Universitat de Vic
Escola Politècnica Superior

Final Master Thesis

Gene Set Analysis for improving genetic association studies

Natàlia Vilor Tejedor

Master of Science program in Omics Data Analysis
Direction: M.Luz Calle Rosingana
Department of Systems Biology, EPS, University of Vic.
January 20th, 2014 Barcelona

Universitat de Vic
Escola Politècnica Superior

Summary of the Final Master Thesis

Master of Science program in Omics Data Analysis

Title: Gene-Set Analysis for improving genetic association studies

Keywords: *Alzheimer disease, Adaptive Rank Truncated Product, Beta Distribution, Extreme Value Theory, Genetic Association Studies, Gene set Analysis, globalTest, globalARTP, globalEVT, Permutational procedure, Reelin signal pathway, single-SNP analysis.*

Author: Natàlia Vilor Tejedor

Direction: M.Luz Calle Rosingana

Date: January 20th, 2014

Abstract

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. The usual statistical approach in genetic epidemiology study is a marginal analysis, where each SNP is analyzed separately for association with the phenotype. **Motivation**. It has been observed, that for common diseases the single-SNP analysis is not very powerful for detecting genetic causing variants. In this work, we consider Gene Set Analysis (GSA) as an alternative to standard marginal association approaches. 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. **Objective.** We present a new optimized implementation of a pair of gene set analysis methodologies for analyze the individual evidence of SNPs in biological pathways. We perform a simulation study for exploring the power of the proposed methodologies in a set of scenarios with different number of causal SNPs under different effect sizes. In addition, we compare the results with the usual single-SNP analysis method. Moreover, we show the advantage of using the proposed gene set approaches in the context of an Alzheimer disease case-control study where we explore the Reelin signal pathway.

Contents

Chapter 1

Introduction and Goals

Genetic epidemiology is focused on the identification of genetic variants that determine health and disease in populations, and also, in the study of how the genetic variants interact with environmental factors. 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. The usual statistical approach in this kind of study is a marginal analysis, where each SNP is analyzed separately for association with the phenotype. We will refer to this as single-SNP analysis. When the number of SNPs to be analyzed is very large, as in Genome Wide Association Studies (GWAS), the multiple testing corrections that are required reduce dramatically the power of the single-SNP strategy.

An alternative to single-SNP analysis is gene-set analysis (GSA) where the joint effect of a set of M SNPs is measured. The set of SNPs that are jointly analyzed may have a biological relationship, for instance, we may test for the joint effect of SNPs within a gene or the joint effect of SNPs within a pathway. Thus, GSA provides a combined association evidence of a set of SNPs (a gene-p-value or a

pathway-p-value) which is meaningful and could be more powerful than single-SNP analysis when the individual effects are small.

Our starting point is the Adaptive Rank Truncated Product method (ARTP) proposed by *Yu et. al, (2009)*[22]. This GSA method consists on the combination of the K smallest marginal p-values, 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 the SNPs in the set (usually, the additive model). But, the most important limitation is computational since the final gene-set p-value relies on the nonparametric null distribution of the ARTP test statistic which is estimated using permutational procedures. The main objective of this work consists in improve these two important limitations.

Summarizing, the scientific archivements of this scientific proposal are the following:

- We propose two alternative algorithms that improve the original ARTP method [*see Chapter* 3]:

  GSA-globalARTP: This method allows different modes of inheritance for each SNPs in the set (max-statistic) using the same permutational proceduce as in ARTP method, improving the first limitation.

  GSA-globalEVT: This method reduces the computational requirements fitting the ARTP statistic using the extreme value theory (EVT), also allowing max-statistic test, improving both limitations mentioned.

- Moreover, we implement the proposed theorical algorithms into a R code package (`globalGSA`[1]) [*see Annex*].

[1] Available at: http://cran.r-project.org/web/packages/globalGSA/index.html

- In addition, we perform a simulation study [*see Chapter* 4] to compare the statistical power and the computational time among the proposed GSA methods including also, the comparison with the results of single-SNP analysis correcting for multiple testing using Benjamini-Hochberg method (*Benjamini-Hochberg, 1995*[3]).

- Finally, we apply these methodologies in the context of Alzheimer disease [*see Chapter* 5] using the public GWAS data of *Reiman et al., 2007*[14] study.

Chapter 2

General concepts of genetic association studies

The objective of Genetic Association Studies is to identify genetic variants that explains the phenotype variability, and concretely, that modifies the risk of disease. The most common genetic variation in the population is called single nucleotide polymorphism (SNP) and the chromosomal location often called a *locus*. SNPs are genetic variations in a DNA sequence that occurs when a single position in a genome is altered. Most SNPs are biallelic polymorphisms, and it means that two possible variants (alleles) are observed in the population at that specific locus. In the majority of scenarios that we will consider, the marker locus has only two distinct alleles, e.g., alleles A and a. Denoting 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), and taking into account that humans are diploid (each cell contains two copies of the genome) each SNP locus can have three possible genotypes: AA for major homozygous, Aa for heterozygous and aa for minor homozygous.

2.1 Statistical genetic principles

Genomic Association studies are typically case–control designs where we consider some individuals that are genotyped to detect nonrandom occurences between each genotype frequency related to the two different stages of the disease. In this context, we distinguish between cases and controls individuals. Such, binary traits can be coded by Y, where $Y = 1$ denotes cases and $Y = 0$ denotes controls, and the penetrance function (*see Equation* 2.1) represents probabilities for each considered genotype G,

$$P(Y = 1 \mid G) + P(Y = 0 \mid G) = 0. \tag{2.1}$$

In a statistical context, SNPs are expressed like categorical variables that can always be coded in the form of numerical or indicator variables. Different codifications of the genotypes correspond to different modes of inheritance as is summarized in Table 2.1. In the dominant model, a single copy of the variant allele

Table 2.1: 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

is sufficient to modify (increase or decrease) the risk of disease,

$$Pr(Y = 1 \mid G = Aa) = Pr(Y = 1 \mid G = aa). \tag{2.2}$$

In contrast, in the recessive model two copies of the variant allele are necessary to

modify the risk,

$$Pr(Y = 1 \mid G = Aa) = Pr(Y = 1 \mid G = AA). \tag{2.3}$$

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 where the three genotypes have different effects on disease risk,

$$Pr(Y = 1 \mid G = AA) \neq Pr(Y = 1 \mid G = Aa) \neq Pr(Y = 1 \mid G = aa). \tag{2.4}$$

2.2 Statistical approaches for disease risk prediction

The usual strategy for considering disease models in Genomic Association Analysis is marginal variable selection, defined in our work as single-SNP analysis. It tests genetic association of individual SNPs and identifies only the most significant subset that captures the majority of the information of genotype-phenotype association.

As we have described, for the human genetic setting, the genotype at a given SNP has three levels: homozygous wildtype, heterozygous, and homozygous rare. Considering a binary outcome, the data can be represented by the 2×3 contingency table, and in this setting, a commonly used measure of association is the odds ratio (OR) defined as the ratio of the odds of disease given a specific genotype to the odds of disease among individuals without the specified genotype. Hence, each locus is evaluated individually for its marginal association with disease performing a marginal chi-square test where the genotypes with a p-value below a specified threshold are included in the prediction model.

Alternatively, a logistic regression model as in (*see Equation* 2.5) can be fitted where G is the codification of the SNP as specified in Table 2.1, and Z represents other non genetic covariates, where $\pi = Pr(Y = 1 \mid G,\ Z)$

$$\text{logit}(\pi) = \beta_0 + \beta_1 G + \delta Z \tag{2.5}$$

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

However, it has been observed that the single-SNP analysis is not very powerful for detecting genetic causing variants of common diseases; concretely most causal SNPs effects are not detectable with the common single-SNP testing procedure followed by correction for multiple comparisons, because the identified SNPs represent only a small fraction of the genetic variants contributing to diseases under study, and the majority of them represent statistical noise. Perhaps, we need to set other focus of interest taking into account other forms of genomic modifications. These drawbacks raise the possibility that genetic variants with a small individual effects can have more jointly significant genetic impact.

