Journal de la Société Française de Statistique
Vol. 159 No. 2 (2018)

Investigating Gene- and Pathway-environment Interaction analysis approaches

Camilo Broc[1], Marina Evangelou[2,3], Pascal Guenel[4], Therese Truong[4] and Benoit Liquet[1,5]

Abstract: Pathway analysis can increase power to detect associations with a gene or a pathway by combining several signals at the single nucleotide polymorphism (SNP)-level into a single test. In this work, we propose to extend two well-known self-contained methods, the Fisher's method (FM) and the Adaptive Rank Truncated Product (ARTP) method to the analysis of gene-environment (GxE) interaction at the gene and pathway-level. It has been previously suggested that the permutation procedures that are usually used to derive the significance of these tests are not appropriate for the analysis of GxE interaction and should be replaced by a bootstrap approach. We analyse and compare the performance of the extension of FM and ARTP using the permutation and the parametric bootstrap procedure in simulation studies. We illustrate its application by analysing the interaction between night work and circadian gene polymorphisms in the risk of breast cancer in a case-control study. The ARTP method, adapted for both gene- and pathway-environment interactions, gives promising results and has been wrapped to the R package PIGE available on the CRAN.

Résumé : Les analyses par pathway permettent d'augmenter la puissance statistique en combinant les signaux au niveau des SNPs pour définir des associations au niveau du gène et/ou du pathway. Dans cette étude, nous proposons d'adapter deux méthodes d'analyse par pathway, la méthode de Fisher (FM) et la méthode ARTP (Adaptive Rank Truncated Product), pour l'analyse des interactions gène-environnement (GxE) au niveau du gène et au niveau du pathway. Il a été précédemment suggéré que les procédures de permutations habituellement utilisées pour estimer la significativité de ces tests ne sont pas appropriées pour l'analyse des interactions GxE et devraient être remplacés par une approche Bootstrap. Ainsi, nous analysons et comparons dans une étude de simulation les performances de l'extension des méthodes FM et ARTP en utilisant une procédure de permutation et une méthode de Bootstrap paramétrique. Ces méthodes sont également appliquées aux données de l'étude cas-témoins CECILE sur les cancers du sein dans laquelle nous avons analysé l'interaction entre le travail de nuit et les polymorphismes des gènes circadiens dans le risque de cancer du sein. La méthode ARTP adaptée aux interactions GxE donne des résultats prometteurs. Un package R PIGE a été développé et est mis à disposition sur le CRAN.

Keywords: Gene-environment interactions, Generalized Linear Models, Pathway analysis, Resampling methods

Mots-clés : Interaction gène-environnement, Modèles linéaire généralisés, Analyse par pathway, Méthodes de ré-échantillonage

AMS 2000 subject classifications: 62F03, 62F40, 62P10

[1] Laboratoire de Mathématiques et de leur Applications, Université de Pau et Pays de l'Adour, UMR CNRS 5142, Pau, France. E-mail: camilo.broc@univ-pau.fr

[2] Department of Mathematics, Faculty of Natural Sciences, Imperial College London, UK

[3] Department of Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine, Imperial College London, UK. E-mail: m.evangelou@imperial.ac.uk

[4] Cancer and Environment team, CESP (Center for Research in Epidemiology and Population Health), INSERM, University Paris-Saclay, University Paris-Sud, Villejuif, FRANCE. E-mail: therese.truong@inserm.fr. E-mail: pascal.guenel@inserm.fr

[5] ARC Centre of Excellence for Mathematical and Statistical Frontiers and School of Mathematical, Sciences at Queensland University of Technology, Brisbane, Australia. E-mail: benoit.liquet@univ-pau.fr

 ISSN: 2102-6238

1. Introduction

During the last decade, genome-wide association studies (GWAS) have been successful in identifying several hundred single nucleotide polymorphisms (SNPs) associated with multiple cancer types (http://www.genome.gov/gwastudies/). However, such findings are not enough for explaining the genetic heritability of these cancers. Several reasons have been discussed that possibly explain the "missing heritability" in complex diseases such as the fact that most of these genetic associations were identified through single-SNP analyses (*each SNP tested independently*). It has been raised that polygenic effects, gene-gene and gene-environment (GxE) interactions are not fully explored in traditional methods (Manolio et al., 2009). Several approaches were developed in order to complete the agnostic GWAS in the discovery of additional genetic risk factors or to provide additional insights into the mechanisms involved in the studied disease.

One such approach is pathway analysis that consists of aggregating signals from SNPs (and/or genes) to pathways. Pathways are sets of genes that work together for the production of a specific biological outcome. Pathway analysis therefore incorporates the available biological knowledge of genes and SNPs for a better understanding of the genetic and biological mechanisms of the studied disease (Mooney et al., 2014, Pers (2016)). One of the main thrust of the statistical analyses will be to gain power and reduce the number of tests by combining weak signals from SNP-level analysis. Over the recent years, numerous pathway analysis for GWAS data have been proposed in the literature for finding pathways associated with the studied disease (a non exclusive list includes the methods proposed by Wang et al., 2007; Yu et al., 2009; Holmans et al., 2009; O'Dushlaine et al., 2009; Shahbaba et al., 2012; Carbonetto and Stephens, 2013; Evangelou et al., 2014a,b; Su et al., 2016). The challenges, properties and statistical methods for conducting pathway (and gene-set) analysis for GWAS data have been discussed and reviewed by Wang et al. (2011); Fridley and Biernacka (2011); de Leeuw et al. (2016).

These methods can be divided by the null hypothesis they test, namely the competitive (enrichment) or self-contained (association) null hypotheses. The self-contained null hypothesis states that no pathway genes are associated with the phenotype. On the other hand, the competitive hypothesis states that the statistics of the pathway genes are no more associated with the phenotype than the statistics of the genes outside of the pathway. A pathway where the competitive null hypothesis is rejected, is said to be an enriched one. The self-contained null hypothesis can be tested in both GWAS and candidate gene analysis, since only the statistics from a selection of genes is required. By contrast, competitive methods are usually used in GWAS data as all pathways are tested simultaneously. As discussed in the literature self-contained methods are generally more powerful than competitive methods (Evangelou et al., 2012).

Self-contained methods could also be classified into marginal approaches, which are based on the combination of p-values of individual SNPs (such as for instance Fisher's Method (FM), Adaptative Rank Truncated Product (ARTP)), or joint approaches that jointly model and test the effect of all the SNPs in the set (such as random and mixed effect models, Sequence Kernel association Test (SKAT) proposed by Ionita-Laza et al., 2013). p-value combination test statistics are usually combined with phenotype permutations for estimating their significance. Phenotype permutations avoid making any assumptions about the distribution of the effect of the genetic variants on the disease.

Another distinction factor is whether the pathway analysis is considered at the gene-level or

 ISSN: 2102-6238

at the SNP-level. Methods that perform a gene-based pathway analysis they first combine the p-values of single-SNP analysis into gene-level test statistics (or p-values) that are subsequently summarised into pathway-level associations. On the other hand, SNP level pathway methods skip the intermediate gene level and map SNPs directly to pathways.

Although a number of statistical approaches have been proposed to test for pathway association with disease, the literature has not been greatly extended for testing for GxE interactions at the pathway level. Lin et al. (2013) proposed a computationally efficient GxE set association test (GESAT), a variance component score test statistic is proposed that extends the SNP-set Kernel association test for GxE testing. The proposed method tests each set of SNPs independently from the other sets and it is a SNP-based pathway analysis approach.

In addition to this, Jiao et al. (2013) proposed the set based gene environment interaction (SBERIA) method and two more extensions that overcome the limitations of SBERIA (Jiao et al., 2015) for both rare and common variants. SBERIA firstly computes the correlation between the environmental factors and all SNPs in the set, where a z-score of correlation is obtained. These scores are translated into weights based on a preselected threshold that are included in a regression model that tests whether they are needed or not in the model. The first extension, named enhanced set-based GxE testing (eSBERIA), is composed of two steps: the first one tests the null hypothesis that the gene-environment weights are not associated with the response. The second step implements the SKAT statistic that accounts for any residual effects that might have been missed by the logistic regression model with the gene-environment interactions. As the two tests are independent, their p-values are combined using Fisher's product statistic. The third proposed approach is coSBERIA which combines SBERIA and SKAT tests for the case-only test. The case-only GxE test for a single SNP has been found to improve the power for testing for GxE under the assumption that G and E are independent (Albert et al., 2001).

In this work we were interested in extension of combination tests to the analysis of set based GxE interactions for which we have looked at replacing the phenotype permutation procedure for testing the significance of each pathway (and/or gene) by the bootstrap approach proposed by Buzkova et al. (2011). We have decided to compare the Fisher's product test statistic and ARTP approach as these two approaches have been discussed in the literature to be the most powerful pathway analyses among combination tests (Evangelou et al., 2012; Su et al., 2016). In contrast to the other proposed approaches GESAT and SBERIA, we are considering the case of gene-based pathway analyses over SNP-based ones. Further, Su et al. (2016) discussed the need for a fast algorithm to test for GxE interactions through pathway analysis and in this conducted work we aim to fill this gap.

A brief description of the context which has first motivated the development of our R package PIGE (Pathway Interaction Gene Environment, Liquet et al., 2017) is presented in Section 2. In Section 3, Fisher's Method (FM) and ARTP approaches are presented in context of gene and pathway-environment interaction. Both permutation and parametric bootstrap resampling methods are presented. In Section 4, a simulation study is presented to analyse the performances of FM and ARTP methods combined with both permutation and parametric bootstrap approaches. The methods are applied on genotype data from the CECILE case-control study in Section 5. Concluding remarks are presented in Section 6.

