Gene-Environment interaction analysis

Tamar Sofer

July 2017

What is a gene-environment interaction?

- We say that gene-environment interaction exists if the genotype effect differs by environment.
- For example, a certain gene may have an effect on a pulmonary outcome in people who smoke, but not in people who do not smoke.
- GxE interaction models are sometimes used for less "environmental" variables, such as sex, or diabetes status.
- Such models may have increased power to detect associated variants, compared to regular association analyses.

The linear model

- The interaction model for a quantitative trait:

$$y_i = \mathbf{x}_i^T \beta + g_i \alpha_g + e_i \alpha_e + e_i g_i \alpha_{ge} + \epsilon_i, i = 1, \ldots, n,$$

- Notation is as before: i indexes participant i.
- e_i is the environmental variable for participant i, its effect is α_e.
- α_{ge} is the interaction effect.
- This models trivially extends to logistic regression.

Tests of interaction

- When testing GxE interaction, there are two common tests.
  - Test α_{ge}, i.e. the null hypothesis $H_0 : \alpha_{ge} = 0$.
  - Test jointly α_{ge}, α_g, i.e. the null hypothesis: $H_0 : \alpha_{ge} = \alpha_g = 0$.
- The latter test is often more powerful than the first.
  - And it is sometimes more powerful than the test of $H_0 : \alpha_g = 0$.

Computational issues

- The interaction model seems like a very simple extension of the linear model (and it is).
- But computationally, it makes matters more complicated.
- When testing only genotypes, our softwares could perform some "tricks" to make computations quick.
  - These "tricks" use linear algebra, performing matrix analysis in chunks of multiple SNPs at a time.
  - And even though the basic number of computations may be the same, there is computing software in place to speed up computations that are performed on matrices.
- These tricks cannot be used the same way when testing GxE interaction terms.
  - So analyses are done one SNP at a time, taking much longer.

Quality issues

- So far we assumed that the variants that we test are common.
- They appear "enough times" in the population, so that statistical properties (also called "asymptotic properties") hold, making statistical tests valid.
- If the environmental variable is relatively rare (e.g. only 100 people in the study may take a specific type of medication), it can generate similar problems in the statistics, as rare variants cause.
- Therefore, we need to make sure we are looking at variants that appear "enough times" in participants in one environmental condition, and in the other.
- E.g. the minor allele may appear at least 50 times in people taking the medication, AND at least 50 times in people who do not take it.

Quality issues

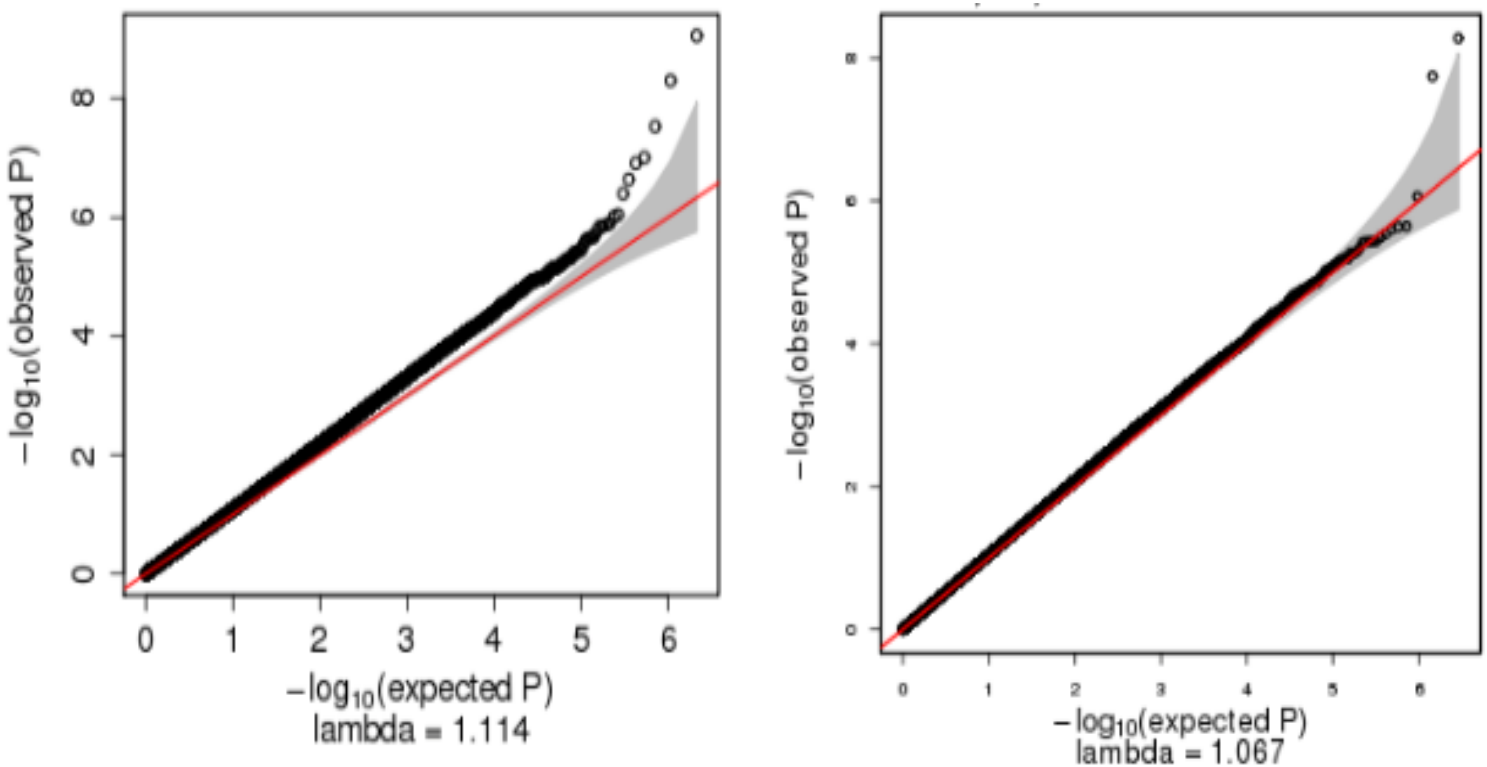
- Heterogeneous variances may also stabilize results.
  - Because the variance of the trait in people taking a medication (say) may be different than that of people who do not.

