PGRNseq Sample and Phenotype Processing

Format the phenotype data for downstream processing.

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

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: 18.99%, 4685/24674
## Chromosome 2: 20.44%, 3514/17190
## Chromosome 3: 19.86%, 2885/14528
## Chromosome 4: 23.03%, 2349/10198
## Chromosome 5: 21.44%, 2426/11314
## Chromosome 6: 18.01%, 2732/15171
## Chromosome 7: 21.46%, 2380/11092
## Chromosome 8: 21.58%, 1925/8921
## Chromosome 9: 20.30%, 2071/10201
## Chromosome 10: 21.83%, 2076/9511
## Chromosome 11: 18.62%, 2891/15528
## Chromosome 12: 19.85%, 2445/12318
## Chromosome 13: 24.95%, 1093/4380
## Chromosome 14: 20.01%, 1545/7721
## Chromosome 15: 19.71%, 1631/8273
## Chromosome 16: 17.79%, 1847/10384
## Chromosome 17: 16.82%, 2181/12963
## Chromosome 18: 24.14%, 905/3749
## Chromosome 19: 15.43%, 2308/14955
## Chromosome 20: 19.24%, 1237/6428
## Chromosome 21: 21.15%, 597/2823
```

```
## Chromosome 22: 18.40%, 940/5109
## 46663 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 35356 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) = 1642908
## PCA: Fri Aug  8 14:04:54 2014    0%
## PCA: Fri Aug  8 14:04:54 2014    100%
## PCA: Fri Aug  8 14:04:54 2014    Begin (eigenvalues and eigenvectors)
## PCA: Fri Aug  8 14:04:54 2014    End (eigenvalues and eigenvectors)
```

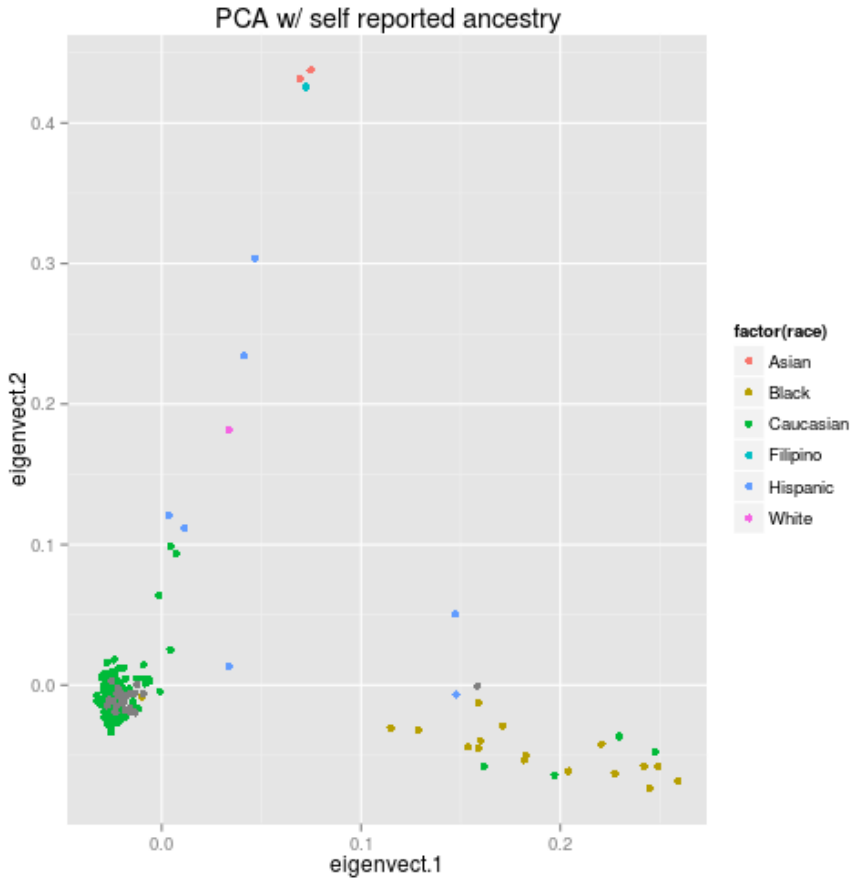
```
pca.df <- data.frame(sample.id=pca$sample.id, eigenvect=pca$eigenvect)
```

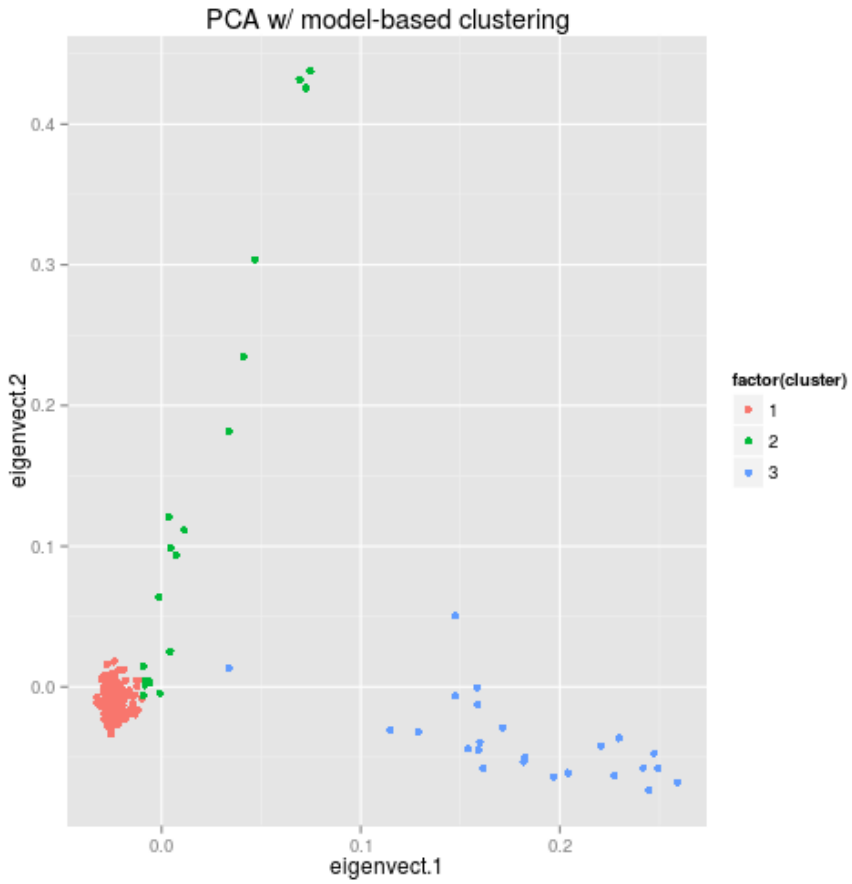
Model-based Clustering

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