 ISSN: 2102-6238

2. Motivation

In a first step, we propose to test the performance of the proposed methods in simulated datasets that mimic the application dataset. In a second step, the methods will be illustrated on genotype data from the CECILE case-control study, in which we are interested in the interaction between nightwork, a binary environmental factor (defined as ever worked at night more than two years: yes/no) and polymorphisms from genes in the circadian rhythm pathway (Truong et al., 2014).

In the case of a binary response Y, the null hypothesis that there is no association between the response and the interaction term between SNP_ℓ and environment is evaluated through the following logistic model:

$$\text{logit}[P(Y=1|SNP_\ell,E)] = \alpha_\ell + \beta_\ell SNP_\ell + \beta_{E,\ell}E + \gamma_\ell E \times SNP_\ell. \quad (1)$$

where E presents the environmental factor. The likelihood ratio test (LRT) could be used to test the evidence of the interaction term ($H_{0,\ell}: \ \gamma_\ell = 0$ versus $H_{1,\ell}: \ \gamma_\ell \neq 0$), resulting to the ℓ-th p-value (p_ℓ).

As discussed earlier, there are usually multiple SNPs within each gene and multiple genes within each pathway. The questions that we will answer through our conducted work are: how to combine these results to get (i) association evidence between gene-environment interaction and the outcome, (ii) association evidence between pathway-environment interaction and the outcome?

In the context of a gene-based pathway analysis, a two-step procedure is needed. At the first level the association evidence between a gene and the response is found and at the second level these gene-level p-values are combined into a test statistic for the disease-pathway association.

Phenotype permutations are usually implemented for computing the null distribution of the test statistic that can be used for obtaining a p-value for the global null hypothesis of no association between the gene with the response. In this work, we are investigating the performance of two alternative resampling approaches one based on phenotype permutations and a second one on the bootstrap approach proposed by Buzkova et al. (2011) that has been proposed for interaction models. Both these resampling approaches are presented in Section 3.3.

3. Methods

In this section, we first present two frequentist approaches for combining p-values under investigation FM, and ARTP methods. We subsequently present the two resampling approaches. Finally, we shortly present an alternative frequentist approach iSKAT.

3.1. Fisher's method

Fisher's method is a well established association method that combines the results from multiple statistical tests. The FM test statistic equals

$$FM = -2\sum_{\ell=1}^{L}\log(p_\ell) = -2\log\left(\prod_{\ell=1}^{L}p_\ell\right), \quad (2)$$

 ISSN: 2102-6238

where L is for example the number of SNPs within a pathway. Under null hypothesis the FM test statistic follows a χ^2 distribution with $2L$ degrees of freedom when the p-values are independent. In the presence of linkage disequilibrium, the correlation between SNPs leads to dependent test-statistics. We have used the resampling approaches (presented in Section 3.3) to approximate the empirical distribution of the FM test statistics.

3.2. Adaptive rank truncated product (ARTP)

The idea behind the ARTP is to truncate the highest p-values in the FM method. The only p-values left are the most significant ones. To simplify the presentation of the ARTP proposed by Yu et al. (2009), we consider a pathway consisting of L SNPs and we want to test the null hypothesis that there is no pathway-environment interaction associated to the disease phenotype. Using model (1), we can perform a LRT test on individual interaction E×SNPs within the considered pathway. We denote the ordered statistics of those p-values $p_{(1)} \leq \ldots \leq p_{(L)}$, with $p_{(\ell)}$ being the ℓ-th smallest p-value. The original RTP statistic given by

$$W_K = \sum_{k=1}^{K} \log\left(p_{(k)}\right) = \log\left(\prod_{k=1}^{K} p_{(k)}\right) \tag{3}$$

combines the K smallest p-values E×SNP statistics of the tested pathway (Dudbridge and Koeleman, 2003). In the adaptive RTP method J different truncation $K_1 \leq \ldots \leq K_J$ are investigated. Let $\hat{s}(K_j)$ be the estimated p-value for W_{K_j}, $(1 \leq j \leq J)$. The ARTP statistic is then defined using the minimum p-value procedure

$$MinP = \min_{1 \leq j \leq J} \hat{s}(K_j). \tag{4}$$

Note that for a single truncation point ($J = 1$), the ARTP method is the RTP method and the RTP statistic simplifies to the FM test statistic when the truncation point K is fixed to L. Two levels of resampling approach are required to get the adjusted p-value for $MinP$: (1) for estimating $\hat{s}(K_j)$, (2) for the adjustment for multiple testing over different truncation points. To avoid this computational issue specially when the number of test L is large, Yu et al. (2009) uses the Ge et al. (2003)'s algorithm which reduces the multiple-level resampling procedure into a single level resampling procedure. In this work, we use the same algorithm.

Let $p_1^{(0)}, \ldots, p_L^{(0)}$ be the p-values for each interaction test on the null hypothesis based on the observed data. We generate B datasets under the complete null hypothesis $H_0 = H_{0,1} \cap \ldots H_{0,L}$ using appropriate resampling procedure (see section 3.3). Let $p_1^{(b)}, \ldots, p_L^{(b)}$ be the p-values for each interaction test on the null hypothesis based on the b-th generated dataset, $1 \leq b \leq B$. The RTP statistic

$$W_j^{(b)} = \sum_{i=1}^{K_j} \log\left(p_{(i)}^{(b)}\right), \quad 0 \leq b \leq B,\ 1 \leq j \leq J \tag{5}$$

is calculated for each truncation point, for both the observed data-set and each of the B simulated datasets. Then Ge's algorithm is used to estimate the p-value

$$\hat{s}_j^{(b)} = \frac{\sum_{b^*=0}^{B} I\left(W_j^{(b^*)} \leq W_j^{(b)}\right)}{B+1}, \quad 0 \leq b \leq B,\ 1 \leq j \leq J \tag{6}$$

 ISSN: 2102-6238

for each W_j. The p-value for the ARTP statistic $MinP^{(0)}$ of the pathway is estimated as

$$\widehat{\mathrm{p}}_{ARTP} = \frac{\sum_{b^*=0}^{B} I\left(MinP^{(b)} \leq MinP^{(0)}\right)}{B+1}, \tag{7}$$

where

$$MinP^{(b)} = \min_{1\leq j\leq J} \hat{s}_j^{(b)}, \quad 0 \leq b \leq B,\ 1 \leq j \leq J. \tag{8}$$

Remark. The adjusted p-value for $MinP^{(b)}$, the ARTP statistic from the b-th dataset, can also be estimated similarly using $\frac{\sum_{b^*=0}^{B} I\left(MinP^{(b^*)} \leq MinP^{(b)}\right)}{B+1}$.

Thus this procedure can give an evidence of association between a pathway-environment interaction and the disease outcome. It is called a SNP-based strategy. We describe in the following the gene-based strategy consisting to used the ARTP method for both derive the gene-environment interaction level summary and to combine gene-environment interaction level p-values across all genes within a pathway. This procedure adapted for interaction investigation is the one described in Yu et al. (2009).

Consider a pathway composed of L genes, with the ℓ-th composed of n_ℓ SNPs, $1 \leq \ell \leq L$. Let $p_{\ell,i}^{(0)}$ be the p-value for the association test on the i-th interaction SNP×environment of the ℓ-th gene based on the observed dataset. We then generate using resampling approach B datasets under the null hypothesis, and define $p_{\ell,j}^{(b)}$ the p-value for the test on the ith interaction SNP×environment of the ℓ-th gene based on the b-th generated dataset, $1 \leq b \leq B$. The ARTP is then applied (with a predetermined set of candidate truncation points, which could be varied from gene to gene) to combine interaction SNPs×environment-level evidence of association within a gene. For the ℓ-th gene, we apply the minimum p-value procedure ($MinP$) given earlier on, $1 \leq i \leq n_\ell$, $0 \leq b \leq B$, to obtain $p_\ell^{*(0)}$, the interaction gene×environment-level p-value for the observed data, and $p_\ell^{*(b)}$, the interaction GxE level p-value for the b-th permuted dataset. Finally in order to get a evidence of interaction pathway×environment the ARTP statistic is used to combine the gene×environment-level p-values for the observed and the resampling "null" data sets. We use the $MinP$ procedure one more time to obtain the adjusted p-value for the pathway×environment-level ARTP statistic. Note that the same set of generated "null" datasets are exploited each time for the $MinP$ procedure to derive interaction gene×environment-level and interaction pathway×environment-level evidence. Thus the full procedure overcomes the expensive computational multi-layer resampling issue. The same procedure is used for FM method.

3.3. Resampling methods

Both ARTP and FM rely on appropriate resampling strategy to generate data set under the null hypothesis considered. For gene- and pathways- environment interaction, we consider the global null hypothesis:

$$H_0 = H_{0,1} \cap \ldots H_{0,\ell} \ldots \cap H_{0,L}, \quad \text{with } H_{0,\ell} :\ \gamma_\ell = 0 \text{ (see equation (1))} \tag{9}$$

where L is the number of SNPs within a considered pathway.

