Generalization analysis

Tamar Sofer

July 2017

What is generalization?

"Generalization" is the replication of an association between a genetic variant and a trait, discovered in one population, to another population.

- Most genetic association studies were performed in populations of European Ancestry (EA)
- These are often detected in very large GWAS (e.g. 100,000 individuals)

Why perform generalization analysis?

There are multiple reasons.

- To know, whether associations that were discovered in one populations exists in another.
  - This may not always be true...
- To gain power by limiting the number of variants tested for associations to those already previously reported.
- Because we need to perform *replication* analysis, but we do not have access to an independent study with the same type of population and/or the same trait.

Generalization analysis

- An intuitive approach to generalization analysis:
  - Take the list of SNP associations reported in a paper
  - Test the same SNPs with the same trait in your data
  - Report the significant associations.
- What should be the p-value threshold to report associations?

Wait for it...

Generalization analysis

- We developed a generalization testing framework that originated in the replication analysis literature.
- We combine test results (p-values) from both the discovery study, and our study (the follow-up)
  - and calculate an r-value.
  - (for every SNP).
- These r-values take into account multiple testing (of both studies),
- And are used like p-values.
- Since they are already adjusted for multiple testing, an association is generalized if the r-value< 0.05.

Generalization analysis

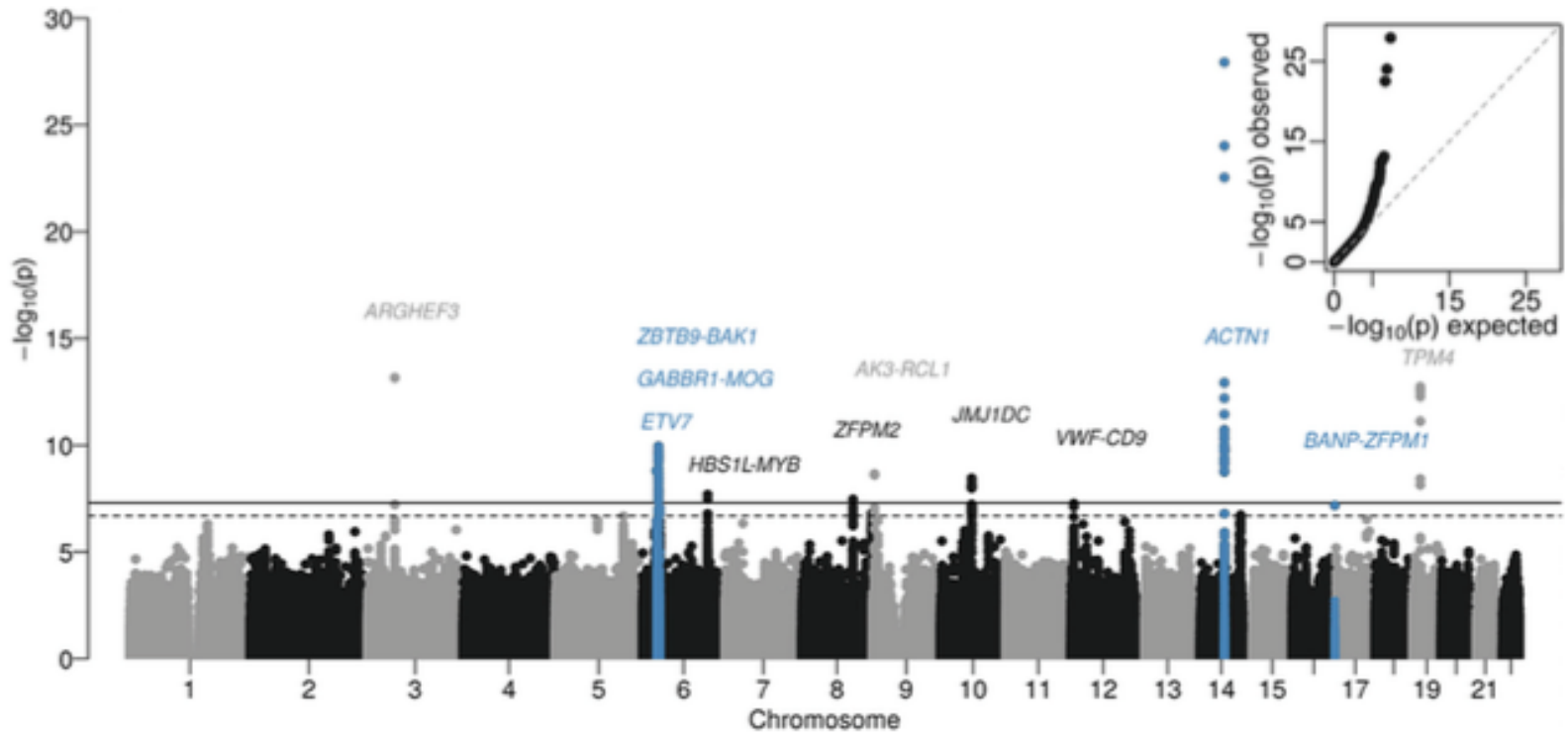
- The generalization framework also takes into account the direction of associations.
- If the estimated association is negative in one study, and positive in the other, the association will not generalize.

