Published in final edited form as:

Biometrics. 2016 June ; 72(2): 629–638. doi:10.1111/biom.12428.

A unified set-based test with adaptive filtering for gene-environment interaction analyses

Qianying Liu¹, Lin S. Chen^{2,*}, Dan L. Nicolae³, and Brandon L. Pierce^{2,4}

¹Sanofi, Cambridge, MA, U.S.A.

²Department of Public Health Sciences, The University of Chicago, Chicago, IL, U.S.A.

³Department of Medicine, Statistics, and Human Genetics, The University of Chicago, Chicago, IL, U.S.A.

⁴Comprehensive Cancer Center, The University of Chicago, Chicago, IL, U.S.A.

Summary

In genome-wide gene-environment interaction (GxE) studies, a common strategy to improve power is to first conduct a filtering test and retain only the SNPs that pass the filtering in the subsequent GxE analyses. Inspired by two-stage tests and gene-based tests in GxE analysis, we consider the general problem of jointly testing a set of parameters when only a few are truly from the alternative hypothesis and when filtering information is available. We propose a unified set-based test that simultaneously considers filtering on individual parameters and testing on the set. We derive the exact distribution and approximate the power function of the proposed unified statistic in simplified settings, and use them to adaptively calculate the optimal filtering threshold for each set. In the context of gene-based GxE analysis, we show that although the empirical power function may be affected by many factors, the optimal filtering threshold corresponding to the peak of the power curve primarily depends on the size of the gene. We further propose a resampling algorithm to calculate *p*-values for each gene given the estimated optimal filtering threshold. The performance of the method is evaluated in simulation studies and illustrated via a genome-wide gene-gender interaction analysis using pancreatic cancer genome-wide association data.

Keywords

adaptive filtering; a unified test; gene-environment interactions; genome-wide interaction studies; set-based (or gene-based) test

* lchen@health.bsd.uchicago.edu.

7. Supplementary Materials

Web Appendices, Tables and Figures, referenced in Sections 2.2, 2.3, 3.2, 4.1 and 5, along with the R code for simulations and data analyses, are available with this paper at the Biometrics website on Wiley Online Library.

The authors declare no conflict of interest.

1. Introduction

The effects of genetic risk variants on complex traits/diseases can be modified by environment, gender, age, or other factors. Those interaction effects play an important role in the disease etiology, and could explain part of the missing heritability of complex traits/diseases. For example, it is known that males tend to have a higher pancreatic cancer incidence rate than females, though there are only limited genetic findings to explain the difference. A good understanding of interactions among different risk factors would contribute to more effective disease prevention, treatment and precision medicine. Although the study of gene-environment interaction (GxE) has been a growing research area in recent years (Thomas, 2010), only a limited number of GxE effects have been identified via genome-wide searches (Rothman et al., 2010), which is likely due to power issues in genome-wide interaction studies.

In contrast to the total number of single nucleotide polymorphisms (SNPs) in the genome, the number of SNPs with GxE effects is expected to be relatively small (McCarthy et al., 2008). In order to improve power for detecting GxE, a common approach is to filter out the majority of irrelevant SNPs and only test GxE on the remaining SNPs. Two-stage SNP-based GxE methods (Kooperberg and LeBlanc, 2008; Murcray et al., 2009) have been proposed: in the first stage, filtering tests are conducted at the SNP level; in the second stage, standard SNP-based GxE tests are conducted and multiple testing corrections are performed only on the remaining SNPs, thus relaxing the overall significance threshold and improving power. The filtering test and the GxE test are required to be asymptotically independent under the null hypothesis to preserve the overall type I error rate (Dai et al., 2012). Popular choices for filtering include testing for gene-environment (G-E) correlation in the combined case-control samples for binary traits (Murcray et al., 2009), and testing for equal variances across genotype groups for quantitative traits (Paré et al., 2010).

In addition to the aforementioned SNP-based GxE tests, gene-based (or set-based) GxE tests have also been proposed. These methods aim to detect individually weak but collectively strong GxE effects in a functional set, i.e., a gene or a gene region. Chatterjee et al. (2006) proposed the Tukey's 1-degree-of-freedom test assuming the interaction effect is proportional to the main effect for each SNP in a gene, an assumption that reduces the complexity of the model but is also quite strong. Jiao et al. (2013) proposed a gene-based GxE test for binary traits in which SNPs with absolute G-E correlations greater than a threshold will have higher weights and the rest SNPs will have very small weights. Additionally, many gene-based association tests, such as the sequence kernel association test (SKAT) (Wu et al., 2011) and the exponential combination procedure (EC) (Chen et al., 2012), could be generalized to test for gene-level GxE. However, due to the sparse number of GxE effects in the genome, the power of such tests would be limited for genes with a small number of SNPs with GxE effects and a large number of SNPs without.

The challenges of genome-wide GxE testing and the existing two-stage or gene-based methods suggest the investigation of a general statistical problem: when only a small number of the parameters are non-null, how does one jointly test a set of parameters using independent filtering and testing statistics? In this work, motivated by a gene-gender

interaction analysis in a pancreatic cancer genome-wide association study (GWAS), we propose a general framework for gene-based interaction test. We consider filtering and testing in a unified framework and propose a set-based test statistic as the sum of statistics for individual parameters passing the filtering. Under the unified framework, the exact distribution of the proposed statistic can be obtained and its power function can be well approximated. Those derivations can be used to calculate the optimal filtering threshold that maximizes the power in simplified settings. We show that although the power function could be affected by some nuisance parameters, the optimal filtering threshold (the peak of the power curve) primarily depends on the size of the set and the number of non-null parameters in the set, and is quite robust to parameter misspecification. In the context of genome-wide gene-based GxE studies, as opposed to the pre-determined fixed filtering threshold employed in Jiao et al. (2013), we suggest adaptively calculating the optimal filtering threshold for each gene, and propose an efficient resampling approach for the p -value calculation given the adaptive filtering threshold. After evaluating the performance of the proposed methods in simulation studies, we apply the method to the motivating pancreatic cancer GWAS data and identify several genes (e.g., *RRP7A*, *ZARIL*, and *ATG16L1*) harboring multiple SNPs with individually weak but collectively strong gene-gender interaction effects on pancreatic cancer risk. We suggest further validations on our findings.

