Prioritizing SNPs for Genetic Interaction Testing with R package '*GEWIST*'

Wei Q. Deng and Guillaume Paré

April 7, 2020

1 Introduction

It is challenging to detect gene-environment interactions in a genome-wide setting because of low statistical power and the heavy computational burden involved. Paré (Paré et al, 2010) proposed a novel method - variance prioritization (VP) - for prioritizing single nucleotide polymorphisms (SNPs) by exploiting the interaction effects on the variance of quantitative traits. The prioritization is achieved by comparing the variance of a quantitative trait conditioned on three possible genotypes using Levene's test (Levene, 1960) for variance inequality. The variance prioritization procedure consists of two steps:

1. Select SNPs with Levene's test p-value lower than their individual optimal variance prioritization thresholds (η_0).
2. Test the selected SNPs against all other SNPs (i.e. gene-gene) or environmental covariates (i.e. gene-environment) using linear regression for interactions while correcting for $\eta_0 * M$ tests (M is the number of total SNPs tested).

We then introduced a fast algorithm - Gene Environment Wide Interaction Search Threshold (GEWIST; Deng & Paré, 2011) - to efficiently and accurately determine the optimal variance prioritization threshold for individual SNPs. The '*GEWIST*' package provides functions to facilitate SNP prioritization using the algorithms described in Deng & Paré.

We will first demonstrate how to compute the optimal variance prioritization p-value threshold η_0 with '*GEWIST*' functions; and provide a working example (simulated genotype and phenotype data) to illustrate the SNP prioritization process.

2 Prioritization thresholds for SNPs of known interaction effect sizes

This section helps to illustrate the prioritization of a single SNP with known (estimated) interaction effect size using the function '*gewistLevene*'.

For example, given the inputs:

- minor allele frequency of the SNP $p = 20\%$
- total number of SNPs to be tested $M = 250{,}000$
- sample size $N = 10{,}000$
- gene-environment interaction explains 0.2% of the quantitative trait variance (*theta_gc*)
- environmental covariate explains 15% of the quantitative trait variance (*theta_c*)
- number of simulations $K = 20{,}000$

```
> library(GEWIST)
> optim_result <- gewistLevene(p=0.2, N=10000, theta_gc=0.002,
+  theta_c=0.15, M = 250000, K = 20000, verbose = FALSE)
> class(optim_result)

[1] "list"

> print(optim_result)

$Conventional_power
[1] 0.3631

$Optimal_VP_power
[1] 0.441

$Optimal_pval_threshold
[1] 0.144
```

The function then returns:

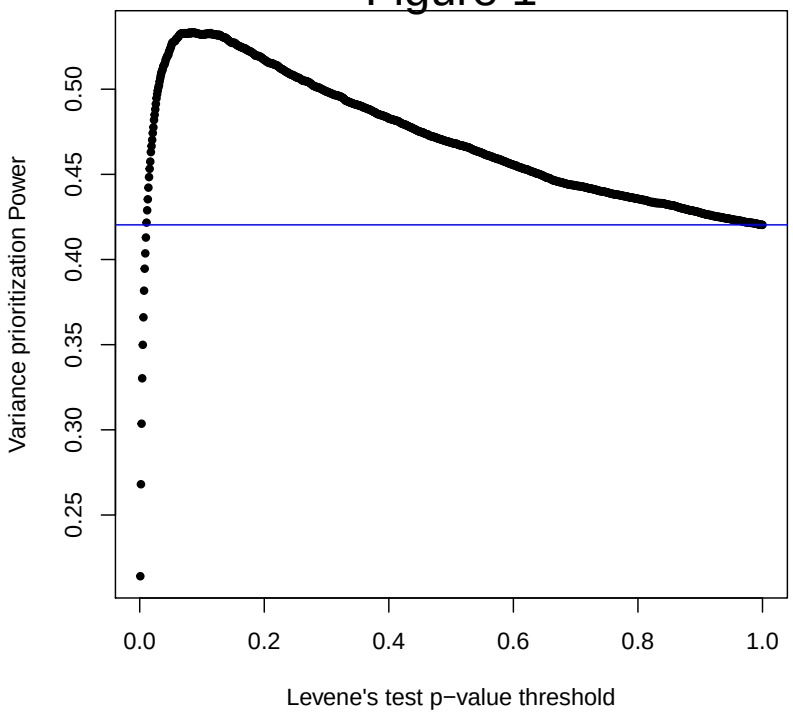
- the optimal VP p-value threshold η_0: *Optimal_pval_threshold*
- the optimal VP power obtained at η_0 while correcting for $\eta_0 * M$ SNPs: *Optimal_VP_power*
- conventional power to detect an interaction while correcting for M SNPs: *Conventional_power*

For a single SNP, the prioritization is done by first applying Levene's test to the variance of quantitative trait conditional on its three genotypes , and comparing the p-value obtained to the optimal VP p-value threshold. Include this SNP for further interaction testing if Levene's test p-value $< \eta_0$.