Follow-up \ Discovery	Left	Null	Right
Left	$(-1,-1)$	$(0,-1)$	$(1,-1)$
Null	$(-1,0)$	$(0,0)$	$(1,0)$
Right	$(-1,1)$	$(0,1)$	$(1,1)$

- Here, the cells in gray represent generalized associations.

Generalization analysis - platelet count example

- Suppose that we ran a GWAS of platelet count in the HCHS/SOL.
- The results are displayed in the Manhattan plot:



Generalization analysis - platelet count example

- The platelet GWAS discovered 5 new associations
  - that were then replicated in independent studies.
  - There was another association that did not replicate.
  - And there were a few additional known associations that were statistically significant.
- What about 55 other associations that were previously reported in other papers, reporting GWAS in other populations?
  - Generalization analysis!

Generalization analysis - platelet count example

- The generalization R package have an example from the HCHS/SOL platelet count paper.
- We first load this package. (Install it if you haven't already!)

```
#library(devtools)
#install_github("tamartsi/generalize@Package_update",
#          subdir = "generalize")
require(generalize)
```

Generalization analysis - let's do it!

- The generalization R package has an example data set.
- It has results reported by Geiger et al., 2011, and matched association results from the HCHS/SOL.
- Generalization analysis is done for one study at a time.

```
# load the data set from the package
data("dat")
# look at the column names:
matrix(colnames(dat), ncol = 3)
```

```
##      [,1]              [,2]              [,3]
## [1,] "rsID"            "study1.beta"     "study2.alleleB"
## [2,] "chromosome"      "study1.se"       "study2.beta"
## [3,] "position"        "study1.pval"     "study2.se"
## [4,] "study1.alleleA" "study1.n.test"   "study2.pval"
## [5,] "study1.alleleB" "study2.alleleA" "Ref"
```

Generalization analysis - let's do it!

- The data.frame with the example provides all information we need for generalization analysis.

```
head(dat)
```

```
##          rsID chromosome  position study1.alleleA study1
## 1  rs2336384          1  12046062              G
## 2 rs10914144          1 171949749              T
## 3  rs1668871          1 205237136              C
## 4  rs7550918          1 247675558              T
## 5  rs3811444          1 248039450              C
## 6  rs1260326          2  27730939              T
##   study1.beta study1.se study1.pval study1.n.test study2
## 1       2.172     0.382    1.25e-08       2710000
## 2       3.417     0.487    2.22e-12       2710000
## 3       2.804     0.368    2.59e-14       2710000
## 4       3.133     0.471    2.91e-11       2710000
## 5       3.346     0.574    5.60e-09       2710000
```