2. Methods

2.1 A general framework for set-based tests with filtering

Consider a set of k parameters $\{\theta_j: j = 1, \dots, k\}$. We are interested in testing

$$H_0: \theta_j = 0 \forall j \text{ v. s. } H_1: \text{at least one } \theta_j \neq 0.$$

Let $\{Z_1, \dots, Z_k\}$ be a set of statistics for testing $\{\theta_j = 0, j = 1, \dots, k\}$, and $\{X_1, \dots, X_k\}$ be another set which is also informative for $\{\theta_1, \dots, \theta_k\}$. For the parametric derivations, we assume $Z_1, \dots, Z_k$ are independent and $X_1, \dots, X_k$ are independent. Without loss of generality, we further assume $Z_j \sim N(\theta_j, 1)$; $X_j \sim N(\nu_j, 1)$, where $\nu_j = 0$ implies $\theta_j = 0$. We propose to jointly consider Z_j 's and X_j 's to test the null hypothesis $H_0: \theta_j = 0$ for all j , versus the alternative, $H_1: \text{at least one } \theta_j \neq 0$. We choose to work with test statistics rather than raw data to keep this framework as general as possible. In the context of gene-based GxE tests, k is the number of SNPs in a gene (referred to as “gene size” henceforth), θ_j models the statistical interaction between the j^{th} SNP and the environment, X_j and Z_j are the filtering and testing statistics with ν_j and θ_j as the corresponding standardized means. One can also adjust covariates in a regression and use the Wald statistics for individual parameters as the filtering or testing statistics. In our simulations and application, we obtain the Wald statistic based on a regression for each SNP.

We consider the set-based test for sparse alternative, i.e. the number of non-null parameters, k_1 , is much smaller than the total number of parameters, k . Chen et al. (2012) has shown that

linear combinations of variant-level statistics, $\sum_{j=1}^k Z_j^2$, has low power when the alternative is very sparse, $k_1 \ll k$. In gene-based GxE studies, large genes tend to harbor more null

variants, and the detection of large genes is more challenging. This motivates us to conduct adaptive filtering in a joint test for a set. In contrast to two-stage methods in which filtering and testing are considered as two separate steps, we propose to jointly consider filtering and testing in a unified statistic:

$$T = \sum_{j=1}^k Z_j^2 \cdot \mathbb{1} \left\{ |X_j| \geq z_{\eta/2} \right\}, \quad (1)$$

where $z_{\eta/2}$ is the $(1 - \eta/2) \times 100$ -th quantile of $N(0, 1)$. Note that filtering can also be performed on the p -value scale with a filtering threshold η . When applying the proposed method, one needs to specify the filtering threshold η .

2.2 The power function and the optimal filtering threshold

In this subsection, we propose a method for calculating the optimal filtering threshold in equation (1) for each set. To do so, we derive the distributions of the proposed test statistic T under the null and alternative hypotheses, and approximate the power function. The proposed optimal filtering threshold will maximize the approximated power function.

In order to derive the null and the alternative distributions of T , we assume Z_j and X_j are independent for all j . Later we will show that this independence assumption can be relaxed to asymptotic independence. For simplicity, we assume the non-null parameters take the same value. If $\theta_j \neq 0$, then $\theta_j = \theta$ and $v_j = v$. To derive the parametric distribution of the proposed statistic, we further assume if $\theta_j = 0$, then $v_j = 0$, i.e., the filtering test is sensitive and specific. Note that we only need these assumptions to approximate the power function in obtaining the optimal filtering threshold. Violations of the assumptions may hurt the optimality of the calculated filtering threshold and subsequently the power, but will not affect the validity of the test since we will calculate the p -values for the statistics in (1) via an efficient resampling approach.

Assuming the filtering test is specific, the probability of a null statistic passing the filtering step equal to the threshold η , and the probability of r null statistics remaining in the set-

based statistic T is $\binom{k}{r} \cdot \eta^r (1 - \eta)^{k-r}$ for r in $\{0, 1, \dots, k\}$. Under the null hypothesis, the

proposed test statistic T follows a chi-bar-squared $\left(\frac{-2}{\chi}\right)$ distribution (Shapiro, 1985)

$$\Pr(T \leq x) = \sum_{r=0}^k \binom{k}{r} \cdot \eta^r (1 - \eta)^{k-r} \cdot \Pr\left(\chi_r^2(0) \leq x\right), \quad (2)$$

where χ_0^2 is set to zero.

Under the alternative hypothesis, k_1 is the number of non-null parameters. Let $k_0 = k - k_1$. The set-based statistic T is the sum of squared individual-level statistics,

$$T = T_0 + T_1 = \sum_{\theta_j=0} Z_j^2 \cdot \mathbb{1} \left\{ |X_j| \geq z_{\eta/2} \right\} + \sum_{\theta_j \neq 0} Z_j^2 \cdot \mathbb{1} \left\{ |X_j| \geq z_{\eta/2} \right\}. \quad (3)$$

The two components in T are independent and each follows a χ^2 distribution

$$\begin{aligned} \Pr(T_0 \leq x) &= \sum_{r=0}^{k_0} \binom{k_0}{r} \cdot \eta^r (1-\eta)^{k_0-r} \cdot \Pr(\chi_r^2(0) \leq x), \text{ and} \\ \Pr(T_1 \leq x) &= \sum_{r=0}^{k_1} \binom{k_1}{r} \cdot \eta_1^r (1-\eta_1)^{k_1-r} \cdot \Pr(\chi_r^2(r\theta^2) \leq x). \end{aligned}$$

Here η_1 is the probability of the individual-level statistic for a non-null parameter retaining in the set-based test statistic, i.e. the power of the filtering test. It depends on the filtering threshold η and the effect size of the filtering statistic ν (the effect size of the G-E correlation in the combined case-control samples in our simulations and application):

$$\eta_1 = 1 - \Phi(z_{\eta/2} + \nu) + \Phi(-z_{\eta/2} + \nu),$$

where $\Phi(\cdot)$ is the distribution function of $N(0, 1)$.

