Is it possible to detect GxE interactions in GWAS when causal exposure is unobserved?

Flora Alarcon, Vittorio Perduca, Gregory Nuel

ORIGINAL ARTICLES

Is it possible to detect G×E interactions in GWAS when causal exposure is unobserved?

Flora Alarcon*[1], Vittorio Perduca[1], Gregory Nuel[2]

[1] Department of Applied Mathematics, Université Paris Descartes, Sorbonne Paris Cité, Paris, France
[2] Institute of Mathematics, National Center for French Research, Laboratory of Probability, Université Pierre et Marie Curie, Sorbonne Université, France

Received: August 30, 2015 **Accepted:** November 12, 2015 **Online Published:** December 2, 2015
DOI: 10.5430/jer.v2n1p109 **URL:** http://dx.doi.org/10.5430/jer.v2n1p109

ABSTRACT

Background: It is generally acknowledged that most complex diseases are affected in part by interactions between genes and genes and/or between genes and environmental factors. Taking into account environmental exposures and their interactions with genetic factors in genome-wide association studies (GWAS) can help to identify high-risk subgroups in the population and provide a better understanding of the disease. For this reason, many methods have been developed to detect gene-environment (G×E) interactions. Despite this, few loci that interact with environmental exposures have been identified so far. Indeed, the modest effect of G×E interactions as well as confounding factors entail low statistical power to detect such interactions. Another potential obstacle to detect G×E interaction is the fact that true exposure is seldom observed: Indeed, only proxy effects are measured in general. Furthermore, power studies used to evaluate a new method often are done through simulations that give an advantage to the new approach over the other methods.

Methods: In this work, we compare the relative performance of popular methods such as PLINK, random forests and linear mixed models to detect G×E interactions in the particular scenario where the causal exposure (E) is unknown and only proxy covariates are observed. For this purpose, we provide an adapted simulated dataset and apply a recently introduced method for H1 simulations called waffect.

Results: When the causal environmental exposure is unobserved but only a proxy of this exposure is observed, all the methods considered fail to detect G×E interaction.

Conclusions: The hidden causal exposure is an obstacle to detect G×E interaction in GWAS and the approaches considered in our power study all have insufficient power to detect the strong simulated interaction.

Key Words: Gene environment interaction, Latent exposure, Power, Simulation, Random forests, Linear mixed models

1. INTRODUCTION

Genome-wide association studies (GWAS) are a standard method to identify common genetic factors that influence health and disease conditions. These methods have improved our understanding of the genetic basis of many complex traits and are among the most used tools for analyzing complex diseases. However, it is known that most complex diseases (*e.g.* diabetes, asthma and cancer) are due to combined effect of genes, environmental factors, as well as their interactions.[1]

Over the last years, considerable efforts have been put to detect gene-environment interactions (G×E) in GWAS and few

*Correspondence: Flora Alarcon; Email: flora.alarcon@parisdescartes.fr; Address: MAP5-UMR CNRS 8145, Sorbonne Paris Cité University, Paris, France.

loci that interact with environmental exposures have been identified.[2–4]

However, this problem is well known to be challenging due to several reasons including the modest effect of such interactions in terms of relative risk or the population structure which can partially explain spurious associations.[5] Another situation when detecting G×E interactions could be difficult is when the causal exposure is unobserved but only proxy covariates are observed. Indeed, an interesting question is to assess whether it is still possible to detect G×E interactions when the environmental factor interacting with the susceptibility gene is unobserved but correlates with one or several observed factors.

Nowadays, several methods are available to perform GWAS. In principle, they could be used to detect G×E interactions. Among them, PLINK[6] can be considered as a gold standard for classical analysis. A major concern in GWAS is the need to account for the complicate dependence structure in the data, between loci as well as between individuals. Effects of population stratification can be easily accounted in PLINK by adding the PCA's first components as covariates.

As an alternative, linear mixed models stem as promising statistical methods to correct for the stratification in the population. A popular implementation of linear mixed models is Fast-LMM.[7]

Furthermore, powerful data mining techniques are being increasingly used. Among them, the application of random forests (RFs) to the discovery of SNPs related to human diseases has grown in recent years.[8]

The importance of empirical power studies based on realistic datasets is fully acknowledged (`http://www.gaworkshop.org/`). In particular, each time a new method is introduced, it is obviously essential to evaluate its performance in comparison with existing techniques through power studies.[9] However, such studies are often conducted by performing H1 simulations under models which are very similar to the ones used to design the new method, thus giving it an obvious advantage over the other methods. It is hence quite common to see many concurrent approaches each claiming to outperform all others.