```
pheno.df <- merge(pca.df, pheno.df, by.x = "sample.id", by.y = "id")
names(pheno.df)[which(names(pheno.df) == "sample.id")] <- "id"
cluster <- Mclust(pheno.df[,c("eigenvect.1", "eigenvect.2")])
pheno.df$cluster <- cluster$classification
```

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






```
## Warning: Removed 34 rows containing missing values (geom_point).
```

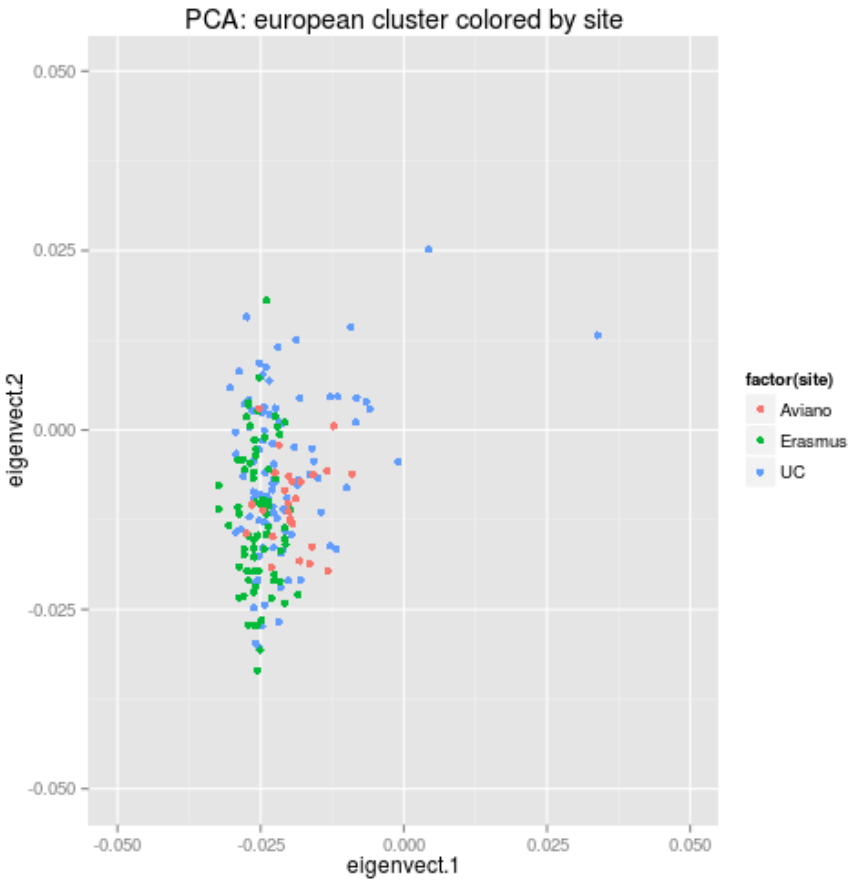


Figure 1: plot of chunk pca_plots