 ISSN: 2102-6238

Yu et al. (2009) use a permutation procedure to evaluate the significance level of the ARTP statistic in the context of disease-pathway association which for example corresponds to the situation of the simplified model:

$$\text{logit}[P(Y=1|SNP_\ell,E)] = \alpha_\ell + \beta_\ell SNP_\ell \tag{10}$$

with $H_{0,\ell}: \ \beta_\ell = 0$. In this situation, there is no difficulty to define the permutation procedure for the complete null hypothesis. One just need to permute the phenotype Y. However, for the interaction model (1), a valid permutation procedure to generate data set under the complete null hypothesis (9) is complex to define. As noted by Buzkova et al. (2011), fixing SNP and E and permuting Y generates data in which the generated phenotype Y^* is independent of SNP and E. This procedure fails to generate data set under the null hypothesis since in model (1) the phenotype Y is not independent of SNP and E. Indeed, for the logistic model there is no permutation procedure which can be used to generate data set for the complete null hypothesis (Edgington, 1987). An alternative to the permutation procedure is to used a parametric bootstrap procedure (Efron and Tibshirani, 1994; Liquet and Riou, 2013) which implies less stringent assumptions (Good, 2000). In our context, the procedure could be defined in the following.

For each SNP ($\ell = 1,\ldots,L$):

1. Fit the model under the null hypothesis $H_{0,\ell}$, using the observed data, and obtain $\hat{\alpha}_\ell$, $\hat{\beta}_\ell$, $\hat{\beta}_{E,\ell}$, the maximum likelihood estimate (MLE) of respectively α_ℓ, β_ℓ and $\beta_{E,\ell}$
2. Generate a new outcome $Y^*_{i,\ell}$ for each subject from the probability measure defined under $H_{0,\ell}$. For example, for model (1), we generate $Y^*_{i,\ell}$ according to:

$$P(Y^*_{i,\ell}=1|SNP_\ell,E) = \frac{\exp\left(\hat{\alpha}_\ell + \hat{\beta}_\ell SNP_\ell + \hat{\beta}_{E,\ell}E\right)}{1+\exp\left(\hat{\alpha}_\ell + \hat{\beta}_\ell SNP_\ell + \hat{\beta}_{E,\ell}E\right)}.$$

   Repeat this for all the subjects to obtain a sample noted $s^*_\ell = \{Y^*_{i,\ell}, SNP_{i,\ell}, E_i\}$ which is related to the ℓ-th SNP.
3. Generate B new datasets $s^*_{b,\ell}$, $b = 1,\ldots,B$ by repeating B times the steps 1, 2 and 3.

Remark: In case of marginal association of both SNP and environmental factor, step 2 might generate unbalanced data which could affect the statistical power of the resampling methods. A screening investigation on the marginal association might be used before using the booststrap method.

3.4. iSKAT

The other frequentist approach is iSKAT proposed by Lin et al. (2016). The method uses the spirit of SKAT-O methods (Wu et al. (2011)) and apply it to an interaction test context. From one side burden tests are know to be an efficient test in many cases but they struggle when rare variants are involved in the data. From the other side kernel test can handle those rare variants. The idea behind the algorithm is to separate from the data the rare variants from the rest and to take advantage of both burden tests and kernel tests. Furthermore, iSKAT offer the possibility of weighting the covariates to take into account extra information. However, no weight have been added in the use of iSKAT in this article. The method GESAT is a particular case of the iSKAT method.

 ISSN: 2102-6238

4. Simulation Study

In this section, we compare the FM and ARTP methods through a simulation study investigating their control of Type-I error and FWER and their power performance. Two resampling approaches (permutation and bootstrap) are compared for a range of sample sizes ($n = 200,\ 500, 1000$). The combination methods FM and ARTP are compared to iSKAT and to the popular frequentist approach MinP which combines p-value by considering only the most signifactive p-value:

$$\text{MinP} = \min_{\ell \in \{1,\dots,L\}} p_\ell.$$

Let's note that MinP method doesn't have the same meaning than the quantity $MinP$ used in the intermediary steps of ARTP (see equation (4)).

4.1. Data Generated

We work on generated data which are supposed to mimic experimental data. The parameters of the generation are inspired from Buzkova et al. (2011). The genetic structure simulated is composed by one pathway containing I genes (genes are called $G_1, G_2, \dots, G_I$). Each gene contains several SNPs. The SNPs are binary variables. In order to generate the i-th gene, composed by k_i SNPs, $SNP_1^i \dots SNP_{k_i}^i$, we use the following procedure:

$$\begin{aligned} S_i &\sim Bern(0.2) \\ \text{logit}(p_j) &= \text{logit}(0.2) + S_i \text{ for } j \in \{1,\dots,k_i\} \\ SNP_j^i | S_i &\sim Bern(p_j) \text{ for } j \in \{1,\dots,k_i\}. \end{aligned}$$

Hence, conditionally on the latent polymorphism S_i, for a given gene i, the individual SNP_j^i are independent and identically distributed, but they are marginally dependent.

A binary environment variable is also simulated that is marginally dependent with one gene i_E and generated with the following procedure:

$$\begin{aligned} \text{logit}(p_E) &= a + bS_{i_E} \\ P[E = 1] &= p_E \end{aligned}$$

Finally, a binary outcome variable is simulated. It is generated from a logistic model using SNPs from gene i_Y. Among those s' SNPs only s SNPs are associated to the response variable Y as specified in the following equations:

$$\lambda_1, \dots, \lambda_{s'} \in \{0,1\} \text{ and } \sum_{l=1}^{s'} \lambda_l = s \tag{11}$$

$$\text{logit}[P(Y = 1 | G_{i_Y}, E)] = \alpha + \beta_E E + \sum_{j=1}^{k_{i_Y}} \lambda_j \beta_{SNP_j^{i_Y}} SNP_j^{i_Y} + \sum_{j=1}^{k_{i_Y}} \lambda_j \gamma_{SNP_j^{i_Y}} SNP_j^{i_Y} \times E \tag{12}$$

 ISSN: 2102-6238

The parameters λ_l control the choice of the SNP involved in the generation of Y and the parameter s controls the number of those SNPs. Different choices for parameters β_j and γ_j are chosen in order to highlight different results. For Type I error results the parameters γ_j are set to 0 whereas for power results parameters β_j and γ_j are chosen in order to evidence the different rejection of the null hypothesis for different parameters and methods used.

4.2. Simulation design

We present nine different simulation models. The first four ones are used to investigate the Type-I error and the Family Wise Error Rate (FWER) results while the others are used to explore the power of the different methods.

As discussed above, different resampling methods are used: (i) permutation that permutes the outcome Y and (ii) parametric bootstrap. We set to 1,000 the number of permutations and bootstrap resampling. The sample size n of the simulation datasets are 200, 500 and 1000. The results we look for are the p-values at gene level that are based on SNP-level information. Then a pathway-level p-values is computed from this information.

We use cases 1, 2, 3 and 4 (defined in the following) to investigate the control of the Type-I error rate for each gene. We also investigate the control of the FWER at the gene level and at the pathway level. An empirical FWER for gene-environment interaction is defined by the number of times the procedure detects wrongly at least one significant gene-environment interaction (from all the genes within the Pathway) over the $N = 500$ simulation replications. We also add the empirical FWER computed using a Bonferroni correction (i.e., each gene-environment interaction p-value is divided by the number of investigated genes). Indeed, the output of each method is a set of p-values (one for each investigating gene-environment interaction). This set of p-values is associated to a set of null hypotheses which define our family of hypotheses. Then it is important to control an overall error for these hypotheses. The empirical FWER for pathway-environment interaction is defined by the number of times the procedure detects wrongly a significant pathway-environment interaction over the 500 replications. As only one pathway is considered in our simulation study, the control of the FWER at the pathway level is similar to the Type-I error rate control of the pathway investigated.

Simulation cases for investigating Type-I error rate and FWER

— **Case 1**: Data is composed of 5 genes with 10 SNPs each. True model is based on the main effect of E and the main effect of 5 randomly selected SNP from the first gene. The environment is marginally correlated with the first gene but not with the other genes. The outcome is therefore generated from the following model:

$$\lambda_1, \ldots, \lambda_{10} \in \{0,1\} \text{ and } \sum_{j=1}^{10} \lambda_j = 5$$

$$\text{logit}[P(Y=1|G_1,E)] = \alpha + \beta_E E + \sum_{j=1}^{10} \lambda_j \beta_{SNP_j^1} SNP_j^1$$

 ISSN: 2102-6238

— **Case 2**: Data is composed of 5 genes with 10 SNPs The true model is based on the main effect of E and the main effect of all SNPs from the second gene. The environment is marginally correlated with first gene but not with other genes. The outcome is therefore generated from the following model:

$$\text{logit}[P(Y=1|G_2,E)] = \alpha + \beta_E E + \sum_{j=1}^{10} \beta_{SNP_j^2} SNP_j^2$$

— **Case 3**: Data is composed of 5 genes, 4 of them with 5 SNPs and one with 50 SNPs (the last one).True model is based on the main effect of E and 2 randomly selected SNPs from the first gene. The environment is marginally correlated with the first gene but not with other genes. The outcome is therefore generated from the following model:

$$\lambda_1,\ldots,\lambda_5 \in \{0,1\} \text{ and } \sum_{j=1}^{5} \lambda_j = 2$$

$$\text{logit}[P(Y=1|G_1,E)] = \alpha + \beta_E E + \sum_{j=1}^{5} \lambda_j \beta_{SNP_j^1} SNP_j^1$$

This model enables us to see how the methods perform when the pathway gene members have different sizes.