Recently, a new method for H1 simulations called waffect (pronounced "double-u affect" for weighted affectations) has been introduced to avoid this issue.[10] Indeed, this method does not make any other assumption than the causal disease model itself, whose choice is completely unconstrained. waffect uses weighted permutations to generate phenotypes conditionally to the genotypes and covariates by taking into account both the causal disease model and the design of the study. With this new approach, it is hence possible to produce non-subjective H1 datasets which do not favor one analysis method over the others.

In this paper, we propose to study the effect of a hidden causal exposure on the power to detect G×E interactions in GWAS. For this purpose, we simulated a dataset mimicking a situation in which the causal exposure is unobserved but some covariates correlating with this hidden exposure are observed. This dataset is based on the publicly available HapMap project datasets[11] for real genotypes with population structures; we used waffect to generate phenotypes for a chosen causal disease model.

We compared four approaches based on three popular methods (PLINK, Fast-LMM and random forests) by performing power analysis on our simulated dataset.

2. SIMULATED DATASET

2.1 Genotypes

The genotypic dataset used in our study was extracted from the HapMap phase III database of genetic variations.[12] This database investigates 11 human populations including 57 unrelated MEX, 146 unrelated YRI, 52 unrelated ASW, 110 unrelated CEU, 154 unrelated MKK, 137 unrelated CHB, 109 unrelated CHD, 101 unrelated GIH, 113 unrelated JPT, 110 unrelated LWK and 102 unrelated TSI (Description of the population codes can be found at `www.broadinstitute.org`). Only the SNPs that are shared by all populations were retained.

Principal component analysis (PCA) was performed on the whole genome, keeping one SNP over 1,000 SNPs; the first five principal components ($pca_i, i = 1, \ldots, 5$) were considered as covariates. The association analysis in our study were conducted on Chromosome 6, after quality control including Hardy-Weinberg equilibrium testing and exclusion of SNPs with a minor allele frequency (MAF) less than 5%.

2.2 Covariates

Covariates and phenotypes were simulated in order to mimic a complex interaction between an arbitrarily chosen causal SNP and an hidden exposure (called *treatment*). The unobserved exposure was defined with high correlation with two observed covariates (*bmi*, for body mass index, and *sex*) as well as with the population of belonging. The idea is that the *treatment* is typically taken by women (and less often by men) trying to loose weight.

bmi was simulated taking into account the five first principal components and another environmental covariate denoted *smoking*. This binary variable was simulated to mimic smoking behaviors with a probability distribution which depends

on the population and *sex*. Indeed, women usually smoke less than men with this difference depending on the population. In order to simulate the smoking covariate, the eleven populations were classified in three sub-populations: European (E); African (Af) and Asian (As).

In the European sub-population, we considered that 32% of individuals were smokers. More specifically, we supposed that 37% of men and 27% of women were smokers. In the African sub-population, the prevalence of smoking was supposed to be 27%: 43.8% among men and 12.9% among women.[13] At last, in the Asian sub-population, we considered that 27% of individuals were smokers: 45.7% among men and 4.8 % among women.[14] Covariate *sex* was obtained from HapMap data.

Specifically, *bmi* was simulated with a regression on the first five principal components in order to have 60% of heritability and with a residual standard deviation of 4.0. To take into account the fact that smokers in average have a lower *bmi* than non-smokers, we simulated a smoker effect in the *bmi* covariate by adding a score of 1.5 to the *bmi* average for non-smokers.[15]

The individual probability to take a *treatment*, P (*treatment*), was correlated with covariates *bmi*, *sex* and the population of belonging as follows:

1/P(*treatment*)=(1+2×$\mathbf{1}_{sex=1}$)×[1+exp(-*bmi*+25+γ)]

where $\gamma \in$ {-inf,-0.1,0,0.15,-0.45,0.35,0.6,-0.4,0.05,0.1} for population 1 to 11.

Table 1. Description of all simulated covariates and real genotypes for 1,191 individuals with 595 cases and 596 controls

Observed covariates	**Values**	**Short description**
genotype	∈ {0, 1}	real genotypes taken from HapMap, length of the genotypic vector is 38634
sex	∈ {1, 2}, factor	sex of each individual
pc	continuous variable	principal components of the PC analysis calculated from genotype
smoking	∈ {0, 1}	depends on population of belonging (pop) and sex
bmi	continuous variable	depends on the five first PC and smoking
Unobserved covariates	**Values**	**short description**
pop	∈ {1, 2,...,11}, factor	population of belonging of individuals
treatment	∈ {0, 1}	depends on sex, bmi and pop
disease	∈ {0, 1}	under *H1* hypothesis, disease depends on causalSNP and treatment

To sum up, in our standard design covariates *sex*, $pca_i (i = 1, \ldots, 5)$, *bmi* and smoking were supposed observed. The population of belonging, covariate *treatment* and the other principal components were supposed unknown (even though these are easily computable), see Table 1.