Generalization analysis - let's do it!

```
dat.matched <- matchEffectAllele(dat$rsID,
                      study2.effect = dat$study2.beta,
                      study1.alleleA = dat$study1.alleleA,
                      study2.alleleA = dat$study2.alleleA,
                      study1.alleleB = dat$study1.alleleB,
                      study2.alleleB = dat$study2.alleleB)
```

```
## passed data entry checks, orienting the effects of study
```

Generalization analysis - let's do it!

```
head(dat.matched)
```

```
##         snpID study2.effect study1.alleleA  flip strand.a
## 1  rs2336384     1.1164496              G FALSE
## 2 rs10914144     1.9402873              T FALSE
## 3  rs1668871    -0.4107451              C  TRUE
## 4  rs7550918     0.9727501              T  TRUE
## 5  rs3811444     3.4528058              C FALSE
## 6  rs1260326     2.5336998              T FALSE
```

Generalization analysis - let's do it!

```
dat$study2.beta <- dat.matched$study2.effect
dat$alleleA <- dat$study1.alleleA
dat$alleleB <- dat$study1.alleleB
dat$study1.alleleA <- dat$study1.alleleB <-
      dat$study2.alleleA <- dat$study2.alleleB <- NULL
head(dat)
```

```
##          rsID chromosome  position study1.beta study1.se
## 1  rs2336384          1  12046062       2.172     0.382
## 2 rs10914144          1 171949749       3.417     0.487
## 3  rs1668871          1 205237136       2.804     0.368
## 4  rs7550918          1 247675558       3.133     0.471
## 5  rs3811444          1 248039450       3.346     0.574
## 6  rs1260326          2  27730939       2.334     0.381
##   study1.n.test study2.beta study2.se  study2.pval
## 1       2710000   1.1164496 0.8084368 0.1672795709 Giege
## 2       2710000   1.9402873 0.9881444 0.0495803692 Giege
## 3       2710000  -0.4107451 0.9386512 0.6616829698 Giege
```

Generalization analysis - let's do it!

- Test for generalization:

```
gen.res <- testGeneralization(dat$rsID, dat$study1.pval,
                    dat$study2.pval, dat$study1.n.test[1],
                    study1.effect = dat$study1.beta,
                      study2.effect = dat$study2.beta,
                    directional.control = TRUE,
                      control.measure = "FDR" )
```

```
## Controlling FDRat the 0.05 level

## Generating one-sided p-values guided by study1's direct

## Calcluating FDR r-values...
```

Generalization analysis - let's do it!

```
head(gen.res)
```

```
##         snpID    gen.rvals generalized
## 1  rs2336384 0.2422669647       FALSE
## 2 rs10914144 0.0867656461       FALSE
## 3  rs1668871 1.0000000000       FALSE
## 4  rs7550918 0.3542344549       FALSE
## 5  rs3811444 0.0005575808        TRUE
## 6  rs1260326 0.0093521516        TRUE
```

Generalization analysis - let's do it!

