
pheno.df$iseuro <- pheno.df$cluster == euroclust
```

Examine the results of our PCA and clustering the first two PCs.



Taking a closer look at the clustering of the Europeans. We may be able to explain some genetic variation just by site of ascertainment.

```
## Warning: Removed 35 rows containing missing values (geom_point).
```

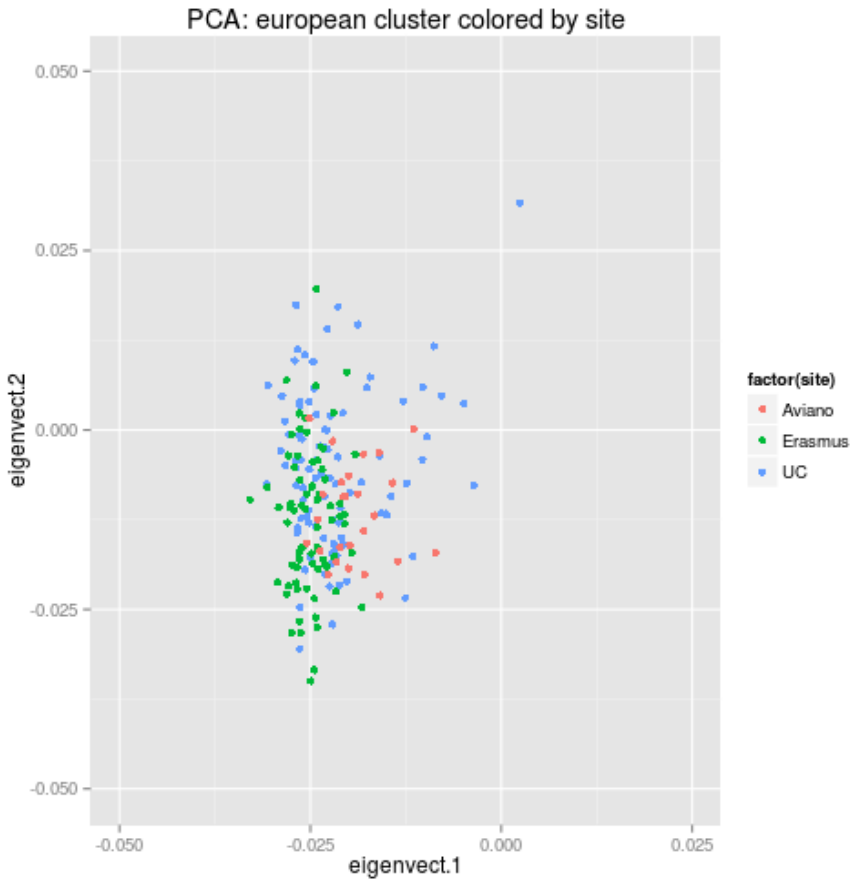


Figure 3: plot of chunk euro_zoom_plot