2.3 Disease model

We arbitrarily chose the SNP in position 22,683,075 in a dense area of chromosome 6 as the binary susceptibility locus, denoted *causalSNP*. Assuming a dominant effect, we encoded *causalSNP* = 1 in presence of at least one minor frequency allele and *causalSNP* = 0 otherwise. We considered a disease model with a very strong G×E interaction (relative risk of 50) with a baseline prevalence of 1%:

P(*disease*)=0.01×(1.0+50.0×$1_{causalSNP=1}$×$1_{treatment}$=1)

It is important to stress that this is not a very realistic model of complex disease due to the lack of genetic marginal effects and also due to the very strong interaction with a relative risk (RR) equal to 50. We chose not to include any marginal effects for sake of clarity and because we were interested on the detection of G×E interactions. Concerning the strong G×E interaction, we chose to have a relative risk as high as 50 in order to increase the chance of detecting the interaction between the *causalSNP* and our hidden causal exposure *treatment*. Indeed, such a strong relative risk is obviously seldom encoutered in genetic epidemiology except for major risk factor like smoking in lung cancer. Our model can therefore be seen as a best case scenario for G×E which should be easily detected considering the right exposure. But what if this causal exposure is unobserved?

Phenotypes were simulated accordingly to the disease models by means of the package waffect[10] publicly available on the CRAN server of R packages.[16] This enabled us to simulate exactly 595 cases and 596 controls for the 1,191 individuals from the HapMap genotypic dataset (see next section for a comprehensive introduction to waffect).

2.4 Dataset availability

Our simulated dataset can be downloaded from `https://www.researchgate.net/FloraAlarcon/`. It comprises (1) the genotypic matrix; (2) a table with the covariates *sex*, $\texttt{pca}_i, (i = 1, \ldots, 5)$, *bmi*, *smoking* (all known in our standard design) as well as *treatment* (unknown in our standard design); (3) a table with 200 replicates of phenotypes under H0 and (4) a table with 200 replicates of phenotypes under H1.

3. POWER ANALYSIS

3.1 Phenotype simulations

In order to assess the empirical statistical power of different tools to detect associations, we simulated 200 phenotypes replicates under the disease model H1 and 200 phenotypes replicates under the null hypothesis H0 of no association. Each replicate consists of 1,191 phenotypes, one for each individual.

The simulations under H0 were obtained by simply permuting the phenotypes, thus breaking potential associations between phenotypes and genotypes. The simulations under the alternative hypothesis H1 were performed using the R package waffect publicly available on CRAN.[10]

The principal function in waffect is based on a backward sampling algorithm which makes it possible to generate weighted permutations. For the purposes of phenotype simulation, the vector of weights is given by the penetrance, that is the probability for each individual to be a case according to the disease model. One crucial consequence of using weighted permutations is that the number of cases and controls is constant across the replicates. This makes it possible to respect the original design in each replicate and therefore to compare the performance of an association method across different replicates.

Simulating phenotypes rather than genotypes, as does the gold standard Hapgen,[17] does not require additional data such as haplotype frequencies or recombination rates and has the obvious advantages of requiring much smaller computational memory and time. Moreover, for the purposes of the present work, the primary benefit of using waffect is that it only requires a vector of probabilities as input. As a result, the choice of the disease model is totally unconstrained; in particular it is possible to include G×E interactions.

In principle, one could achieve the same result by simply using a rejection algorithm which samples the phenotype of each individual according to the probability to be a case and then accepts the resulting replicate only if there are *enough* cases. Because the probability of obtaining a full configuration of phenotypes with the correct number of cases is extremely low, this approach cannot be used in practice. In order to overcome this problem, a solution often applied in practice is to increase the prevalence in the disease model, maintaining unchanged the relative risks. However it can be proved that adjusting the prevalence creates bias in the empirical power estimate.[10]

3.2 Statistical analysis

In this section, we briefly describe the four popular approaches adopted in our study to perform GWA analysis. The gold standard PLINK (individual SNP logistic regression)[6] was applied in two alternative approaches: 1- analysis performed regardless of G×E interactions and considering only genetic effects; and, since it is easy to consider interaction terms with PLINK, 2- analysis performed taking into account G×E interactions. The other two approaches are 3- the linear mixed model for population structure correction implemented in Fast-LMM,[7] and 4- the random forest data mining technique, RandomForest R package.[18]

3.2.1 *PLINK*

PLINK implements a logistic regression approach[6] allowing for multiple binary or continuous covariates when testing for disease trait SNP association and interactions with covariates. PLINK provides *p*-values for significance coefficients in the logistic model. In this work, we considered two approaches using PLINK.