The analytical power function for the proposed test statistic is intractable. Instead of deriving the density function, we use a scaled and shifted central chi-squared distribution to approximate the χ^2 distribution by moment matching (Zhang, 2005). We match the first three moments of the null and alternative distributions of the proposed test statistic T in (2) and (3) by $A_0 \cdot \chi_{B_0}^2(0) + C_0$ and $A_1 \cdot \chi_{B_1}^2(0) + C_1$, respectively. Web Figure 1 shows that the approximations match well with the null and alternative χ^2 distributions. The power of the set-based test in (1) can be approximated by

$$\beta(\eta; \alpha, A_i, B_i, C_i) = 1 - \mathcal{F}_{B_1} \left(\frac{A_0}{A_1} \cdot \mathcal{F}_{B_0}^{-1}(1 - \alpha) + \frac{C_0 - C_1}{A_1} \right), \quad (4)$$

where $\mathcal{F}_{B_i}(\cdot)$ is the cumulative distribution function of a central χ^2 distribution with B_i degrees of freedom, α is the significance level with Bonferroni correction, and the formulae of A_i , B_i and C_i ($i = 0, 1$) are given in Web Appendix A.

With the power approximation $\beta(\eta)$ in (4), we estimate the optimal filtering threshold by $\widehat{\eta}_{\text{opt}} = \arg \max_{\eta} \beta(\eta)$. The power function depends on many factors, e.g. the number of parameters k , the number of non-null parameters k_1 , the filtering power η_1 , the GxE effect size θ_j , and the correlations among the k statistics. Those factors vary for different sets and we will show that adaptively choosing the optimal filtering threshold would improve the power compared to using a fixed filtering threshold. Interestingly, the proposed test statistic

T is comparable to a Bayesian statistic derived from a hierarchical Bayesian framework – the Bayesian statistic is also a weighted sum of Z_j^2 with adaptive filtering (see Web Appendix B).

2.3 The assumption on asymptotic independence between filtering and testing

When deriving the power approximation, it is assumed that the individual-level testing and filtering statistics, Z_j and X_j , are independent. We will show that the independence assumption can be relaxed to asymptotic independence. Considering a study with n independent subjects, assume the asymptotic marginal distributions of Z_j and X_j are $N(\theta, 1)$ and $N(\nu, 1)$, and $\nu = 0$ implies $\theta_j = 0$. Here, Z_j and X_j are the Wald statistics for testing $\theta_j = 0$ and $\nu = 0$, i.e. $Z_j = \hat{\theta}_j / \text{s.e.}(\hat{\theta}_j)$ and $X_j = \hat{\nu}_j / \text{s.e.}(\hat{\nu}_j)$, where $\hat{\theta}_j$ and $\hat{\nu}_j$ are the maximum-likelihood estimates and $\text{s.e.}(\hat{\theta}_j)$ and $\text{s.e.}(\hat{\nu}_j)$ are the estimated standard errors.

Proposition 1—*Suppose $\hat{\theta}_1, \dots, \hat{\theta}_k$ are independent and $\hat{\nu}_1, \dots, \hat{\nu}_k$ are independent.*

Assuming $\theta_j \sim O(n^{-\frac{1}{2}})$, $\nu \sim O(n^{-\frac{1}{2}})$, and $(\hat{\theta}_j, \hat{\nu}_j)$ follows a Gaussian distribution with $\text{cov}(\hat{\theta}_j, \hat{\nu}_j) = o(1/n)$, for all $j \in \{1, \dots, k\}$, the proposed test statistic T converges in distribution to the χ^2 distributions in (2) and (3) under the null and the alternative hypotheses, respectively.

The proof of Proposition 1 is given in Web Appendix C. Under the assumptions in Proposition 1, the power of the proposed unified test can be approximated by the function in (4), and the type I error rate based on the parametric p -value calculation is controlled. This is because type I error rate and power are essentially functions based on the null and the alternative distributions. It also follows that the proposed filtering threshold obtained by maximizing the power function $\beta(\eta)$ is asymptotically optimal under the asymptotic independence assumption.

In genome-wide analysis, existing two-stage SNP-based GxE testing methods only consider the remaining SNPs in the second stage testing. In accounting for multiple testing, these methods would require certain independence assumption on filtering and testing to relax the genome-wide significance threshold from $\alpha/\{\text{\# total SNPs}\}$ to $\alpha/\{\text{\# remaining SNPs}\}$ (Dai et al., 2012). In contrast, the proposed test in (1) is a unified test that jointly considers filtering and testing in one statistic. The overall genome-wide significance threshold is not relaxed, still at $\alpha/\{\text{\# total genes}\}$. When p -values are calculated via resampling approaches, the overall type I error rate can be controlled even when the assumptions in Proposition 1 are violated. Those assumptions ensure the optimality of the filtering threshold but are not required for the validity of the proposed test.

The proposed unified test is distinct from a two-stage gene-based test by the adaptive filtering strategy and required assumptions. Moreover, the null distribution of the proposed test is a chibar-squared distribution given in (2). In a two-stage test, gene-based tests are performed by only considering the remaining SNPs after filtering. The null distribution of statistics in the second stage is a chi-squared distribution with the degree of freedom being

the number of remaining SNPs in the gene. Since the null distributions are different, even if the observed statistics are the same at the same filtering threshold, the proposed unified test would produce a different p -value than the two-stage procedure. We will compare the power of the two approaches in later sections.

3. Gene-based GxE tests with adaptive filtering

3.1 The options of filtering

In the context of genome-wide GxE tests, let Y be the phenotype, E be the environmental factor, $(G_1, \dots, G_k)$ be the genotypes for the k SNPs in a gene, θ_j be a parameter modeling the interaction between E and G_j ($j = 1, \dots, k$). The individual-level testing statistic could be chosen as the Wald statistic for GxE derived from a regression adjusting for covariates,

$Z_j = \hat{\theta}_j / \text{s.e.}(\hat{\theta}_j)$, though the choices of the filtering statistic X_j often depend on the study design.

