Efficient and powerful testing for gene set analysis applied to Genome-wide Association Studies.

N. Vilor-Tejedor*, JR. González and ML. Calle

Abstract— The goal of Genome-wide Association Studies (GWAS) is the identification of genetic variants, usually Single Nucleotide Polymorphisms (SNPs), that are associated with disease risk. However, SNPs detected so far with GWAS for most common diseases only explain a small proportion of their total heritability. Gene Set Analysis (GSA) has been proposed as an alternative to single-SNP analysis with the aim of improving the power of genetic association studies. Nevertheless, most GSA methods rely on expensive computational procedures that make unfeasible their implementation in GWAS. We propose a new GSA method, referred as globalEVT, that uses the extreme value theory to derive gene set p-values. GlobalEVT reduces dramatically the computational requirements compared to other GSA approaches. In addition, this new approach improves the power by allowing different inheritance models for each genetic variant as illustrated in the simulation study performed and allows the existence of correlation between the SNPs. Real data analysis of an Attention-deficit/hyperactivity disorder (ADHD) study illustrates the importance of using GSA approaches for exploring new susceptibility genes. Specifically, the globalEVT method is able to detect genes related to Cyclophilin A like domain proteins which is known to play an important role in the mechanisms of ADHD development.

Index Terms— Attention-deficit/hyperactivity disorder, Adaptive Rank Truncated Product method, Cyclophilin domain, Extreme value theory, globalEVT.

1 Introduction

Genetic epidemiology focuses on the identification of genetic variants that are associated with health and disease in populations, and also, in the study of how these genetic variants interact with environmental factors [1]. A common strategy is to explore differences in genetic variability between diseased and non-diseased individuals using single nucleotide polymorphisms (SNPs) as markers of the variability in a genome region. Single-SNP analysis, where each SNP is individually tested, is the usual statistical approach in Genome-wide Association Studies (GWAS). However, the main limitation of this strategy is the multiple testing correction that reduces dramatically the statistical power [2]. Gene set analysis (GSA) is an alternative approach that provides the association between a set of SNPs (i.e., gene sets or pathways) and the trait. These strategies meant to improve the power of single-SNP analysis when the marginal effect of each SNP is small.

A number of GSA methods have been proposed to analyze aggregated association evidences across SNPs. These include the Fisher's combination product method [3], the Stouffer's method [4] and the Wilkinson procedure [5]. Based on these, more recent methods were proposed such as the Threshold Truncated Product (TPM) method [6], the Rank Truncated Product (RPT) method [7], the Adaptive Rank Truncated Product (ARTP) method [8], the Sequential Test (SEQ) method [9] and the global Adaptive Rank Truncated Product (globalARTP) method [10]. Nevertheless, theoretical distributions of these techniques assume independence on the p-values while this is not likely to be true because of the existence of Linkage Disequilibrium (LD). Therefore, significance is usually obtained through permutational approaches, such as randomly permuting the phenotype among individuals several times (at least 100,000,000 permutations for GWAS), which requires expensive computations.

To overcome these problems, we propose the globalEVT algorithm, a new implementation of the ARTP test statistic, where the theoretical distribution is obtained based on the extreme value theory. The main advantages of the proposed approach are that (1) it considers different inheritance models as proposed in [10], (2) it allows for correlation between the SNPs as suggested in [11], and more importantly, (3) it reduces dramatically the required computational time making possible its application in the context of GWAS.

This article is organized as follows. In section 2, we present the proposed globalEVT algorithm. In section 3, we explore the performance and the computational requirements of the

- *N. Vilor-Tejedor, Center for Research in Environmental Epidemiology (CREAL), Universitat Pompeu Fabra (UPF) and CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain. E-mail:* ***nvilor@creal.cat***
- *JR. González, Center for Research in Environmental Epidemiology (CREAL), Universitat Pompeu Fabra (UPF) and CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain. E-mail: jrgonzalez@creal.cat*
- *ML. Calle, Department of Systems Biology, Bioinformatics and Medical Statistics Group, Universitat de Vic - Universitat Central de Catalunya, CO 08570-Vic, Spain. E-mail: malu.calle@uvic.cat*

******Please provide a complete mailing address for each author, as this is the address the 10 complimentary reprints of your paper will be sent***

*Please note that all acknowledgments should be placed at the end of the paper, before the bibliography (**note that corresponding authorship is not noted in affiliation box, but in acknowledgment section**).*

globalEVT. For this purpose, we simulated different scenarios where we compare the results for globalEVT with ARTP and globalARTP. In section 4, we applied single-SNP analysis and the globalEVT in the context of a medical study on Attention-deficit/hyperactivity disorder (ADHD). As a result, we obtained some significant genes involved in the Cyclophilin-like protein receptor which was previously suggested as an important biological regulator for adult ADHD [12]. Finally, the paper ends with a discussion.