Example: thiazide-genotype interaction and QT interval

- The HCHS/SOL Genetic Analysis Center (GAC) executed an interaction GWAS, with outcome QT interval, and environmental variable Thiazide medications.
  - These types of medications are used to treat hypertension and edema.
- About 1,050 individuals of the $\sim 12,000$ participants in the analysis used Thiazides.

Example: thiazide-genotype interaction and QT interval

- The following qq-plots compare results from analysis with (right) and without (left) the use of heterogeneous variances (between drug users and non-users).

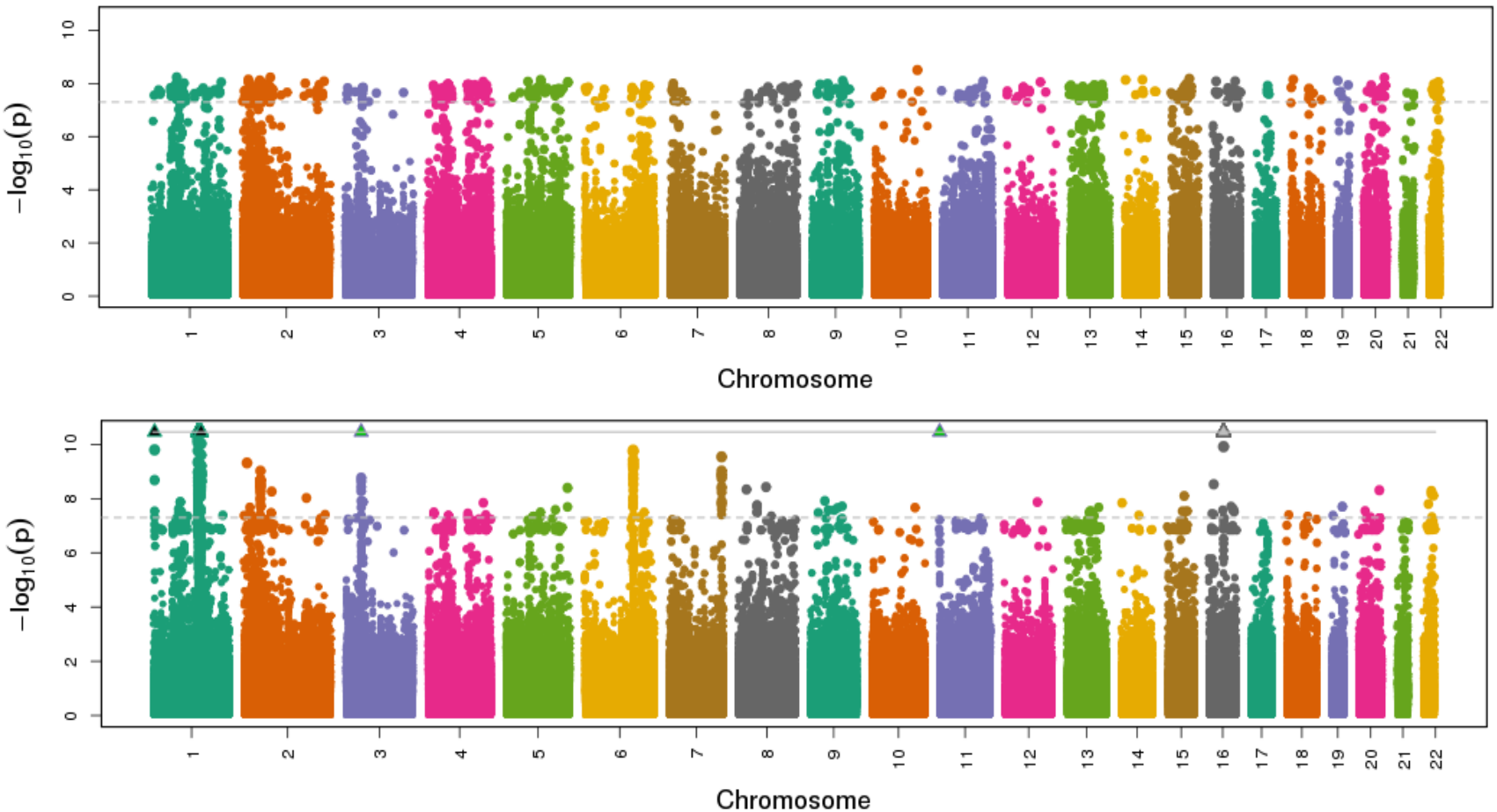


Example: TCA-genotype interaction and QT interval

