Session 07 - Exercises

Deepika Dokuru

workflowr

The R template to do the exercises is here (https://github.com/joellembatchou/SISG2022_Association_Mapping/tree/master/code).

Set your working directory to your home directory using in R*

The data files are in the folder `/data/SISG2022M15/data/`.

Rare Variant Analysis

Introduction

We will look into a dataset collected on a quantitative phenotype which was first analyzed through GWAS and a signal was detected in chromosome 1. Let's determine whether the signal is present when we focus on rare variation at the locus. In our analyses, *we will define rare variants as those with* $MAF \leq 5\%$.

The file "rv_pheno.txt" (https://github.com/joellembatchou/SISG2022_Association_Mapping/tree/master/data)" contains the phenotype measurements for a set of individuals and the file "rv_geno_chr1.bed" is a binary file in PLINK BED format with accompanying BIM and FAM files which contains the genotype data.

Exercises

Here are some things to try:

1. Using PLINK, extract **rare variants** in a new PLINK BED file. (Hint: use options `--max-maf` to select rare variants and `--maj-ref force` so that the minor allele is the effect allele)
2. Load the data in R:

- Load the phenotype data from `rv_pheno.txt`

```
      vars    n mean   sd median trimmed  mad   min  max range  skew
kurtosis
Pheno    3 9949 0.01 1.01   0.02    0.01 1.02 -3.95 3.66  7.61 -0.01
-0.05
        se
Pheno 0.01
```

- Keep only samples who are present both in the genotype as well as phenotype data and who don't have missing values for the phenotype

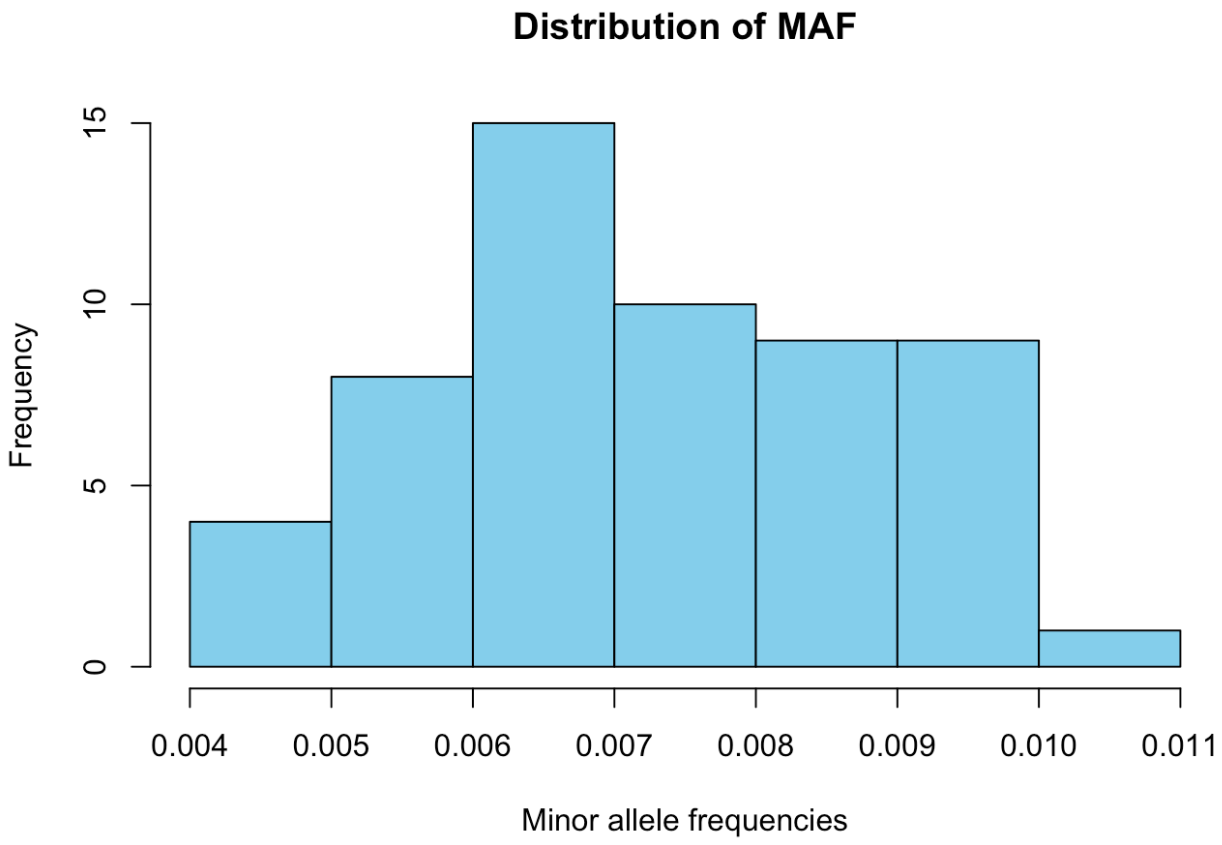
```
[1] "Initial Number of samples: 9953"
```

```
[1] "Initial Number of samples: 9953"
```

```
[1] "Number of samples filtered for missing data: 9949"
```

3. Examine the genotype data:

- Compute the minor allele frequency (MAF) for each SNP and plot histogram. (hint: use `na.rm=TRUE` when calling `mean()`)



- Check for missing values (hint: use function `is.na()` which returns TRUE/FALSE value for missing status)

```
[1] 0
```

4. Run the single variant association tests in PLINK (only for the extracted variants).

- What would be your significance threshold after applying Bonferroni correction for the multiple tests (assume the significance level is 0.05)? Is anything significant after this correction?