The first approach, which we referred to as "PLINK SNP", consisted in performing analysis regardless of G×E interactions by looking at the *p*-value associated to the significant coefficients for the SNPs.

We referred to the second approach as "PLINK SNP × COV", where COV is the environmental covariate under consideration (either *bmi* in our standard design or *treatment* in our further analysis, see below). PLINK SNP × COV accounted for all the interactions between the analyzed SNPs and the environmental factor considered through the *p*-values associated to the significance coefficients of such interactions.

Correction for population structure was taken into account by considering the five first principal components resulting from the PCA performed on the whole genome.

3.2.2 *Random forests*

Random forests (RFs) have been introduced by Purcell.[19] The general principle consists in building repeatedly classification and regression trees (CART) from bootstrapping of the original data. This process produces a forest of classification trees which are statistically analyzed to produce importance measures of the covariates (*e.g.* a variable belonging to many trees probably plays a key role in the classification).

Random forests are a popular way to perform data mining on GWAS data. Despite the fact that they exploit heavily marginal linear regressions, random forests are able to detect interactions between variables (see[20] for an overview of random forests in the GWAS context). Recently, a regularized version of random forests has been proposed to deal with high dimensional data.[21] In this work we decided to disregard this approach because it was too slow on our data to be practical.

For our random forests analysis, we used the *randomForest* package (version 4.6-7) from R.[16] We simply used the default parameters of the method with the disease status as a binary outcome and with all observed covariates and SNPs as explanatory variables. Once the forest computed, we extracted for each variable its importance measure using the default approach of the package (normalized difference between out-of-bag proportion error using original data or a permuted version). We hence obtained for each variable and each replicate a real value which reflects the importance of the variable for discriminating between cases and controls. The higher this importance value, the stronger the association with the disease.

3.2.3 *Fast-LMM*

It is a well known problem that in GWAS confounding effects of population structure lead to false positive and therefore need to be taken into account. An alternative to including the first principal components in linear or logistic regression models in order to correct for confounding factors are Linear Mixed Models LMMs.[22] LMMs generalize linear models by introducing random effects as predictors, in addition to the usual fixed effects. Indeed, LMMs are known to be effective when observations are not independent but rather involve related individuals.

In LMMs for GWAS, the random effect is expressed by a multivariate normal distribution whose variance-covariance matrix measures the genetic similarity between individuals. Recently the algorithm Fast-LMM has been introduced to perform efficiently exact inference for LMMs.[7] Roughly speaking, Fast-LMM (for Factored spectrally transformed Linear Mixed Models) is based on a spectral decomposition of the genetic similarity matrix which rotates the phenotypes into uncorrelated phenotypes thus converting the original estimation problem into the maximization of the likelihood of a linear regression model.

A drawback of the current implementation of Fast-LMM is that it does not allow to consider explicit interaction terms between genotypic variables and covariates in the linear mixed model (We unsuccessfully tried to contact the authors of Fast-LMM on this matter.). A possible solution to overcome this limitation is to code directly such interactions in the covariate file, thus adding new columns. However, this solution was not appropriate in this context because it would have required to magnify several times the size of the variables file in order to consider the cartesian product of all the SNPs with all the covariates. We then remove simply applied Fast-LMM to the original genotypic and covariate datasets.

3.3 Power, ROC curves and AUC

Instead of computing the power of our detection methods for a given controlled type-I error rate, we used the Area under the Curve (AUC) corresponding to the receiver operating characteristic (ROC) curves (AUROC curves) which has the benefit to avoid choosing a type-I error rate and having to control the procedure (ex: adjusting for multiple testing). Our AUROC results directly reflect the overall discriminative power of the chosen statistic. Therefore, the overall performance of each of the four methods described above was assessed by simply looking at a summary statistics.[10] These global statistics were then used to estimate the AUROC curves of the four methods, each expressing the performance of the corresponding method.

More specifically, for PLINK SNP we took as simple global statistics the smallest among all the p-values associated to the significance coefficients of the SNPs, similarly for Fast-LMM. For PLINK SNP $\times$ COV, we took the smallest among all the p-values associated to significance coefficients of the terms coding for the interactions between the SNPs and the covariate (*bmi* or *treatment*). At last, the summary statistics for the random forest-based method was defined as the maximum of the importance statistics over all considered SNPs.

