Global Adaptive Rank Truncated Product method for gene set analysis in Association Studies

N. Vilor-Tejedor* and ML. Calle

¹ Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain.

² Bioinformatics and Medical Statistics group, Department of Systems Biology, University of Vic.

Received 15 November 2004, revised 30 November 2004, accepted 2 December 2004

Published online 3 December 2004

Gene set analysis (GSA) aims to assess the overall association of a set of genetic variants with a phenotype and has the potential to detect subtle effects of variants in a gene or a pathway that might be missed when assessed individually. We present a new implementation of the Adaptive Rank Truncated Product method (ARTP) for analyzing the association of a set of Single Nucleotide Polymorphisms (SNPs) in a gene or pathway. The new implementation, referred as **globalARTP**, improves the original one by allowing the different SNPs in the set to have different modes of inheritance. We perform a simulation study for exploring the power of the proposed methodology in a set of scenarios with different number of causal SNPs with different effect sizes. Moreover, we show the advantage of using the gene set approach in the context of an Alzheimer disease case-control study where we explore the endocytosis pathway.

The new method is implemented in the R function **globalARTP** of the **globalGSA** package available at <http://cran.r-project.org/>

Key words: Adaptive method, Alzheimer disease, ARTP, endocytosis pathway, genetic susceptibility, gene-set analysis, globalARTP, single-SNP analysis

1 Introduction

Genetic epidemiology is focused on the study of the genetic causes that determine health and diseases in populations. To achieve this goal a common strategy is to explore differences in genetic variability between diseased and non-diseased individuals. Usual markers of genetic variability are single nucleotide polymorphisms (SNPs) which are changes in just one base in the genome. SNPs are responsible of approximately 90% of human genetic variation. In candidate gene studies, a set of genes are selected for their presumed relationship with the disease, and the SNPs on these genes are genotyped. In contrast, Genome Wide Association Studies (GWAS) are large scale hypothesis free case-control studies where SNPs are genotyped throughout the genome in order to cover most of the human genetic variability.

The usual statistical approach in genetic epidemiology study is a marginal univariate analysis, referred in our work as single-SNP analysis, where each SNP is analyzed separately for association with the phenotype. It has been observed, however, that the single-SNP analysis is not very powerful for detecting genetic causing variants of common diseases. Indeed, extremely large sample sizes would be needed to reliably detect the individual risk SNPs. Moreover, genetic variants with small individual effect sizes but jointly significant genetic effects would be missed by single-SNP analysis. Thus, the identified SNPs represent only a small part of all genetic variants related with the disease while the majority of the variants remain hidden within the statistical 'noise'. As a result, the identified genetic variants only explain a small fraction of the genetic variability (heritability) for most studied traits and it produces an important limitation in this type of studies.

*Corresponding author: e-mail: nvilor@creal.cat

In this work, we consider Gene-Set Analysis (GSA) as an alternative to single-SNP analysis that could improve the power of genetic association studies by exploring functionally and biologically meaningful sets of SNPs, corresponding to genes or pathways. This strategy aims to provide an overall measure of association of a set of genetic variants with a phenotype, such as disease status or a quantitative trait.

This approach provides the potential to detect subtle effects of multiple SNPs in a gene or a pathway that might be missed when assessed individually. Moreover, it reduces the multiple testing burden that appears when performing a large number of single-SNP tests and it incorporates biological knowledge in the statistical analysis, improving the statistical power to detect causal genes Maciejewski (2013).

One such GSA approach is the Adaptive Rank Truncated Product method Yu et al. (2009) that provides an algorithm for combining marginal p-values to a gene or pathway p-value level using a rank truncated statistic. One important limitation of the current implementation of this method is that it is only possible to combine p-values that have been obtained from the same inheritance model (additive), while the different SNPs in a gene or pathway may follow different modes of inheritance which require different modeling techniques.

