qQTL exercise

Mahdi

September 25, 2020

Part 1

Question 1

In the `sub_geno.tab` file, 0, 1 and 2 most likely represent the two homozygous and hgeteroguzous genotypes. -1 probably means missing data.

Question 2

The `design.tab` file contains information about each column of the `sub_expr.tab` file. It says which population they belong to and other characteristics.

Question 3

Gene expression levels can be explained by the first SNP but not the second. This is because the first SNP has a very small p value for its coefficient meaning the the relationship did not occur by chance. In contrast the second SNP has a high P value so it most likely occured by chance.

Question 4 Do a linear regression for snp__22__43336231 on ENSG00000100266.11

Without covariates

```
gene2 <- "ENSG00000100266.11"
snp2 <- "snp_22_43336231"

gene2_col <- t(gene_expr)[,gene2]
gene2_snp <- t(snps_filtered)[,snp2]
lm_no_cov <- lm(gene2_col ~ gene2_snp)
summary(lm_no_cov)
```

```
##
## Call:
## lm(formula = gene2_col ~ gene2_snp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.367  -5.791  -0.774   4.563  41.890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.8641     0.5297   45.05  < 2e-16 ***
## gene2_snp     3.3238     0.6121    5.43 9.13e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.746 on 460 degrees of freedom
## Multiple R-squared:  0.06024,    Adjusted R-squared:  0.0582
## F-statistic: 29.49 on 1 and 460 DF,  p-value: 9.131e-08
```

Using the genotype PCs from pc_cvrt.tab as covariates

```
pc <- read.table("pc_cvrt.tab")
lm_pc <- lm(gene2_col ~ pc$PC1 + pc$PC2 + pc$PC3 + pc$PC4 + pc$PC5)
summary(lm_pc)
```

```
##
## Call:
## lm(formula = gene2_col ~ pc$PC1 + pc$PC2 + pc$PC3 + pc$PC4 +
##     pc$PC5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.711  -5.961  -0.875   4.280  44.762
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.705859   0.416127  61.774  < 2e-16 ***
## pc$PC1       0.002991   0.004319   0.693  0.48888
## pc$PC2       0.022313   0.014529   1.536  0.12528
## pc$PC3      -0.040613   0.014734  -2.756  0.00608 **
## pc$PC4      -0.011114   0.015847  -0.701  0.48346
## pc$PC5       0.016945   0.016530   1.025  0.30587
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.944 on 456 degrees of freedom
## Multiple R-squared:  0.0256, Adjusted R-squared:  0.01491
## F-statistic: 2.396 on 5 and 456 DF,  p-value: 0.03672
```

Separately for african and non-africans without covariates. Hint: Use the information in the design.tab

```
get_pop_gene <- function(genes, pop, gene_mat, design_mat, inv = F){
  genes_table <- filter_snp_population(pop, gene_mat, design_mat, inv)
  genes <- t(genes_table)[,genes]
  return(genes)
}

#make african model
gene2_africa <- get_pop_gene(gene2, "YRI", gene_expr, design)
snp2_africa <- t(african_snps)[,snp2]
lm_africa <- lm(gene2_africa ~ snp2_africa)
summary(lm_africa)
```

```
##
## Call:
## lm(formula = gene2_africa ~ snp2_africa)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0137  -4.1504  -0.3292   5.0336  19.5839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.3095     0.7353  35.781   <2e-16 ***
## snp2_africa  -0.7181     2.8319  -0.254      0.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.699 on 87 degrees of freedom
## Multiple R-squared:  0.0007385,  Adjusted R-squared:  -0.01075
## F-statistic: 0.0643 on 1 and 87 DF,  p-value: 0.8004
```

```
#make non african model
gene2_nonafrica <- get_pop_gene(gene2, "YRI", gene_expr, design, T)
snp2_nonafrica <- t(non_african_snps)[,snp2]
lm_nonafrica <- lm(gene2_nonafrica ~ snp2_nonafrica)
summary(lm_nonafrica)
```

```
##
## Call:
## lm(formula = gene2_nonafrica ~ snp2_nonafrica)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.922  -5.727  -0.700   4.583  42.142
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     22.8046     0.6598  34.562  < 2e-16 ***
## snp2_nonafrica   4.1310     0.6911   5.978 5.32e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.075 on 371 degrees of freedom
## Multiple R-squared:  0.08785,    Adjusted R-squared:  0.08539
## F-statistic: 35.73 on 1 and 371 DF,  p-value: 5.321e-09
```