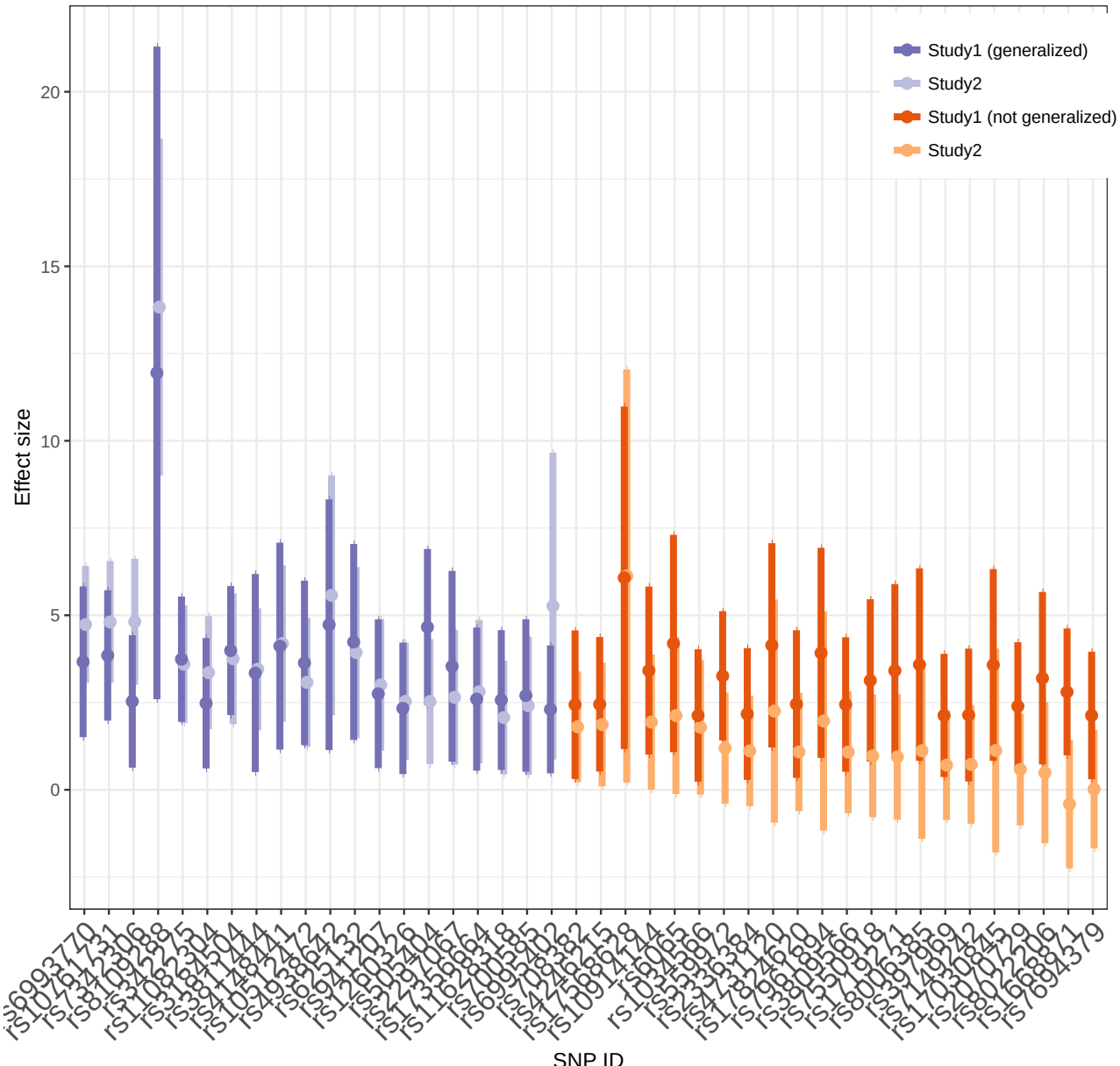
- Create a figure:

```
require(ggplot2,quietly = TRUE)
require(gridExtra,quietly = TRUE)
require(RColorBrewer,quietly = TRUE)
figure.out <-  paste0(getwd(),
                      "/Generalization_example.pdf")

prepareGenResFigure(dat$rsID, dat$study1.beta,
          dat$study1.se, dat$study2.beta, dat$study2.se,
          gen.res$generalized, gen.res$gen.rvals,
          dat$study1.n.test[1],
          output.file = figure.out,
            study1.name = "Study1",
          study2.name = "Study2")
```

Generalization analysis - let's do it!

- Look look at our figure!