In this work we propose a new GSA method, called global Adaptive Rank Truncated Product method (globalARTP) (Subsection 2.2), that extends the ARTP method by allowing different modes of inheritance for each SNP in a gene. We explore the performance of the globalARTP algorithm for identifying causal genes in different simulated scenarios in comparison with the usual single-SNP analysis method (Section 3). Finally, we apply both methods in the context of a real study on Alzheimer disease (Section 4).

2 Materials and methods

2.1 Single-SNP analysis

Most SNPs are biallelic polymorphisms, this means that two possible variants (alleles) are observed in the population at that specific locus. Let's denote by a the allele that is more frequent in the population (wild-type or major allele) and by A the less frequent allele (minor or variant allele). Since humans are diploid, that is, each cell contains two copies of the genome, at each SNP locus three possible genotypes are observed: AA for major homozygous, Aa for heterozygous and aa for minor homozygous. Different codifications of the genotypes correspond to different modes of inheritance as is summarized in Table 1. In the dominant model, a single copy of the variant allele is sufficient to modify (increase or decrease) the risk of disease. In contrast, in the recessive model two copies of the variant allele are necessary to modify the risk. In the additive model, each copy of the variant allele confers an additive increase (or decrease) in risk (in the appropriate disease risk scale). In this case, disease risk is linearly related with the number of minor alleles. And finally, the most general model is the codominant model where the three genotypes confer different effects on disease risk. This model requires the specification of two dummy variables in the model.

Table 1: SNP codification under different inheritance modes.

	Dominant	Recessive	Additive	Codominant	
Genotype	G	G	G	G_1	G_2
AA	0	0	0	0	0
Aa	1	0	1	1	0
aa	1	1	2	0	1

In case-control studies, where the response Y is a binary variable representing disease status, the association of a SNP with the phenotype can be performed by analyzing a 2×3 contingency table with a chi-squared or Fisher test. Alternatively, a logistic regression model as in equation (1) can be fit where G is the codification of the SNP as specified in Table 1 and Z represents other non genetic covariates, where

$$\pi = \Pr(Y = 1 \mid G, Z)$$

$$\text{logit}(\pi) = \beta_0 + \beta_1 G + \delta Z \quad (1)$$

In this model, $\exp(\beta_1)$ is the odds-ratio of the group with $G = 1$ with respect to the reference group.

2.2 Global Adaptive Rank Truncated Product method (globalARTP)

We consider a case-control genetic association study where Y denotes disease status and $G_1, \dots, G_M$ are the genotypes for a set of M genotyped SNPs within a gene or pathway. For easy of explanation we will describe the methodology assuming that the M SNPs belong to a gene. In this context, the goal is to combine the individual SNP p-values into a gene-level p-value. Yu et al. (2009) proposed a GSA method, the Adaptive Rank Truncated Product method (ARTP), for combining the individual evidence of association over the different SNPs. The test statistic for the ARTP method is the Rank Truncated Product statistic, which combines the k smallest marginal p-values:

$$W_k = \prod_{j=1}^k p_{(j)}. \quad (2)$$

The RTP statistic W_k was initially proposed by Zaykin et al. (2002) with a fixed truncation point k specified in advance. In the ARTP method, k is obtained in an adaptive way by optimizing the association evidence obtained for a set of candidate truncation points (described below in step 4).

One limitation of the ARTP algorithm, and also of other GSA methods, is that they assume the same mode of inheritance for all the SNPs in the set (usually, the additive model). For this reason, we propose a new gene-set method, the Global Adaptive Rank Truncated Product method (globalARTP), that improves the original ARTP algorithm by introducing an additional step where a global test for the best mode of inheritance of each SNP is performed.

The globalARTP algorithm is described below, where step 1 performs the selection of the best inheritance model for each SNP and steps 2 to 5 correspond to the standard ARTP method with small modifications.