— **Case 4**: Data is composed of 20 genes with 20 SNPS for the first gene and 10 SNPs for the others. True model is based on: the main effect of E; the main effect of 10 randomly selected SNPs from the first gene and 5 from the second genes. The environment is marginally correlated with first gene but not with other genes. The outcome is therefore generated from the following model:

$$\lambda_1^1,\ldots,\lambda_{20}^1 \in \{0,1\} \text{ with } \sum_{j=1}^{20} \lambda_j^1 = 10 \text{ and } \lambda_1^2,\ldots,\lambda_{10}^2 \in \{0,1\} \text{ with } \sum_{j=1}^{10} \lambda_j^2 = 5$$

$$\text{logit}[P(Y=1|G_1\,,G_2,E)] = \alpha + \beta_E E + \sum_{j=1}^{20} \lambda_j^1 \beta_{SNP_j^1} SNP_j^1 + \sum_{j=1}^{10} \lambda_j^2 \beta_{SNP_j^2} SNP_j^2$$

Simulation cases for power performance

— **Case 5**: Data is composed of 5 genes with 20, 10, 10, 10, 10 SNPs. True model is based on the main effect of E and the main effect of 10 randomly selected SNPs from the first gene and the interaction between the environment with each of the selected SNPs. The environment is marginally correlated with first gene but not with other genes. The outcome is therefore generated from the following model:

$$\lambda_1,\ldots,\lambda_{20} \in \{0,1\} \text{ and } \sum_{j=1}^{20} \lambda_j = 10$$

$$\text{logit}[P(Y=1|G_1,E)] = \alpha + \beta_E E + \sum_{j=1}^{20} \lambda_j \beta_{SNP_j^1} SNP_j^1 + \sum_{j=1}^{20} \lambda_j \gamma_{SNP_j^1} SNP_j^1 \times E$$

 ISSN: 2102-6238

— **Case 6**: Data is composed of 5 genes with 20, 10, 10, 10, 10 SNPs. True model is based on the main effect of E and the main effect of 2 randomly selected SNPs from the first gene and the interaction between the environment with each of the selected SNPs. The environment is marginally correlated with first gene but not with other genes. The outcome is therefore generated from the following model:

$$\lambda_1,\dots,\lambda_{20} \in \{0,1\} \text{ and } \sum_{j=1}^{20} \lambda_j = 2$$

$$\text{logit}[P(Y=1|G_1,E)] = \alpha + \beta_E E + \sum_{j=1}^{20} \lambda_j \beta_{SNP_j^1} SNP_j^1 + \sum_{j=1}^{20} \lambda_j \gamma_{SNP_j^1} SNP_j^1 \times E$$

— **Case 7**: Data is composed of 20 genes with 20 SNPS for the first gene and 10 SNPs for the others. True model is based on: the main effect of E; the main effect of 10 randomly selected SNPs from the first gene and 5 from the second genes; the interactions between the environment with each of the selected SNPs. The environment is marginally correlated with first gene but not with other genes. The outcome is therefore generated from the following model:

$$\lambda_1^1,\dots,\lambda_{20}^1 \in \{0,1\} \text{ with } \sum_{j=1}^{20} \lambda_j^1 = 10 \text{ and } \lambda_1^2,\dots,\lambda_{10}^2 \in \{0,1\} \text{ with } \sum_{j=1}^{10} \lambda_j^2 = 5$$

$$\begin{aligned}\text{logit}[P(Y=1|G_1,G_2,E)] = \alpha + \beta_E E + \sum_{j=1}^{20} \lambda_j^1 \beta_{SNP_j^1} SNP_j^1 + \sum_{j=1}^{10} \lambda_j^2 \beta_{SNP_j^2} SNP_j^2 \\ + \sum_{j=1}^{20} \lambda_j^1 \gamma_{SNP_j^1} SNP_j^1 \times E + \sum_{j=1}^{10} \lambda_j^2 \gamma_{SNP_j^2} SNP_j^2 \times E\end{aligned}$$

— **Case 8**: Data is composed of 20 genes with 20 SNPS for the first gene and 10 SNPs for the others. True model is based on: the main effect of E; the main effect of 2 randomly selected SNPs from the first gene and 2 from the second genes; the interactions between the environment with each of the selected SNPs. The environment is marginally correlated with first gene but not with other genes. The outcome is therefore generated from the following model:

$$\lambda_1^1,\dots,\lambda_{20}^1 \in \{0,1\} \text{ with } \sum_{j=1}^{20} \lambda_j^1 = 2 \text{ and } \lambda_1^2,\dots,\lambda_{10}^2 \in \{0,1\} \text{ with } \sum_{j=1}^{10} \lambda_j^2 = 2$$

$$\begin{aligned}\text{logit}[P(Y=1|G_1,E)] = \alpha + \beta_E E + \sum_{j=1}^{20} \lambda_j^1 \beta_{SNP_j^1} SNP_j^1 + \sum_{j=1}^{10} \lambda_j^2 \beta_{SNP_j^2} SNP_j^2 \\ + \sum_{j=1}^{20} \lambda_j^1 \gamma_{SNP_j^1} SNP_j^1 \times E + \sum_{j=1}^{10} \lambda_j^2 \gamma_{SNP_j^2} SNP_j^2 \times E\end{aligned}$$

— **Case 9**: Data is composed of 2 pathways with 10 genes in each pathways. Each genes includes 10 SNPS. True model is based on: the main effect of E; the main effect of 2 randomly selected SNPs from the first gene and 2 from the second genes of each pathways;

 ISSN: 2102-6238

the interactions between the environment with each of the selected SNPs. The outcome is therefore generated from the following model:

$$\lambda_1^1,\ldots,\lambda_{10}^1 \in \{0,1\} \text{ with } \sum_{j=1}^{10}\lambda_j^1 = 2 \text{ and } \lambda_1^2,\ldots,\lambda_{10}^2 \in \{0,1\} \text{ with } \sum_{j=1}^{10}\lambda_j^2 = 2$$

$$\lambda_1^{11},\ldots,\lambda_{10}^{11} \in \{0,1\} \text{ with } \sum_{j=1}^{10}\lambda_j^{11} = 2 \text{ and } \lambda_1^{12},\ldots,\lambda_{10}^{12} \in \{0,1\} \text{ with } \sum_{j=1}^{10}\lambda_j^{12} = 2$$

$$\text{logit}[P(Y=1|G_1,E)] = \alpha + \beta_E E + \sum_{k\in K}\sum_{j=1}^{10}\lambda_j^k \beta_{SNP_j^k} SNP_j^k + \sum_{k\in K}\sum_{j=1}^{10}\lambda_j^k \gamma_{SNP_j^k} SNP_j^k \times E$$

where $K = \{1,2,11,12\}$.

Simulation parameters

The different coefficient used in our cases are gathered in table 1. The notation used refer to part 4.1. For each simulation case, FM, and ARTP methods are applied for the 9 simulation cases to investigate the presence of interaction effects of gene- and pathway- environment based on a gene-based strategy (see end of Section 3.2).

For all of the 9 cases the truncation points of the ARTP parameters are optimized like in previous ARTP results Yu et al. (2009). Let m be an integer; k_G and k_{SNP} be real numbers; k_i be the number of SNP in the i-th gene; I the number of genes. Let $\lfloor k_G \times I \rfloor, \lfloor 2 \times k_G \times I \rfloor, \ldots, \lfloor m \times k_G \times I \rfloor$ be a set of truncation points for genes and, for each gene i, let $\lfloor k_{SNP} \times k_i \rfloor, \lfloor 2 \times k_{SNP} \times k_i \rfloor, \ldots, \lfloor m \times k_{SNP} \times k_i \rfloor$ be a set of truncation points for the SNPs of this gene. The notation $\lfloor x \rfloor$ gives the largest integer that does not exceed x (if $\lfloor x \rfloor = 0$ we set the value to 1) . We define $p_{k_{SNP},k_G,m}$ the p-value of the ARTP computed with this set of truncation points. The optimal p-value of the ARTP is defined as:

$$\min_{k_{SNP}\in\mathscr{A},k_G\in\mathscr{A}} p_{k_{SNP},k_G,m} \quad \text{with} \quad \mathscr{A} = \{2\%,4\%,\ldots,20\%\}.$$

The parameter m is fixed to 5 in our study.

4.3. Type I error rate and FWER

Type I error rate and FWER of the methods are computed in cases 1, 2, 3 and 4. The data are generated under the null hypothesis (i.e. no interaction). The SNP-level tests are performed under the interaction assumption with a significance level of 0.05. Hence, the expected value of all p-values at gene-level and pathway level are 0.05. In the section, we study the behavior of the different methods for this case. Tables 2, 3, 4 , 5 , 6 and 7 present the results. The expected value of the average p-values is 0.05. Computing a binomial model, the average of the p-values on the 500 iterations should be between 3% and 7%.

 ISSN: 2102-6238

TABLE 1. *Generating parameters for cases 1 to 9. The notation used refer to part 4.1.*

	p_E	a	b	α	β_E	$\beta_{SNP_j^{i_Y}}$ $i_Y \in \{1,\dots,k_{i_Y}\}$	$\gamma_{SNP_j^{i_Y}}$ $i_Y \in \{1,\dots,k_{i_Y}\}$
case 1	0.2	logit(2)	log(2)	-2	2	$\in \{3,2,1\}$	$=0$
case 2	0.2	logit(2)	log(2)	-2	2	$\in \{1.5,1.0,0.5\}$	$=0$
case 3	0.2	logit(2)	log(2)	-2	2	$\in \{3,2,1\}$	$=0$
case 4	0.5	logit(2)	2	-1	1	$\in \{0.3,0.2,0.1\}$	$=0$
case 5	0.2	logit(2)	log(2)	-2	0.4	$\in \{0.06,0.04,0.02\}$	$=0.5$
case 6	0.2	logit(2)	log(2)	-2	2	$\in \{0.3,0.2,0.1\}$	$=1.5$
case 7	0.5	logit(2)	2	-1	0.1	$\in \{0.03,0.02,0.01\}$	$=0.3$
case 8	0.5	logit(2)	2	-1	0.1	$\in \{0.075,0.050,0.025\}$	$=0.7$
case 9	0.5	logit(2)	2	-1	0	$=0$	$=0.5$

The permutation approach obtained very low error rate for both approaches (FM, MinP, ARTP). The bootstrap approach gives good results for controlling the Type-I error rate for both FM and ARTP methods. As expected the FWER at the gene level is not controlled. The FWER computed can then be corrected using the Bonferroni method which is known to be conservative and more trustable. Finally, a pathway p-value is given by each combining method (FM, MinP, ARTP). When the number of genes is low (cases 1 to 3), the Type-I error rate at the pathway level is well controlled using the bootstrap approach for both FM, MinP and ARTP methods while permutation approach give conservative results. When the number of genes is higher (case 4), the type-I error rate at the pathway level of the ARTP and iSKAT is slightly inflated.