There is also the '*verbose*' option if the prioritization power to detect an interaction at p-value thresholds other than the optimal p-value is desired. In the following example, the VP power of a single SNP with known interaction effect size, is graphed against p-value thresholds from 0.001 to 1 with 0.001 incremental increase (Figure 1). The blue line represents the power to detect an interaction correcting for all $M = 250{,}000$ SNPs (*Conventional_power*).

```
> optim_ver <- gewistLevene(p=0.2, N=10000, theta_gc=0.002,
+ theta_c=0.2, M = 250000, K = 20000, verbose = TRUE)
```

Figure 1



3 Prioritization thresholds for SNPs of unknown interaction effect sizes

This section helps to illustrate the prioritization of a single SNP with unknown interaction effect size using the function '*effectPDF*'.

When it is reasonable to assume that multiple SNP-Covariate interactions of small effect sizes rather than a few interactions of large effect are present, the effect sizes can be described by the Weibull distribution. Other available distributions include: beta distribution, normal distribution and uniform distribution.

For example, given the inputs:

- minor allele frequency of the SNP $p = 10\%$
- total number of SNPs to be tested $M = 350{,}000$
- sample size $N = 10{,}000$

- the interaction effect sizes range from 0.025% to 0.3%
- gene-environment interaction effect size follows a Weibull distribution ($k = 0.8$, $\lambda = 0.3$)
- environmental covariate explains 10% of the quantitative trait variance (*theta_c*)
- number of simulations $K = 10{,}000$
- the number of intervals for numerical integration *nb_incr* = 50

Note that the computational time is proportional to the number of intervals (*nb_incr*) selected.

```
> weibull_exp1 <- effectPDF(distribution = "weibull", parameter1 = 0.8, parameter2 = 0.3,
+ parameter3 = NULL, p = 0.1 ,N = 10000, theta_c = 0.1, M = 350000, K = 10000, nb_incr = 50, ra

############ Interaction Testing For GenexEnvironment ############

 Weibull Distribution

> print(weibull_exp1)

$Conventional_power
[1] 0.1585007

$Optimal_VP_power
[1] 0.1803865

$Optimal_pval_threshold
[1] 0.231
```

The function returns the following:

- the optimal VP p-value threshold η_0: *Optimal_pval_threshold*
- the expected optimal VP power obtained at η_0: *Optimal_VP_power*
- the expected power assuming no prior information about the interaction effect size (i.e. uniform distribution) *Conventional_power*

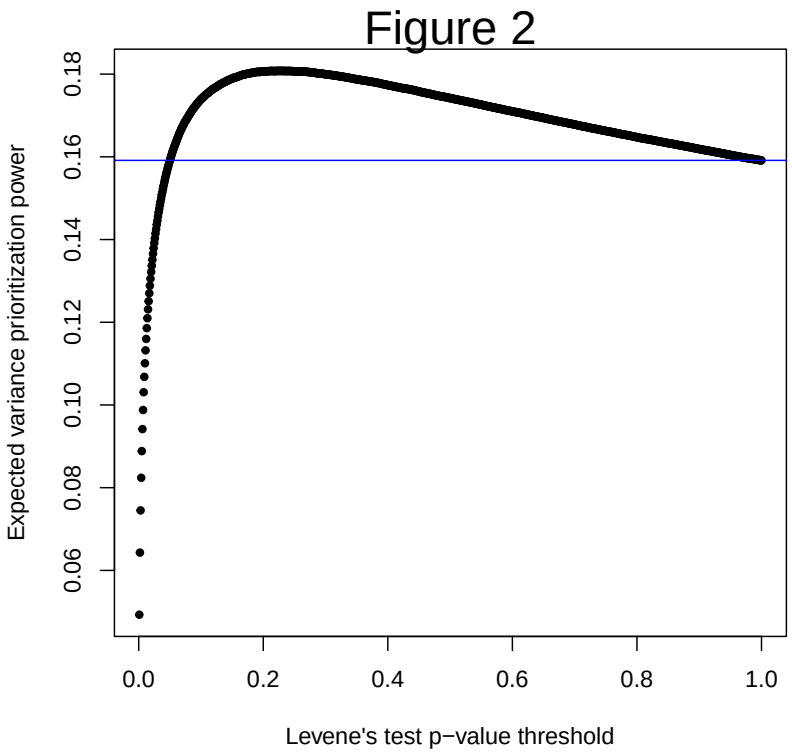
Similarly, if the VP power at p-value thresholds other than the optimal p-value is of interest, the '*verbose*' option will be useful. In the following example, the VP power of a single SNP with unknown interaction effect size, is graphed against p-value thresholds from 0.001 to 1 with 0.001 incremental increase (Figure 2). The blue line represents the power to detect an interaction correcting for all $M = 350{,}000$ SNPs (*Conventional_power*).

```
> weibull_exp2 <- effectPDF(distribution = "weibull", parameter1 = 0.8, parameter2 = 0.3,
+ parameter3 = NULL, p = 0.1 ,N = 10000, theta_c = 0.1, M = 350000,
+ K = 10000, nb_incr = 50, range = c(0.025/100,0.3/100), verbose = T)
```

```
############ Interaction Testing For GenexEnvironment ############

 Weibull Distribution

>
```