- We worked on another pharmacogenomic GWAS, of interaction effects between genotypes and antidepressants of type TCA, on QT interval.
- Only 133 of the $\sim 12,000$ participants in the analysis used TCAs.
- Therefore, many of the variants are effectively rare in the participants treated by TCAs.
- In the following slides, we compare between Manhattan plots when filtering results by count of alleles in the entire study population, and when filtering results by count of alleles in the treated.
    - The filter: the minor allele appears at least 30 times in the group of interest.

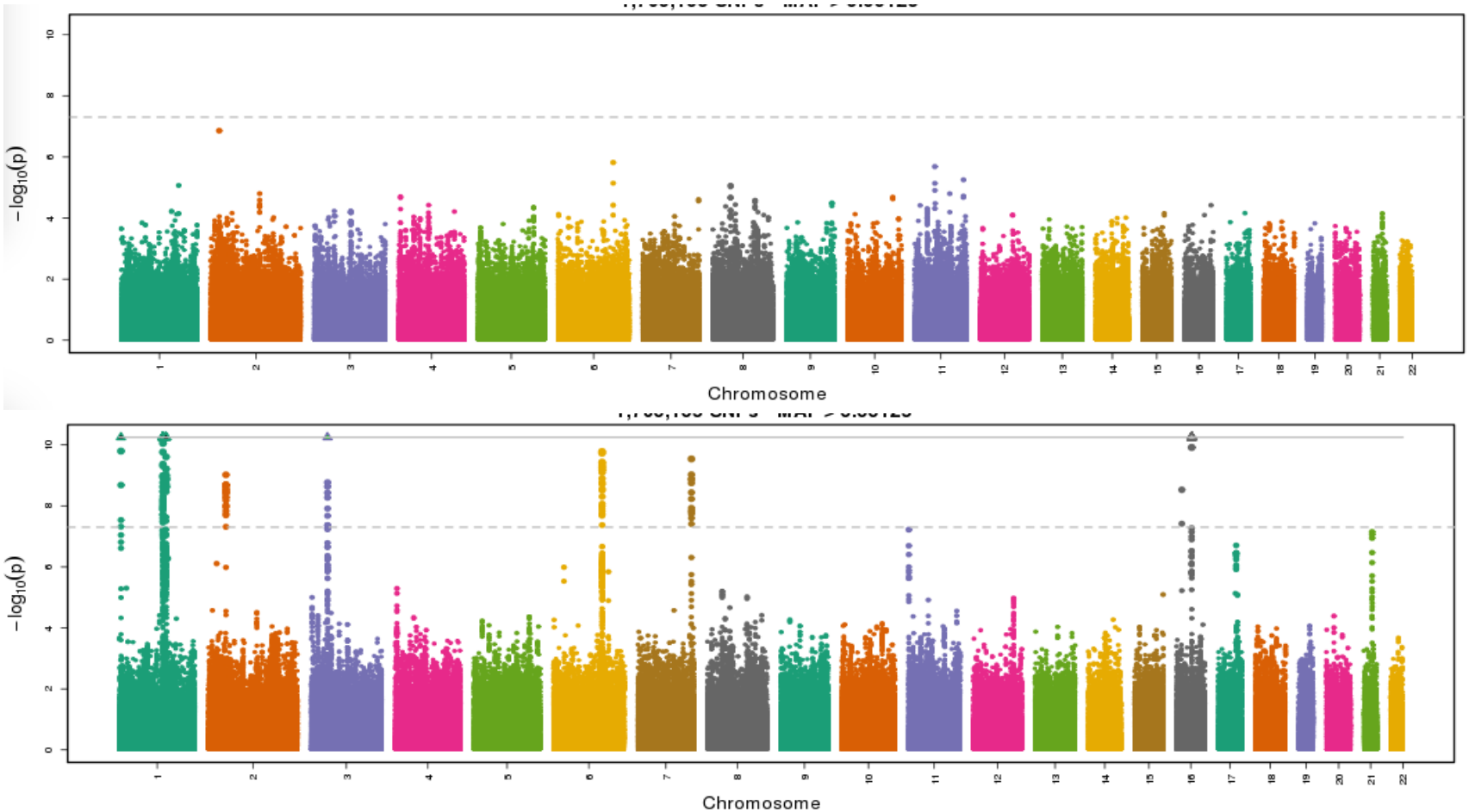
Example: TCA-genotype interaction and QT interval