Then, for each method, we obtained two vectors of length 200, one under H0 and one under H1. These two vectors of comprehensive signals were used to estimate the ROC curves of the four methods using the R package pROC.[23]

We recall that ROC curves provide a graphical representation of the specificities and sensitivities (*i.e.* values of statistical power) that can be obtained for all possible values of the threshold of significance.[24] An informative summary of the ROC curve information is the area under the ROC curve (*i.e.* AUC). The AUC can be qualitatively interpreted as follows: AUC ≤ 0.6 means "fail"; $0.6 <$ AUC ≤ 0.70 means "poor"; $0.7 <$ AUC ≤ 0.80 means "fair"; $0.8 <$ AUC ≤ 0.9 means "good"; $0.9 <$ AUC ≤ 1.0 means "excellent".

4. RESULTS AND DISCUSSION

Association analysis were adjusted on covariates *sex*, *smoking* together with either *bmi* (our standard design) or *treatment*. Moreover, for PLINK and random forests the five first principal components $pc_i, i \in \{1, \ldots, 5\}$ were included as predictors. We recall that in order to evaluate empirically the detection power of the four approaches in presence of interaction with an hidden exposure, analysis were performed on 200 + 200 phenotypic replicates under H0 and H1. Investigations were performed on chromosome 6 and subregions.

4.1 The causal exposure is observed

At first, the causal exposure *treatment* was supposed to be observed. Given the very strong interaction simulated, we expected to have a good power to detect the G×E interaction. Table 2 shows the estimated AUC together with 95% confidence intervals obtained in this context. While AUC estimated with PLINK SNP×*treatment* are equal to 1 (as expected given the strength of the simulated G×E interaction), the AUC estimated with the other methods are poor. These results confirm that the approach accounting for the interaction is better than the approach accounting only for the SNP and demonstrate the importance of accounting for G×E interaction in a process of consideration of an hidden exposure.

Table 2. Association analysis performed on chromosome 6 with the four approaches, when the causal exposurer *treatment* is considered observed. Restricted regions are centered on causal SNP

AUC (%)	PLINK SNP	PLINK SNP × bmi	RF	Fast-LMM
whole chromosome 6	63.35 (57.84-68.85)	99.97 (99.93-100.0)	67.66 (62.44-72.88)	61.54 (56.02-67.06)
8,000 SNPs region	70.67 (65.58-75.75)	100.0 (99.99-100.0)	78.30 (73.88-82.73)	68.02 (62.82-73.23)
2,000 SNPs region	73.91 (68.95- 78.88)	100.0 (99.99-100.0)	82.57 (78.60-86.54)	72.03 (67.02-77.04)
800 SNPs regions	80.24 (75.78- 84.69)	100.0 (100.0-100.0)	86.07 (82.54-89.61)	80.22 (75.95-84.48)
200 SNPs region	87.62 (84.26- 90.98)	100.0 (100.0-100.0)	89.88 (86.87-92.90)	86.65 (83.16-90.14)
causal SNP	99.03 (98.39-99.67)	100.0 (100.0-100.0)	92.01 (89.13-94.89)	99.08 (98.47-99.69)

4.2 The causal exposure is not observed

In our original design, the causal exposure was supposed to be unobserved: the covariate *bmi* was observed instead of *treatment* and considered, mistakenly, as the environmental exposure of interest.

Table 3 shows the estimated AUC together with 95% confidence intervals for the four approaches. Obviously, the performance of each method increases when the region under consideration decreases and reach a good power when the region is restricted to the causal SNP. However power is low (fail or poor) when estimation is done on whole chromosome 6.

Table 3. Association analysis performed on chromosome 6 with the four approaches with an hidden causal exposure (*i.e.* covariate bmi is observed but covariate treatment is unknown). Restricted regions are centered on causal SNP

AUC (%)	PLINK SNP	PLINK SNP × bmi	RF	Fast-LMM
whole chromosome 6	64.69 (59.26-70.13)	56.39 (50.69-62.08)	66.23 (62.44-72.88)	61.91 (56.42-67.41)
8 000 SNPs region	72.04 (67.03-77.05)	55.32 (49.6-61.04)	71.99 (67.07-76.91)	68.96 (63.84-74.07)
2 000 SNPs region	74.44 (69.52-79.35)	58.05 (52.41-63.70)	76.15 (71.49-80.82)	71.36 (66.35-76.36)
800 SNPs regions	81.78 (77.54-86.02)	60.24 (54.66-65.81)	79.48 (75.10-83.87)	80.65 (76.48-84.82)
200 SNPs region	85.50 (85.27-91.72)	68.62 (63.38-73.85)	84.71 (80.77-88.66)	86.72 (83.28-90.17)
causal SNP	99.15 (98.57-99.73)	88.67 (85.48-91.86)	89.03 (85.75-92.32)	99.09 (98.49-99.70)

Surprisingly, the PLINK SNP×*bmi* approach exhibits a drastic loss of performance when the true causal exposure is hidden: the drop in the AUC is as high as 43.58% with respect to the case when *treatment* is observed. PLINK SNP, Fast-LMM and RF provide comparable results even if PLINK SNP seems to be a little more efficient.