4 Prioritizing SNPs for genetic interactions from scratch

Here we provide an example to help demonstrate SNP selection, from raw SNPs to prioritized SNPs. Instead of using genome-wide datasets, which could be time-consuming and computationally heavy, we will simulate a small dataset comprises of 100 SNPs and one quantitative phenotype collected from 10,000 individuals.

A list of inputs to prioritize SNPs for GxE interactions includes (not limited to):

- *Trait*: quantitative phenotype collected from n individuals $\{y_1, y_2, y_3 ... y_n\}$
- *GenoSet*: genotype data of m SNPs for n individuals in an n by m array
- *theta_c*: estimated total quantitative trait variance explained by environmental covariate

- *theta_gc*: estimated total quantitative trait variance explained by GxE interaction $\{theta\_gc_1, theta\_gc_2, theta\_gc_3 ... theta\_gc_m\}$
- *Cov*: covariate measurements collected from n individuals $\{c_1, c_2, c_3 ... c_n\}$

4.1 Step 1: Levene's test p-values

The first task is to obtain variance inequality p-value by performing Levene's test on the quantitative trait variance conditional on the genotypes for individual SNPs. We recommend using '*leveneTest*' with option '*center = mean*' from '*car*' package [Fox J & Weisberg S].

INPUTS: *Trait* and *GenoSet*

```
> n  <- dim(GenoSet)[1]
> m <- dim(GenoSet)[2]
> library(car)
> levene_pval <- NA
>         for (i in 1: m) {
+
+                 levene_pval[i] <- leveneTest(Trait, as.factor(GenoSet[,i]),center = mean)[1,3
+
+                           }
```

OUTPUT: *levene_pval*: a vector of length $m = 100$.

4.2 Step 2: Optimal Variance Prioritization P-value Threshold

We then need to calculate the optimal VP p-value threshold for individual SNPs using the '*gewistLevene*' function.

INPUTS: *Trait*, *GenoSet*, *theta_c* and *theta_gc*

```
> optimal_pval <- NA
> for ( i in 1: m){
+
+ optimal_pval[i] <- gewistLevene(p = SNPset[i,2], N = n, theta_gc = theta_gc,
+ theta_c = theta_c, M = m )$Optimal_pval
+
+                            }
>
```

OUTPUT: *optimal_pval*: a vector of length m

The optimal VP p-value threshold is expected to change under the influence of many factors, namely, sample size, number of SNPs, minor allele frequency (MAF) of SNPs, and the proportion of variance explained by both the covariate and the interaction. However, the sample size and number of SNPs and the variance explained by the environmental covariate are fixed for a given study. Thus, for a genome-wide dataset, it is sensible to calculate

an optimal VP p-value threshold matrix, where each entry corresponds to the optimal VP p-value for SNPs with different combinations of MAF and estimated interaction effect sizes.

4.3 Step 3: Interaction testing using prioritized SNPs

The SNPs are selected such that their Levene's test p-values from **Step 1** are lower than the optimal VP p-value threshold from **Step 2**. The subset of prioritized SNPs are tested for gene-environment interaction with the measured environmental covariate (or all other SNPs for gene-gene interaction) while correcting for only the chosen SNPs.

INPUTS: *Trait*, *GenoSet*, *theta_c*, *theta_gc* and *COV*

```
> SNPind <- which(levene_pval < optimal_pval)
>  Reduced <- GenoSet[,SNPind]
>  intPval <- NA
>          for (j in 1: length(SNPind)) {
+
+                  intPval[j] <- summary(lm(Trait~ Reduced[,j] * COV ))$coef[4,4]
+
+                                                  }
```

OUTPUTS: *SNPind* (the index of prioritized SNPs), *intPval* (interaction p-values of the prioritized SNPs)

References

1. Paré, G., Cook, N. R., Ridker, P. M., & Chasman, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS genetics*, 6(6), e1000981. doi:10.1371/journal.pgen.1000981

2. Levene H. 1960. Robust tests for equality of variances. In: Olkin I, editor. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling.* Stanford, CA: Stanford University Press. p 278-292

3. Deng, W. Q., & Paré, G. (2011). A fast algorithm to optimize SNP prioritization for gene-gene and gene-environment interactions. *Genetic epidemiology*, 35: 729-738. doi: 10.1002/gepi.20624

4. John Fox and Sanford Weisberg (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`