LD Thresholding

Christopher B. Cole

June 6, 2016

All of the following uses coRge v0.5.6.9000 as tagged [here](https://github.com/Chris1221/coR-ge/releases/tag/v0.5.6.9000), SHA commit 7148e9e.

## Introduction

We take the list of generated causal SNPs (, ) and use plink 1.9 to find all SNPs within a 500kb region which have with any of the causal SNPs. We take the unique entries in this list and note . We then sweep through all levels of LD c(0, 0.2, 0.4, 0.6, 0.8, 0.9, 1) and correct accordingly:

Correction function:

for(th in levels(as.factor(strata$ld))){  
  
 message("Calculating sFDR and FDR")  
  
 strata %>%  
 filter(s == 1) %>%  
 mutate(p.adj = p.adjust(P, method = "BH")) %>%  
 fdr(., mode = "ld", level = th) ->  
 s1  
  
 strata %>%  
 filter(s == 2) %>%  
 mutate(p.adj = p.adjust(P, method = "BH")) %>%  
 fdr(., mode = "ld", level = th) ->  
 s2  
  
 strata %>%  
 mutate(p.adj = p.adjust(P, method = "BH")) %>%  
 fdr(., mode = "ld", level = th) ->  
 agg  
  
 sfdr <- (s1[1]+s2[1]) / (s1[1]+s2[1] + s1[2]+s2[2])  
 fdr <- agg[3]  
  
 out[nrow(out)+1,] <- c(sfdr, fdr, "all", th)  
 }

and fdr():

else if(mode == "ld"){  
  
 fp <- sum(!is.na(df$p.adj[(df$p.adj < 0.05 & !(df$h1)) | (df$p.adj < 0.05 & df$ld > level)]))  
 tp <- sum(!is.na(df$p.adj[(df$p.adj < 0.05 & df$h1) | (df$p.adj < 0.05 & df$ld > level)]))  
 fdr <- fp / (tp + fp)  
  
 }  
  
 return(c(fp, tp, fdr))

## Results

#### Omnibus

After 30 runs (not all of which finished), we see the following:

ld <- fread("~/repos/coR-ge/data/test\_run2.txt", h = T)  
head(ld)

## sfdr fdr k th  
## 1: 0.5683837 0.5681416 all 0.0  
## 2: 0.5906433 0.5903614 1 0.0  
## 3: 0.5033259 0.5031447 2 0.0  
## 4: 0.6226415 0.6226415 3 0.0  
## 5: 0.6015038 0.6030534 4 0.0  
## 6: 0.6286837 0.6281800 all 0.2

We reformat the data from wide to long:

ld %<>% gather(key = "s", value = "fdr", sfdr, fdr)  
  
ld$s <- as.factor(ld$s)  
ld$k <- as.factor(ld$k)  
ld$th <- as.factor(ld$th)  
  
head(ld)

## k th s fdr  
## 1 all 0 sfdr 0.5683837  
## 2 1 0 sfdr 0.5906433  
## 3 2 0 sfdr 0.5033259  
## 4 3 0 sfdr 0.6226415  
## 5 4 0 sfdr 0.6015038  
## 6 all 0.2 sfdr 0.6286837

And perform a regression to examine the differences:

ld %>% lm(fdr ~ s + k + th, .) %>% summary

##   
## Call:  
## lm(formula = fdr ~ s + k + th, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.89540 -0.05118 -0.00104 0.05188 0.37957   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.546641 0.011960 45.706 < 2e-16 \*\*\*  
## ssfdr 0.004501 0.007373 0.610 0.542   
## k2 -0.006208 0.012350 -0.503 0.615   
## k3 0.081366 0.011562 7.037 3.85e-12 \*\*\*  
## k4 0.012644 0.011093 1.140 0.255   
## kall 0.016002 0.011093 1.443 0.150   
## th0.2 0.061146 0.012768 4.789 1.95e-06 \*\*\*  
## th0.4 0.112063 0.012768 8.777 < 2e-16 \*\*\*  
## th0.6 0.187109 0.012768 14.655 < 2e-16 \*\*\*  
## th0.8 0.241544 0.012768 18.919 < 2e-16 \*\*\*  
## th0.9 0.336113 0.012768 26.325 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.112 on 913 degrees of freedom  
## (96 observations deleted due to missingness)  
## Multiple R-squared: 0.5228, Adjusted R-squared: 0.5175   
## F-statistic: 100 on 10 and 913 DF, p-value: < 2.2e-16

We can see that stratefied / non stratefied only leads to a small 0.5% decrease in FDR.

#### Pairwise

Despite the fact that as a whole the differences were not significant, in some subsets, the differences are much more pronounced, leading to a difference of between 2 and 3% reduction in FDR. With increased number of trials, this will probably become significant.

ld %>% filter(k == 3) %>% filter(th == 0.4) %>% t.test(fdr ~ s, data = .)

##   
## Welch Two Sample t-test  
##   
## data: fdr by s  
## t = -0.29649, df = 26.876, p-value = 0.7691  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.1388806 0.1038184  
## sample estimates:  
## mean in group fdr mean in group sfdr   
## 0.7412383 0.7587694

In general, I think that we have too low power to reliably see the difference between FDR and sFDR. Either that, or the 50/50 split is not enough to actually prioritize the results and see a difference.

This is , split = 50 50 causal, 50 50 non-causal.