For binary traits in case-control studies, the filtering statistic X_j could be chosen as the statistic for testing the association between genotype G_j and the environmental factor E in the combined case-control samples (Murcray et al., 2009), assuming the disease is rare and gene and environment (G-E) are independent in the general population. The rationale is that under the assumption of G-E independence in the population, gene and environment would be independent in the healthy controls if the disease is rare. In the presence of GxE, there would be G-E dependence in the cases and thus G-E association in the combined case-control samples. The G-E association test has been shown to be asymptotically independent of the GxE test (Murcray et al., 2009; Dai et al., 2012) and is powerful for filtering. It is also the filtering test used in our simulations and application.

For quantitative traits, if gene and environment are independent, it has been shown that the variance of the trait conditional on the genotype are different in the presence of GxE (Paré et al., 2010). It follows that the filtering statistic X_j could be chosen as the Levene's statistic (Levene, 1960) for testing equal variances across genotype groups. Without GxE effects, X_j follows a central F -distribution with 2 and $n - 3$ degrees of freedom. The proposed framework can be easily extended to F -distributed filtering statistics by defining the set-

based test statistic as $T = \sum_{j=1}^k Z_j^2 \cdot \mathbb{1}\{X_j \geq F(\eta, 2, n - 3)\}$, where $F(\eta, 2, n - 3)$ is the $(1 - \eta) \times 100$ -th quantile of the F -distribution with 2 and $n - 3$ degrees of freedom.

Accordingly, in the power approximation $\beta(\eta)$, η_1 is the power of Levene's test.

3.2 An efficient re-sampling approach for calculating p-values

The power approximation in (4) is derived based on the assumption that the individual-level statistics are independent. In the presence of linkage disequilibrium (LD), SNPs in a gene are correlated, and the corresponding statistics are correlated as well. Strong LD can affect both the effective number of SNPs and the effective number of SNPs with GxE. As a result, this may affect the optimality of the calculated filtering threshold. Therefore, we propose to first conduct a within-gene pruning to eliminate SNPs in high LD ($r^2 > 0.9$), and use the remaining number of SNPs as k in estimating the optimal filtering threshold. In the

simulation sections, we will show that the proposed optimal filtering threshold is quite robust to the presence of moderate-to-low LD.

In order to calculate a p -value in the presence of LD, one could permute the phenotype a large number of times to obtain the null statistics in equation (1) for each gene. For candidate GxE analysis, it is computationally feasible to evaluate the significance for a few genes with a sufficiently large number of permutations. However, for genome-wide GxE testing, performing millions of permutations for every gene is computationally demanding. Here we propose to obtain the null statistics in the presence of LD by approximating the joint distributions of $(X_1, \dots, X_k)$ and $(Z_1, \dots, Z_k)$ under the null hypothesis, and sampling individual-level statistics X_j 's and Z_j 's from the joint distribution. Specifically, when the filtering statistics and testing statistics are chosen as the Wald statistics for the interaction terms in linear or logistic regressions, both $(X_1, \dots, X_k)$ and $(Z_1, \dots, Z_k)$ are asymptotically multivariate normally distributed. Existing literature has shown that the correlation among association statistics in a gene can be well approximated by the correlation among genotypes (Yang et al., 2012). Through simulations (see Web Table 1), we found that this conclusion also applies to GxE filtering and testing statistics. With centered genotypes and a centered environmental factor, the correlation among filtering (or testing) statistics can be approximated by the correlation matrix of the genotypes, i.e. $\text{cor}(X_i, X_j) \simeq \text{cor}(Z_i, Z_j) \simeq \text{cor}(G_i, G_j) = R_{ij}$ for any i, j . To obtain sets of filtering and testing statistics under the null, and subsequently the null distribution of the proposed unified statistic, we can sample from the multivariate normal distribution $N(\mathbf{0}, \mathbf{R})$, where $\mathbf{R} = (R_{ij})_{k \times k}$.

In genome-wide GxE testing, we further adopt a sequential sampling strategy to relieve the computational burden (Chen et al., 2012). Specifically, for a given gene, if no SNP passes the filtering, the gene-based test statistic is 0, the p -value is 1 and no sampling is needed. For genes with at least one SNP passing the filtering, we first draw $S = 100$ sets of the filtering and testing statistics from the multivariate normal distribution $N(\mathbf{0}, \mathbf{R})$. If the p -value calculated based on these 100 null statistics is smaller than 0.1, we draw 10 times more sets (i.e. $S = 1000$) to obtain a p -value with higher precision. We repeat this procedure until the p -value is greater than $10/S$ or a desired precision level is reached. This proposed sequential resampling procedure is much more efficient than performing millions of permutations and is used in real data analysis in Section 5.

4. Simulations

4.1 Power of the proposed unified test with different filtering thresholds

We first examined the power of the proposed method for independent SNPs and compared the optimal filtering thresholds for genes of different sizes. We simulated case-control studies with 5000 cases and 5000 controls from a population with disease prevalence of 0.05. SNPs are independent with minor allele frequencies (MAFs) of 0.2. We also simulated a binary environmental factor, which is independent of the SNPs in the population and is distributed as Bernoulli(0.5). We expect that GxE effects have similar or higher effect sizes and are less likely to occur than associations. As such, we simulated $k_1 = 3$ independent SNPs in each gene with log-additive GxE effects being $\log(1.3)$. We simulated 10,000 genes each of size 6, 30, 100 and 200. We then calculated the SNP-level G-E association statistics

using the combined case-control samples as the filtering statistics and the SNP-level GxE interaction statistics as the testing statistics. We compared the power of the proposed statistic with different filtering thresholds at the significance level of 10^{-5} , which corresponds to a genome-wide error rate of 0.1 ($10^{-5} = 0.1/10,000$ genes).

Figure 1 showed that larger genes have smaller optimal filtering thresholds. Based on the proposed unified test, the adaptive filtering threshold $\widehat{n}_{\text{opt}}$ outperformed no filtering and the fixed filtering threshold of 0.1. The estimated $\widehat{n}_{\text{opt}}$ was close to the true optimal n_{opt} , which was obtained by a greedy search in the range of (0,1) with a grid as small as 0.0001. We also compared our method with a two-stage method that filters at 0.1 and only considers the remaining SNPs in the second stage set-based tests with a slightly relaxed significance threshold. At the same genome-wide error rate of 0.1, the power of a two-stage fixed-filtering method were 0.903, 0.836, 0.654, and 0.456 for $k = 6, 30, 100$, and 200, respectively, and were 7–76% lower than the proposed test with adaptive filtering. Simulations for quantitative traits reached the same conclusion, and the results were presented in Web Appendix D.