Generalization analysis - more considerations

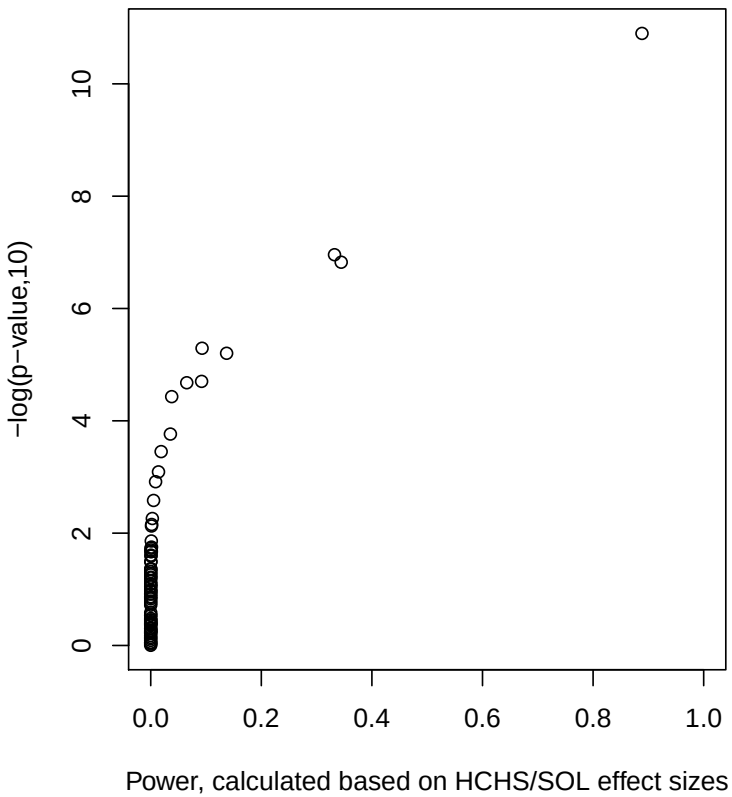
- Coverage of the confidence intervals... depends on the number of tests!
  - e.g. $(1 - \alpha/10) \times 100\%$ for 10 tests in a study for Bonferroni-type coverage.
  - There are other options, controlling "False coverage rate", more complicated.
- Generalization of only "lead SNPs" compared to all SNPs with *p*-value below some threshold.
  - Lead SNP in EA GWAS may be correlated with the causal SNP in EA, but not with Hispanics/Latinos!
- Non-generalization due to lack of power.
  - Summarize information across non-generalized associations, e.g.:
  - Test consistency of direction of associations between the discovery study and HCHS/SOL;
  - Test trait association with Genetic Risk Score (GRS) - GRS can be generated as the sum of reported trait-increasing alleles. Test a GRS composed solely of SNP alleles of non-generalized associations.

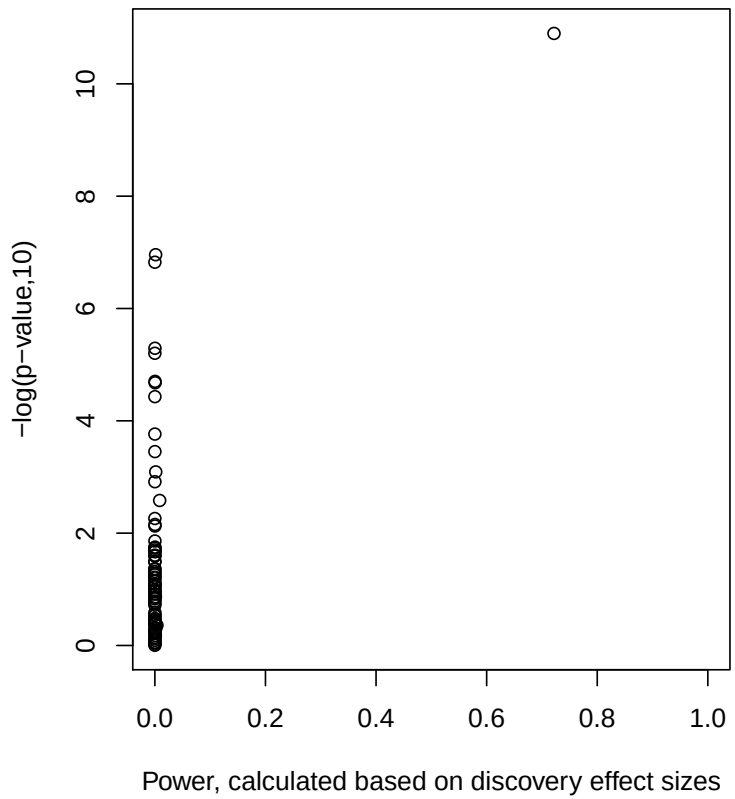
Examples from our work - diabetes

- We ran a GWAS of Diabetes in the HCHS/SOL.
  - Reported in Qi et. al. (2017) "Genetics of Type 2 Diabetes in US Hispanic/Latino Individuals: Results from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL)", *Diabetes*.
- The GWAS identified two genome-wide significant associations (p-value$< 5 \times 10^{-8}$) in known regions.
  - There were 76 known independent associations at the time.
  - The power to detect these associations at the p-value$< 5 \times 10^{-8}$ was low.