Chapter 3

Gene Set Analysis

We consider Gene-set analysis to try to solve common limitations of single-SNP analysis. 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 obtain a more accurate measurement of association of a set of genetic variants with a phenotype, and also provides the potential to detect combined effects of SNPs in a gene or a pathway that might be missed doing a marginal single analysis. 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.

3.1 Methods

In this work we consider two different approaches in order to combine the statistical information obtained from the single-SNP tests starting from the idea of the Adaptative Rank Truncated Product method (*Yu, et al. 2009*[22]).

3.1.1 Adaptive rank truncation product method

Adaptive Rank Truncated Product method (ARTP), is a GSA method for combining the individual evidence of association over different SNPs within a gene or pathway using the product of the K smallest marginal p-values

$$W_K^{(b)} = \prod_{i=1}^{K_j} p_{(i)}^{(b)},\ 1 \leq j. \tag{3.1}$$

In the initial method RTP (*Zaykin et al., 2002*[23]), the value K was fixed and specified in advance, while in the ARTP method K is obtained in an adaptive way and the gene-p-value is obtained from the permutational null distribution of W_K.

The main goal of this work consists in improve some limitations of the original ARTP algorithm. The first improvement consists in taking into account the inheritance information of genetic variants, because, this an other GSA methods, only consider p-values assuming an additive model. Following this idea, we propose an improvement of the ARTP method by combining the p-values obtained from the max-statistic test. We will refer to this as the globalARTP method.

3.1.2 Combining statistical tests by permutation procedure

The proposed methodology improves the original algorithms by introducing an additional step where a global test for the best mode of inheritance of each SNP is performed. Using the global adaptation it can be determined whether the global pattern of a group of SNPs is significantly related to some phenotype of interest. For easy of explanation we describe the proposed algorithms in the simplest case where all M SNPs belong to the same set (gene). In this case, the algorithms provide a gene-p-value which indicates if variation within the gene is associated

with the phenotype. However, the implemented approach is more general and allows the SNPs in the study to belong to different sets.

Algorithm

Step 1. Best genetic model: In terms of our proposed algorithm, the first step performs an association analysis of each SNP with the phenotype, considering the three different modes of inheritance (dominant, recessive, and additive) taking the minimum of the three p-values, based on the likelihood ratio test. So, this first step provides M p-values, one for each SNP in the gene, that are sorted increasingly:

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(M)}. \tag{3.2}$$

Step 2. Rank truncated product statistic: Given a value $K \leq M$, we would compute the rank truncated product statistic W_k for each candidate truncation point $k \leq K$ as defined in *Equation* 3.3. Indeed, in order to improve computational efficiency and avoid computational problems we will work with the log transformation of W_k, denoted by

$$V_k = \sum_{j=1}^{k} -\log(p_{(j)}), \quad 1 \leq k \leq K. \tag{3.3}$$

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 permutated datasets. From step 1, we obtain and sort the M single-SNP p-values corresponding to the best inheritance mode of each SNP:

$$p^b_{(1)} \leq p^b_{(2)} \leq \ldots \leq p^b_{(M)}, \quad 0 \leq b \leq B, \tag{3.4}$$

where $b = 0$ corresponds to the original dataset. When performing step 2 we obtain the test statistic, $V_k^{(b)},\ 1 \leq k \leq K,\quad 0 \leq b \leq B$. Significance of the original test statistics can be explored by comparing $V_k^{(0)},\ 1 \leq k \leq K$ with $V_k^{(b)},\ 1 \leq k \leq K,\ \ 1 \leq b \leq B$. The following expression provides the permutational p-values for statistics $V_k^{(0)},\ 1 \leq k \leq K$ under the null hypothesis:

$$\hat{S}_k^{(0)} = \frac{\sum_{l=0}^{B} I(V_k^{(l)} \geq V_k^{(0)})}{B+1}, \quad 1 \leq k \leq K. \tag{3.5}$$

In fact, the algorithm requires the computation of the p-values not only for the original statistics, but also for the permuted statistics, which are given by

$$\hat{S}_k^{(b)} = \frac{\sum_{l=0}^{B} I(V_k^{(l)} \geq V_k^{(b)})}{B+1}, \quad 1 \leq k \leq K,\ \ 1 \leq b \leq B. \tag{3.6}$$

Step 4. Best truncated point: An additional step is to optimize the number k of SNPs that are combined for each gene. For this we define $k_{opt}^{(b)},\ 0 \leq b \leq B$ as the number $k \in \{1, \ldots, K\}$ that minimizes $\hat{S}_k^{(b)}$, and this minimum is denoted as $minP^{(b)}$:

$$minP^{(b)} = min_{1 \leq k \leq K} S_k^{(b)},\ \ 0 \leq b \leq B. \tag{3.7}$$

Step 5. Gene-p-value: Finally, we estimate the gene-p-value by comparing the original dataset $minP^{(0)}$ with the permuted datasets $minP^{(b)}$, $1 \leq b \leq B$:

$$gene - p - value = \frac{\sum_{l=0}^{B} I(minP^{(l)} \leq minP^{(0)})}{B+1},\ \ 0 \leq b \leq B. \tag{3.8}$$

Still, an important limitation of both, the ARTP and the globalARTP methods, is computational. Both rely on permutational procedures for estimating the non parametric null distribution of the test statistic. *Dudbridge et. al., (2004)*

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 tales of the distribution are estimated more accurately.

Hence, we also propose an alternative algorithm, referred as globalEVT, for estimating the null distribution of the ARTP statistic using the extreme-value theory (EVT). This proposed method reduces importantly the computational requirements since only one-parameter distributions are to be fitted. In addition, we improve the statistical power of the globalEVT approach allowing different modes of inheritance for each SNP in the set by using the Max-statistic test (*Gonzalez et al., 2008*[8]) as in the previous proposed algorithm.

3.1.3 Combining statistical tests using Extreme-value theory

Considering the same notation as in the previous method, the proposed algorithm is based on the following result:

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

$$f_{U_{(l)}}(u) = \frac{M!}{(l-1)!(M-l)!} u^{l-1}(1-u)^{u-l} \tag{3.9}$$

Assumption: We will also assume that when independence does not hold, that is, when $U_1, U_2, \cdots, U_M$ are dependent variables with standard Uniform distribution, it is possible to find a number $m^* < M$ so that the distribution of the lth 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:

Algorithm

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. So, this first step provides M p-values, one for each SNP in the gene:

$$p_j^{min} = min\{p_j^{dom}, p_j^{rec}, p_j^{add}\},\ j = 1, \ldots, M, \tag{3.10}$$

where p_j^{dom}, p_j^{rec} and p_j^{add} are the p-values of j-SNP assuming a dominant, a recessive and an additive model respectively. If the three test were independent the distribution of p^{min} would follow a $Beta(1, 3)$ distribution (*see Proposition* 1 with $l = 1$ and $M = 3$), but, since the three tests are performed on the same SNP, the three p-values 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 (*Sladek et al., 2007*[17]). We transform $p_j^{min}, j = 1, \ldots, M$ into values from a standard Uniform distribution by applying the inverse distribution function:

$$r_j = F^{-1}_{Beta(1,x=2.2)}(p_j^{min}),\ j = 1, \ldots, M \tag{3.11}$$

Step 2. Summarizing the k most associated SNPs: We sort increasingly the uniformly distributed p-values obtained in step 1, considering the k best results for