4.4. Power performance

Tables 8, 9, 10, 11 and 12 present the results for the power of the methods. As expected for all methods power performances increase with larger sample size.

In those results the proportion of significant SNP in the true model have a huge importance on the performances. In cases 5 and 7 the proportions of significant SNPs are high whereas in cases 6, 8 and 9 they are low. For higher proportions bootstrap is slightly but consistently more powerful than the permutation. For lower proportions permutation and bootstrap results are equivalent. We can see that when the proportion is high FM and ARTP have equivalent results and MinP has abysmal results. This is due to the fact that MinP truncates too much of the information contained in the data. When the proportion are lower, ARTP and then MinP have a good performance but FM have lower ones. This is due to the fact that FM results take into account too much irrelevant SNPs in its combination. The ARTP have the merit of having good power whatever is the proportion of significant SNP in the true model. FM and MinP seems to detect different patterns but ARTP can detect both. When we compare the combining methods with iSKAT we can see that the level of performance of ARTP and iSKAT is similar. We notice that in general, ARTP is more powerful on small sample sizes (200 and 500).

 ISSN: 2102-6238

TABLE 2. *Simulation of case 1 with 1000 permutations and 1000 bootstrap resampling with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0	0.03	0.036	0.04	0.024	0.038	0.024
Gene2	10	0.016	0.08	0.04	0.04	0.038	0.076	0.052
Gene3	10	0.022	0.074	0.048	0.064	0.042	0.078	0.054
Gene4	10	0.008	0.056	0.04	0.052	0.028	0.05	0.05
Gene5	10	0.02	0.07	0.04	0.05	0.042	0.066	0.038
$FWER_{BF}$		0.014	0.066	0.046	0.052	0.042	0.068	0.038
$FWER$		0.066	0.282	0.188	0.23	0.166	0.29	0.2
Type-I Error: Pathway		0.014	0.066	0.048	0.052	0.028	0.078	0.052

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.002	0.016	0.008	0.01	0.01	0.02	0.03
Gene2	10	0.028	0.062	0.052	0.062	0.044	0.064	0.044
Gene3	10	0.028	0.046	0.028	0.044	0.028	0.05	0.038
Gene4	10	0.018	0.034	0.03	0.044	0.028	0.038	0.03
Gene5	10	0.028	0.054	0.036	0.042	0.042	0.054	0.066
$FWER_{BF}$		0.028	0.052	0.038	0.042	0.042	0.064	0.042
$FWER$		0.098	0.188	0.148	0.188	0.142	0.208	0.186
Type-I Error: Pathway		0.03	0.052	0.038	0.042	0.042	0.078	0.04

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.008	0.032	0.026	0.028	0.018	0.028	0.032
Gene2	10	0.032	0.04	0.03	0.048	0.036	0.052	0.042
Gene3	10	0.022	0.032	0.036	0.044	0.04	0.05	0.038
Gene4	10	0.044	0.054	0.046	0.064	0.06	0.076	0.054
Gene5	10	0.036	0.036	0.032	0.046	0.036	0.042	0.042
$FWER_{BF}$		0.026	0.04	0.036	0.04	0.04	0.048	0.03
$FWER$		0.138	0.178	0.16	0.21	0.178	0.228	0.196
Type-I Error: Pathway		0.026	0.04	0.036	0.04	0.042	0.048	0.044

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 3. *Simulation of case 2 with 1000 permutations and 1000 bootstrap resampling with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.018	0.09	0.032	0.046	0.03	0.094	0.046
Gene2	10	0.002	0.028	0.054	0.044	0.036	0.032	0.04
Gene3	10	0.008	0.062	0.03	0.046	0.02	0.054	0.038
Gene4	10	0.012	0.084	0.042	0.056	0.04	0.104	0.062
Gene5	10	0.016	0.076	0.03	0.038	0.032	0.07	0.038
$FWER_{BF}$		0.008	0.084	0.03	0.034	0.04	0.076	0.038
$FWER$		0.056	0.304	0.172	0.212	0.152	0.31	0.204
Type-I Error: Pathway		0.008	0.086	0.03	0.036	0.03	0.09	0.028

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.02	0.046	0.028	0.05	0.022	0.05	0.032
Gene2	10	0.002	0.032	0.036	0.038	0.03	0.03	0.058
Gene3	10	0.028	0.07	0.038	0.046	0.034	0.066	0.054
Gene4	10	0.028	0.062	0.042	0.054	0.044	0.076	0.044
Gene5	10	0.018	0.05	0.036	0.052	0.026	0.046	0.026
$FWER_{BF}$		0.016	0.046	0.042	0.042	0.042	0.06	0.036
$FWER$		0.09	0.232	0.164	0.214	0.146	0.238	0.196
Type-I Error: Pathway		0.016	0.046	0.048	0.044	0.032	0.066	0.036

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.056	0.07	0.044	0.064	0.056	0.078	0.046
Gene2	10	0.014	0.03	0.022	0.03	0.024	0.026	0.04
Gene3	10	0.038	0.054	0.03	0.042	0.042	0.06	0.038
Gene4	10	0.032	0.042	0.028	0.036	0.036	0.046	0.062
Gene5	10	0.042	0.048	0.028	0.038	0.038	0.046	0.038
$FWER_{BF}$		0.032	0.048	0.032	0.038	0.034	0.07	0.038
$FWER$		0.164	0.216	0.144	0.194	0.178	0.226	0.204
Type-I Error: Pathway		0.032	0.048	0.032	0.04	0.042	0.076	0.044

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 4. *Simulation of case 3 with 1000 permutations and 1000 bootstrap resampling with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	5	0.02	0.064	0.042	0.05	0.034	0.054	0.044
Gene2	5	0.026	0.07	0.048	0.068	0.038	0.07	0.07
Gene3	5	0.03	0.068	0.04	0.06	0.044	0.074	0.054
Gene4	5	0.018	0.04	0.026	0.036	0.02	0.034	0.04
Gene5	50	0.01	0.144	0.066	0.066	0.062	0.16	0.192
$FWER_{BF}$		0.02	0.144	0.044	0.046	0.052	0.11	0.072
$FWER$		0.096	0.32	0.19	0.242	0.176	0.326	0.348
Type-I Error: Pathway		0.022	0.146	0.046	0.048	0.058	0.132	0.008

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	5	0.026	0.048	0.036	0.05	0.032	0.048	0.058
Gene2	5	0.028	0.042	0.038	0.052	0.036	0.056	0.042
Gene3	5	0.038	0.056	0.048	0.056	0.044	0.068	0.06
Gene4	5	0.042	0.062	0.032	0.062	0.04	0.06	0.052
Gene5	50	0.032	0.098	0.032	0.038	0.06	0.11	0.088
$FWER_{BF}$		0.038	0.08	0.042	0.05	0.044	0.092	0.088
$FWER$		0.156	0.278	0.172	0.232	0.198	0.304	0.278
Type-I Error: Pathway		0.038	0.084	0.044	0.052	0.034	0.096	0.02

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	5	0.016	0.03	0.034	0.042	0.022	0.028	0.04
Gene2	5	0.046	0.05	0.04	0.054	0.046	0.054	0.046
Gene3	5	0.04	0.05	0.042	0.056	0.046	0.054	0.052
Gene4	5	0.03	0.042	0.038	0.046	0.032	0.04	0.04
Gene5	50	0.042	0.05	0.036	0.05	0.06	0.098	0.05
$FWER_{BF}$		0.024	0.044	0.044	0.054	0.044	0.056	0.038
$FWER$		0.16	0.208	0.176	0.222	0.188	0.248	0.21
Type-I Error: Pathway		0.024	0.044	0.046	0.056	0.04	0.06	0.012

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 5. *Simulation of case 4 for sample size of 200 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200								
Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.046	0.062	0.036	0.05	0.058	0.076	0.028
Gene2	10	0.05	0.062	0.044	0.056	0.05	0.064	0.034
Gene3	10	0.038	0.042	0.034	0.036	0.046	0.058	0.024
Gene4	10	0.034	0.044	0.046	0.052	0.04	0.046	0.032
Gene5	10	0.048	0.066	0.038	0.042	0.048	0.062	0.042
Gene6	10	0.058	0.07	0.062	0.064	0.06	0.068	0.056
Gene7	10	0.044	0.042	0.026	0.034	0.036	0.038	0.036
Gene8	10	0.054	0.068	0.044	0.04	0.044	0.06	0.04
Gene9	10	0.046	0.064	0.052	0.062	0.06	0.066	0.044
Gene10	10	0.032	0.036	0.026	0.042	0.026	0.04	0.032
Gene11	10	0.042	0.056	0.038	0.044	0.046	0.058	0.034
Gene12	10	0.054	0.072	0.054	0.062	0.06	0.072	0.046
Gene13	10	0.038	0.052	0.036	0.052	0.048	0.058	0.026
Gene14	10	0.062	0.078	0.044	0.048	0.066	0.076	0.056
Gene15	10	0.032	0.044	0.05	0.06	0.04	0.048	0.038
Gene16	10	0.038	0.044	0.042	0.05	0.038	0.048	0.034
Gene17	10	0.062	0.076	0.028	0.038	0.048	0.07	0.056
Gene18	10	0.058	0.07	0.032	0.042	0.044	0.054	0.05
Gene19	10	0.038	0.042	0.032	0.048	0.042	0.046	0.028
Gene20	10	0.06	0.076	0.04	0.05	0.072	0.086	0.048
$FWER_{BF}$		0.02	0.056	0.024	0.026	0.054	0.08	0.018
$FWER$		0.612	0.704	0.564	0.642	0.646	0.732	0.556
Type-I Error: Pathway		0.028	0.07	0.028	0.04	0.084	0.152	0.004