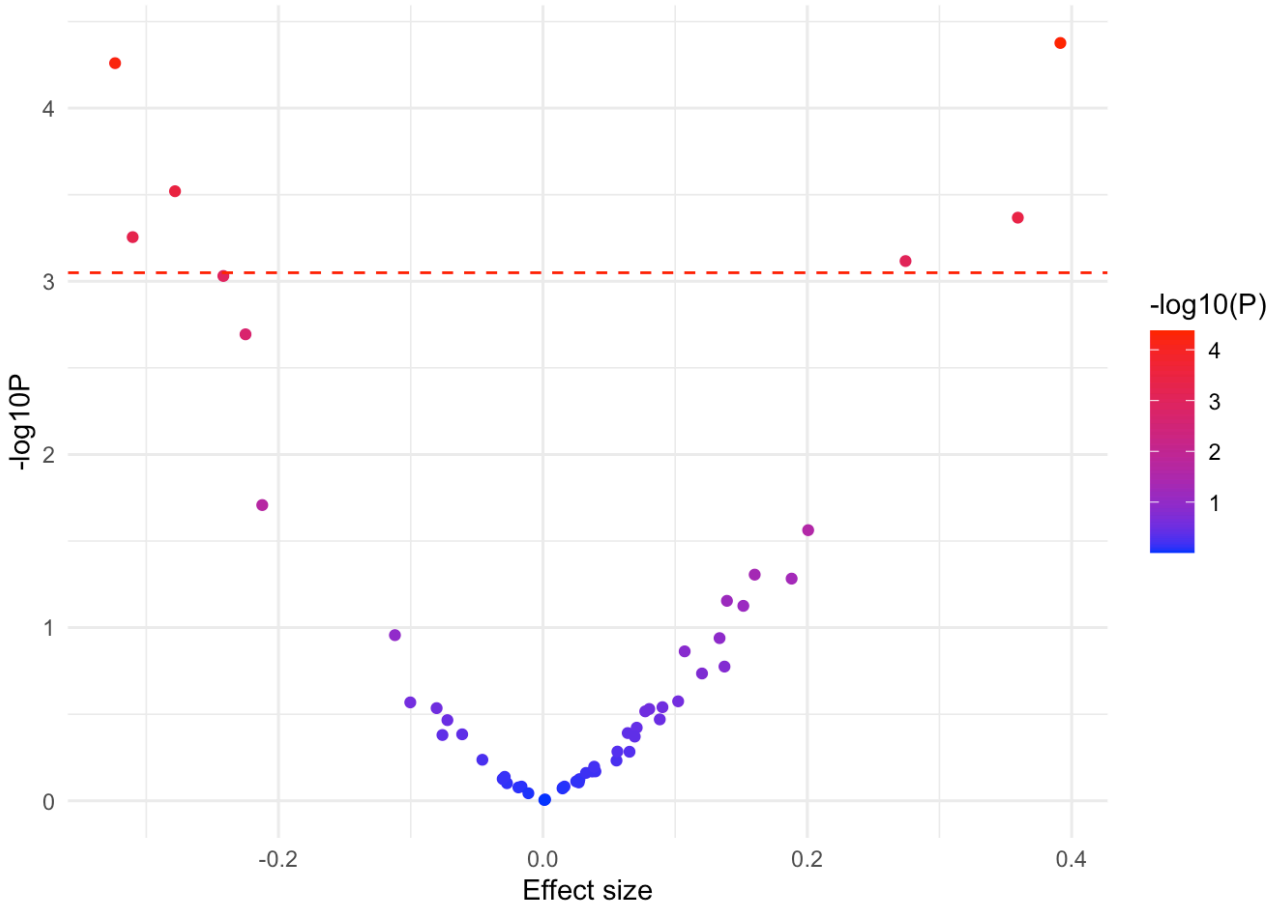
```
[1] "Number of SNPs: 56"
```

```
[1] "alpha= 0.05"
```

```
[1] "Boferonni corrected p: 0.000892857142857143"
```

```
   #CHROM      POS             ID    REF    ALT PROVISIONAL_REF?
A1 OMITTED
    <int>    <int>         <char> <char> <char>           <char> <cha
r>  <char>
1:      1 12639385 1:12639385:G:A      A      G                Y
G       A
2:      1 12057950 1:12057950:C:T      T      C                Y
C       T
3:      1 12734720 1:12734720:A:C      C      A                Y
A       C
4:      1 12405413 1:12405413:T:C      C      T                Y
T       C
5:      1 12360016 1:12360016:G:A      A      G                Y
G       A
6:      1 12183493 1:12183493:G:A      A      G                Y
G       A
      A1_FREQ   TEST OBS_CT      BETA        SE   T_STAT           P
ERRCODE
        <num> <char>  <int>     <num>     <num>    <num>       <num>
<char>
1: 0.00567896    ADD   9949  0.391421 0.0955243  4.09761 4.20759e-05
.
2: 0.00789024    ADD   9949 -0.323574 0.0801934 -4.03492 5.50288e-05
.
3: 0.00849332    ADD   9949 -0.278236 0.0769709 -3.61482 3.02042e-04
.
4: 0.00487486    ADD   9949  0.359249 0.1019850  3.52258 4.29282e-04
.
5: 0.00643281    ADD   9949 -0.310260 0.0898433 -3.45335 5.55975e-04
.
6: 0.00773947    ADD   9949  0.274311 0.0814842  3.36643 7.64368e-04
.
```