4.2 Power with misspecification of nuisance parameters

In Section 4.1, the proposed filtering threshold $\widehat{n}_{\text{opt}}$ was calculated based on the true effect size and the true number of SNPs with GxE effects. When analyzing real data, these parameters are unknown and need to be specified. Here, we evaluated the performance of the proposed adaptive filtering when these parameters were misspecified.

We simulated case-control studies by varying GxE effect sizes and the number of SNPs with GxE, k_1 . In Figure 2(a), we evaluated the power when the simulated log odds ratio of GxE was $\log(1.25)$, $\log(1.3)$, and $\log(1.35)$, as well as when the log odds ratio for each SNP was sampled from a uniform distribution between $\log(1.2)$ to $\log(1.4)$. We again calculated the optimal filtering threshold by specifying $k_1 = 3$ and the effect size as $\log(1.3)$. Although the power curve changed substantially with different effect sizes, the optimal filtering threshold corresponding to the peak of the power curve was robust. The optimal filtering threshold was close to the proposed one even when GxE effect size was misspecified. For the same reason, although the MAFs of SNPs in a set may have varied and would have affected the power of GxE filtering and testing, they barely affected the optimal filtering threshold (results not shown). In Figure 2(b), we evaluated the power when the true k_1 was 2, 3 and 4. Then, when calculating the proposed filtering threshold, we specified k_1 as 3. The simulated log odds ratio of GxE effect was $\log(1.3)$ in all settings, and was correctly specified in our calculation. The true optimal filtering thresholds were close to the proposed ones when k_1 was misspecified in a certain range. The simulations in Figure 2 were based on genes with 100 SNPs. We repeated the simulations for other gene sizes, and the conclusions were the same.

In practice, we suggest specifying k_1 by the nature of the trait/disease of interest: if the trait/disease has moderate-to-low heritability and only dozens of SNPs in the genome are expected to show GxE effects, e.g. pancreatic cancer, we propose a fixed k_1 of 3 to 5. Note that since the power curve is steep on the left (near 0), an under-specified k_1 would result in

an underspecified optimal filtering threshold and may reduce the power. For highly polygenic traits, a larger number of SNPs in the genome or in a gene may show GxE effects. One can specify k_1 as a function of gene size k and this function can be chosen based on the literature or with some exploratory data analyses extracting common patterns from known genes.

Additionally, we evaluated the performance of the proposed method on polygenic traits with true $k_1 = \max(2, 0.05 \times k)$. Table 1 shows the power comparison for different gene sizes based on 10,000 simulated case-control studies each with 3,000 cases and 3,000 controls, at the significance level of 10^{-5} . Misspecification of k_1 within a reasonable range (reducing by half or doubling the true k_1) may affect the optimality of the calculated filtering threshold, but could still improve power as compared to a fixed threshold, particularly for moderate- to large-sized genes. As we expected, an underspecified k_1 may reduce the power due to overly-stringent filtering, and a slightly overspecified k_1 has little impact on the results. In the last two columns of Table 1, we compared the power at the same significance threshold for our proposed test with filtering at 0.1 versus a two-stage gene-based test with the same filtering. It can be seen that even though the observed statistics are the same for the two approaches, because the null distributions are different, they yield different p -values.

4.3 Power and Type I error rate in the presence of LD

To evaluate the performance of the proposed method when applied to SNPs in LD, we simulated genotypes with realistic LD structures based on the Cancer Genetic Markers of Susceptibility (CGEMS) PanScan-I and PanScan-II GWAS data (Amundadottir et al., 2009; Petersen et al., 2010). With genotyped SNPs from 5,805 individuals, we randomly paired haplotypes from 10 large genes to simulate a population size of 1,000,000. After LD pruning at $r^2 > 0.9$, we took 100 adjacent SNPs from each LD structure to simulate genes of size $k = 100$. For each simulated gene, we repeated the following procedure 1,000 times: We first sampled $k_1 = 3$ SNPs to be the ones with GxE effects and then simulated the disease status for each individual using a logistic regression model; the disease prevalence was about 0.05. The log odds ratio of a binary environmental factor with 50% prevalence was set to $\log(1.3)$; the log odds ratio of GxE could have been either $\log(1.3)$ or $-\log(1.3)$, each with a probability of 50%. We then randomly sampled 5,000 cases and 5,000 controls from the population and conducted gene-based GxE tests with different filtering thresholds in (0,1). For a given LD structure, the true optimal filtering threshold η_{opt} was determined as the one that maximized the power based on a greedy search averaging over 1,000 simulations.

We compared the power of the proposed unified test with 1) no filtering, 2) a fixed filtering threshold of 0.1, 3) the proposed filtering threshold $\widehat{\eta_{\text{opt}}}$, and 4) the true optimal filtering threshold η_{opt} at the significance level of 10^{-5} . Without filtering, the average power was 0.338. With filtering, the average power based on the fixed, the proposed, and the true optimal thresholds were 0.584, 0.656, and 0.672, respectively. Compared to no filtering, adaptive filtering in gene-based tests almost doubled the power. In addition, the proposed adaptive filtering threshold $\widehat{\eta_{\text{opt}}}$ outperformed fixed filtering threshold with a performance comparable to the true optimal.

Additionally, we set the gene size to $k = 30$ and compared the power when $k_1 = 3$ without LD, and when $k_1 = 1$ with 4 other SNPs in moderate LD with the causal GxE variant. Figure 2(c) shows that the optimal threshold is robust to LD and misspecification of k_1 in a certain range.

We evaluated the type I error rate in the presence of LD when p -values were calculated based on resampling. We took the genotype data from the PanScan GWAS and simulated phenotypes under the null. We applied the proposed method to the null data set. Table 2 presents the type I error rates for different gene sizes at different p -value significance thresholds, averaged over 100 repeated simulations. Since the proposed test calculates the p -values based on resampling, the type I error rate is well controlled.