Step 1. Best genetic model: The first step of the globalARTP performs a test for association of each SNP with the phenotype, considering four different modes of inheritance (dominant, recessive, additive and codominant). The minimum of the four likelihood ratio test p-values is taken:

$$p_i = \min\{p_i^{dom}, p_i^{rec}, p_i^{add}, p_i^{cod}\}, \quad i \in \{1, \dots, M\}. \quad (3)$$

Step 2. Rank truncated product statistic: The M p-values, one for each SNP in the gene, are sorted increasingly:

$$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(M)}. \quad (4)$$

Given a maximum truncation value $K \leq M$, the rank truncated product statistic W_k , as defined in (2), would be computed for each candidate truncation point $k \in \{1, \dots, K\}$. Indeed, in order to improve computational efficiency and avoid computational problems we will work with the log transformation of W_k , denoted by

$$V_k = \sum_{j=1}^k -\log(p_{(j)}), \quad 1 \leq k \leq K. \quad (5)$$

Step 3. Permutational null distribution of statistics V_k : We obtain the permutational null distribution of V_k , $1 \leq k \leq K$, under the null hypothesis that none of the M SNPs in the gene are associated with the

disease, by resampling the phenotype variable B times and performing steps 1 and 2 on each permuted datasets. From step 1, we obtain and sort the M single-SNP p-values corresponding to the best inheritance mode of each SNP:

$$p_{(1)}^b \leq p_{(2)}^b \leq \dots \leq p_{(M)}^b, \quad 0 \leq b \leq B. \quad (6)$$

We include in (2.2) the value $b = 0$ which corresponds to the original dataset. When performing step 2 we obtain the test statistic for each candidate truncation point k and each permuted sample b :

$$V_k^{(b)} = \sum_{j=1}^k -\log(p_{(j)}^b), \quad 1 \leq k \leq K, \quad 0 \leq b \leq B. \quad (7)$$

For each truncation point $k \in \{1, \dots, K\}$, we can obtain the significance of the observed test statistic $V_k^{(0)}$ by comparing its value with the values under the permutational null distribution $V_k^{(b)}$, $1 \leq b \leq B$. The following expression provides the permutational p-values for statistics $V_k^{(0)}$, $1 \leq k \leq K$ under the null hypothesis:

$$\hat{S}_k^{(0)} = \frac{\sum_{l=0}^B I(V_k^{(l)} \geq V_k^{(0)})}{B+1}, \quad 1 \leq k \leq K. \quad (8)$$

Step 4. Best truncated point: An additional step is to optimize the number k of SNPs that are combined for each gene. We define the optimal truncation point k_{opt} as the number $k \in \{1, \dots, K\}$ that minimizes $\hat{S}_k^{(0)}$, and this minimum is denoted as $\min P^{(0)}$:

$$\min P^{(0)} = \min_{1 \leq k \leq K} S_k^{(0)}. \quad (9)$$

However, this step introduces a bias in favor of the alternative hypothesis that has to be adjusted. As discussed in Yu et al. (2009), the possibility of performing this adjustment with an additional permutational procedure would become computationally infeasible in many situations. For this reason they proposed an alternative algorithm that provides an adjusted gene-p-value but only requires one permutational step (step 3). The procedure relies on obtaining the p-values S_k not only for the original statistics, but also for the permuted statistics:

$$\hat{S}_k^{(b)} = \frac{\sum_{l=0}^B I(V_k^{(l)} \geq V_k^{(b)})}{B+1}, \quad 1 \leq k \leq K, \quad 1 \leq b \leq B. \quad (10)$$

Then, for each permuted dataset $0 \leq b \leq B$ we obtain the minimum p-value denoted as $\min P^{(b)}$:

$$\min P^{(b)} = \min_{1 \leq k \leq K} S_k^{(b)}, \quad 1 \leq b \leq B. \quad (11)$$

Step 5. Gene-p-value: Finally, the gene-p-value is obtained by comparing the minimum p-value in the original dataset $\min P^{(0)}$ with the minimum p-values in the permuted datasets $\min P^{(b)}$, $1 \leq b \leq B$:

$$\text{gene-p-value} = \frac{\sum_{l=0}^B I(\min P^{(l)} \leq \min P^{(0)})}{B+1}. \quad (12)$$

3 Simulation study

3.1 Simulation design

The purpose of this simulation study is two-fold: The first goal is to compare the power of the single-SNP analysis and the GSA approach using the globalARTP algorithm, and the second goal is to prove the improvement of the globalARTP over the ARTP when the underlying inheritance model is not additive.