Manhattan plots for the GxE (top) and joint (bottom) tests, when filtering by count of alleles in the entire study population.



Example: TCA-genotype interaction and QT interval

Manhattan plots for the GxE (top) and joint (bottom) tests, when filtering by count of alleles in the treated.



Example: TCA-genotype interaction and QT interval

- When filtering variants by count of alleles in the entire study population, there were 2.8 million SNPs plotted.
- When filtering by count of alleles in the treated, there were 1.7 million SNPs plotted.

GxE interaction - let's do it!

```
library(GWASTools)
library(GENESIS)
dir <- paste0("/home/postdoc/tsofer/SISG/",
    "Preparing_simulated_data_2")

scanAnnot <- getobj(file.path(dir,
                              "SISG_phenotypes.RData"))
```

GxE interaction - let's do it!

```
covariates <- c("EV1", "EV2", "sex", "age", "group")
outcome <- "trait"
HH.mat <- getobj(file.path(dir,
                           "SISG_houshold_matrix.RData"))
kin.mat <- getobj(file.path(dir,
                           "SISG_relatedness_matrix.RData"))
covMatList <- list(HH = HH.mat, kinship = kin.mat)
nullmod <- fitNullMM(scanData = scanAnnot,
                     outcome = outcome, covars = covariates,
                     covMatList = covMatList, verbose = FALSE)
```

GxE interaction - let's do it!

```
gds <- GdsGenotypeReader(file.path(dir,
                                   "SISG_snp_dosages.gds"))
snpAnnot <- getobj(file.path(dir,
                             "SISG_snp_dosages_snpAnnot.RData"))
genoData <- GenotypeData(gds,
                  snpAnnot=snpAnnot, scanAnnot = scanAnnot)
system.time(assoc <- assocTestMM(genoData =
                                 genoData, nullMMobj = nullmod))
```

```
## Running analysis with 500 Samples and 7463 SNPs

## Beginning Calculations...

## Block 1 of 2 Completed - 1.254 secs

## Block 2 of 2 Completed - 0.604 secs

##    user  system elapsed
##   1.777   0.070   1.867
```

GxE interaction - let's do it!

```
system.time(assoc.ge <- assocTestMM(genoData =
                    genoData, nullMMobj = nullmod,
                    ivars = "sex"))
```

```
## Running analysis with 500 Samples and 7463 SNPs

## Beginning Calculations...

## Block 1 of 2 Completed - 2.564 secs

## Block 2 of 2 Completed - 1.255 secs

##    user  system elapsed
##   3.746   0.082   3.829
```

GxE interaction - let's do it!