2 METHODS

Our starting point is the Adaptive Rank Truncated Product method (ARTP) proposed by Yu et al., [8]. This GSA method consists on the combination of the k smallest marginal p-values using a rank truncated statistic, where k is determined in an adaptive way. One limitation of this approach, and also of other GSA methods, is that they assume the same mode of inheritance for all SNPs in the set, usually, the additive model. GlobalARTP proposed by Vilor-Tejedor et al., [10] is an extension of ARTP that allows for different modes of inheritance of the SNPs. Another problem of using this approach is the computational requirements since the final gene set p-value relies on the nonparametric null distribution of the ARTP test statistic which is estimated using permutational procedures. [13] proposed the use of the generalized extreme value distribution for estimating the null distribution of this statistic. The maximum likelihood estimation of the three parameters (location, scale, and shape parameters) of the generalized extreme value distribution also requires the performance of a large number of permutations, but much less than the nonparametric estimation and the tails of the distribution are estimated more accurately.

In this work we propose an alternative algorithm, referred to as the globalEVT, for estimating the null distribution of the ARTP test statistic using the extreme-value theory and a limited number of permutations. Our method reduces dramatically the computational requirements since only one-parameter distributions have to be fitted. We also improve the power of the proposed GSA approach by allowing different modes of inheritance for each SNP in the set and using the Max-statistic test [14]. Moreover, the proposed method accounts for correlation between the SNPs.

Considering a 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, the proposed algorithm is based on the following result:

Proposition. If $U_1, U_2, \dots, U_M$ are independent and identically distributed uniform random variables in the interval $[0,1]$ then the l-th order statistic, denoted by $U_{(l)}$, follows a Beta distribution Beta(l, M + 1-l) with density given by

$$f_{U_l}(u) = \frac{M!}{(l-1)!\,(M-l)!} u^{l-1}(1-u)^{u-l}$$

We will also assume that when independence does not hold, that is, when $U_1, U_2, \dots, U_M$ are dependent variables with standard Uniform distribution, it is possible to find a number $m^* < M$, so that the distribution of the *l-th* order statistic, $U_{(l)}$, is approximately a Beta distribution, $Beta(l,\ m^* + 1 - l)$, where m^* is interpreted as the effective number of independent tests. Taking these considerations, we propose the following algorithm for obtaining the combined effect of a set of *M* SNPs. For easy of explanation we describe the method in the case where the *M* SNPs belong to the same gene and thus, the combined p-value corresponds to the gene p-value.

Step 1. Best genetic model and transformation to uniformly distributed p-values: The first step performs an association analysis of each SNP with the phenotype, considering three different modes of inheritance (dominant, recessive and additive) and takes the minimum of the three likelihood ratio test p-values. This first step provides *M* p-values, one for each SNP in the gene:

$$\mathrm{p_j^{min}} = \min\{\mathrm{p_j^{dom}}, \mathrm{p_j^{rec}}, \mathrm{p_j^{add}}\}, \quad j = 1, \dots, M$$

where $p_j^{dom}, p_j^{rec}, p_j^{add}$ are the p-values of *j*-SNP assuming a dominant, a recessive and an additive model, respectively. If the three tests were independent the distribution of p^{min} would follow a $Beta(1,3)$ distribution (*see Preposition*, considering $l = 1$, and $M = 3$) but, since the three tests are performed on the same SNP, the three tests are dependent and p^{min} follows a $Beta(1, x)$ where x, the effective number of tests, has been estimated to be equal to 2.2 [14].

We transform p_j^{min}, $\mathrm{j} = 1, \dots, M$ into values from a standard Uniform distribution by applying the distribution function:

$$r_j = F_{Beta(1,2.2)}(p_j^{min}), \qquad j = 1, \dots, M$$

Step 2. Summarizing the k most associated SNPs: We sort increasingly the uniformly distributed p-values, r_j, obtained in step 1, and considers the k best results, for $k \in$ between $\{1, \dots, M\}$.

$$r_{(1)} < r_{(2)} < \cdots < r_{(k)}, \qquad 1 \le k < M$$

Our goal is to summarize these k first order statistics into a unique statistic but, for this, we first transform these values into uniformly distributed values. If the SNPs are not correlated, the order statistics $r_{(j)}$, $j = 1, \dots, k$, would follow a Beta distribution $\mathrm{Beta}(1, M - j + 1)$, but if the SNPs were correlated, the distribution is $\mathrm{Beta}(1, y - j + 1)$, $j = 1, \dots, k$, where y is the effective number of tests calculated through the approximation approach proposed by Li et al., [11]. This method estimates y from the eigenvalues of the correlation matrix of the *M* SNPs, as

$$y = M - \sum [\, I(\boldsymbol{\lambda_j} > 1)(\boldsymbol{\lambda_j} - 1)]), \qquad j = 1, \dots, M$$

where $\boldsymbol{\lambda_j}$ are the eigenvalues and $I(x)$ is an indicator function.