For this, we generated different scenarios corresponding to balanced case-control studies with sample size $N = 2,000$ (1,000 cases and 1,000 controls) following the *Odds of risk model* as described in Wray and Goddard (2010). We considered groups of SNPs (genes) containing M independent SNPs ($M = 10, 50, 100$) with a random minor allele frequency following a Beta distribution restricted to the interval $[0, 0.5]$ with mean equal to 0.2. For generating the disease status we considered a disease prevalence equal to 0.2 and assumed that the first c SNPs in the dataset were causal SNPs sharing the same inheritance model (additive or recessive) and with the same effect size. We also considered the dominant model but the results are very similar to the additive model and are not reported. We explored the size of the test in the case where none of the SNPs in the gene is causal ($c = 0$) and the power of the test for detecting association at the gene-level when the number of causal SNPs within the gene is $c = 10$. For the additive genetic model we considered $RR = 1.2$ and $RR = 1.1$, where $RR = P(Y = 1|Aa)/P(Y = 1|AA)$ is the relative risk of the heterozygous group versus major homozygous group and the relative risk of the minor homozygous group versus major homozygous group is RR^2 . For the recessive genetic model we considered again $RR = 1.2$ and $RR = 1.1$, but now $RR = P(Y = 1|aa)/P(Y = 1|AA \cup Aa)$ is the relative risk of the minor homozygous group versus the group containing the major homozygous and heterozygous individuals.

For each scenario we computed the gene-p-value using the ARTP and globalARTP algorithms allocating the number of permutation data sets equal to $B = 10,000$ and the maximum truncation point value equal to $K = 5$. We repeated this process a hundred times for each scenario and we averaged the results over the hundred replications. We provide the percentage of times that the gene is significant at a significance level $\alpha = 0.05$ (gene-p-value < 0.05).

The GlobalARTP and single-SNP analysis are difficult to compare since one is a gene set approach providing just one p-value for the gene while the single-SNP analysis provides several p-values, one for each SNP in the gene. In order to be able to compare both approaches, we performed single-SNP analysis and declared that a gene was significantly associated with the disease when at least one SNP in the gene was significant after adjustment for the M univariate tests performed in each gene. The multiple testing adjustment is performed with the Bonferroni correction that controls the familywise error rate (FWER) and also by the method of Benjamini and Hochberg that controls the false discovery rate (FDR). For both methods we considered a threshold level equal to 0.05.

3.2 Simulation Results

The results of the simulation study are summarized in Tables from 2 to 4.

Table 2 provides the size of the tests, that is the percentage of significant results when there is no causal SNP. While globalARTP and ARTP controls the size around the specified significance level (5%), both, single-SNP analysis with Bonferroni adjustment and FDR adjustment are rather conservative.

Tables 3 and 4 provide the power of the test, that is, the percentage of significant results when there are $c = 10$ causal SNPs within the M available SNPs in the gene. Table 3 corresponds to the additive model while Table 4 corresponds to the recessive model.

For an additive model and a high individual effect of each causal SNP ($RR = 1.2$) (first part of Table 3), both GSA methods, ARTP and globalARTP, have maximum power (100%), while the power of the single-SNP approach reduces when the number of non-causal SNPs in the gene, $M - c$, increases. Specifically, when all SNPs in the gene are causal ($M = c = 10$), all methods have maximum power (100%). When the number M of SNPs in the gene increases to 50 and 100 the globalARTP and ARTP methods still have maximum power (100%) while the single-SNP analysis reduces the power due to the multiple testing adjustments: 96% and 90% with Bonferroni correction and 99% and 94% with FDR correction, respectively.