Surprisingly, the estimated performance is better with PLINK SNP than with PLINK SNP× *bmi*. For example, applying PLINK SNP on a region with less than 800 SNPs around the causal SNP provides good to excellent power while analysis with PLINK SNP×*bmi* needs to be restricted to the causal SNP to reach similar power. For the RF, we can only observe a slight improvement (1% to 3% of AUC) when the *treatment* is observed. This is due to the fact that, like for Fast-LMM or PLINK SNP, the RF approach does not consider explicitly interactions with the covariates. However, RF are known to be able to capture complex non-linear interaction between covariates which tend to be co-selected in the same trees. In our example, this alleged feature clearly shows its limits. Moreover, results obtained observing the causal exposure have shown that PLINK SNP (as LMM and RF) were methods not suited to this context. Table 4 presents a summary of advantages and disadvantages of each method.

Table 4. Advantages and disadvantages of studied methods when the SNP to detect interacts with an unobserved environmental exposure

Methods	Strengths	Weakness
PLINK	Easy to use and possibility to consider explicitly interactions with covariates in the model	Provides a bad power to detect the interaction with an unobserved environmental exposure
RF	Able to capture complex non-linear interactions between covariates	Does not consider explicitly interactions with the covariates
Fast-LMM	Able to account for population stratification from random effects	Does not consider explicitly interactions with the covariates

Furthermore, we note that AUC estimate applying PLINK SNP×*bmi* when the analysis is performed on the whole chromosome 6 appears higher than on a region of 4,000 SNPs around the causal SNP. This result could be explained by sample variability.

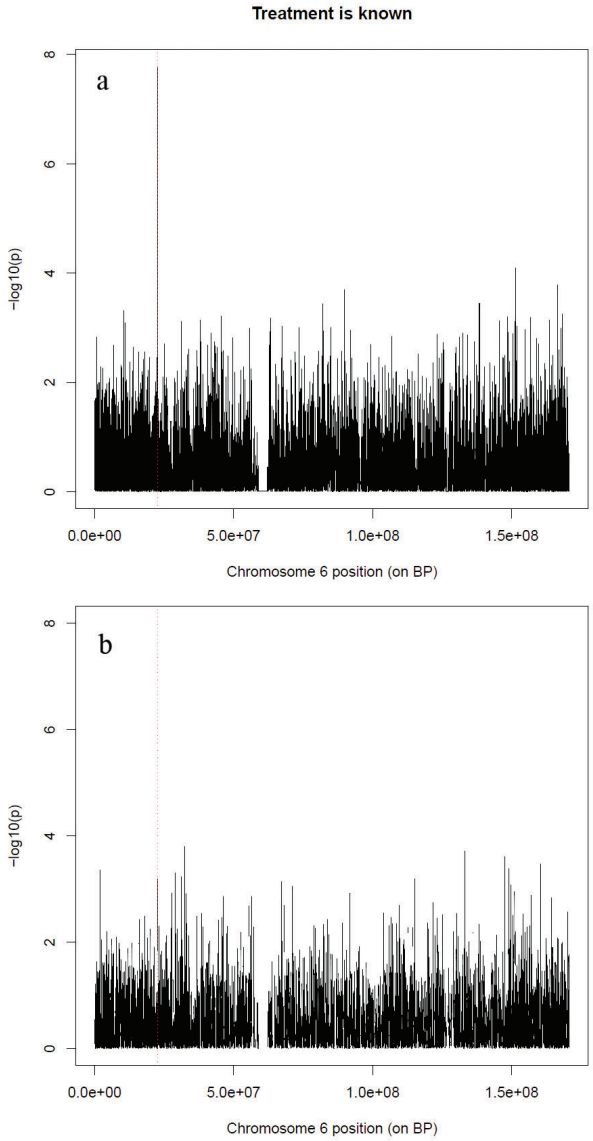

Figure 1. Manhattan plots considering chromosome 6 and either the covariate treatment or the covariate *bmi*. The red vertical line indicates the causal SNP

4.3 Further considerations

An interesting property of Fast-LMM is its ability to account for population structure by introducing random effects as predictors. In this context, analysis were performed with Fast-LMM, with no principal components as covariates. We verified (results not shown) that accounting explicitly for principal components gives similar results. In the same way, we performed analysis with PLINK SNP×*bmi* considering explicitly the population of belonging instead of principal components and found similar results than considering principal components (results not shown).