As in the previous step, we transform the order statistics r_j, tics r_j, $j = 1, \dots, k$ into values from a standard Uniform distribution by applying their distribution function:

$$t_j = F_{Beta(j,y-j+1)}(r_{(j)}), \qquad j = 1, \dots, k$$

As a summary statistic of the k best results, we consider the Fisher's combination approach:

$$S_k = -2\sum_{j=1}^{k} \log t_j$$

Due to t_j are uniformly distributed, then $-2\ log\ t_j$ follows a chi-squared distribution with 2 degrees of freedom and, if the k SNPs were uncorrelated the summary statistic S_k would follow a chi-squared distribution with $2k$ degrees of freedom. Since the SNPs may be correlated, the distribution of S_k is a chi-squared distribution with γ degrees of freedom. A permutational procedure is performed to estimate γ as the mean from the permuted values. A small number of permutations, for instance, a hundred, are enough to obtain a good estimate of γ.

We calculate the p-value corresponding to the statistic, S_k, using the reference chi-squared distribution,

$$U_k = 1 - F_{Chi}(S_k)$$

Step 3. Adaptive step: selection of the best truncation point: We repeat Step 2 for every k from 1 to K, where $K \leq M$ is the maximum truncation point to be explored. As a final gene set statistic we take the best of all,

$$W = \min\{U_1, \dots, U_K\}$$

Since $U_1, \dots, U_K$ are correlated, the distribution of W can be approximated by a $Beta(1, z)$, where z can be estimated from a permutational procedure as $z = \frac{\bar{W}-1}{\bar{W}}$, using a small number of permutations. As previously, a hundred, are enough to obtain a good estimate of z.

Finally, the transformation of W to a uniformly distributed value provides the adjusted p-value for the set of M SNPs:

$$gene_{pvalue} = F_{Beta(1,z)}(W).$$

3 Simulation Studies

3.1 Simulation Design

We performed a simulation study to evaluate the performance of the globalEVT algorithm compared with the ARTP and globalARTP methods, in terms of type I error, power and computational time. To cover these objectives we simulated 36 different scenarios summarized in Table 1.

We simulated balanced case-control datasets with sample size N = 2000 (1000 cases and 1000 controls), one binary response Y indicating disease status and M variables (M = 10, 50, 100) representing the genotypes of a set of SNPs within a gene. The genotypes were simulated assuming independence between the SNPs and also assuming a Linkage Disequilibrium (LD) structure. For this, we used HAPGEN version 2.1.0 [15], and took the M SNPs genotypes within a 700kb region on chromosome 21 with CEU HapMap as the reference panel. A subset of c SNPs were considered to be causal, with c=0 (non causal SNPs), c=5 and c=10. For the independent SNPs scenarios, disease status was generated assuming an odds of risk model as described in [16] with a prevalence equal to 0.2. We examined the effect of the inheritance model. We assumed that the causal SNPs followed an additive model with heterozygote relative risk equal to 1.2 or 1.12 and a recessive inheritance model with minor homozygote relative risk equal to 1.2. For the LD scenarios, disease status was generated using HAPGEN considering an additive inheritance model with heterozygote relative risk equal to 1.2 or 1.12.

Table 1. Characteristics of the simulated scenarios.

Total number of SNPs: *M=10, 50, 100*				
Disease Model	**Relative Risk**	**N cSNPs**	**SI**	**SLD**
--	--	*c*=0	1:3	4:6
Additive	*RR(Aa|AA)=1.2, RR(aa|AA)=1.4*	*c*=5	7:9	13:15
		c=10	10:12	16:18
Additive	*RR(Aa|AA)=1.12, RR(aa|AA)=1.25*	*c*=5	19:21	25:27
		c=10	22:24	28:30
Recessive	*RR(aa|AA)=1.2*	*c*=5	31:33	
		c=10	34:36	

N cSNPs: Number of causal SNPs; SI: Scenarios with independent SNPs; SLD: Scenarios with SNPs in LD

Gene set p-values and effective computational times were computed for each scenario, setting the truncation point value equal to K = 5. For the permutational procedures, ARTP and globalARTP algorithms, the total number of permutations was B = 1,000. We repeated the process a hundred times for each scenario. As a summary result, we provided the empirical type I error, the power of the tests as the percentage of significant results at a nominal significance level (gene p-value < 0.05) and the mean effective computational time for each methodology based on the different gene sizes. The simulation procedure was executed using a Linux platform Centos 5 (64-bit) 24 x Intel Xeon CPU with 2.4GHz and 64 Gb of RAM Memory, and using version 3.1.0 of R software.