$k \in \{1, ..., M\}$. That is, we want to summarize the k first order statistics into a unique statistic,

$$r_{(1)} \leq r_{(2)} \leq \ldots \leq r_{(M)} \tag{3.12}$$

If the SNPs were not correlated, the order statistics, $r_{(j)}$, follows a Beta distribution $Beta(j, M - j + 1), j \in \{1, \ldots, k\}$, but if the SNPs are correlated, the distribution is $Beta(j, y - j + 1), j \in \{1, \ldots, k\}$, where y is the effective number of tests. We estimate y through a permutational approach. As in the previous step, we transform the order statistics $r_{(j)}, j = 1, \ldots, k$ into values from a standard Uniform distribution by applying their inverse distribution function:

$$t_j = F^{-1}_{Beta(j,y-j+1)}(r_{(j)}),\ j = 1, \ldots, k \tag{3.13}$$

As a summary statistic we take:

$$S_k = -2 \sum_{j=1}^{k} \log t_j. \tag{3.14}$$

As in Fisher's method (*Fisher, 1925*[7]), since t_j are uniformly distributed, then $-2 \log t_j$ follow 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 chi-squared distribution with ν degrees of freedom where ν should be estimated through a permutational approach. We transform the sum statistic S_k into a uniformly distributed value by applying its inverse distribution function:

$$U_k = F^{-1}_{Chi(\nu)}(S_k) \tag{3.15}$$

Step 3. Adaptive step: selection of the best truncation point: We repeat Step 2 for every k from 1 to K, where K is a specified truncation parameter. As a final statistic gene set statistic we take the best of all:

$$W = min\{U_1, \ldots, U_K\}. \tag{3.16}$$

If the values U_k were independent and identically distributed (i.i.d), W follows a Beta distribution $Beta(1, K)$ but the U_k are correlated because they are calculated as a cumulative sum of values, thus, it is necessary to approximate its distribution by $Beta(1, z)$, where z is the effective number of tests and is estimated using a permutational procedure. Finally, the transformation of W to a uniformly distributed valued provides the adjusted p-value for the set of M SNPs:

$$\text{geneset}p_{adjust} = F^{-1}_{Beta(1,z)}(W) \tag{3.17}$$

Model fitting: The proposed model requires the estimation of three different parameters; y, z, ν. We apply a permutational approach to estimating the first two parameters taking into account the relationship between the mean and the second shape parameter of a Beta distribution. Concretely, we reproduce a hundred permutations of a Beta distribution, $Beta(a, b)$, where the first shape parameter, a, is known. Our purpose is estimate the second shape parameter, b, that is the total number of effective tests (denoted by y and z in each case), as $\hat{b} = \frac{\hat{\mu}-a}{\hat{\mu}}$.

On the other hand, in order to estimate ν, that is defined as the degrees of freedom from a Chi squared fitted distribution, we also reproduce a permutational procedure considering a hundred permutations, taking the mean from the permuted values.

Chapter 4

Simulation study

4.1 Simulation design

We performed a simulation study with the goal of exploring the power and performance of the proposed globalGSA methodology for detecting genes associated with a phenotype. 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). We consider 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 with the same effect size ($RR = 1.2,\ 1.1$), where RR 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 (*Urrea, et al., 2014* [20]).

We explore the size of the test in the case where there is no causal SNP ($c = 0$) and the power of the test for detecting association at the gene-level in the scenarios with $c = 10$ causal SNPs within the gene-set. This produces a total of twelve

scenarios. For each scenario we compute the gene p-value using the globalARTP algorithm allocating the number of permutations to $B = 10,000$ and the truncation point value equal to $K = 10$, and also, we compute the gene-p-value using the globalEVT method considering $B = 100$ and $K = 10$. We repeat this process a thousand of times for each scenario and we averaged the results over the thousand replications providing the percentage of times (P_c) that the gene is significant (gene p-value< 0.05).

Notice that GSA methods and single-SNP analysis are difficult to compare since one is a gene-set approach providing just one p-value of the gene while the single-SNP analysis provides several p-values related to the gene. But the comparison is very important since it will indicate whether the gene-set analysis is more powerful than the standard single-SNP approach or not. With this comparative goal, we perform single-SNP analysis and declar that a gene was significantly associated with the disease when at least one SNP in the gene was significant at the usual 0.05 level after Benjamini and Holchberg multiple testing correction (*Benjamini-Holchberg, 1995*[3]) for the M univariate tests performed in each gene.

4.2 Simulation results

Results are summarized in Tables 4.1 to 4.3. Table 4.1 provides the size of the tests, that is the percentage of significant results when there is no causal SNP. While both globalGSA methodologies control the size around the specified significance level (5%), both, single-SNP results are rather conservative. Tables 4.2 and 4.3 provide the power of the test, that is, the percentage of significant results (P_c) when there are $c = 10$ causal SNPs within the M available SNPs in a gene. In Table 4.2 we can compare the performance of the gene set analysis and single-SNP analysis when the effect of the 10 causal SNPs is relatively high ($RR = 1.2$). When all SNPs in

the gene are causal ($M = c = 10$), all the methods cosidered are very powerful to detect association of the gene. When the number M of SNPs in the gene increases to 50 and 100 the globalARTP methods are still very powerful ($P_c = 100\%$) while we can observe a slight decrease in the power of single-SNP analysis due to the multiple testing correction: $P_c = 96\%$ and $P_c = 93\%$. If we focus on the globalEVT results we can observe that there are similar to those arising from the single-SNP analysis, and lower than the globalARTP results. However, if we consider the computational time, globalEVT becomes more quicker.

In Table 4.3 we can compare the results when the marginal risk effect of each individual causal SNPs is very small ($RR = 1.1$). The gene-based approaches are clearly more powerful than the single-SNP analysis. The increase in power is very evident when all SNPs in the gene are causal. In this case the advantage of globalGSA over the single-SNP analysis is approximately larger than 30% in globalARTP method, and larger than 15% in globalEVT method. However, when the relative risks are so small, the inclusion of noise (null SNPs in the gene when $M = 50$ and $M = 100$) reduces the power of all considered approaches, although globalGSA methods are still above single-SNP analysis results.

In summary, globalGSA methods are more powerful than single-SNP analysis in all different considered scenarios. Furthermore, globalGSA adapted methods reduce the lost of power produced by the multiple testing correction, and allows the incorporation of biological knowledge too. Results obtained by comparing globalGSA methods suggest that the adapted approaches have a similar behavior. On the other hand, if we compare the two different consireded GSA strategies, we can see that globalARTP is slightly above globalEVT as far as association risk detection is concerned. However, in order to obtain an acceptable level of statistical

Table 4.1: Size of the tests

Methodology		$M = 10$	$M = 50$	$M = 100$
globalARTP	$c = 0$	5.3%	4.5%	2.4%
globalEVT	$c = 0$	3.5%	5.1%	2.8%
FDR	$c = 0$	4.1%	1.1%	1.8%

Table 4.2: Power of the tests when $RR = 1.2$

Methodology		$M = 10$	$M = 50$	$M = 100$
globalARTP	$c = 10$	100%	100%	100%
globalEVT	$c = 10$	100%	98.5%	97.1%
FDR	$c = 10$	100%	96.2%	93.4%

Table 4.3: Power of the tests when $RR = 1.1$

Methodology		$M = 10$	$M = 50$	$M = 100$
globalARTP	$c = 10$	79.7%	34.6%	33.2%
globalEVT	$c = 10$	55.7%	34.2%	33.1%
FDR	$c = 10$	42.9%	20.1%	11.8%

significance, permutational procedures require a high number of permutations (at least 10,000 permutations to obtain a significance level of $1e - 04$). Even if we are rigorous, as we are working in a genetic context, the required significance level should be $1e - 07$ needing in this case a total of 10,000,000 permutations, that are very expensive (maybe impossible) to compute. So, if we taking into account the computational time, in order to obtain a suitable level of statistical significance, globalEVT becomes much more efficient.

In conclusion, the GSA proposed methods increases the statistical power in genetic association studies compared with single-SNP analysis. Moreover, the use of extreme-value distribution (EVT) produce a reduction in computation compared with a standard permutation test, and this can be translate to significant time savings.

Chapter 5

Alzheimer disease application

We apply the proposed methodologies to an Alzheimer disease study for determining which genes are associated with Reelin signal, a protein that is thought to be related with an increase risk of Alzheimer disease (*Rice et al., 2001*[15]; *Tissir et al., 2003*[19]). In this context we compare the gene set analysis proposed approaches with the usual single-SNP analysis results.