- Make a volcano plot (i.e. log10 p-values vs effect sizes). Which of the Burden/SKAT/ACAT tests do you expect will give us most power?




```
[1] "P values lower than Boferonni corrected p: c(FALSE, TRUE)"
[2] "P values lower than Boferonni corrected p: c(50, 6)"
```

5. We will first compare three collapsing/burden approaches:

- CAST (Binary collapsing approach): for each individual, count where they have a rare allele at any of the sites

```
burden.cast
   0    1
4352 5597
```

```
Call:
lm(formula = pheno$Pheno ~ burden.cast)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9531 -0.6880  0.0012  0.6822  3.6581

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.002423   0.015317   0.158    0.874
burden.cast 0.017757   0.020422   0.870    0.385

Residual standard error: 1.01 on 9947 degrees of freedom
Multiple R-squared:  7.6e-05,   Adjusted R-squared:  -2.452e-05
F-statistic: 0.7561 on 1 and 9947 DF,  p-value: 0.3846
```

- MZ Test/GRANVIL (Count based collapsing): for each individual, count the total number of sites where a rare allele is present

```
burden.mz
   0    1    2    3    4    5
4352 3690 1438  373   77   19
```

```
Call:
lm(formula = pheno$Pheno ~ burden.mz)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9521 -0.6894 -0.0013  0.6805  3.6591

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.001444   0.013700   0.105    0.916
burden.mz   0.013492   0.011346   1.189    0.234

Residual standard error: 1.01 on 9947 degrees of freedom
Multiple R-squared:  0.0001421, Adjusted R-squared:  4.162e-05
F-statistic: 1.414 on 1 and 9947 DF,  p-value: 0.2344
```

- Weighted burden test: for each individual, take a weighted count of the rare alleles across sites (for the weights, use `weights <- dbeta(MAF, 1, 25)`)

```
Call:
lm(formula = pheno$Pheno ~ burden.weighted)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9519 -0.6896 -0.0014  0.6804  3.6593

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.0012244  0.0136778   0.090    0.929
burden.weighted 0.0006577  0.0005402   1.217    0.223

Residual standard error: 1.01 on 9947 degrees of freedom
Multiple R-squared:  0.000149,  Adjusted R-squared:  4.846e-05
F-statistic: 1.482 on 1 and 9947 DF,  p-value: 0.2235
```

For each approach, first generate the burden scores vector then test it for association with the phenotype using `lm()` R function.

6. Now use SKAT to test for an association. The basic command would look like

```
[1] "skat.null <- SKAT_Null_Model( <phenotype_vector> ~ 1 , out_type
= 'C')"
```

```
[1] "SKAT( <genotype_matrix>, skat.null )"
```

```
[1] 8.745405e-11
```

7. Run the omnibus SKAT, but consider setting ρ (i.e. `r.corr`) to 0 and then 1.

- Compare the results to using the CAST,MZ/GRANVIL and Weighted burden collapsing approaches in Question 5 as well as SKAT in Question 6. What tests do these ρ values correspond to?

```
        rho0         rho.5         rho1
8.745405e-11 7.405918e-02 2.234603e-01
```

```
[1] "Predictor p value CAST: 0.384579707780373"
```

```
[1] "Predictor p value MZ Test/GRANVIL: 0.234411045400139"
```

```
[1] "Predictor p value Weighted Burden Test: 0.223477961578584"
```

```
[1] "Predictor p value SKAT O at rho 1: 0.223460298922282"
```

8. Now the omnibus version of SKAT, but use the “optimal.adj” approach which searches across a range of rho values.

```
[1] 6.121784e-10
```

9. Run ACATV on the single variant p-values.

```
[1] 0.00112117
```

10. Run ACATO combining the SKAT (only rho 0 and 1) and BURDEN p-values (from Question 7) with the ACATV p-value (from Question 9).

```
[1] 2.623621e-10
```

Session information