3.2 Simulation Results

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

Table 2 provides the size or type I error of the tests (scenarios from 1 to 6). All considered methods controls type I error around the specified significance level.

Tables 3 and 4 provide the power of the tests, that is, the percentage of significant results when the causal SNPs follow an additive inheritance model. In Table 3 we considered a stronger effect of each causal SNP with $RR = P(Y = 1|Aa)/P(Y$

$= 1|AA) = 1.2$ and $RR^2 = P(Y = 1|aa)/P(Y = 1|AA) = 1.44$ (scenarios from 7 to 18) while in table 4 we considered a weaker effect of $RR = P(Y = 1|Aa)/P(Y = 1|AA) = 1.12$ and $RR^2 = P(Y = 1|aa)/P(Y = 1|AA) = 1.25$ (scenarios from 19 to 30). Additionally, Table 5 provides the power of the tests when causal SNPs follow a recessive inheritance model with $RR = P(Y = 1|aa)/P(Y = 1|AA \cup Aa) = 1.2$ (scenarios from 31 to 36).

Table 2. Type I error of the tests.

		Scenarios independent SNPs			Scenarios SNPs in LD		
		M=10	*M=50*	*M=100*	*M=10*	*M=50*	*M=100*
globalEVT	*c=0*	5%	3%	4%	4%	6%	3%
globalARTP	*c=0*	6%	5%	1%	5%	7%	1%
ARTP	*c=0*	3%	6%	1%	7%	5%	3%

As it was to be expected, the strongest the effect, the larger the power of the test. Indeed, in table 3 all methods reach a power near 100%.

Table 3. Power of the tests with $RR = P(Y = 1|Aa)/P(Y = 1|AA) = 1.2$ and $RR^2 = P(Y = 1|aa)/P(Y = 1|AA) = 1.44$

	Scenarios independent SNPs						Scenarios SNPs in LD					
	c=5			*c=10*			*c=5*			*c=10*		
	M=10	*M=50*	*M=100*	*M=10*	*M=50*	*M=100*	*M=10*	*M=50*	*M=100*	*M=10*	*M=50*	*M=100*
globalEVT	100%	85%	89%	100%	99%	97%	100%	98%	97%	100%	100%	100%
globalARTP	100%	91%	94%	100%	100%	100%	100%	100%	98%	100%	100%	100%
ARTP	100%	94%	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%

When the individual marginal effects are not so strong (Table 4) the power depends mainly on the number of causal SNPs (larger powers for c=10 than for c=5) and on the number of non-causal SNPs: The larger the number of noncausal SNPs (M-c), the lower the power of the tests.

We compare the performances of the global methods

Table 4. Power of the tests with $RR = P(Y = 1|Aa)/P(Y = 1|AA) = 1.12$ and $RR^2 = P(Y = 1|aa)/P(Y = 1|AA) = 1.25$

	Scenarios independent SNPs						Scenarios SNPs in LD					
	c=5			*c=10*			*c=5*			*c=10*		
	M=10	*M=50*	*M=100*	*M=10*	*M=50*	*M=100*	*M=10*	*M=50*	*M=100*	*M=10*	*M=50*	*M=100*
globalEVT	61%	33%	28%	87%	45%	42%	67%	58%	56%	88%	79%	79%
globalARTP	63%	33%	25%	92%	58%	50%	69%	59%	51%	88%	82%	83%
ARTP	79%	43%	37%	91%	75%	62%	73%	67%	65%	93%	88%	83%

(globalEVT and globalARTP) which allow different inheritance models, and the standard ARTP method that assumes the additive model for all SNPs. In the scenarios where data was generated under an additive model (Table 3 and Table 4), the power of the global methods is very similar and the power of the ARTP method is slightly larger. However, when the recessive model was used for simulations (Table 5), the GlobalEVT and globalARTP are clearly more powerful than the ARTP algorithm. These results were to be expected since the globalEVT and globalARTP algorithms take account of the best inheritance model for each SNP while, as mentioned before, the ARTP method only considers the additive model.

Hence, the results of this simulation study suggest that globalEVT and globalARTP algorithms have a similar performance and clearly outperform the ARTP method which is strongly penalized by the inheritance model.