5.1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder without cure that afflects an increasing part of our ageing population. It was described by Alois Alzheimer in 1906 (*Alzheimer, 1906*[1]) and it is characterized by amyloid plaques, neurofibrillary tangles and loss of synapses (*Berchtold et al., 1998*[4]). Alzheimer's disease is usually diagnosed clinically based on the presence of neurological characteristics and neuropsychological features. However an accurate diagnosis can only be obtained post-mortem when brain material is available and can be examined histologicaly, as is extensively explained in (*Nussbaum et al., 2003*[13].

We still know very little about the etiology of Alzheimer's disease but it is clear that there is a genetic component. Some genes have been associated with AD

as amyloid precursor protein (APP) and presenilins 1 (PSEN1) and 2 (PSEN2) (*Waring et al., 2008*[21]). But, in fact, the best known genetic risk factor is the inheritance of the e4 allele of the apolipoprotein E gene (APOE). It has been demonstrated that between $50\% - 80\%$ of people with Alzheimer disease carry at least one APOE-e4 allele (*Mahley et al., 2006*[12]); *Strittmatter et al., 1993*[18]). APOE occurs in 3 common isoforms (E2, E3, E4) in the human population and APOE-e4 is the primary genetic risk factor for late-onset Alzheimer's disease. This strong genetic association suggest that APOE receptors are very related with the Alzheimer's Disease patogenesis (*Herz et al., 2000*[10]; *Herz et al., 2006*[11]). In addition to these well known genes, other genetic pathways, as the Reelin pathway, are currently investigated for their association with the risk of AD.

5.2 Importance of Reelin in Alzheimer disease

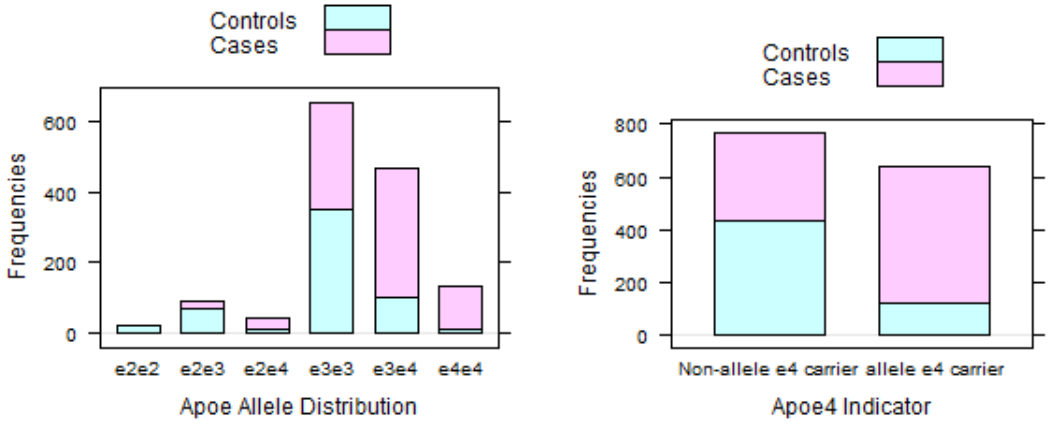
Many studies connect Reelin protein with Alzheimer disease (*Botella-López et al., 2006*[5]; *Baloyannis, 2005*[2]; *Saez-Valero et al., 2003*[16]). Clinical investigations have shown that Reelin plays an eminent role at the most active neurogenesis sites in interaction with APOE protein. According with many studies, Reelin expression is altered in Alzheimer's disease. In the cortex of the patients, Reelin levels were 40% higher compared with controls, but the cerebella levels of the protein remain normal in the same patients. These evidences drives to the hypothesis that an inappropriate activation of Reelin signal can be associated with cellular harm and cellular death.

The objective and motivation of this application is to use our new implented GSA approaches, globalARTP and globalEVT, to derive gene-level association signals of the Reelin pathway with AD and compare these results with single-SNP analysis.

5.3 Descriptive Information

The data for this study is extracted from a public GWAS from *Reiman et al., 2007*[14], reporting 312,316 SNPs in a case-control study with 1411 individuals (861 cases, 550 controls). An exhaustive analysis about the Reelin signal pathway was carried out using Biomart website (http://www.biomart.org), identifying 32 genes in this pathway (682 SNPs). Data information can be consulted on Table 5.1. It contains the gene's name, the chromosome, the strand, the staffed position, the length and the promotor's position. Also, we have information about APOE genotypes. We can observe in Table 5.2 and in Figure 5.1 that 80% of individuals with at least one copy of the e4 allele were affected by the disease.

Figure 5.1: Cases and controls Apoe genotypes distribution.



This percentage increase more than 90% for individuals with two e4 alleles and it decreases until 50%, for non APOE-e4 carriers.

Table 5.1: Genomic data information.

Gene	Chr	nº SNPs	Str	bp1	bp2	length
Abl1	9	27	+	132578987	132752883	173896
Abl2	1	8	-	177335085	177465155	130070
ApoE	19	11	+	50100879	50104489	3610
APP	21	50	-	26174733	26465003	290270
Bdnf	11	6	-	27633016	27700181	67165
CDC42	1	12	+	22235157	22292024	56867
Cdk5	7	4	-	150381832	150385929	4097
CNR1	6	8	-	88906302	88932385	26083
Dab1	1	252	-	57233039	58488763	1255724
Emx2	10	1	+	119291946	119299043	7097
EPHA1	7	3	-	142798331	142816107	17776
Fyn	6	37	-	112089190	112301320	212130
GSK3B	3	7	-	121028238	121295954	267716
itga3	17	10	+	45488488	45522842	34354
LDLR	19	4	+	11061132	11105490	44358
LRP2	2	44	-	169691865	169927368	235503
TP73	1	4	+	3558989	3639716	80727
AKT1	14	3	-	104306734	104333125	26391
PLK2	5	1	-	57785571	57791670	6099
PSEN1	14	4	+	72672908	72756862	83954
PSEN2	1	6	+	225124896	225150429	25533
RAC1	7	5	+	6380651	6410123	29472
Reln	7	83	-	102899473	103417198	517725
Rho	3	3	+	130730172	130736867	6695
RHOA	3	4	-	49371585	49424530	52945
INPP5D	2	18	+	233633433	233781287	147854
Src	20	4	+	35407971	35467867	59896
MAPT	17	31	+	41327624	41461547	133923
VLDLR	9	7	+	2611793	2644485	32692

Table 5.2: Cases and controls Apoe genotypes distribution.

	Controls	Cases
Allele e2e2	85.71	14.29
Allele e2e3	71.43	28.57
Allele e2e4	19.51	**80.49**
Allele e3e3	53.44	46.56
Allele e3e4	21.70	**78.30**
Allele e4e4	5.30	**94.70**
Non-e4 carriers	56.45	43.55
e4 carriers	18.20	**81.80**

5.4 Statistical analysis

5.4.1 single-SNP analysis

In this step we perform a marginal association analysis of each SNP with the phenotype (where we consider $Y = 1$ as an affected individual, and $Y = 0$ as a control individual). Since carriers of APOE-e4 variant have an increased risk of disease, we should consider this in the marginal analysis. Specifically, we define the APOE indicator variable (IndApoe) as the indicator for those individuals carrying at least one copy of APOE-e4. Then, we analyze three different datasets; all individuals adjusting by APOE Indicator variable, non APOE carriers, and APOE carriers. Since the response is dichotomous (status: case/control) we adjust a logistic regression model using the `GWassociation` function from `SNPassoc` R package (*Gonzalez, et al., 2007*[9]). This function provides SNPs' p-values considering different inheritance modes (dominant, codominant, recessive and additive).

With a significance level equal to 1%, we obtain 13 significant SNPs for the adjusted model without multiple testing correction (*see Table* 5.3), 11 significant

SNPs for carriers (*see Table* 5.4) and 20 significant SNPs for non-carriers (*see Table* 5.5). But, if we correct the results using Benjamini-Holchberg method (*Benjamini-Holchberg, 1995*[3]), all the SNPs become non-significant. Hence, in summary, single-SNP analysis is not able to identify any genetic variant in the Reelin pathway that is significantly associated with Alzheimer's disease.

Figure 5.2: Manhattan plots.