For an additive model and a small individual effect of each causal SNP ($RR = 1.1$) (second part of Table 3), all methods experience a reduction in their power. Still, there is a clear superiority of GSA methods over the single-SNP approach. The best situation is when all SNPs in the gene are causal ($M = c = 10$). In

this case the GSA approaches have a power around 80% while the single-SNP analysis have a power around 50%. When the number M of SNPs in the gene increases to 50, the globalARTP and ARTP methods reduces their power to values around 40% while the power of the single-SNP analysis reduces to 26%. When the number M of SNPs in the gene increases to 100, the globalARTP and ARTP methods have a power around 30% while the power of the single-SNP analysis is around 25%. In this settings where the underlying inheritance model is additive, the ARTP, that assumes an additive model, is more powerful than the globalARTP that explores among the four possible genetic models. The ARTP has a power that is around 8% larger than the power of the globalARTP.

For a recessive model and a high individual effect of each causal SNP ($RR = 1.2$) (first part of Table 4), the most powerful approach is clearly the globalARTP method. When all SNPs in the gene are causal ($M = c = 10$), the power of the globalARTP is 91% while the power of the ARTP is only 43%. When the number M of SNPs in the gene increases to 50, the globalARTP reduces its power to 47%, but the ARTP reduces the power to 17%. For $M = 100$, the power of the globalARTP is 38% while the ARTP method has a power of 24%.

For a recessive model and a small individual effect of each causal SNP ($RR = 1.1$) (second part of Table 4), all methods have a poor behavior. When all SNPs are causal ($M = c = 10$), the power of the globalARTP is 25% while the power of the ARTP is only 15% and, when the number of noncausal SNPs in the gene increases ($M = 50$ and $M = 100$) all methods have a power below 10%.

Table 2: Size of the tests

	$M = 10$	$M = 50$	$M = 100$
<i>GSA: globalARTP</i>	5%	4%	2%
<i>GSA: ARTP</i>	5%	5%	2%
<i>Single-SNP: Bonferroni adj.</i>	5%	3%	2%
<i>Single-SNP: FDR adj.</i>	5%	3%	2%

Table 3: Power of the tests when the 10 causal SNPs follow an additive inheritance model: $RR = P(Y = 1|Aa)/P(Y = 1|AA)$ and $RR^2 = P(Y = 1|aa)/P(Y = 1|AA)$

$RR = 1.2$	$M = 10$	$M = 50$	$M = 100$
<i>GSA: globalARTP</i>	100%	100%	100%
<i>GSA: ARTP</i>	100%	100%	100%
<i>Single-SNP: Bonferroni adj.</i>	100%	96%	90%
<i>Single-SNP: FDR adj.</i>	100%	99%	94%
$RR = 1.1$	$M = 10$	$M = 50$	$M = 100$
<i>GSA: globalARTP</i>	79%	35%	29%
<i>GSA: ARTP</i>	87%	43%	37%
<i>Single-SNP: Bonferroni adj.</i>	54%	26%	24%
<i>Single-SNP: FDR adj.</i>	55%	26%	25%

In summary, first we observe that as it was expected the GSA approaches are more powerful than single-SNP analysis in all settings considered in this simulation study. The advantage of the GSA approach over the single-SNP approach is very important in some situations, for instance when the underlying genetic model is recessive and $M = c = 10$, the difference in power between the globalARTP and the single-SNP approach is more than 25%.

On the other hand, with this simulation study we can compare the behavior of the original ARTP algorithm and the new globalARTP method. When the true genetic model is additive, the ARTP method is more efficient than the globalARTP. In this setting, the additional step for searching the best genetic model reduces the power of the globalARTP in about 8% with respect to the ARTP algorithm. However, when

Table 4: Power of the tests when the 10 causal SNPs follow a recessive inheritance model:
 $RR = P(Y = 1|aa)/P(Y = 1|AA \cup Aa)$