5. Application

We applied the proposed unified test to the CGEMS PanScan-I and PanScan-II GWAS data of pancreatic cancer, which consist of 2,850 cases and 2,955 controls of European ancestry. There is a known gender difference in the risk of pancreatic cancer (Michaud, 2004), and males tend to have a higher incidence rate than females. Here, we treated gender as the “environmental factor” and the goal was to identify SNPs showing interaction effects with gender. We imputed missing genotypes using data from The 1000 Genomes Project as the reference. We analyzed 14,066 protein-coding genes with at least 5 SNPs with MAF > 0.05 after LD pruning at $r^2 > 0.9$. We used the SNP-level G-E association statistics in the combined case-control samples as the filtering statistics. To estimate the optimal filtering threshold at the genome-wide significance level of 10^{-5} (~ a genome-wide error rate of 0.15), we specified the number of SNPs with GxE effects in a gene as $k_1 = 3$, the log odds ratio of the additive GxE effect as $\log(1.3)$, and the disease prevalence as 0.01. We used a logistic regression with the genotype of each SNP, gender and the interaction term as predictors to obtain the SNP-level Wald statistic for the interaction term as the testing statistic, adjusting for age and the top five principal components. Web Figure 2 shows the QQ plot of the p -values based on the proposed test, and there is no inflation in the p -values.

Table 3 lists the top 30 genes identified by our gene-gender interaction test, as well as their p -values, gene sizes, the numbers of SNPs after filtering and the calculated optimal filtering thresholds. We also calculated the p -values of these genes based on the same tests with filtering thresholds of 1 (no filtering). It can be seen that these genes would not be identified as the top ones based on tests with no filtering, indicating that our proposed test at least improves power for some genes. Web Table 2 lists the SNP-level GxE odds ratios and p -values for the SNPs passing the filtering in the top 10 genes. We also reevaluated the p -values of the top 10 genes with k_1 being specified as 2 and 5 in our filtering threshold calculation (shown in Web Table 3). The results were largely unchanged, though a smaller k_1 may hurt the power. This again suggests that parameter misspecification only has a minor impact on our test and recommends a slightly over-specified k_1 .

The gene *RRP7A* (ribosomal RNA processing 7 homolog A) was ranked first with a gene-level GxE p -value of 0.00012. Among its 7 SNPs, 4 were retained after filtering with individual SNP-level GxE p -values of 0.001 to 0.003 and odds ratios of 1.37 to 1.48. This

gene was ranked 19th based on the test with no filtering. *RRP7A* has been reported to be associated with gastric cancer risk (Lin et al., 2002). Gastric cancer can develop in any part of the stomach and can spread to nearby organs such as pancreas. Further validation may establish the potential pleiotropy effects.

The gene *ZARIL* (zygote arrest 1-like) was ranked second in our analysis with a gene-level GxE *p*-value of 0.00021. Three out of its six SNPs were retained after filtering, each showing moderate to strong SNP-level GxE effects. *ZARIL* has been reported to be binding to *BRCA2* promoter in vivo and be responsible for the silencing of *BRCA2* gene expression in the G0/G1 phase in human breast cells (Misra et al., 2010). Since germline mutations in the breast cancer gene *BRCA2* is associated with an increased risk of pancreatic cancer (Mocci et al., 2013), we conjecture that the gene *ZARIL* may play a role in the observed gender risk difference in pancreatic cancer.

The gene *ATG16L1* was ranked sixth in our analysis. It has been associated with Crohn's disease risk, and there was evidence of gender-specific Crohn's disease risk-association for the SNP rs3792106 in *ATG16L1* (Liu et al., 2012). In our analysis, this SNP not only passed the filtering test, but its GxE testing statistic also supported a potential interaction with gender for pancreatic cancer risk. Since Crohn's disease may impair pancreatic function, we suggest further validation of this gene. Additionally, we used the gene *ATG16L1* as an example to illustrate that the proposed test is powerful for identifying individually weak but collectively strong GxE effects. Figure 3 shows the SNP-level gene-gender interaction effects for the 20 SNPs in the gene and the LD heatmap for the 7 SNPs that passed the filtering. There were multiple SNPs showing moderate GxE effects and not all of them were in high LD. None of the SNP-level GxE effects can be detected in a SNP-based analysis. The gene was ranked 519 based on the test with no filtering. With our test, the adaptive filtering greatly increased the proportion of SNPs with potential GxE effects in the gene from 35% before filtering to 71% after filtering. In addition, the proposed adaptive filtering $\widehat{n}_{\text{opt}}$ outperformed the fixed threshold of 0.1 and led to a higher proportion of potential SNPs with GxE (71% versus 55%) and a more significant gene-based *p*-value (4.5×10^{-4} versus 0.001).

6. Discussion

We propose a unified set-based testing framework that jointly considers individual-level filtering and set-based testing. As opposed to existing two-stage tests with fixed filtering thresholds, our test chooses an adaptive filtering threshold by maximizing the approximated power function. Moreover, instead of relaxing the genome-wide significance threshold, we consider the filtered genes in the null distribution of the unified statistic and calculate *p*-values via an efficient resampling approach. With the unified test, the independence assumption of filtering and testing is only required to ensure the optimality of the calculated filtering threshold and is not needed to preserve the type I error rate.

In simulations and real data analyses, we showed that although many factors may affect the power function of our test, the peak of the power curve (the optimal filtering threshold) primarily depends on gene size (*k*), the number of SNPs with GxE effects (*k*₁), and the LD

structure. We showed that misspecification of k_1 in a reasonable range and/or the presence of moderate LD would only slightly affect the optimality of filtering and power of the proposed test.

The proposed framework is widely applicable and only requires individual-level filtering and testing statistics. It can be applied to both binary and quantitative outcomes; it can be applied to genome-wide family-based association studies with genetic effect filtering and family-based testing; and it can be extended to meta-analysis of genome-wide GxE studies. Here, we briefly describe the extension to meta-analysis of GxE: for genes with the same SNPs or LD blocks showing GxE effects in multiple studies in a meta-analysis, one may greatly improve the power by conducting meta-filtering, i.e., combining the filtering p -values across studies using Fisher's method. The meta-filtering strategy improves the filtering power, and subsequently the results from different studies can be combined based on a standard meta-analysis. For genes with different SNPs or LD blocks showing GxE effects across studies, the meta-filtering strategy may not improve power. However, the proposed adaptive filtering threshold would be close to 1 and the proposed test would reduce to a standard meta-analysis without filtering.