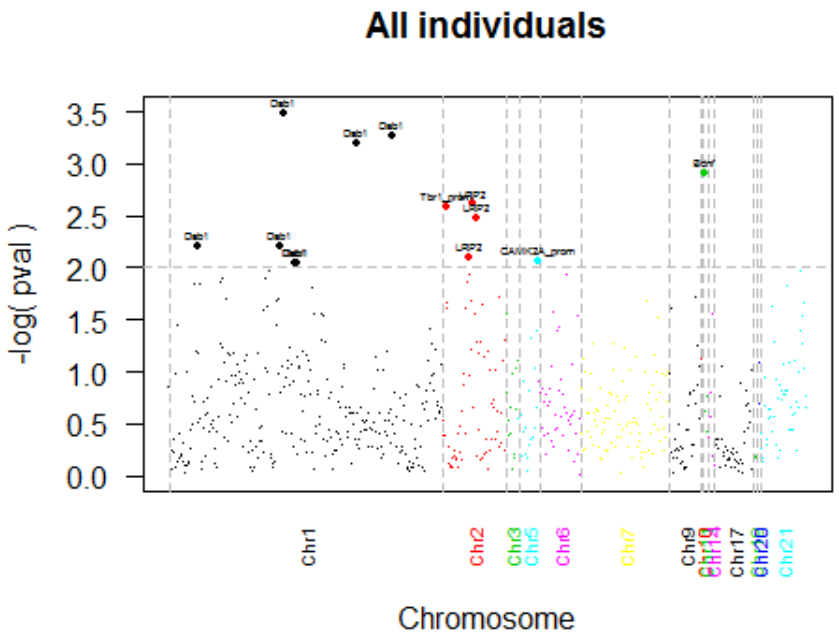


Figure 5.3: Manhattan plots.

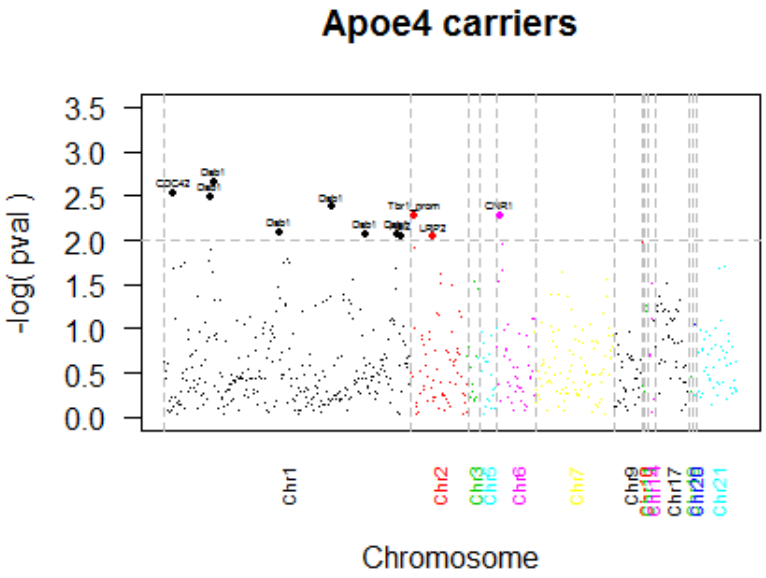


Figure 5.4: Manhattan plots.

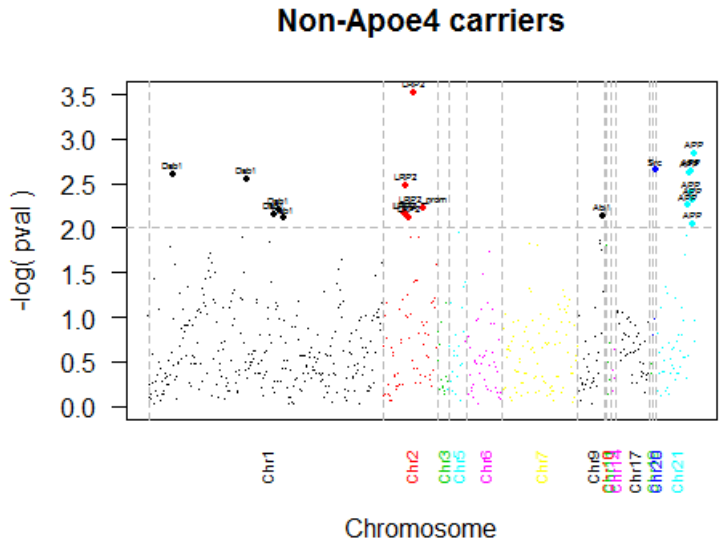


Table 5.3: Significant SNPs without multiple correction all individuals.

	Gene	codominant	dominant	recessive	additive	p_min	p_adjust
rs17416642	DAB1	1.35e-02	6.30e-03	5.30e-01	1.08e-02	6.30e-03	0.387
rs17115767	DAB1	2.32e-02	5.07e-01	6.16e-03	2.18e-01	6.16e-03	0.387
rs10493223	DAB1	3.30e-04	2.21e-01	7.38e-02	5.11e-02	3.30e-04	0.047*
rs1404388	DAB1	9.08e-03				9.08e-03	0.387
rs10493230	DAB1	8.88e-03	1.13e-02	3.38e-01	1.03e-01	8.88e-03	0.387
rs1202774	DAB1	2.17e-03	7.50e-01	6.47e-04	8.23e-01	6.47e-04	0.147
rs4448540	DAB1	5.44e-04				5.44e-04	0.147
rs16845844	PSMD14	1.05e-02	4.22e-01	2.58e-03	1.42e-01	2.58e-03	0.292
rs2193193	LRP2	2.75e-02	2.35e-01	7.88e-03	1.89e-02	7.88e-03	0.387
rs2239594	LRP2	9.28e-03	2.18e-02	7.87e-03	2.44e-03	2.44e-03	0.292
rs830959	LRP2	1.12e-02	5.31e-03	5.19e-02	3.34e-03	3.34e-03	0.324
rs17111118	ARSI	1.94e-02	8.74e-03	7.67e-01	2.42e-02	8.74e-03	0.387
rs11030102	BDNF	1.24e-03	3.22e-01	2.22e-03	6.99e-01	1.24e-03	0.211

Table 5.4: Significant SNPs without multiple correction Apoe4 carriers.

	Gene	codominant	dominant	recessive	additive	p_min	p_adjust
rs10917139	CDC42	6.16e-03	6.95e-01	2.96e-03	6.01e-01	2.96e-03	0.551
rs6680219	DAB1	1.34e-02	3.24e-01	3.32e-03	1.80e-02	3.32e-03	0.551
rs17482980	DAB1	1.05e-02	2.16e-03	1.00	1.05e-02	2.16e-03	0.551
rs10493230	DAB1	8.06e-03	6.94e-02	5.35e-02	5.26e-01	8.06e-03	0.551
rs1202774	DAB1	4.26e-03	1.22e-01	3.13e-02	4.26e-03	4.26e-03	0.551
rs4448540	DAB1	8.51e-03				8.51e-03	0.551
rs6663243	DAB1	3.07e-02	1.20e-02	1.27e-01	8.64e-03	8.64e-03	0.551
rs6668200	ABL2	2.61e-02	9.65e-01	9.07e-03	4.35e-01	9.07e-03	0.551
rs16845844	PSMD14	9.20e-03	2.11e-02	1.17e-02	5.24e-03	5.24e-03	0.551
rs11689553	LRP2	9.21e-03	2.17e-02	1.60e-01	1.54e-01	9.21e-03	0.551
rs7752758	CNR1	1.98e-02	5.83e-03	2.77e-01	5.35e-03	5.35e-03	0.551

Table 5.5: Significant SNPs without multiple correction non-Apoe4 carriers.