$FWER_{BF}$ stands for FWER results using Bonferroni correction

TABLE 6. *Simulation of case 4 for sample size of 500 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 500								
Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.054	0.060	0.046	0.056	0.068	0.073	0.038
Gene2	10	0.048	0.053	0.048	0.053	0.064	0.070	0.056
Gene3	10	0.044	0.043	0.046	0.049	0.052	0.051	0.032
Gene4	10	0.056	0.062	0.042	0.043	0.052	0.060	0.052
Gene5	10	0.048	0.043	0.052	0.049	0.052	0.053	0.038
Gene6	10	0.046	0.043	0.046	0.049	0.04	0.047	0.028
Gene7	10	0.054	0.058	0.042	0.053	0.05	0.049	0.042
Gene8	10	0.038	0.030	0.044	0.041	0.052	0.041	0.02
Gene9	10	0.066	0.073	0.052	0.053	0.064	0.071	0.054
Gene10	10	0.038	0.039	0.052	0.053	0.052	0.058	0.036
Gene11	10	0.036	0.045	0.038	0.043	0.054	0.058	0.04
Gene12	10	0.054	0.062	0.054	0.064	0.064	0.066	0.056
Gene13	10	0.05	0.058	0.04	0.047	0.06	0.075	0.042
Gene14	10	0.038	0.036	0.05	0.049	0.04	0.051	0.026
Gene15	10	0.048	0.053	0.06	0.062	0.056	0.058	0.044
Gene16	10	0.06	0.053	0.058	0.062	0.06	0.066	0.048
Gene17	10	0.056	0.060	0.074	0.068	0.058	0.056	0.044
Gene18	10	0.052	0.056	0.042	0.045	0.056	0.068	0.048
Gene19	10	0.046	0.043	0.07	0.071	0.07	0.079	0.036
Gene20	10	0.032	0.032	0.03	0.041	0.04	0.047	0.028
$FWER_{BF}$		0.046	0.056	0.042	0.058	0.078	0.111	0.032
$FWER$		0.626	0.637	0.628	0.658	0.676	0.712	0.544
Type-I Error: Pathway		0.056	0.068	0.056	0.071	0.122	0.156	0.02

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 7. *Simulation of case 4 for sample size of 1000 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 1000								
Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.038	0.038	0.05	0.052	0.052	0.054	0.04
Gene2	10	0.066	0.074	0.058	0.054	0.076	0.07	0.054
Gene3	10	0.052	0.054	0.042	0.044	0.046	0.056	0.04
Gene4	10	0.044	0.054	0.038	0.038	0.042	0.044	0.03
Gene5	10	0.058	0.06	0.038	0.038	0.044	0.052	0.046
Gene6	10	0.052	0.058	0.066	0.068	0.06	0.06	0.036
Gene7	10	0.058	0.06	0.058	0.054	0.064	0.066	0.046
Gene8	10	0.048	0.052	0.044	0.044	0.046	0.046	0.034
Gene9	10	0.036	0.036	0.038	0.038	0.032	0.036	0.028
Gene10	10	0.062	0.06	0.064	0.064	0.07	0.072	0.052
Gene11	10	0.052	0.054	0.042	0.042	0.046	0.056	0.046
Gene12	10	0.044	0.048	0.032	0.036	0.038	0.03	0.034
Gene13	10	0.046	0.044	0.062	0.07	0.052	0.058	0.036
Gene14	10	0.05	0.048	0.054	0.046	0.05	0.058	0.036
Gene15	10	0.06	0.054	0.054	0.056	0.062	0.06	0.052
Gene16	10	0.034	0.036	0.052	0.048	0.054	0.05	0.028
Gene17	10	0.034	0.032	0.048	0.05	0.042	0.044	0.028
Gene18	10	0.056	0.06	0.038	0.032	0.052	0.048	0.048
Gene19	10	0.048	0.048	0.052	0.046	0.044	0.046	0.038
Gene20	10	0.044	0.048	0.06	0.066	0.056	0.06	0.036
$FWER_{BF}$		0.044	0.056	0.044	0.05	0.076	0.092	0.046
$FWER$		0.614	0.61	0.63	0.622	0.634	0.67	0.536
Type-I Error: Pathway		0.058	0.058	0.054	0.066	0.12	0.13	0.104

$FWER_{BF}$ stands for FWER results using Bonferroni correction

TABLE 8. *Simulation of case 5 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.1	0.122	0.054	0.052	0.1	0.122	0.21
Power: Pathway		0.062	0.096	0.044	0.042	0.084	0.098	0.012

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.512	0.536	0.162	0.17	0.496	0.526	0.532
Power: Pathway		0.308	0.326	0.078	0.078	0.328	0.35	0.356

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene1	20	0.876	0.884	0.432	0.42	0.86	0.872	0.872
Power: Pathway		0.722	0.744	0.184	0.196	0.722	0.726	0.836

$FWER_{BF}$ stands for FWER results using Bonferroni correction

TABLE 9. *Simulation of case 6 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.076	0.09	0.046	0.052	0.094	0.086	0.186
Power: Pathway		0.056	0.068	0.054	0.04	0.1	0.098	0.048

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.264	0.254	0.222	0.19	0.364	0.348	0.382
Power: Pathway		0.162	0.152	0.11	0.094	0.226	0.192	0.212

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.62	0.616	0.73	0.7	0.798	0.782	0.752
Power: Pathway		0.406	0.426	0.448	0.408	0.616	0.598	0.684

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 10. *Simulation of case 7 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.108	0.118	0.060	0.080	0.108	0.134	0.096
Gene2	10	0.072	0.084	0.054	0.072	0.062	0.088	0.072
Power: Pathway		0.060	0.062	0.040	0.034	0.086	0.128	0.004

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.272	0.278	0.144	0.148	0.270	0.272	0.268
Gene2	10	0.158	0.170	0.100	0.100	0.138	0.144	0.132
Power: Pathway		0.112	0.132	0.074	0.070	0.214	0.240	0.168

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.572	0.580	0.262	0.268	0.562	0.578	0.592
Gene2	10	0.318	0.324	0.180	0.186	0.290	0.288	0.3
Power: Pathway		0.298	0.334	0.080	0.090	0.416	0.444	0.464

$FWER_{BF}$ stands for FWER results using Bonferroni correction

TABLE 11. *Simulation of case 8 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.102	0.114	0.108	0.110	0.134	0.153	0.124
Gene2	10	0.092	0.094	0.104	0.102	0.128	0.125	0.120
Power: Pathway		0.044	0.060	0.050	0.048	0.142	0.157	0.016

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.224	0.230	0.160	0.160	0.256	0.250	0.302
Gene2	10	0.310	0.318	0.274	0.272	0.340	0.332	0.398
Power: Pathway		0.108	0.116	0.096	0.100	0.262	0.280	0.168

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	20	0.376	0.372	0.436	0.438	0.476	0.467	0.550
Gene2	10	0.538	0.548	0.544	0.544	0.602	0.618	0.700
Power: Pathway		0.266	0.284	0.292	0.288	0.516	0.526	0.568

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

TABLE 12. *Simulation of case 9 with 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP. 500 replications of the synthetic data are performed.*

sample size 200

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.070	0.080	0.040	0.045	0.070	0.065	0.090
Gene2	10	0.095	0.105	0.070	0.075	0.095	0.095	0.080
Power: Pathway1		0.055	0.065	0.04	0.030	0.110	0.130	0.012
Gene3	10	0.110	0.125	0.095	0.090	0.100	0.120	0.120
Gene4	10	0.085	0.100	0.085	0.100	0.110	0.115	0.118
Power: Pathway2		0.065	0.095	0.04	0.050	0.105	0.105	0.016
Power: Pathway all		0.055	0.070	0.05	0.025	0.115	0.155	0.012

sample size 500

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.150	0.145	0.130	0.135	0.150	0.170	0.194
Gene2	10	0.155	0.160	0.125	0.125	0.160	0.160	0.190
Power: Pathway1		0.085	0.10	0.075	0.075	0.145	0.150	0.132
Gene3	10	0.145	0.150	0.100	0.105	0.150	0.145	0.204
Gene4	10	0.135	0.145	0.125	0.110	0.140	0.145	0.166
Power: Pathway2		0.100	0.09	0.080	0.085	0.155	0.165	0.084
Pathway all		0.065	0.10	0.075	0.070	0.260	0.230	0.148

sample size 1000

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	iSKAT
Gene1	10	0.245	0.245	0.260	0.260	0.295	0.290	0.366
Gene2	10	0.235	0.260	0.225	0.220	0.275	0.270	0.364
Power: Pathway1		0.145	0.150	0.135	0.12	0.240	0.255	0.324
Gene3	10	0.230	0.235	0.240	0.240	0.270	0.285	0.364
Gene4	10	0.295	0.305	0.270	0.265	0.345	0.335	0.392
Power: Pathway2		0.160	0.195	0.150	0.14	0.250	0.265	0.304
Pathway all		0.160	0.220	0.170	0.18	0.425	0.465	0.484