$RR = 1.2$	$M = 10$	$M = 50$	$M = 100$
<i>GSA: globalARTP</i>	91%	47%	38%
<i>GSA: ARTP</i>	43%	17%	24%
<i>Single-SNP: Bonferroni adj.</i>	65%	30%	33%
<i>Single-SNP: FDR adj.</i>	66%	32%	33%
$RR = 1.1$	$M = 10$	$M = 50$	$M = 100$
<i>GSA: globalARTP</i>	25%	8%	9%
<i>GSA: ARTP</i>	15%	7%	8%
<i>Single-SNP: Bonferroni adj.</i>	18%	7%	8%
<i>Single-SNP: FDR adj.</i>	19%	7%	8%

the true underlying genetic model is not additive, the globalARTP is more powerful than the ARTP. In some situations, this advantage is very large, like in the scenario with $M = c = 10$ where the power of the globalARTP is 91% while the power of the ARTP is only 43%.

4 Statistical Evidence of Neuronal Endocytosis in Alzheimer's disease

To demonstrate the application of the proposed methodology, we analyze the association between the endocytic pathway, conformed by *Cdk5*, *LRP2*, *APP* and *Abl1* genes, with Alzheimer's disease (Figure 1). The neuronal endocytic pathway is an intracellular system for processing extracellular nutrients and trophic factors, for recycling or catabolizing receptors and other integral membrane proteins after neurotransmitter release and for directing information to intracellular biosynthetic pathways. Endosomal alterations were evident in most neurons in individuals with Alzheimer's disease, absent in several other neurodegenerative disorders examined Cataldo et al. (2000). Although, the genetic causes for most Alzheimer's cases are still essentially unknown, clinical investigations showed that *APOE* protein plays an eminent role at the most active neurogenesis sites. *APOE* occurs in 3 common isoforms (*e2*, *e3* and *e4*) in the human population and the *APOE-e4* variant is the primary genetic risk factor for late-onset Alzheimer's disease Herz and Beffert (2000). In this example, we restrict the analysis to *APOE-e4* non-carrier individuals. We compare the usual single-SNP strategy that analyzes marginal association of each SNP with the phenotype with the proposed gene-set analysis method. The genotype information about the considered genes is extracted from a public GWAS Reiman et al. (2007), for 1411 individuals (861 cases, 550 controls). The number of genotyped SNPs in each gene is given in Table 5.

Table 5: Number of SNPs in each gene of the endocytosis pathway.

Gene	number of SNP
<i>LRP2</i>	44
<i>APP</i>	50
<i>Abl1</i>	27
<i>Cdk5</i>	4

4.1 Single-SNP analysis

In the context of this application, we focused our analysis on *APOE-e4* non-carrier individuals (334 cases / 433 controls). We analyzed the univariate association between each SNP with the response variable (status: case/control). A total of 125 SNPs in the four genes were analyzed. For each gene we considered four

Figure 1: Endocytosis pathway Network.

different inheritance modes (dominant, codominant, recessive and log-additive) and performed a logistic regression with the corresponding codification of the SNPs. Table 6 provides the list of SNPs that were significant at a significance level of 0.01. We obtained 16 significant SNPs before multiple testing adjustment. After multiple testing correction (Bonferroni and FDR) none of the SNPs remained significant.

4.2 GlobalARTP analysis

We performed the proposed gene-set analysis on the four considered genes in the Endocytosis pathway for association with Alzheimer disease. The global-p-value for each gene is computed considering a truncation point equal to 5 and using 1,000 permutations. The results are given in Table 7. In contrast to the single-SNP approach where none SNPs in the pathway were identified as associated with disease, the gene-set analysis identifies 2 genes, *LRP2* and *APP*, with a gene-set p-value smaller or very close to the significance level of 0.05. This indicates that the genetic variability within *LRP2* and *APP* genes might be associated with Alzheimer disease risk. Of course, this preliminary result should be taken with caution until the result is validated in an independent dataset.