	Gene	codominant	dominant	recessive	additive	p_min	p_adjust
rs17416642	DAB1	2.47e-03				2.47e-03	0.274
rs10493223	DAB1	4.85e-03	1.46e-01	2.82e-03	4.85e-03	2.82e-03	0.274
rs17472030	DAB1	7.08e-03	1.34e-02	2.55e-01	1.08e-01	7.08e-03	0.279
rs11207103	DAB1	6.70e-03	6.19e-03	4.96e-01	4.56e-02	6.19e-03	0.279
rs12143653	DAB1	7.77e-03	9.33e-03	3.89e-01	6.95e-02	7.77e-03	0.279
rs2052297	LRP2	2.51e-02	2.78e-01	6.86e-03	2.06e-02	6.86e-03	0.279
rs2193193	LRP2	9.06e-03	6.96e-02	3.36e-03	3.64e-03	3.36e-03	0.279
rs2268370	LRP2	2.51e-02	3.28e-02	2.29e-02	7.03e-03	7.03e-03	0.279
rs2239594	LRP2	2.25e-02	1.28e-01	7.58e-03	1.10e-02	7.58e-03	0.279
rs16856748	LRP2	6.73e-04	3.03e-04	6.98e-01	1.62e-03	3.03e-04	0.206
rs830955	LRP2	2.16e-02	2.06e-02	3.29e-02	5.99e-03	5.99e-03	0.279
rs11792273	ABL1	2.58e-02	9.90e-03	1.99e-01	7.20e-03	7.20e-03	0.279
rs6018100	SRC	7.03e-03	1.00e+00	2.25e-03	3.91e-01	2.25e-03	0.274
rs2830073	APP	1.73e-02	5.59e-03	1.18e-01	7.01e-03	5.59e-03	0.279
rs2830075	APP	8.47e-03	2.65e-03	1.37e-01	2.41e-03	2.41e-03	0.274
rs2830076	APP	9.41e-03	6.42e-03	3.49e-02	2.26e-03	2.26e-03	0.274
rs432766	APP	1.51e-02	8.90e-03	4.05e-02	4.02e-03	4.02e-03	0.279
rs375369	APP	1.72e-02	9.84e-03	5.01e-02	4.79e-03	4.79e-03	0.279
rs2186302	APP	3.12e-02	1.19e-02	1.46e-01	8.91e-03	8.91e-03	0.303
rs436011	APP	5.93e-03	7.00e-03	1.46e-02	1.48e-03	1.48e-03	0.274

5.4.2 Gene Set Analysis

The proposed globalARTP and globalEVT methods, as GSA approaches, estimate the joint effect of all genetic variants in each gene. So, we use the to obtaining significative genes associated with Alzheimer disease. For globalARTP method, we fix $B = 10,000$ permutated data sets and $K = 10$ as the truncation point. For globalEVT we fix $B = 100$ permutations and also $K = 10$ as a truncation point. The results are given in Tables 5.6, 5.7 and 5.8. Table 5.9 provides a summary of the significatively associated genes identified with both methodologies.

Applying globalARTP we obtain that the most significative genes are `Bdnf` and `Tbr1`, for APOE-e4 carriers the only significative gene is `CNR1`, while, for non-APOE-e4 carriers model the most important gens is `Src`. Applying globalEVT we obtain that the most significative genes are `Dab1`, `Bdnf`, `AKT1` and `Cdk5` for all individuals, `Dab1` for carriers model, and `LRP2`, `Src` for non carriers model.

Table 5.9: GlobalGSA Results

	Methodology	
Model	globalARTP	globalEVT
All individuals	`Bdnf`,`Tbr1`	`Dab1`, `Bdnf`, `AKT1`, `CDK5`
Apoe4 carriers	`CNR1`	`Dab1`
Non-Apoe4 carriers	`Src`	`Src`, `LRP2`

In conclusion, we can observe while using the single-SNP analysis we don't find statistical significance after multiple testing correction, our proposed GSA methodologies get to capture some genes that are associated with Alzheimer disease.

Table 5.6: GSA adjusted model results

Gene	globalARTP_pvalue	globalEVT_pvalue
TP73	0.75	0.994
CDC42	0.53	1
ApoER2	0.88	1
Dab1	0.07	2.184e-41
Abl2	0.17	0.991
PSEN2	0.47	0.999
Tbr1	0.04	0.954
LRP2	0.13	0.999
SHIP	0.48	0.999
RHOA	0.21	0.999
GSK3B	0.58	0.999
Rho	0.30	0.661
PIK3R1	0.91	0.999
CAMK2A	0.27	0.997
CNR1	0.62	0.999
Fyn	0.48	0.999
RAC1	0.33	0.932
Reln	0.94	0.999
EPHA1	0.80	0.999
Cdk5	0.28	0.031
VLDLR	0.43	0.993
Abl1	0.44	0.999
Bdnf	0.02	0.008
PSEN1	0.68	0.999
AKT1	0.25	0.039
Tau	0.54	0.999
itga3	0.69	0.999
LDLR	0.64	0.771
Src	0.31	0.999
APP	0.45	0.999

Table 5.7: GSA APOE-e4 carriers model results

Gene	globalARTP_pvalue	globalEVT_pvalue
TP73	0.49	0.999
CDC42	0.09	0.703
ApoER2	0.25	0.999
Dab1	0.51	0.0295
Abl2	0.13	0.999
PSEN2	0.61	0.999
Tbr1	0.08	0.623
LRP2	0.58	1
SHIP	0.73	0.999
RHOA	0.49	0.993
GSK3B	0.31	0.929
Rho	0.13	0.057
PIK3R1	0.84	1
CAMK2A	0.92	0.999
CNR1	0.05	0.561
Fyn	0.87	0.999
RAC1	0.63	0.999
Reln	0.89	0.999
EPHA1	0.26	0.824
Cdk5	0.34	0.501
VLDLR	0.74	0.999
Abl1	0.95	1
Bdnf	0.21	0.994
PSEN1	0.26	0.921
AKT1	0.26	0.155
Tau	0.33	0.999
itga3	0.91	0.999
LDLR	0.93	0.999
Src	0.51	0.999
APP	0.70	0.999

Table 5.8: GSA APOE-e4 non carriers model results

Gene	globalARTP_pvalue	globalEVT_pvalue
TP73	0.52	0.664
CDC42	0.55	0.999
ApoER2	1.00	1
Dab1	0.58	0.999
Abl2	0.29	0.994
PSEN2	0.14	0.901
Tbr1	0.51	0.999
LRP2	0.06	5.38e-13
SHIP	0.40	0.999
RHOA	0.60	0.991
GSK3B	0.56	0.999
Rho	0.85	0.973
PIK3R1	0.72	0.999
CAMK2A	0.26	0.986
CNR1	0.36	0.904
Fyn	0.64	0.999
RAC1	0.52	0.997
Reln	0.83	1
EPHA1	0.93	0.781
Cdk5	0.50	0.982
VLDLR	0.67	0.999
Abl1	0.22	0.986
Bdnf	0.19	0.855
PSEN1	0.98	0.999
AKT1	0.84	0.991
Tau	0.52	1
itga3	0.84	1
LDLR	0.42	0.564
Src	0.04	9.16e-06
APP	0.05	0.915

Chapter 6

Discussion

In this project we center on Gene Set Analysis (GSA), a strategy for combining the effects of many genetic variants within a gene. We propose the algorithms globalARTP and globalEVT, as a new implementations of the ARTP method that are specifically designed for genetic association studies involving SNPs. New implementation incorporates the selection of the best inheritance model for each SNP as a first step of the algorithm, and in the case of globalEVT, the computational time required is improved.

Through a simulation study we proved that Gene Set Analysis proposed approaches increase 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 SNPs 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. 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 some interesting genes that are worth further investigating.

Hence, in conclusion, obtained results show that the two proposed new methodologies increase significantly statistical power opposite to single-SNP analysis and, concretely, the second proposed method (globalEVT) reduces importantly the computational requirements since only one-parameter distributions are to be fitted. But, 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., 2010*[6]).

Acknowledgments

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.

Bibliography

[1] A. Alzheimer. Über eine eigenartige erkrankung der hirnrinde. *Allgemeine Z Psychiatrie Psychisch-Gerichtliche Medizin*, pages 146–148, 1906.