- We can allow for different residuals variances by interaction terms groups

```
nullmod.hetvars <- fitNullMM(scanData = scanAnnot,
                outcome = outcome, covars = covariates,
                covMatList = covMatList, verbose = FALSE,
                group.var = "sex")
assoc.ge.hetvars <- assocTestMM(genoData = genoData,
                  nullMMobj = nullmod.hetvars,
                  ivars = "sex")
```

```
## Running analysis with 500 Samples and 7463 SNPs

## Beginning Calculations...

## Block 1 of 2 Completed - 2.615 secs

## Block 2 of 2 Completed - 1.328 secs
```

GxE interaction - let's do it!

- What if we wanted to extract the genotype p-value?
- Consider the model under GxE interaction:

$$y_i = \mathbf{x}_i^T \beta + g_i \alpha_g + e_i \alpha_e + e_i g_i \alpha_{ge} + \epsilon_i, i = 1, \ldots, n.$$

- And compare it to the marginal model:

$$y_i = \mathbf{x}_i^T \beta + g_i \alpha_g \epsilon_i, i = 1, \ldots, n,$$

- Note that the β in the first and second models have different interpretations
- So be careful in reporting the p-value of β obtained from the interaction model!

Question: what are the interpretations of β in the two models?

GxE interaction - extracting SEs of various parameters

- Sometimes we do want to reported the SEs of the parameters α_g, α_{ge} from this model

$$y_i = \mathbf{x}_i^T \beta + g_i \alpha_g + e_i \alpha_e + e_i g_i \alpha_{ge} + \epsilon_i, i = 1, \ldots, n,$$

  - and/or the covariance between them.
- Because they are needed for meta-analysis of GxE interaction studies!

GxE interaction - extracting SEs of various parameters

```
assoc.ge <- assocTestMM(genoData = genoData,
                        nullMMobj = nullmod, ivars = "sex",
                        ivar.return.betaCov = TRUE,
                        verbose = FALSE)
names(assoc.ge)
```

```
## [1] "results" "betaCov"
```

- Now instead of a data.frame with results, we have a list. The first entry is the results data.frame, the second entry is a list of covariance matrices.

GxE interaction - extracting SEs of various parameters

```
assoc.ge$results[1:3,]
```

```
##   snpID chr   n   MAF minor.allele    Est.G Est.G:sexM
## 1     1   1 500 0.000            A       NA         NA
## 2     2   1 500 0.001            A       NA         NA
## 3     3   1 500 0.008            A 11.99249  -10.99661
##    GxE.Stat  GxE.pval Joint.Stat Joint.pval
## 1        NA        NA         NA         NA
## 2        NA        NA         NA         NA
## 3 0.7629439 0.3824088   1.250611   0.535098
```

```
assoc.ge$betaCov[1:3]
```

```
## $`1`
## NULL
##
## $`2`
## NULL
```

GxE interaction - extracting SEs of various parameters

```
assoc.ge$betaCov[3]
```

```
## $`3`
##             [,1]       [,2]
## [1,]  117.5367 -117.9719
## [2,] -117.9719  158.4984
```

- The joint test reported by GENESIS is the Wald test. It uses these matrices.

GxE interaction - extracting SEs of various parameters

- Here the [1,1] entry in the *j* matrix is the variance of Est.G for the *j* SNP
- The [2,2] entry in the matrix is the variance of Est.G:sexM
- And the [1,2], [2,1] entries in the matrix are the covariance between the estimates Est.G and Est.T:sexM.

Proof:

```
(assoc.ge$results$SE.G[3] ==
      sqrt(assoc.ge$betaCov[[3]][1,1]))
```

```
## [1] TRUE
```

```
(assoc.ge$results$"SE.G:sexM"[3] ==
      sqrt(assoc.ge$betaCov[[3]][2,2]))
```

```
## [1] TRUE
```

GxE interaction - extracting SEs of various parameters

- So if you are participating in a data analysis and need to add the cov(β, α_{g_j}) to your results:

```
assoc.ge$results$cov_estG_est_GE <-
    unlist(lapply(assoc.ge$betaCov, function(x)
        {if (is.null(x)) return(NA);  x[1,2] }))
```

Finish - close genotype file

```
close(gds)
```

Additional comments

- The various additions to the basic GWAS often cannot be combined together.
- For example, GxE interaction is not implemented for stratified analysis, or for admixture mapping.
- Heterogeneous variances are only defined for one factor variable, not for multiple of them.
  - Etc.
- These things were not prioritized in the preparation of software packages.

Exercises

1. Compare the p-values of SNPs in the interaction analysis and the marginal analysis.
2. Run an interaction analysis with "disease" as the environmental variable.