Another question that may arise is about the signal localization. Indeed, presence of hidden causal exposure makes very difficult to identify the causal SNP. Figure 1 shows the Manhattan plots obtained performing PLINK SNP × COV analysis on Chromosome 6 from one simulation under H1. The vertical red line indicates the causal SNP location. On figure 1a analysis was performed observing the causal exposure *treatment* (*i.e.* using PLINK SNP×*treatment* approach) and on figure 1b analysis was performed with the hidden exposure (*i.e.* using PLINK SNP×*bmi* approach). When the causal exposure is observed, the location detected by the signal is the same as the causal SNP location. In contrast, when the causal exposure is hidden, no signal is clearly detected.

5. Conclusion

The aim of this article is to study the power to detect G×E interactions in the particular case where the causal exposure is hidden (*i.e.* non observed) and instead, proxy covariates are observed. In order to mimic this typical design we simulated a dataset: real genotypes with population structure were obtained from the HapMap project dataset and phenotypes were simulated using waffect according to a disease model.

The disease model was chosen without marginal effect for sake of clarity and with a very strong G×E interaction. Despite its simulated strength, we showed that usual methods not accounting for interactions are not able to detect any association at all.

Moreover, we show that when analyses are done without searching for interactions, observing or not the causal exposure has no impact on the detection power. These results

highlight the importance of taking into account G×E interactions at the risk of finding no signal at all.

Furthermore, when a method accounting for interactions is applied to detect G×E interactions, the fact that the causal exposure is unobserved causes a drastic loss of detection power. In our simulation study this was true even though we simulated a very strong G×E effect!

By using HapMap genotypes, our dataset has indeed a very strong population structure that has to be accounted for (ex: by using principal components as covariates). Performing H0 simulations without population stratification (ex: populate specific prevalence) might clearly generate spurious association. Howaver, in real life GWAS, H0 simulations seldom account for possible population structure, and it is moreover clear that this possible source of bias fail to favor the detection of the targeted G×E effect in our design. Further investigation will definitely include population stratified H0 and H1 simulations.

In this work, we chose to focus on three popular methods belonging to different families of statistical techniques. Unfortunately, the current implementation of Fast-LMM does not allow to account for G×E interactions. As an alternative, we considered the idea to precompute a full covariate matrix including interactions as an input to Fast-LMM, but this approach was finally discarded as Fast-LMM would not perform the appropriate significance tests in any case. The proposed dataset has the potential to provide a good framework to develop further features of Fast-LMM enabling it to account for such interactions. Similar consideration hold for the RF which are not specifically designed to deal with explicit interactions.

At the best of our knowledge, this is the first empirical study of the performance of methods for detecting gene-environment interactions when the exposure is not observed. Indeed, previous works compare the performance of methods when the exposure is observed. Typically these comparison studies are done when a new method to detect a G×E interaction is introduced for case control data. Kraft *et al*.[25] proposed a powerful 2-df joint test of marginal association and G×E interaction. Shortly after, Mucray *et al*.[1] proposed a 2-step approach for detecting G×E interaction in GWA studies. Dai *et al*.[26] proposed a new way to combine the test of marginal genetic effect and the test of G×E interaction, by exploiting the independence between the two tests. While these methods have demonstrated their efficiency, their performance was assessed through simulations that do not account for realistic complexity such as the inclusion of confounding factors, hidden causal exposures and/or of complicated dependance structures between individuals as well as between loci. Testing the detection power of these tests on our simulated dataset could then be an interesting development. Results in this paper already seem to show that methods based on the linear model have poor power to detect G×E interaction when the causal exposure is not observed, therefore we expect that the tests mentioned above will not perform well either.

In conclusion, efforts should be put in developing standard methods in order to detect G×E interactions as well. Moreover, it would be of interest to develop a logistic regression with latent exposure in order to gain power in detecting G×E interactions when the causal exposure is unobserved or partially observed.

ACKNOWLEDGEMENTS

We acknoledge funding from ANR SAMOGWAS.

CONFLICTS OF INTEREST DISCLOSURE

The authors declare that they have no competing interests.

REFERENCES

[1] Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. American journal of epidemiology. 2009; 169(2): 219-226. PMid:19022827 http://dx.doi.org/10.1093/aje/kwn353

[2] Rothman N, Garcia-Closas M, Chatterjee N, *et al*. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nature genetics. 2010; 42(11): 978-984. PMid:20972438 http://dx.doi.org/10.1038/ng.687

[3] Hamza TH, Chen H, Hill-Burns EM, *et al*. Genome-wide gene-environment study identifies glutamate receptor gene grin2a as a parkinson's disease modifier gene via interaction with coffee. PLoS genetics. 2011; 7(8): e1002237. PMid:21876681 http://dx.doi.org/10.1371/journal.pgen.1002237

[4] Garcia-Closas M, Jacobs K, Kraft P, *et al*. Analysis of epidemiologic studies of genetic effects and gene-environment interactions. IARC scientific publications. 2010; 163(163): 281-301.