Table 6 provides the computational times for the three different methods. Notice that we based the computational procedure on only B = 1,000 permutations, although at least 10^7 would be required in GWAS. Larger values than B = 1,000 were unfeasible for this simulation study due to the computational time required for the permutational approaches, ARTP and globalARTP algorithms. From this results we see that although globalEVT and globalARTP approaches have a similar power to detect significant genes, globalARTP takes around 10 times more computational time than globalEVT.

These results reinforce that the globalEVT method has a good performance taking into account not only independent structures, but also the LD structures between SNPs.

Table 5. Power of the tests with $RR = P(Y = 1|aa)/P(Y = 1|AA \cup Aa) = 1.2$

	Scenarios with independent SNPs					
	c = 5			*c = 10*		
	M=10	*M=50*	*M=100*	*M=10*	*M=50*	*M=100*
globalEVT	45%	22%	23%	89%	31%	35%
globalARTP	16%	26%	20%	91%	48%	40%
ARTP	14%	16%	14%	45%	19%	27%

Table 6. Mean computational time.

	M=10	*M=50*	*M=100*
globalEVT	2min 2sec	6min 29sec	11min 38sec
globalARTP*	19min 15sec	54min 49sec	1h 31min 57sec
ARTP*	5min 22sec	17min 46sec	22min 24sec

**considering 1,000 permutations.*

We can conclude that in all scenarios, globalEVT decreases manifestly the computational time required to compute gene set p-values while keeping the statistical power to assess gene set associations. These clearly improve the permutational gene set methods. The data sets supporting the results of these simulation studies are included within the article.

4 APPLICATION ON ATTENTION-DEFICIT/HYPERACTIVITY DISORDER

Attention-deficit/hyperactivity disorder (ADHD) is an important and common childhood disorder characterized by three important neurological aspects, hyperactivity, impulsivity and inattention, which affects children and can continue through adolescence and adulthood [17]. Results from recognized international GWAS studies suggest several genes that may be associated with the development of this disorder [18]. However, the polygenetic characterization of ADHD is still incompletely understood. We proposed the application of the globalEVT algorithm in order to improve the available information of gene set effects.

4.1 BREATHE project

The BRrain dEvelopment and Air pollution ultrafine particles in school childrEn (BREATHE) project is a longitudinal study conducted from January 2012 to March 2013 in 39 schools in Barcelona (Catalonia, Spain) to study the association between air pollution and cognitive development of school children [19]. From the total of 2,904 children who participated in BREATHE, a subsample consisting of 1,648 (154 cases and 1,494 controls) children aged 7 to 10 years was selected for the present study (Table 7) based on the available genetic and neurobehavioral information.

Table 7. Descriptive characteristics for the variables of the study.

	ADHD cases	*controls*
Total, *n*	154 (9.3%)	1,494 (90.7%)
Gender:		
Boys, *n (%)*	111 (72.1%)	751 (50.3%)
Girls, *n (%)*	43 (27.9%)	743 (49.7%)
Age, mean (s.d.: years)	9.27 (0.98)	9.22 (0.86)

Main Outcome: child ADHD

The ADHD outcome was collected using the ADHD criteria of Diagnostic and Statistical Manual of Mental Disorders, fourth edition [20] and was dichotomized as 0 (*ADHD symptom absent*), and 1 (*ADHD symptom present*) as described in [21].

Genomic sample

Genome-wide genotyping was performed using the HumanCore BeadChip WG-330-1101 (Illumina) at the Spanish National Genotyping Center (CEGEN). A total of 298,930 SNPs coded in b37 and positive strand were genotyped. PLINK was used for the data quality control following [22]. The quality control criteria excluded 58,827 SNPs based on Hardy Weinberg Equilibrium (p < 10-e06), minor allele frequency (<1%) and call rate information (95%). The final genotyped data set consisted of 240,103 SNPs within 14,662 different genes.