5 Discussion

In this paper we focus on Gene Set Analysis (GSA), a strategy for combining the effects of many genetic variants within a gene. We propose the algorithm globalARTP, a new implementation of the ARTP method that is specifically designed for genetic association studies involving SNPs. The new implementation incorporates the selection of the best inheritance model for each SNP as a first step of the algorithm.

Table 6: Significant SNPs at 1% significance level before multiple testing adjustment ordered by p-value.

	codominant	dominant	recessive	log.additive	min	adj.Bonf	adj.FDR	Gene
<i>rs16856748</i>	0.0017	0.0003	0.4209	0.0007	0.0003	0.2380	0.2278	<i>LRP2</i>
<i>rs436011</i>	0.0047	0.0053	0.0134	0.0011	0.0011	0.7617	0.2278	<i>APP</i>
<i>rs2830076</i>	0.0051	0.0041	0.0203	0.0012	0.0012	0.7881	0.2278	<i>APP</i>
<i>rs2830075</i>	0.0063	0.0021	0.1052	0.0017	0.0017	1.0000	0.2278	<i>APP</i>
<i>rs830955</i>	0.0114	0.0140	0.0185	0.0031	0.0031	1.0000	0.2428	<i>LRP2</i>
<i>rs11792273</i>	0.0096	0.0074	0.0434	0.0034	0.0034	1.0000	0.2428	<i>Ab11</i>
<i>rs2193193</i>	0.0132	0.1074	0.0043	0.0063	0.0043	1.0000	0.2428	<i>LRP2</i>
<i>rs432766</i>	0.0201	0.0120	0.0457	0.0055	0.0055	1.0000	0.2597	<i>APP</i>
<i>rs2830073</i>	0.0173	0.0056	0.1184	0.0070	0.0056	1.0000	0.2597	<i>APP</i>
<i>rs2268370</i>	0.0272	0.0313	0.0268	0.0075	0.0075	1.0000	0.2597	<i>LRP2</i>
<i>rs2052297</i>	0.0296	0.2677	0.0083	0.0220	0.0083	1.0000	0.2597	<i>LRP2</i>
<i>rs830959</i>	0.0318	0.0385	0.0277	0.0088	0.0088	1.0000	0.2597	<i>LRP2</i>
<i>rs6597644</i>	0.0289	0.0179	0.0687	0.0089	0.0089	1.0000	0.2597	<i>Ab11</i>
<i>rs2186302</i>	0.0312	0.0119	0.1457	0.0089	0.0089	1.0000	0.2597	<i>APP</i>
<i>rs2239594</i>	0.0234	0.0936	0.0092	0.0094	0.0092	1.0000	0.2597	<i>LRP2</i>
<i>rs2234988</i>	0.0338	0.0229	0.0561	0.0093	0.0093	1.0000	0.2597	<i>APP</i>

Table 7: globalARTP results for non-ApoE4 carriers.

Gene	nSNP	Permutations	Truncation point	gene p-value
<i>LRP2</i>	44	1,000	5	0.0549
<i>APP</i>	50	1,000	5	0.0499
<i>Ab11</i>	27	1,000	5	0.1808
<i>Cdk5</i>	4	1,000	4	0.5254

Through a simulation study we proved that Gene Set Analysis increases the power to detect genetic associations when the individual effects are very small, which is the usual case in complex diseases. In this situation, most causal SNP effects are not detectable with the common single-SNP testing procedure followed by correction for multiple comparisons. By combining the p-values of a set of SNPs in a gene we reduce the number of tests, and thus the multiple testing corrections needed. Moreover, in many cases, the association results given at the gene level may be more biologically interpretable.

With the simulation study we also compared the original ARTP algorithm with the proposed globalARTP method. While the ARTP method assumes the same inheritance model (additive) for all the SNPs in the gene, the new globalARTP method searches for the best inheritance model for each SNP. As it was to be expected, the ARTP method performs slightly better than the globalARTP when the true genetic model is additive, with a difference in power around 8%. This is the consequence of the multiple testing adjustment that intrinsically the globalARTP permutational procedure performs for the four genetic tests considered. In contrast, when the true underlying genetic model is not additive, the globalARTP is more powerful than the ARTP. The difference in power in favor of the globalARTP can be very large, ranging from 48% to 14% when the model is recessive and the relative risk is $RR = 1.2$. Thus, we conclude that the globalARTP is preferable since what we can lose by using it is much less than what we can lose by using the ARTP method when the additive model is not correct.