$FWER_{BF}$ stands for FWER results using Bonferroni correction

 ISSN: 2102-6238

5. Application: Breast cancer and night work

Circadian rhythm is a roughly 24 hours cycle of biological processes that are synchronized by external cues such as light or temperature, and regulated endogenously by periodic transcription of a set of genes that form a network of self-regulated feedback loop. The circadian rhythm pathway plays a key role in the maintenance of various endocrine, physiological factors and behavioral functions including cell cycle regulation, hormone secretion, body temperature and sleep/wake cycle. Shift work that involves circadian disruption was classified as probably carcinogenic to humans (group 2A) by the International Agency for Research on Cancer in 2007 (Straif et al., 2007). An increased risk of breast cancer was reported in women working at night by several studies (Hansen and Lassen, 2012; Menegaux et al., 2013) and it was hypothesized that this association could be modulated by polymorphisms in the circadian pathway genes. As the circadian pacemaker requires multiple molecular interactions to generate the circadian rhythms, single-SNP analyses may not be sufficient to analyze the association between circadian genes and breast cancer. Therefore, we have investigated the role of circadian clock gene polymorphisms and their interaction with nightwork in breast cancer risk using a pathway analysis. This work was previously described in more details using only the ARTP method with a modified permutation procedure that permutes the outcome, the environmental factor and the adjustment variables together (Truong et al., 2014). Here, we present the results using FM, MinP and ARTP methods using permutation and Bootstrap resampling procedures as well as iSKAT method for comparison.

Briefly, the analyses are conducted in a population-based case-control study from France including 1126 breast cancer cases and 1174 controls.

We considered the circadian pathway as defined in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database that included 23 genes (*CLOCK*, *ARNTL*, *NPAS2*, *CRY1*, *CRY2*, *PER1*, *PER2*, *PER3*, *RORA*, *RORB*, *RORC*, *BHLHE40*, *BHLHE41*, *SKP1*, *FBXW11*, *CUL1*, *TIMELESS*, *FBXL3*, *NR1D1*, *CSNK1D*, *CSNK1E*, *RBX1*, and *BTRC*). These genes constitute a complex regulatory network with multiple negative and positive feedback loops. A selection of tag SNPs from these genes were selected in order to capture SNPs within 5 kb of each genes (pairwise approach with a square of correlation coefficient $r^2 > 0.8$) with a minimum minor allele frequency of 0.05 in the CEU population from HapMap project. After quality controls, we have included 577 SNPs from the 23 genes. The circadian pathway was additionnally divided into two subpathways: the core circadian genes which are involved in the same transcriptional feedback loop (*CLOCK*, *ARNTL*, *NPAS2*, *CRY1*, *CRY2*, *PER1*, *PER2*, *PER3*,*CSNK1E*) and the other genes that are involved in other auxiliary loops.

Odds ratios (OR) and corresponding 95% confidence intervals (CI) were calculated using unconditional logistic regression models adjusted for the matching factors (age, area of residence) and for established risk factors of breast cancer (age at menarche, age at first full-term pregnancy, parity, current use of menopausal hormone therapy, body mass index, alcohol consumption and tobacco consumption). An OR of 1.42 (95% CI: 1.08-1.88) (p=0.01) was observed in women that have a lifetime duration of nightwork greater than 2 years compared to less. The interaction between the polymorphisms in circadian genes and nightwork were first analysed using a SNP by SNP approach and no interaction term was statistically significant after correction for multiple testing (results not shown). Gene-level and pathway-level interaction p-values obtained by

 ISSN: 2102-6238

the FM, MinP, ARTP and iSKAT are shown in Table 13 for 1000 resampling. The parameters of the ARTP are calibrated in the same way than in the simulation part (see section 4.2).

TABLE 13. *Results of the investigation of gene-environment interaction of Circadian Pathway using 1000 permutations and 1000 bootstrap resamplings with FM, MinP and ARTP.*

Circadian Pathway		FM		MinP		ARTP		iSKAT
Gene	Size	Permutation	Bootstrap	Permutation	Bootstrap	Permutation	Bootstrap	
ARNTL	24	**0.04**	**0.001**	0.1359	0.1578	**0.0539**	**0.001**	**0.0078**
PER1	5	**0.0849**	**0.035**	**0.032**	**0.04**	**0.049**	**0.0619**	0.1132
NPAS2	62	0.2647	0.1249	0.0879	0.1309	0.2957	0.1608	0.4338
CSNK1E	9	0.3337	0.3427	0.4166	0.4965	0.4655	0.5265	0.6011
CRY1	7	0.4935	0.5235	0.4985	0.5504	0.5295	0.5694	0.6081
CRY2	9	0.5425	0.7722	0.6603	0.9111	0.6374	0.8761	0.3475
PER2	11	0.9071	0.967	0.9401	0.983	0.8721	0.9491	0.734
PER3	15	0.8901	0.984	0.8571	0.976	0.8941	0.995	0.5695
CLOCK	11	0.9181	0.99	0.982	1	0.964	0.999	0.8104
subpathway		0.2967	**0.007**	0.2537	0.2997	0.2957	**0.0020**	0.2095
FBXL3	7	0.1658	0.0769	0.1239	0.1848	0.1628	0.1918	0.2986
SKP1	4	0.4046	0.4236	0.2128	0.2318	0.2827	0.3057	0.4056
CSNK1D	3	0.3706	0.3906	0.4256	0.4655	0.3736	0.4056	0.4412
RBX1	2	0.4146	0.3876	0.5005	0.4825	0.4505	0.4296	0.4977
BHLHE40	9	0.3277	0.3237	0.5864	0.7113	0.4126	0.4396	0.7204
RORA	288	0.3836	0.3027	0.6184	0.7612	0.4515	0.4456	0.4576
NR1D1	8	0.4476	0.4985	0.3906	0.4945	0.4206	0.4825	0.5439
RORC	14	0.1958	0.1139	0.4226	0.6434	0.2687	0.5055	0.4815
CUL1	23	0.4585	0.5814	0.1009	0.1578	0.3916	0.5105	0.1035
TIMELESS	7	0.6643	0.7632	0.4406	0.5504	0.5395	0.6513	0.7471
BTRC	13	0.977	0.998	0.4236	0.6064	0.7153	0.7013	0.8507
RORB	34	0.5974	0.7163	0.5994	0.7972	0.5614	0.7293	0.7152
FBXW11	8	0.8741	0.952	0.8711	0.9361	0.9121	0.962	0.7715
BHLHE41	4	0.959	0.983	0.8891	0.957	0.9211	0.97	0.8227
subpathway		0.9231	0.6474	0.7682	0.9101	0.8042	0.8951	0.5552
circadien		0.6054	**0.009**	0.5085	0.6114	0.6374	**0.02**	0.4166

At the gene level, we observed that both methods FM and ARTP highlight the same two genes *PER1* and *ARNTL* in the interaction analysis with nightwork, while only *PER1* is significant with the MinP method and only *ARNTL* is significant with the iSKAT method. Bootstrap resampling method tends to give lower p-values than permutations for these two genes in particular. This is in accordance with the simulation section in which we shown that the parametric bootstrap method is more powerful for large sample size.

At the pathway level, a significant interaction p-value (see Table 13) was observed for the overall circadian pathway for both FM and ARTP methods when parametric bootstrap is used while no association is observed using permutation resampling approach. This association is observed only for the core circadian genes subpathway that includes the genes *PER1* and *ARNTL*. No significant association was observed while using the methods MinP and iSKAT.

To summarize, FM and ARTP gave similar results in our data. Significant interaction p-values were observed at the gene and pathway levels using the boostrap resampling method, while only

significant results at the gene level were observed using the permutation resampling method. MinP and iSKAT methods highlighted only part of the genes that were found significant by FM and ARTP methods and reported non-significant interaction at the pathway level.

PER1 and *ARNTL* which are highlighted in the gene level analysis, are important components of the circadian system which is regulated by molecular feedback loops. Heterodimers composed of ARNTL and either of the two related proteins CLOCK or NPAS2 are transcriptional factors that induce the expression of *PER* and *CRY* genes by binding to their promoters, which in turn will act on the ARNTL-CLOCK/NPAS2 complex to repress their own transcription.

Variants in both genes has been previously associated to breast cancer risk (Hansen and Lassen, 2012; Zienolddiny et al., 2013). The finding with *PER1* from the interaction analysis may be of particular interest, as a variant in *PER1* (rs2735611) was previously associated with an extreme morning preference (Carpen et al., 2006), a condition that was associated with an increased breast cancer risk among Danish military women working in night shifts (Hansen and Lassen, 2012).

5.1. Running time performance

The most demanding part of the p-value algorithms in terms of time computation is the resampling part. All p-value combining methods have been ran with the same resampling samples. We focus on the mesure of the running time related to this part of the algorithm. Tables 14 and 15 presents the running time performances of permutation and bootstrap approaches. The results given are computed on one standard core, and results are running times on the application data (see table 14) and on 500 iterations of simulation case 5 (see table 15). We can see that bootstrap and permutation have similar running times. The running time of iSKAT is added in comparison. p-value combining methods have a much higher running time than iSKAT. Hopefully it can be computed in parallel, whereas iSKAT cannot.