Make a dotplot of PC1 vs PC2 and color the dots by population

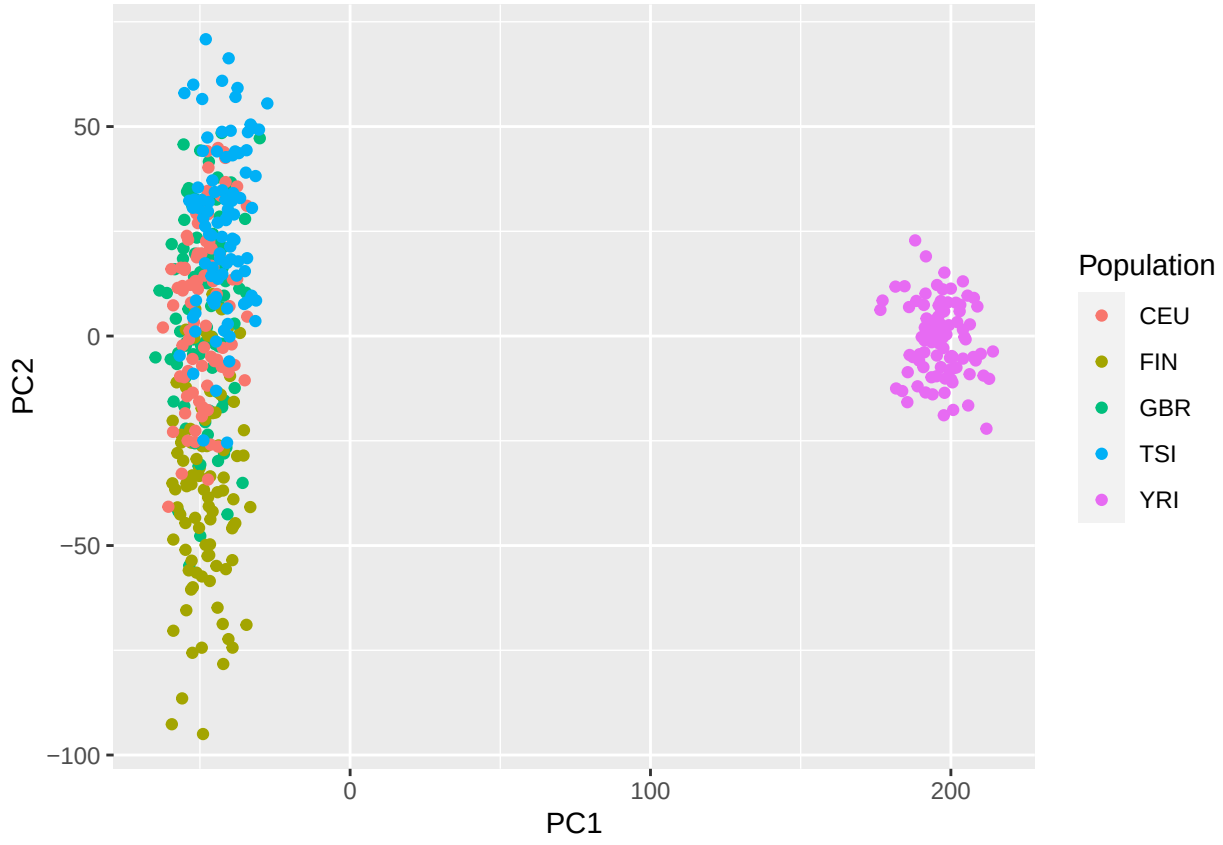
```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
pc_df <- data.frame(PC1=pc$PC1, PC2=pc$PC2, Population=design$Characteristics.population.)
pc_df %>%
  #gather(-Population, key="PC", value="Value") %>%
  ggplot(aes(x=PC1, y=PC2, col=Population)) +
  geom_point()
```



Question 5