[5] Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. Statistical Science. 2009; 24(4): 451-471. http://dx.doi.org/10.1214/09-STS307

[6] Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007; 81(3): 559-575. PMid:17701901 http://dx.doi.org/10.1086/519795

[7] Lippert C, Listgarten J, Liu Y, *et al*. Fast linear mixed models for genome-wide association studies. Nature Methods. 2011; 8(10): 833-835. PMid:21892150 http://dx.doi.org/10.1038/nmeth.1681

[8] Goldstein BA, Hubbard AE, Cutler A, *et al*. An application of Random Forests to a genome-wide association dataset: Methodolog-

ical considerations & new findings. BMC genetics. 2010; 11(1): 49. PMid:20546594 http://dx.doi.org/10.1186/1471-2156-11-49

[9] Spencer CCA, Su Z, Donnelly P, *et al*. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS genetics. 2009; 5(5): e1000477. PMid:19492015 http://dx.doi.org/10.1371/journal.pgen.1000477

[10] Perduca V, Sinoquet C, Mourad R, *et al*. Alternative Methods for H1 Simulations in Genome-Wide Association Studies. Human Heredity. 2012; 73(2): 95-104. PMid:22472690 http://dx.doi.org/10.1159/000336194

[11] Thorisson GA, Smith AV, Krishnan L, *et al*. The international hapmap project web site. Genome research. 2005; 15(11): 1592-1593. PMid:16251469 http://dx.doi.org/10.1101/gr.4413105

[12] Gibbs RA, Belmont JW, Hardenbol P, *et al*. The international hapmap project. Nature. 2003; 426(6968): 789-796. PMid:14685227 http://dx.doi.org/10.1038/nature02168

[13] Christopoulou R, Lillard DR. The role of culture in smoking behavior: evidence from british immigrants in australia, south africa, and the us. Technical report, Cornell University. 2011.

[14] Tsai YW, Tsai ZT, Yang CL, *et al*. Gender differences in smoking behaviorsin an asian population. Journal of Women's Health. 2008; 17(6): 971-978. PMid:18681817 http://dx.doi.org/10.1089/jwh.2007.0621

[15] Chiolero A, Faeh D, Paccaud F, *et al*. Consequences of smoking for body weight, body fat distribution, and insulin resistance. The American journal of clinical nutrition. 2008; 87(4): 801-809. PMid:18400700

[16] Team RC. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.

[17] Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. Bioinformatics. 2011; 27(16): 2304-2305. PMid:21653516 http://dx.doi.org/10.1093/bioinformatics/btr341

[18] Liaw A, Wiener M. Classification and regression by randomforest. R News. 2002; 2(3): 18-22.

[19] Breiman L. Random forests. Machine learning. 2001; 45(1): 5-32. http://dx.doi.org/10.1023/A:1010933404324

[20] Boulesteix AL, Janitza S, Kruppa J, *et al*. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012; 2(6): 493-507. http://dx.doi.org/10.1002/widm.1072

[21] Deng H, Runger G. Feature selection via regularized trees. The 2012 International Joint Conference on Neural Networks (IJCNN). 2012.

[22] Hoffman GE. Correcting for population structure and kinship using the linear mixed model: Theory and extensions. PLoS ONE. 2013; 8(10): e75707-10. PMid:24204578 http://dx.doi.org/10.1371/journal.pone.0075707

[23] Robin X, Turck N, Hainard A, *et al*. Proc: an open-source package for r and s+ to analyze andcompare roc curves. BMC bioinformatics. 2011; 12(1): 77. PMid:21414208 http://dx.doi.org/10.1186/1471-2105-12-77

[24] Metz CE. Basic principles of roc analysis. In Seminars in nuclear medicine. 1978; 8: 283-298. http://dx.doi.org/10.1016/S0001-2998(78)80014-2

[25] Kraft P, Yen YC , Stram DO, *et al*. Exploiting gene-environment interaction to detect genetic associations. Human heredity. 2007; 63(2): 111-119. PMid:17283440 http://dx.doi.org/10.1159/000099183

[26] Dai JY, Logsdon BA, Huang Y, *et al*. Simultaneously testing for marginal genetic association and geneenvironment interaction. American journal of epidemiology. 2012; 176(2): 164-173. PMid:22771729 http://dx.doi.org/10.1093/aje/kwr521