[2] S. Baloyannis. Morphological and morphometric alterations of cajal-retzius cells in early cases of alzheimer's disease. *International Journal of Neuroscience*, 115:965–980, 2005.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society,*, B 57 (1):289–300, 1995.

[4] N. Berchtold and C. Cotman. Evolution in the conceptualization of dementia and alzheimer's disease: Greco-roman period to the 1960s. *Neurobiology of Aging*, 19(3):173–189, 1998.

[5] A. Botella-López, F. Burgaya, R. Gavín, M. García-Ayllón, and E. Gómez-Tortosa. Reelin expression and glycosylation patterns are altered in alzheimer's disease. *Proceedings of the National Academy of Sciences*, 103:5573–5578, 2010.

[6] M. Calle, V. Urrea, N. Malats, and K. Steen. mbmdr: an r package for exploring gene-gene interactions associated with binary or quantitative traits. *Bioinformatics*, 26(17):2198–2199, 2010.

[7] R. Fisher. *Statistical Methods for Research Workers*. ISBN 0-05-002170-2, 1925.

[8] J. González, J. Carrasco, F. Dudbridge, L. Armengol, X. Estivill, and V. Moreno. Maximizing association statistics over genetic models. *Genetic Epidemiology*, 32(3):246–54, 2008.

[9] J. R. González, L. Armengol, X. Solé, E. Guinó, J. M. Mercader, X. Estivill, and V. Moreno. Snpassoc: an r package to perform whole genome association studies. *Bioinformatics*, 23(5):654–655, 2007.

[10] J. Herz and U. Beffert. Apolipoprotein e receptors: linking brain development and alzheimer's disease. *Nature Reviews Neuroscience*, 1(1):51–58, 2000.

[11] J. Herz and Y. Chen. Reelin, lipoprotein receptors and synaptic plasticity. *Nature Reviews Neuroscience*, 7(11):850–859, 2006.

[12] R. Mahley, K. Weisgraber, and Y. Huang. Apolipoprotein e4: A causative factor and therapeutic target in neuropathology, including alzheimer's disease. *Proceedings of the National Academy of Sciences (PNAS)*, 103(15):5644–51, 2006.

[13] R. Nussbaum and C. Ellis. Alzheimer's disease and parkinson's disease. *The New England Journal of Medicine*, 348:1356–1364, 2003.

[14] E. Reiman, J. Webster, A. Myers, J. Hardy, T. Dunckley, V. Zismann, K. Joshipura, J. Pearson, D. Hu-Lince, M. Huentelman, D. Craig, K. Coon, W. Liang, R. Herbert, K. Roher, A. Zhao, D. Leung, L. Bryden, L. Marlowe, M. Kaleem, D. Mastroeni, A. Grover, C. Heward, R. Ravid, J. Rogers, M. Hutton, S. Melquist, R. Petersen, G. Alexander, R. Caselli, A. Papassotiropoulos,

and D. Stephan. Gab2 alleles modify alzheimer's risk in apoe epsilon4 carriers. *Neuron*, 54(5):713–720, 2007.

[15] D. Rice and T. Curran. Role os the reelin signaling pathway in central nervous system development. *Annual Review of Neuroscience*, 24:1005–39, 2001.

[16] J. Sáez-Valero, M. Costell, M. Sjögren, N. Andreasen, and K. Blennow. Altered levels of cerebrospinal fluid reelin in frontotemporal dementia and alzheimer's disease. *Journal of Neuroscience Research*, 72:132–136, 2003.

[17] R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T. J. Hudson, A. Montpetit, A. V. Pshezhetsky, M. Prentki, B. I. Posner, D. J. Balding, D. Meyre, C. Polychronakos, and P. Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, 2007.

[18] W. Strittmatter. Apolipoprotein e: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial alzheimer disease. *Proceedings of the National Academy of Sciences (PNAS)*, 90(5):1977–1981, 1993.

[19] F. Tissir and A. Goffinet. Reelin and brain development. *Nature Review Neuroscience*, 4:496–505, 2003.

[20] V. Urrea and M. Calle. Simulation of genetic risk profiles for binary, continuous and time-to-event phenotypes. *(submitted)*, 2014.

[21] S. Waring and R. Rosenberg. Genome-wide association studies in alzheimer disease. *Archives of Neurology*, 65(3):329–334, 2008.

[22] K. Yu, Q. Li, A. Bergen, R. Pfeiffer, P. Rosenberg, N. Caporaso, P. Kraft, and N. Chatterjee. Pathway analysis by adaptive combination of P-values. *Genetic epidemiology*, 33(8):700–709, Dec. 2009.

[23] D. Zaykin, L. Zhivotovsky, P. Westfall, and B. Weir. Truncated product method for combining p-values. *Genetic Epidemiology*, 22(2):170–185, 2002.

Package 'globalGSA'

January 12, 2014

Type Package

Title Global Gene-Set Analysis for Association Studies.

Version 1.1

Date 2014-01-12

Author Natalia Vilor, M.Luz Calle

Maintainer Natalia Vilor <natalia.vilor@uvic.cat>

Description Implementation of four different Gene set analysis (GSA) algorithms for combining the individual pvalues of a set of genetic variats (SNPs) in a gene level pvalue. The implementation includes the selection of the best inheritance model for each SNP.

License GPL (>= 2)

LazyLoad yes

R topics documented:

`globalGSA-package` *Gene-set analysis for combining p-values in a joint test of association between a phenotype and a set of genetic variants (SNPs). Previously, a global test for the best inheritance model of each SNP is performed.*

Description

This package implements four different Gene-set analysis (GSA) methods for combining individual p-values of a set of SNPs using two different strategies. Each method provides a p-value for a joint test of association between the phenotype and the specified set of genetic variants. The four implemented methods are: Fisher method [1], Simes method [2], ARTP method [3] and EVT method.

Since the SNPs in a set may follow different modes of inheritance, previously to the GSA, a global test for the best inheritance model (dominant, recessive, log-additive and co-dominant) is performed on every SNP. The permutational p-value of the best model is obtained.

Details

Package:	globalGSA
Type:	Package
Version:	1.0
Date:	2013-09-22
License:	GPL (>= 2)

Author(s)

Natalia Vilor, M.Luz Calle

Maintainer: natalia.vilor@uvic.cat

References

[1] Fisher, R.A. (1925). Statistical Methods for Research Workers. ISBN 0-05-002170-2.

[2] Simes, R.J. (1986). An Improved Bonferroni Procedure for Multiple Tests of Significance. Biometrika, 73, 751-754.

[3] 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. December; 33(8): 700-709.

`globalARTP` *Global Adaptive Rank Truncated Product method.*

Description

This function provides the p-value for a joint test of association between a phenotype and a set of genetic variants (SNPs) using the Adaptive Rank Truncated Product method [1] after a global test for the best mode of inheritance of every SNP. The final gene-p-value is obtained from the permutational null distribution of the test statistic.

Usage

```
globalARTP(data, B, K, gene_list, Gene = "all", addit = FALSE,
covariable = NULL, family = binomial)
```

Arguments

data	Data frame containing the variables in the model. The first column is the dependent variable which must be a binary variable defined as factor (in case-control studies, the usual codification is 1 for cases and 0 for controls). SNP values may be codified in a numerical form (0,1,2) denoting the number of minor alleles, or using a character form where the two alleles are specified, without spaces, tabs or any other symbol between the two alleles.
B	Number of permutations considered in the permutational procedure.
K	Integer that indicates the maximum truncation point.
gene_list	File that provides the name of the set (for instance, gene) where each SNP belongs. This file has two columns: the SNP-Id ("Id"), and the Gene-Id ("Gene"). The SNP-Id must have the same label as the colnames of the data file.
Gene	Name of the gene that we want to analyze. The default value is Gene= "all" that indicates that the p-values of all SNPs in the database are to be combined. In this case it is not necessary to specify the gene_list file. In other case, we need to specify the name of the gene, for instance, Gene = "Gene1", and also the gene_list file.
addit	logical to determine if only an additive inheritance model should be considered in the global Test or, conversely, if we want to consider all possible inheritance models (dominant, recessive, log-additive and co-dominant). By default, addit = FALSE.
covariable	Data frame containing the covariables in the model. Each column represents one covariable. By default, covariable=NULL.
family	This can be a character string naming a family distribution. By default, family=binomial.