We also applied GSA in a real case in the context of Alzheimer's disease. While the single-SNP analysis does not detect any association at the univariate level, GSA detects two interesting genes that are worth further investigating. GSA has also some limitations. This strategy will only be useful in the presence of marginal effects. However, it will not be effective when the genetic association is due to gene interactions

without marginal individual effects. This will require specific methods for gene-gene interaction detection such as the MB-MDR method Calle et al., (2007).

The new method is implemented in the R function `globalARTP` of the `globalGSA` package available at <http://cran.r-project.org/>.

The data for the simulation study is available at <http://campus.omicsuvic.com/globalARTP/>.

Acknowledgements This research was partially supported by grant MTM2012-38067-C02-02 from the Ministerio de Economía e Innovación (Spain), grant 2009SGR-581 from Generalitat de Catalunya and Beca UNNIM en Ciències de la Salut 2011.

References

- Calle, M.L. Urrea, V. Malats, N. and Van Steen, K. (2007). MB-MDR: Model-Based Multifactor Dimensionality Reduction for detecting interactions in high-dimensional genomic data. *Ann. Hum. Genetics.* **75**(1), 78-89.
- Cataldo, A.M. Peterhoff, C.M. Troncoso, J.C. Gomez-Isla, T. Hyman, B.T. and Nixon, R.A. (2000). Endocytic pathway abnormalities precede amyloid beta deposition in sporadic Alzheimer's disease and Down syndrome: differential effects of APOE genotype and presenilin mutations. *Am J Pathol.* **157**(1), 277-86.
- Fridley, B.L. and Biernacka, J.M. (2011). Gene set analysis of SNP data: benefits, challenges, and future directions. *European Journal of Human genetics.* **19**, 837-843.
- Wray, N.R. and Goddard M.E. (2010). Multi-locus models of genetic risk of disease. *Genome Medicine.* **2**:10.
- Herz, J. and Beffert, U. (2000). Apolipoprotein E receptors: linking brain development and Alzheimer's disease. *Nature Reviews Neuroscience* . **Oct.1**, 51-58.
- Maciejewski, H. (2013). Gene set analysis methods: statistical models and methodological differences. *Briefing in Bioinformatics.* doi: 10.1093/bib/bbt002.
- Reiman, E.M. Webster, J.A. Myers, A.J. Hardy, J. Dunckley, T. Zismann, V.L. Joshipura, K.D. Pearson, J.V. Hu-Lince, D. Huentelman, M.J. Craig, D.W. Coon, K.D. Liang, W.S. Herbert, R.H. Beach, T. Roher, K.C. Zhao, A.S. Leung, D. Bryden, L. Marlowe, L. Kaleem, M. Mastroeni, D. Grover, A. Heward, C.B. Ravid, R. Rogers, J. Hutton, M.L. Melquist, S. Petersen, R.C. Alexander, G.E. Caselli, R.J. Papassotiropoulos, A. and Stephan D.A. (2007). GAB2 Alleles Modify Alzheimer's Risk in APOE e4 Carriers. *Neuron.* **Jun 7; 54**(5), 713-20.
- Yu, K. Li, Q. Bergen, A.W. Pfeiffer, R.M. Rosenberg, P.S. Caporaso, N. Kraft, P. and Chatterjee, N. (2009). Pathway analysis by adaptive combination of P-values. *Genet Epidemiol.* **33**(8), 700-9.
- Zaykin, D.V. Zhivotovsky, L.A. Westfall, P.H. and Weir, B.S. (2002). Truncated product method for combining P-values. *Genet Epidemiol.* **22**(2), 170-85.