TABLE 14. *Running time: permutation and bootstrap performances using 1000 resampling related to the application. Results are in seconds.*

Running time		
permutation	bootstrap	iSKAT
20187.5	20264.8	486.3

TABLE 15. *Running time: permutation and bootstrap performances using 1000 resampling related to the simulation case 5. Results are the average time in seconds over 500. replications.*

Running time								
size 200			size 500			size 1000		
Permutation	Bootstrap	iSKAT	Permutation	Bootstrap	iSKAT	Permutation	Bootstrap	iSKAT
488	491	2.37	611	709	2.81	1170	1376	3.58

 ISSN: 2102-6238

6. Concluding Remark

Based on the work of Yu et al. (2009), we have proposed an efficient practical tool for investigating gene- and pathway-environment interaction. Both FM and ARTP methods are extended in this context and available through our R package `PIGE` (Liquet et al., 2017). Permutation and parametric bootstrap approaches have been implemented. Our simulation study suggests sightly better results from bootstrap compared to permutation, especially when the number of significant SNP is high. Furthermore we have shown that our proposed methods can be competitive and even slightly more powerful then the cutting edge methods like iSKAT.

The cornerstone of the implemented approaches are the running time of the resampling approaches which could be problematic in presence of large data set (i.e., large sample size and large number of genetic information). To overcome this issue, PIGE offers a parallel implementation of these approaches. As an example, our application on interaction between circadian genes and night work in breast cancer risk which includes $n = 2300$ subjects and $p = 577$ SNPs took 45 minutes with the permutation procedure and 1 hours 5 minutes using 4 cores and 1000 resampling.

In this application study, using ARTP method with the parametric bootstrap approach, we highlighted significant interactions at the pathway-level which were missed when using the permutation procedures. Our results suggest that polymorphisms in the circadian rhythms pathway could modulate the association between night work and breast cancer. This association seems to be driven mostly by the genes *PER1* and *ARNTL*.

Note that our approaches can deal in the context of $p > n$ as the methods are based on combining individual p-values. Finally, our proposed approaches are not restricted to a binary case-control outcome. In this study, we focus the presentation on an binary environment variable which was motivated by binary environment data of our application. The method is not restricted to binary environment variable and has been extended and implemented in our R package PIGE (Liquet et al., 2017) to any quantitative environment variable. Further, our package also offers the possibility to deal with survival outcome variable or quantitative outcome in general. It is also possible to investigate gene- and pathway-environment interaction for more than one pathway during the same analysis.

Acknowledgement

Most of the computations presented in this paper were performed thanks to the computing facilities the Direction du Numï¿½rique of the Universitï¿½ de Pau et des Pays de l'Adour provided us.

References

Albert, P., Ratnasinghe, D., Tangrea, J., and S., W. (2001). Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology*, 154:687–693.

Buzkova, P., Lumley, T., and Rice, K. (2011). Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *American Journal of Human Genetics*, 75(1):36–45.

Carbonetto, P. and Stephens, M. (2013). Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central for for il-2 signaling genes in type 1 diabetes and cytokine signaling genes in crohn's disease. *PLoS Genetics*, 9:e1003770.

 ISSN: 2102-6238

Carpen, J., von Schantz, M., Smits, M., Skene, D., and Archer, S. (2006). A silent polymorphism in the per1 gene associates with extreme diurnal preference in humans. *Journal of Human Genetics*, 51:1122–1125.

de Leeuw, C., Neale, B., Heskes, T., and Posthuma, D. (2016). The statistical properties of gene-set analysis. *Nature Reviews Genetics*, 17:353–364.

Dudbridge, F. and Koeleman, B. (2003). Rank truncated product of p-values, with application to genomewide association scans. *Genetic Epidemiology*, 25:360–366.

Edgington, E. (1987). *Randomization tests*. Statistics, textbooks and monographs. M. Dekker.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, London.

Evangelou, M., Dudbridge, F., and Wernisch, L. (2014a). Two novel pathway analysis methods based on a hierarchical model. *Bioinformatics*, 30(5):690–697.

Evangelou, M., Rendon, A., Ouwehand, W. H., Wernisch, L., and Dudbridge, F. (2012). Comparison of methods for competitive tests of pathway analysis. *PloS one*, 7(7):e41018.

Evangelou, M., Smyth, D., Fortune, M., Burren, O., Walker, N., Guo, H., Onengut-Gumuscu, S., Chen, W., Concannon, P., Rich, S., Todd, J., and Wallace, C. (2014b). A method for gene-based pathway analysis using genomewide association study summary statistics reveals nine new type 1 diabetes associations. *Genetic Epidemiology*, 38(8):661–670.

Fridley, B. and Biernacka, J. (2011). Gene set analysis of snp data: benefits, challenges and future directions. *European Journal of Human Genetics*, 19:837–843.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77.

Good, P. (2000). *Permutation Tests: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag, New-York.

Hansen, J. and Lassen, C. (2012). Nested case control study shift work and breast cancer risk among women in the danish military. *Occupational and Environmental Medicine*, 69:551–556.

Holmans, P., Green, E. K., Pahwa, J. S., Ferreira, M. A., Purcell, S. M., Sklar, P., Consortium, W. T. C.-C., Owen, M. J., MC'Donovan, O., and Carddock, N. (2009). Gene ontology analysis of gwa study data sets provides insights into the biology of bipolar disorder. *American Journal of Human Genetics*, 85:13–24.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics*, 92:841–853.

Jiao, S., Hsu, L., Bezleau, S., Brenner, H., Chan, A., Chang-Claude, J., Le Marchand, L., Lemire, M., Newcomb, P., Slattery, M., and Peters, U. (2013). Sberia: set-based gene-environment interaction test for rare and common variants in complex diseases. *Genetic Epidemiology*, 37(5):452–464.

Jiao, S., Peters, U., Berndt, S., Bezleau, S., Brenner, H., Campbell, P., Chan, A., Chang-Claude, J., Lemire, M., Newcomb, P., Potter, J., Slattery, M., Woods, M., and Hsu, L. (2015). Powerful set-based gene-environment interaction testing framework for complex diseases. *Genetic Epidemiology*, 39:609–618.

Lin, X., Lee, S., Christiani, D., and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, 14(4):667–681.

Lin, X., Lee, S., Wu, M. C., Wang, C., Chen, H., Li, Z., and Lin, X. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics*, 72(1):156–164.

Liquet, B. and Riou, J. (2013). Correction of the significance level when attempting multiple transformations of an explanatory variable in generalized linear models. *BMC Medical Research Methodology*, 13(1):75.

Liquet, B., Truong, T., and Broc, C. (2017). *PIGE: Self Contained Gene Set Analysis for Gene- And Pathway-Environment Interaction Analysis*. R package version 1.1.

Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., DJ, H., McCarthy, M., Ramos, E., Cardon, L., Chakravarti, A., Cho, J., Guttmacher, A., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C., Slatkin, M., Valle, D., Whittemore, A., Boehnke, M., Clark, A., Eichler, E., Gibson, G., Haines, J., Mackay, T., McCarroll, S., and Visscher, P. (2009). Functional and genomic context in pathway analysis of gwas data. *Nature*, 461(7265):747–53.

Menegaux, F., Truong, T., Anger, A., Cordina-Duverger, E., Lamkarkach, F., Arveux, P., Kerbrat, P., Fevotte, J., and Guenel, P. (2013). Night work and breast cancer: a population-based case control study in france (the cecile study). *International Journal of Cancer*, 132:924–931.

Mooney, M., Nigg, J., McWeeney, S., and Wilmot, B. (2014). Functional and genomic context in pathway analysis of gwas data. *Trends in Genetics*, 30(9):390–400.

O'Dushlaine, C., Heron, E. A., Segurado, R., Gill, M., Morris, D. W., and Corvin, A. (2009). The snp ratio test:

 ISSN: 2102-6238

pathway analysis of genome-wide association datasets. *Bioinformatics, Application Note*, 25:2762–2763.

Pers, T. (2016). Gene set analysis for interpreting genetic studies. *Human molecular genetics*, 25(R2):R133–R140.

Shahbaba, B., Shachaf, C. M., and Yu, Z. (2012). A pathway analysis method for genome-wide association studies. *Statistics in Medicine*, 31:988–1000.

Straif, K., Baan, R., Grosse, Y., Secretan, B., El Ghissassi, F., Bouvard, V., Altieri, A., Benbrahim-Tallaa, L., and Cogliano, V. (2007). Carcinogenicity of shift-work, painting, and fire-fighting. *Lancet Oncology*, 8:1065–1066.

Su, Y., Gauderman, W. J., Berhane, K., and Lewinger, J. P. (2016). Adaptive set-based methods for association testing. *Genetic Epidemiology*, 40:113–122.

Truong, T., Liquet, B., Menegaux, M., Plancoulaine, S., Laurent-Puig, P., Mulot, C., Cordina-Duverger, E., Sanchez, M., Arveux, P., Kerbrat, P., Richardson, S., and Guenel, P. (2014). Breast cancer risk, nightwork and circadian clock gene polymorphisms. *Endocrine-related cancer*, 21(4):629–38.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *American Journal of Human Genetics*, 81:1278–1283.

Wang, L., Jia, P., Wolfinger, R., Chen, X., and Zhao, Z. (2011). Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics*, 98(1):1–8.

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89:82–93.

Yu, K., Li, Q., andrew W., B., Pfeiffer, R. M., Rosenberg, P. S., Caporaso, N., Kraft, P., and Chatterjee, N. (2009). Pathway analysis by adaptive combination of p-values. *Genetic Epidemiology*, 33(8):700–709.

Zienolddiny, S., Haugen, A., Lie, J., Kjuus, H., K.H., A., and KjÃ¦rheim, K. (2013). Analysis of polymorphisms in the circadian-related genes and breast cancer risk in norwegian nurses working night shifts. *Breast Cancer Research*, 15:R53.

 ISSN: 2102-6238