Examples from our work - diabetes

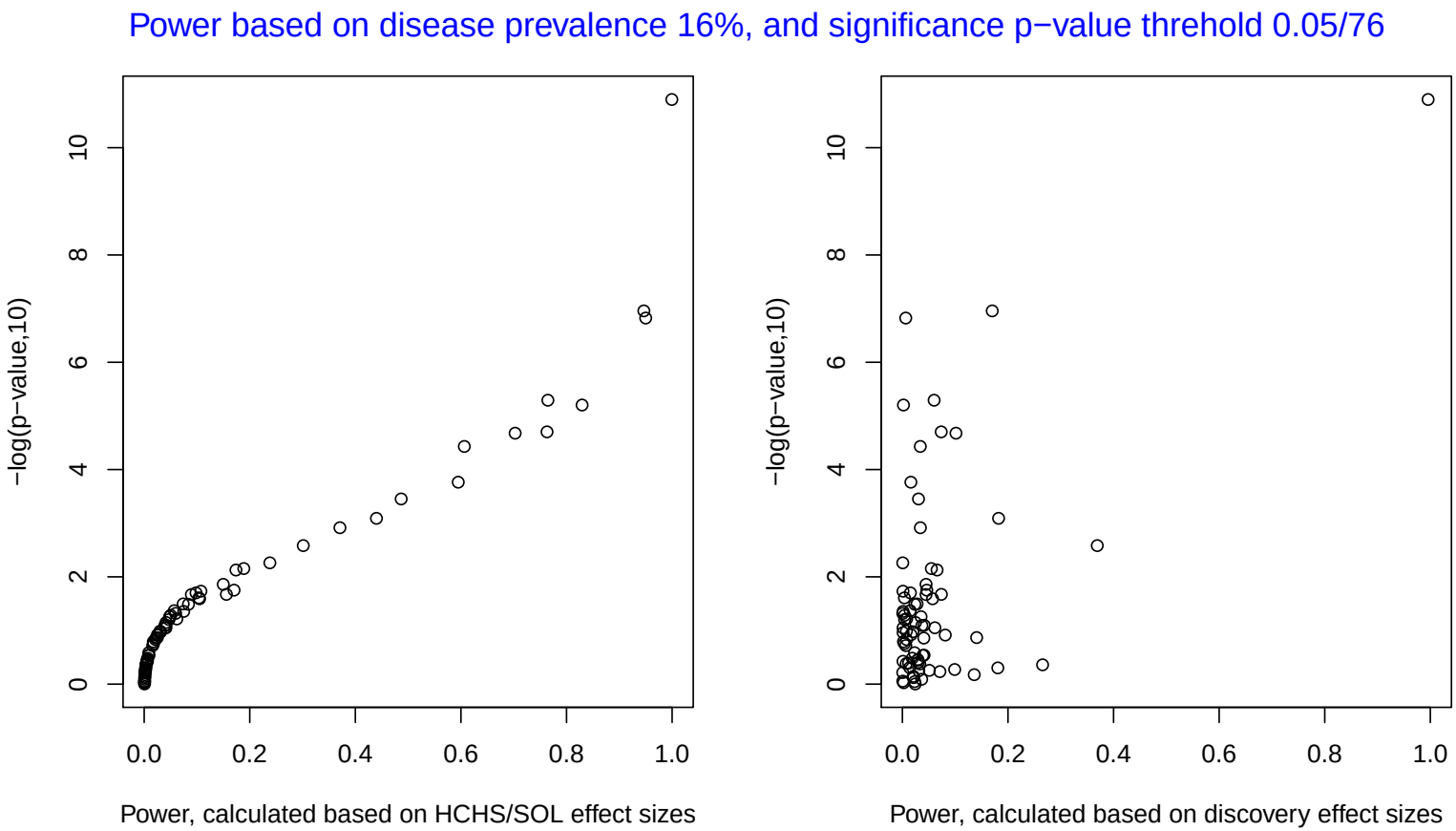
Power based on disease prevalence 16%, and significance p−value threhold 5e−8