4.2 Statistical Analysis

Single-SNP analysis

A logistic marginal regression analysis adjusting by gender and age was performed, resulting in 6 significant SNPs considering a suggestive level of significance (P < e-05) (Figure 1, Table 8). However, none of them were significant after multiple testing correction using the False Discovery Rate procedure at a 5% FDR level **[**23**].**

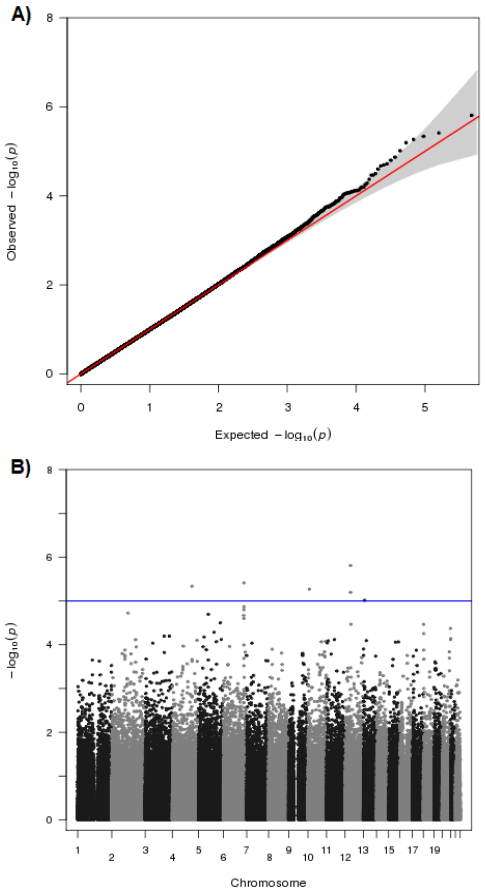

Figure 1. A) Quantile-quantile plot comparing empirical (−log10) p-values against those expected under the null p-value distribution. B) Manhattan plot for the significance of the marginal results.

Table 8. Marginal results at suggestive significant level (P<e-05) for ADHD.

SNP	CHR	POS	E_ALL	EAF	BETA	SE	P	P_ADJ	GENE
rs11564252	*12*	40,813,733	*T*	0.816	-0.75	0.156	1.55e-06	0.306	*MUC19*
rs504985	*6*	149,658,978	*T*	0.644	0.568	0.123	3.87e-06	0.306	*TAB2*
rs1425533	*4*	143,421,175	*G*	0.811	-0.704	0.154	4.61e-06	0.306	*INPP4B*
rs7090158	*10*	10,136,768	*G*	0.831	-0.74	0.163	5.40e-06	0.306	*Intergenic*
rs2098963	*12*	40,780,270	*G*	0.816	-0.695	0.154	6.36e-06	0.306	*Intergenic*
rs2481962	*13*	28,531,385	*T*	0.78	-0.63	0.142	9.69e-06	0.388	*Intergenic*

SNP, single nucleotide polymorphism; CHR, chromosome; POS, position; E_ALL, effect allele; EAF, effect allele frequency; BETA, regression coefficient; SE, standard error; P, unadjusted p-value; P_ADJ, adjusted p-value; GENE, gene.

GSA Analysis

To explore associations at a gene set level, we mapped all SNPs from BREATHE project using the information provided by the HumanCore BeadChip WG-330-1101 of Illumina, and then, we grouped all different SNPs by gene.

Table S1 shows the 21 significant genes found using globalEVT: the first two columns provide the name of the genes that are statistically significant at a 5% FDR level and the chromosome. The next two columns show the original p-value provided by globalEVT and the adjusted p-value after multiple comparison correction. Last column provide the

description of the gene. These results suggest an interesting relationship with several neurological disorders genetically close to the ADHD, as *HTN1* for Autism [24], *EPHX2* for Cerebrovascular function [25] and *TRAPPC8* for congenital intellectual disability [26]. Moreover, some regions of *GAPVD1*, *TRAPPC8*, *CMC2*, *FAM168A*, *C6*, *C11orf30* and *SLU7* have been previously reported as significant in the Genetic online Database for ADHD available at http://adhd.psych.ac.cn/index.do. Nonetheless, the most important result was found in a more complex biological level. Given the list of significant genes obtained from globalEVT (Table S1), we assessed the induced functional network by application of ConsensusPathDB [27-28]. ConsensusPathDB interconnects globalEVT significant genes through different types of biological interactions. This search identified three important biochemical reactions between a set of Cyclophilin A like domain proteins [Figure 2], which were revealed as an important regulator of the Ubiquitin-proteasome system and ADHD development mechanisms [29]. Hence, results suggest that the significant genes obtained from globalEVT may encode a relevant functional protein complex that has a high influence in ADHD.

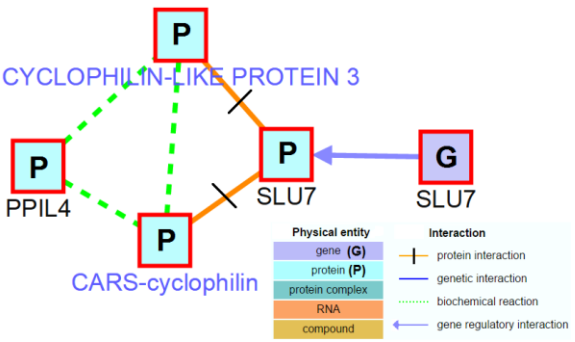

Figure 2. Functional induced network for significant gene sets from globalEVT.