Value

List with the following components:

nPerm	Number of permutations.
Gene	Considered Gene.
Trunkpoint	Considered truncation point.
Kopt	Optimal truncation point.
genevalue	gene-pvalue.

References

[3] 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. December; 33(8): 700-709.

Examples

```
# load the included example dataset.
# This is a simulated case/control study data set
# with 2000 patients (1000 cases / 1000 controls)
# and 10 SNPs, where all of them have
# a direct association with the outcome:
data(data)
```

```
#globalARTP(data, B=1000, K=10, Gene="all", addit = FALSE)

# it may take some time,
# hence the result of this example is included:
data(ans11)

# You can test:
globalARTP(data, B=1, K=10, Gene="all", addit = FALSE)

# We consider that the first four SNPs
# are included in "Gene1",
# and the other six SNPs
# are included in "Gene2":
data(gene_list)
#globalARTP(data, B=1000, K=10, gene_list=gene_list, Gene="Gene1", addit = FALSE)

# it may take some time,
# hence the result of this example is included:
data(ans1)

# You can test:
globalARTP(data, B=1, K=10, gene_list=gene_list, Gene="Gene1", addit = FALSE)
```

`globalEVT` *Global Adaptive Extreme Value Distribution method.*

Description

This function provides the p-value for a joint test of association between a phenotype and a set of genetic variants (SNPs) using an Adaptive Extreme Value Distribution after a global test for the best mode of inheritance of every SNP. The final gene-p-value is obtained from

Usage

```
globalEVT(data, K)
```

Arguments

`data`	Data frame containing the variables in the model. The first column is the dependent variable which must be a binary variable defined as factor (in case-control studies, the usual codification is 1 for cases and 0 for controls). SNP values may be codified in a numerical form (0,1,2) denoting the number of minor alleles, or using a character form where the two alleles are specified, without spaces, tabs or any other symbol between the two alleles.
`K`	Integer that indicates the maximum truncation point.

Value

List with the following components:

`genevalue`	gene-pvalue.

Examples

```
# load the included example dataset.
# This is a simulated case/control study data set
# with 2000 patients (1000 cases / 1000 controls)
# and 10 SNPs, where all of them have
# a direct association with the outcome:
data(data)
globalEVT(data, K=10)
```

`globalFisher` *Global Fisher combination method.*

Description

This function provides the p-value for a joint test of association between a phenotype and a set of genetic variants (SNPs) using the Fisher method [1] after a global test for the best mode of inheritance of every SNP. The final gene-p-value is obtained from the permutational null distribution of the test statistic

Usage

```
globalFisher(data, B, gene_list, Gene = "all", addit = FALSE,
covariable = NULL, family = binomial)
```

Arguments

`data`	Data frame containing the variables in the model. The first column is the dependent variable which must be a binary variable defined as factor (in case-control studies, the usual codification is 1 for cases and 0 for controls). SNP values may be codified in a numerical form (0,1,2) denoting the number of minor alleles, or using a character form where the two alleles are specified, without spaces, tabs or any other symbol between the two alleles.
`B`	Number of permutations considered in the permutational procedure.
`gene_list`	File that provides the name of the set (for instance, gene) where each SNP belongs. This file has two columns: the SNP-Id ("Id"), and the Gene-Id ("Gene"). The SNP-Id must have the same label as the colnames of the data file.
`Gene`	Name of the gene that we want to analyze. The default value is Gene= "all" that indicates that the p-values of all SNPs in the database are to be combined. In this case it is not necessary to specify the gene_list file. In other case, we need to specify the name of the gene, for instance, Gene = "Gene1", and also the gene_list file.
`addit`	logical to determine if only an additive inheritance model should be considered in the global Test or, conversely, if we want to consider all possible inheritance models (dominant, recessive, log-additive and co-dominant). By default, addit = FALSE.
`covariable`	Data frame containing the covariables in the model. Each column represents one covariable. By default, covariable=NULL.
`family`	This can be a character string naming a family distribution. By default, family=binomial.

Value

List with the following components:

nPerm	Number of permutations.
Gene	Considered Gene.
genevalue	gene-pvalue.

References

[1] Fisher, R.A. (1925). Statistical Methods for Research Workers. ISBN 0-05-002170-2.

Examples

```
# load the included example dataset.
# This is a simulated case/control study data set
# with 2000 patients (1000 cases / 1000 controls)
# and 10 SNPs, where all of them have
# a direct association with the outcome:
data(data)
#globalFisher(data, B=1000,  Gene="all", addit=FALSE)

# it may take some time,
# hence the result of this example is included:
data(ans21)

# You can test:
globalFisher(data, B=1,  Gene="all", addit=FALSE)

# We consider that the first four SNPs
# are included in "Gene1",
# and the other six SNPs
# are included in "Gene2":
data(gene_list)
#globalFisher(data, B=1000, gene_list=gene_list, Gene="Gene1", addit=FALSE)

# it may take some time,
# hence the result of this example is included:
data(ans2)

# You can test:
globalFisher(data, B=1, gene_list=gene_list, Gene="Gene1", addit=FALSE)
```

---

globalSimes *Global Simes' combination method.*

Description

This function provides the p-value for a joint test of association between a phenotype and a set of genetic variants (SNPs) using the Simes method [1] after a global test for the best mode of inheritance of every SNP. The final gene-p-value is obtained from the permutational null distribution of the test statistic

Usage

```
globalSimes(data, B, gene_list, Gene = "all", addit = FALSE,
covariable = NULL, family = binomial)
```

Arguments

data	Data frame containing the variables in the model. The first column is the dependent variable which must be a binary variable defined as factor (in case-control studies, the usual codification is 1 for cases and 0 for controls). SNP values may be codified in a numerical form (0,1,2) denoting the number of minor alleles, or using a character form where the two alleles are specified, without spaces, tabs or any other symbol between the two alleles.
B	Number of permutations considered in the permutational procedure.
gene_list	File that provides the name of the set (for instance, gene) where each SNP belongs. This file has two columns: the SNP-Id ("Id"), and the Gene-Id ("Gene"). The SNP-Id must have the same label as the colnames of the data file.
Gene	Name of the gene that we want to analyze. The default value is Gene= "all" that indicates that the p-values of all SNPs in the database are to be combined. In this case it is not necessary to specify the gene_list file. In other case, we need to specify the name of the gene, for instance, Gene = "Gene1", and also the gene_list file.
addit	logical to determine if only an additive inheritance model should be considered in the global Test or, conversely, if we want to consider all possible inheritance models (dominant, recessive, log-additive and co-dominant). By default, addit = FALSE.
covariable	Data frame containing the covariables in the model. Each column represents one covariable. By default, covariable=NULL.
family	This can be a character string naming a family distribution. By default, family=binomial.

Value

List with the following components:

nPerm	Number of permutations.
Gene	Considered Gene.
genevalue	gene-pvalue.

References

[1] Simes, R.J. (1986). An Improved Bonferroni Procedure for Multiple Tests of Significance. Biometrika, 73, 751-754.

Examples

```
# load the included example dataset.
# This is a simulated case/control study data set
# with 2000 patients (1000 cases / 1000 controls)
# and 10 SNPs, where all of them have
# a direct association with the outcome:
data(data)
```

```
#globalSimes(data, B=1000,  Gene="all", addit=FALSE)

# it may take some time,
# hence the result of this example is included:
data(ans31)

# You can test:
globalSimes(data, B=1,  Gene="all", addit=FALSE)

# We consider that the first four SNPs
# are included in "Gene1",
# and the other six SNPs
# are included in "Gene2":
data(gene_list)
#globalSimes(data, B=1000, gene_list=gene_list, Gene="Gene1", addit=FALSE)

# it may take some time,
# hence the result of this example is included:
data(ans3)

# You can test:
globalSimes(data, B=1, gene_list=gene_list, Gene="Gene1", addit=FALSE)
```

Index