Examples from our work - diabetes

- We approximated the power to detect the associations in generalization analysis using Bonferroni threshold.



- The post-hoc power (left) was higher because actual effect sizes in the HCHS/SOL were higher than those reported in the (mainly) European ancestry discovery population.

Examples from our work - diabetes

- 14 of the associations generalized in generalization analysis.

Question: could other associations generalize if we had more power?

- To address this, we constructed a GRS by summing all non-generalized diabetes risk-alleles for all participants in the analysis.
- And tested the association of this GRS with diabetes.
- The resulting p-value$=6.12 \times 10^{-14}$.

Examples from our work - total cholesterol (TC)

- In the generalization manuscript we investigated approaches for generalization when entire GWAS is available
  - Compared to the case where only lead SNPs are available.
  - Reported in Sofer et. al. (2017), "A powerful statistical framework for generalization testing in GWAS, with application to the HCHS/SOL", *Genetic Epidemiology*.
- The GLGC consortium published a list of 74 lead SNPs, from 74 genomic regions, in Willer et al. (2013).
  - European Ancestry (EA); $\sim 190,000$ individuals.
- In addition, the complete results from Willer et al.'s analysis are freely available online.
- In generalization analysis applied on these 74 SNPs 33 SNPs generalized.

Examples from our work - total cholesterol (TC)

- In generalization analysis applied on 4,106 SNPs SNPs with p-value$< 5 \times 10^{-8}$ in the Willer et al. GWAS 2,206 SNPs generalized.
  - These SNPs were from 42 distinct genomic regions.
  - 34 of the lead SNPs reported by Willer et al. generalized (only 33 of these generalized in the "usual" generalization analysis)
  - And also non-lead SNPs from 8 additional genomics regions.
- In generalization analysis applied on 5,399 SNPs SNPs with p-value$< 1 \times 10^{-6}$ in the Willer et al. GWAS 2,418 SNPs generalized.
  - These SNPs were from 43 distinct genomic regions.

Examples from our work - total cholesterol (TC)

The TC example demonstrates that

- Due to differences in LD structure, there are instances where the lead EA SNP is different than the lead SNP in HCHS/SOL.
  - Applying generalization testing on more SNPs (not just the lead SNPs) is useful.
- Considering SNPs with higher p-value than the commonly-used 5×10^{-8} can increase power.

Exercise

- I generated a data set based on generalization analysis that I have done for the diabetes GWAS manuscript in HCHS/SOL.
- The following exercise will take you through generalization analysis based on this data set.

1. Use the command read.csv() to read the files dscvr_diabetes_res.csv and sol_diabetes_res.csv with
   - Association results published in a Mahajan et al. (2014) paper with results of diabetes GWAS in the DIAGRAM consortium (altered a bit).
   - Association results of a few more variants in the HCHS/SOL (also altered a bit).

More in the next slide. . .

Exercise

2. Use the function match() to subset the results from HCHS/SOL to those from Mahajan et al.
3. How would you know if variants have the same direction of association in the HCHS/SOL and in the DIAGRAM consortium?
4. Use the function matchEffectAllele() to match the effect sizes in the HCHS/SOL to correspond the same effect allele as in the DIAGRAM.
5. Test which associations generalize to the HCHS/SOL.
    - Take the number of tested associations in the DIAGRAM to be 10^6.
6. How many associations generalized?
7. Compare the effect allele frequencies between the two studies using plot() command.