These results together with the computational efficiency confirm that globalEVT is able to analyze GWAS providing genes which may play an important role in the mechanisms of the development of complex diseases, while the considered permutational GSA procedures (globalARTP and ARTP) are unfeasible.

5 Conclusions

We proposed a new algorithm in the context of GSA, the globalEVT, that reduces dramatically the computational time and the requirement of a large sample size with respect to the other GSA methods. The new approach improves power by allowing different inheritance models for each genetic variant as illustrated in the simulation study performed and also, it allows the existence of correlation between the SNPs computing the total number of effective tests based on the idea of [11]. For illustrative purposes, we applied our proposed algorithm in a clinical context of ADHD. While marginal results are not conclusive and GSA-permutational procedures are not feasible to compute, the proposed globalEVT method improves the efficiency and allows identifying significant signals of association at a gene-level. Hence, the application to ADHD study proved that the use of globalEVT in GWAS is feasible while permutation GSA methods are not. In addition, using the set of causal genes obtained from globalEVT for the ADHD study, we obtained a functional network configuration that reinforce the performance of our proposed approach. They reveal a strong relationship between some genes and several neuronal disorders suggesting new biological mechanisms linked to childhood-ADHD development which have not been described yet. However, it is important to take into account that the proposed method, and in general, the existing GSA approaches were designed to detect nominal effects and, for that, the main limitation is that its use is not appropriate when the genetic association is due to epistasis and not to marginal effects.

The proposed algorithm is implemented in the R function globalEVT within the globalGSA package, available at CRAN (http://cran.r-project.org/web/packages/globalGSA/index.html).

Acknowledgment

Natalia Vilor-Tejedor is funded by a pre-doctoral grant from the Agència de Gestió d'Ajuts Universitaris i de Recerca (2015 FI_B 00636), Generalitat de Catalunya.

This research was also supported by grants MTM2011-26515 and MTM2012-38067-C02-02 from the Ministerio de Economía e Innovación (Spain) and the European Research Council under the ERC Grant Agreement number 268479.

References

[1] Morton NE. Genetic epidemiology. Ann Hum Genet. 1997; 61(Pt 1):1-13.

[2] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009; 8;461(7265):747-53. doi: 10.1038/nature08494.

[3] Fisher RA. Statistical methods for research workers. Oliver and Boyd,

London. 1932. ISBN 005-002170-2.

[4] Stouffer SA, Suchman EA, DeVinney LC, et al. The American Soldier. Vol. 1. Adjustment During Army Life. 1949. Princeton Univ. Press, Princeton.

[5] Wilkinson B. A statistical consideration in psychological research. Psychological Bull. 1951; 48:156–158.

[6] Zaykin DV, Zhivotovsky LA, Westfall PH and Weir BS. Truncated product method for combining P-values. Genet Epidemiol. 2002; 22:170-185.

[7] Dudbridge F and Koeleman BPC. Rank truncated product of P values, with application to genomewide association scans. Genet Epidemiol. 2003; 25:360-366.

[8] Yu K, Li Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, Kraft P, Chatterjee N. Pathway analysis by adaptive combination of Pvalues. Genet. Epidemiol. 2009; 33:700[9].

[9] Chen HS, Pfeiffer RM and Zhang S. A powerful method for combining P-Values in genomic studies. Genet Epidemiol. 2013; 37(8):814-9.

[10] Vilor-Tejedor N and Calle ML. Global adaptive rank truncated product method for gene-set analysis in association studies. Biom. J. 2014; 56: 901–911.

[11] Li MX, Yeung JM, Cherny SS and Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum Genet. 2012; 131(5):747-56. doi: 10.1007/s00439-011-1118-2.

[12] Sánchez-Mora C, Ramos-Quiroga JA, Bosch R, Corrales M, Garcia-Martínez I, Nogueira M, Pagerols M, Palomar G, Richarte V, Vidal R, Arias-Vasquez A, Bustamante M, Forns J, Gross-Lesch S, Guxens M, Hinney A, Hoogman M, Jacob C, Jacobsen KK, Kan CC, Kiemeney L, Kittel-Schneider S, Klein M, Onnink M, Rivero O, Zayats T, Buitelaar J, Faraone SV, Franke B, Haavik J, Johansson S, Lesch KP, Reif A, Sunyer J, Bayés M, Casas M, Cormand B, Ribasés M. 2014. Case-Control Genome-Wide Association Study of Persistent Attention-Deficit Hyperactivity Disorder Identifies FBXO33 as a Novel Susceptibility Gene for the Disorder. Neuropsychopharmacology. 2015; 40(4):915-26. doi: 10.1038/npp.2014.267.