The optimal filtering threshold depends on the power of the filtering test. When the filtering power is low, SNPs with moderate GxE effects may be filtered out with a stringent threshold. If using a fixed and stringent filtering threshold, the power can be reduced substantially. In contrast, the proposed method will adjust the optimal filtering thresholds to values close to 1, i.e., no filtering. When the filtering power is high, the optimal filtering thresholds will be stringent.

One caveat of the proposed method is that the optimality of the proposed filtering threshold, i.e., the power of the test, depends on the LD structure. Specifically, when SNPs with GxE effects are in LD with others, correctly specifying k_1 would be a challenge. We suggest specifying k_1 as either a fixed number or as a function of k , depending on the nature of the trait of interest. In general, we recommend a slightly over-specified k_1 to avoid over-filtering.

The proposed method can be further improved by imposing more informative weights than the current binary weights, or incorporating other types of set-based tests in constructing the set-based statistics. The distribution of the set-based test statistic is more complicated with informative weighting or other combination methods. Further investigation is needed to efficiently approximate the power and choose the corresponding optimal filtering threshold. The proposed framework will serve as a complement approach to SNP-based GxE testing and help identify individually weak but collectively strong GxE effects in the genome-wide interaction analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported in part by R01GM108711, R03CA174984, U01CA161032, and R01MH101820.

References

- Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, Bueno-de-Mesquita HB, Gross M, Helzlsouer K, Jacobs EJ, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nature Genetics*. 2009; 41:986–990. [PubMed: 19648918]
- Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *The American Journal of Human Genetics*. 2006; 79:1002–1016. [PubMed: 17186459]
- Chen LS, Hsu L, Gamazon ER, Cox NJ, Nicolae DL. An exponential combination procedure for set-based association tests in sequencing studies. *The American Journal of Human Genetics*. 2012; 91:997–986.
- Dai JY, Kooperberg C, Leblanc M, Prentice RL. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika*. 2012; 99:929–944. [PubMed: 23843674]
- Jiao S, Hsu L, Béziau S, Brenner H, Chan AT, Chang-Claude J, Le Marchand L, Lemire M, Newcomb PA, Slattery ML, Peters U. SBERIA: Set-based gene-environment interaction test for rare and common variants in complex diseases. *Genetic Epidemiology*. 2013; 37:452–464. [PubMed: 23720162]
- Kooperberg C, LeBlanc M. Increasing the power of identifying gene×gene interactions in genome-wide association studies. *Genetic Epidemiology*. 2008; 32:255–263. [PubMed: 18200600]
- Levene, H. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Olkin, I., et al., editors. Stanford University Press; Stanford, California: 1960.
- Lin A, Stenger A, Slucka Z, Li G, Jankevics E, Rees RC. Serological identification and expression analysis of gastric cancer-associated genes. *British Journal of Cancer*. 2002; 86:1824–1830. [PubMed: 12087473]
- Liu LY, Schaub MA, Sirota M, Butte AJ. Transmission distortion in Crohn's disease risk gene ATG16L1 leads to sex difference in disease association. *Inflammatory Bowel Diseases*. 2012; 18:312–322. [PubMed: 21618365]
- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Review Genetics*. 2008; 9:356–369.
- Michaud DS. Epidemiology of pancreatic cancer. *Minerva Chirurgica*. 2004; 59:99–111. [PubMed: 15238885]
- Misra S, Sharma S, Agarwal A, Khedkar SV, Tripathi MK, Mittal MK, Chaudhuri G. Cell cycle-dependent regulation of the bi-directional overlapping promoter of human *BRCA2/ZAR2* genes in breast cancer cells. *Molecular Cancer*. 2010; 9:50. [PubMed: 20202217]
- Mocci E, Milne RL, Méndez-Villamil EY, Hopper JL, John EM, Andrulis IL, Chung WK, Daly M, Buys SS, Malats N, Goldgar DE. Risk of pancreatic cancer in breast cancer families from the breast cancer family registry. *Cancer Epidemiol Biomarkers Prev*. 2013; 22:803–811. [PubMed: 23456555]
- Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology*. 2009; 169:219–226. [PubMed: 19022827]
- Paré G, Cook NR, Ridker PM, Chasman DI. 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*. 2010; 6:e1000981. [PubMed: 20585554]
- Petersen GM, Amundadottir L, Fuchs CS, Kraft P, Stolzenberg-Solomon RZ, Jacobs KB, Arslan AA, Bueno-de-Mesquita HB, Gallinger S, Gross M, et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nature Genetics*. 2010; 42:224–228. [PubMed: 20101243]

- Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, Figueroa JD, Real FX, Van Den Berg D, Matullo G, Baris D, et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nature Genetics*. 2010; 42:978–984. [PubMed: 20972438]
- Shapiro A. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*. 1985; 72:133–144.
- Thomas D. Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics*. 2010; 11:259–272.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*. 2011; 89:82–93. [PubMed: 21737059]
- Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden PA, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*. 2012; 44(4): 369–375. [PubMed: 22426310]
- Zhang JT. Approximate and asymptotic distributions of chi-squared–type mixtures with applications. *Journal of the American Statistical Association*. 2005; 100:273–285.

Figure 1.

Power versus filtering thresholds for different gene sizes, based on the proposed unified test at the significance level of 10^{-5} . The number of SNPs with GxE effects, $k_1 = 3$, for all of the gene sizes. The optimal filtering threshold depends on the gene size: given k_1 , larger genes have more stringent optimal filtering thresholds. The solid black dots show the proposed adaptive filtering thresholds, which are often close to the true optimal ones (the peaks of the power curves). Especially for large genes, power is greatly improved with adaptive filtering compared to filtering with a fixed threshold of 0.1. At the same genome-wide error rate, the power of a two-stage method with the filtering threshold of 0.1 are 0.903, 0.836, 0.654, and 0.456 for $k = 6, 30, 100$, and 200 , respectively. These are 7–76% lower than the proposed test.