[13] Dudbridge F and Koeleman BPC. Efficient Computation of Significance Levels for Multiple Associations in Large Studies of Correlated Data, Including Genome wide Association Studies. The Am. Journ. of Hum. Gen. 2004; Volume 75, Issue 3, 424-435, 1.

[14] Gonzalez JR, Carrasco JL, Dudbridge F, Armengol L, Estivill X, Moreno V. Maximizing association statistics over genetic models. Gen. Epidemiology. 2008; Volume 32, Issue 3, 246-254.

[15] Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. Bioinformatics. 2011; 27(16), 2304–2305. doi:10.1093/bioinformatics/btr341

[16] Wray NR and Goddard ME. Multi-locus models of genetic risk of disease. Genome Medicine. 2010; 2;2(2):10.

[17] Ramos-Quiroga JA, Picado M, Mallorquí-Bagué N, Vilarroya O, Palomar G, Richarte V, Vidal R, Casas M. The neuroanatomy of attention deficit hyperativity disorder in adults: Structural and functional neuroimaging findings. revue Neurologique. 2013; 56, S93-S106.

[18] Zhang, L., Chang, S., Li, Z., Zhang, K., Du, Y., Ott, J., & Wang, J. ADHDgene: a genetic database for attention deficit hyperactivity disorder. Nucleic Acids Research. 2012; 40(Database issue), D1003–D1009. doi:10.1093/nar/gkr992

[19] Sunyer J, Esnaola M, Alvarez-Pedrerol M, Forns J, Rivas I, López-Vicente M, Suades González E, Foraster M, Garcia-Esteban R, Basagaña X, Viana M, Cirach M, Moreno T, Alastuey A, Sebastian-Galles N, Nieuwenhuijsen M, Querol X. Association between Traffic-Related Air Pollution in Schools and Cognitive Development in Primary School Children: A Prospective Cohort Study. PLoS Med. 2015; 12(3):e1001792. doi: 10.1371/journal.pmed.1001792.

[20] American Psychiatric Association. Manual diagnóstico y estadístico de los trastornos mentales. 2002. Barcelona, Spain: Masson.

[21] Forns J, Esnaola M, López-Vicente M, Suades-González E, Alvarez-Pedrerol M, Julvez J, Grellier J, Sebastián-Gallés N, Sunyer J. The n-back Test and the Attentional Network Task as Measures of Child Neuropsychological Development in Epidemiological Studies. Neuropsychology. 2014; 28(4):519-29.

[22] 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; Volume 81, Issue 3 , 559 - 575.

[23] Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: a Practical Approach to Multiple Testing. J.R.Statist. Soc B. 1995; 57, No 1, pp. 289-300

[24] Castagnola M, Messana I, Inzitari R, Fanali C, Cabras T, Morelli A, Pecoraro AM, Neri G, Torrioli MG, Gurrieri F. Hypo-phosphorylation of salivary peptidome as a clue to the molecular pathogenesis of autism spectrum disorders. J Proteome Res. 2008; 7(12):5327-32.

[25] Zhang W, Davis CM, Edin ML, Lee CR, Zeldin DC, Alkayed NJ. Role of endothelial soluble epoxide hydrolase in cerebrovascular function and ischemic injury. PLoS One. 2013; 9;8(4):e61244.

[26] Zong M, Wu XG, Chan CW, Choi MY, Chan HC, Tanner JA, Yu S. The adaptor function of TRAPPC2 in mammalian TRAPPs explains TRAPPC2-associated SEDT and TRAPPC9-associated congenital intellectual disability. PLoS One. 2011; 6(8):e23350.

[27] Kamburov A, Wierling C, Lehrach H and Herwig R. ConsensusPathDB–a database for integrating human functional interaction networks. Nucleic Acids Res. 2009; 37(Database issue):D623-8. doi: 10.1093/nar/gkn698.

[28] Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R. ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Res. 2011; 39(Database issue):D712-7. doi: 10.1093/nar/gkq1156.

[29] Bousman CA, Chana G, Glatt SJ, Chandler SD, Lucero GR, Tatro E, May T, Lohr JB, Kremen WS, Tsuang MT, Everall IP. Preliminary evidence of ubiquitin proteasome system dysregulation in schizophrenia and bipolar disorder: convergent pathway analysis findings from two independent samples. Am J Med Genet B Neuropsychiatr Genet. 2010; 153B(2):494-502.