Figure 2.

The performance of the proposed filtering threshold with parameter misspecification and/or in the presence of LD, at the significance level of 10^{-5} . (a) Power with different GxE effect sizes when $k_1 = 3$ and $k = 100$; (b) Power when true $k_1 = 2, 3, 4$ and $k = 100$; (c) Power with or without LD for genes with 30 SNPs. In the simulation without LD, the true $k_1 = 3$; in the simulation with LD, the true $k_1 = 1$ with 4 other SNPs in moderate LD with the true one. The solid dots show the proposed filtering thresholds by specifying $k_1 = 3$ and log odds ratio of GxE as $\log(1.3)$. When effect size or k_1 is misspecified and/or in the presence of LD, the proposed filtering threshold is still close to the true optimal with ignorable power loss.

Figure 3. SNP-level gene by gender interactions for the 20 SNPs in the gene *ATG16L1* and the LD heatmap for the 7 SNPs that passed filtering. The blue dots correspond to the SNPs that passed filtering using the proposed threshold. The upper dashed line corresponds to a SNP-level GxE p -value of 0.05. The filtering effectively increases the proportion of SNPs with potential GxE: among the 7 SNPs with SNP-level GxE p -values ≤ 0.05 , 5 passed the filtering; among the 13 SNPs with SNP-level GxE p -values > 0.05 , only 2 passed the filtering. The LD heatmap shows that not all of the 7 SNPs passing filtering are in high LD.

Table 1

Power of the proposed method with misspecification of k_1 for polygenic traits at the significance threshold of 10^{-5} . We simulated the true $k_1 = \max(2; 0.05 \times k)$ and evaluated the power of the proposed test when k_1 is over- or underspecified. The power is calculated by averaging the results of 10,000 simulations, each with 3,000 cases and 3,000 controls.

k	k_1	Optimal empirical power	Power with true k_1	Under-specifying k_1		Over-specifying k_1		Power with fixed $\eta = 0.1$	Power by a two-stage method
				Specified k_1	Power	Specified k_1	Power		
6	2	0.329	0.283	1	0.218	3	0.248	0.246	0.203
30	2	0.210	0.201	1	0.136	3	0.198	0.122	0.121
100	5	0.634	0.632	2	0.560	10	0.580	0.507	0.503
200	10	0.956	0.948	4	0.929	20	0.934	0.938	0.919

Table 2

The proposed method controls the type I error rate for different gene sizes in the presence of LD. The numbers in the table are averaged over the genes in the corresponding gene size ranges. The simulation is repeated 100 times.

Significance threshold	$k = 1-10$	$k = 11-30$	$k = 30-100$	$k > 100$	Overall
0.05	0.050	0.050	0.050	0.050	0.050
0.01	0.009	0.009	0.009	0.009	0.009
0.001	8.5×10^{-4}	8.7×10^{-4}	9.2×10^{-4}	9.3×10^{-4}	8.8×10^{-4}
10^{-4}	1.1×10^{-4}	8.2×10^{-5}	1.2×10^{-4}	1.6×10^{-4}	10^{-4}
10^{-5}	1.7×10^{-5}	1.2×10^{-5}	1.5×10^{-5}	1.7×10^{-5}	1.4×10^{-5}

Table 3

The top 30 genes in the gene-gender interaction analysis with the proposed unified test based on the PanScan GWAS data.

Rank	Gene Name	Chr	P-value	P-value without filtering	Rank without filtering	Gene size after pruning (<i>k</i>)	# SNPs after filtering	The calculated optimal filtering threshold
1	<i>RRP7A</i>	22	0.00012	0.00078	19	7	4	0.071
2	<i>ZARIL</i>	13	0.00021	0.028	431	6	3	0.047
3	<i>TBCK</i>	4	0.00024	0.015	247	42	7	0.017
4	<i>UFSP2</i>	4	0.00025	0.027	418	8	4	0.055
5	<i>EED</i>	11	0.00029	0.018	297	9	3	0.039
6	<i>ATG16L1</i>	2	0.00045	0.034	519	20	7	0.025
7	<i>AIMP1</i>	4	0.00055	0.037	579	11	2	0.041
8	<i>XYLB</i>	3	0.00056	0.011	158	46	6	0.015
9	<i>SPICE1</i>	3	0.00056	0.016	263	18	5	0.026
10	<i>MBNL2</i>	13	0.00057	0.056	830	87	7	0.0089
11	<i>AHSA2</i>	2	0.00062	0.0043	70	10	3	0.035
12	<i>UGT2B11</i>	4	0.00065	0.0040	67	8	2	0.043
13	<i>KSR2</i>	12	0.00067	0.26	3655	575	10	0.0020
14	<i>OSGEP</i>	14	0.00076	0.096	1235	7	4	0.042
15	<i>ARSK</i>	5	0.00087	0.10	1408	9	6	0.048
16	<i>RAD54B</i>	8	0.00098	0.048	726	31	13	0.019
17	<i>ACOT6</i>	14	0.0011	0.0066	100	7	1	0.042
18	<i>IRAK1BP1</i>	6	0.0011	0.049	741	12	2	0.036
19	<i>HECTD2</i>	10	0.0011	0.012	178	6	2	0.11
20	<i>TTLL7</i>	1	0.0012	0.0038	65	18	6	0.025
21	<i>GIMD1</i>	4	0.0012	0.058	863	8	2	0.046
22	<i>GABRA1</i>	5	0.0012	0.080	1109	21	6	0.025
23	<i>SUN2</i>	22	0.0012	0.050	760	5	4	0.050
24	<i>ERO1LB</i>	1	0.0013	0.17	2398	43	6	0.015
25	<i>ROGDI</i>	16	0.0013	0.11	1449	6	3	0.045
26	<i>VAMP1</i>	12	0.0014	0.024	382	11	2	0.033
27	<i>EXD2</i>	14	0.0014	0.073	1031	15	4	0.028
28	<i>GXYLT1</i>	12	0.0015	0.092	1202	20	3	0.024
29	<i>FRMD5</i>	15	0.0015	0.026	402	58	6	0.015
30	<i>NIM1K</i>	5	0.0017	0.32	4498	51	18	0.014