Evaluating De Novo Locus-Disease Discoveries in GWAS Using the Signal-to-Noise Ratio

Xia Jiang, PhD[1], M. Michael Barmada, PhD[2], Michael J. Becich, MD, PhD[1]

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
[2]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA

Abstract

A genome-wide association study (GWAS) involves examining representative SNPs obtained using high throughput technologies. A GWAS data set can entail a million SNPs and may soon entail many millions. In a GWAS researchers often investigate the correlation of each SNP with a disease. With so many hypotheses, it is not straightforward how to interpret the results. Strategies include using the Bonferroni correction to determine the significance of a model and Bayesian methods. However, when we are discovering new locus-disease associations, i.e., so called de novo discoveries, we should not just endeavor to determine the significance of particular models, but also concern ourselves with determining whether it is likely that we have any true discoveries, and if so how many of the highest ranking models we should investigate further. We develop a method based on a signal-to-noise ratio that targets this issue. We apply the method to a GWAS Alzheimer's data set.

Introduction

A *genome-wide association study* (*GWAS*) involves examining a large number of representative single-nucleotide polymorphisms (SNPs) obtained using high throughput technologies. A typical GWAS data set entails up to a million SNPs. Often a GWAS is conducted using cases and controls, where *cases* are individuals with a disease and *controls* are individuals without the disease. We investigate the statistical dependence of each SNP with the disease. In the process, many hypotheses are investigated.

With the maturation of next generation sequencing technology, the data set involved in a GWAS could entail many millions of SNPs[1], gene-environment-wide association studies are increasing in number[2], and researchers are investigating epistasis (gene-gene interactions) using GWAS data sets[3]. These investigations will significantly increase the number of hypotheses that are evaluated.

This unprecedented opportunity to learn potential disease risk from high-dimensional data sets does not come without difficulty. Owing to the vast number of hypotheses, the interpretation of the results is not straightforward. Strategies for handling multiple hypotheses testing include computing the significance or the posterior probability of a model using the Bonferroni correction and recently developed Bayesian methods[4,5]. Researchers have noted problems with using the Bonferroni correction in studies such as GWAS[6, 7,8], and the Bayesian methods are also not without difficulty.

A GWAS analysis is ordinarily an agnostic study. By an *agnostic study* we mean an explorative study in which we have no special prior belief concerning any particular locus. In such a study, our purpose is to discover new locus-disease associations, i.e., so called *de novo* discoveries. Therefore, besides striving to determine the significance or probability of particular models, we should concern ourselves with 1) determining whether it is likely that we have any true discoveries; and if so; 2) how many of the highest ranking models it would be prudent (cost-effective) to investigate further. For example, suppose that based on an analysis of a given data set, we find that it is likely that 15 of the top 20 models will be true discoveries, and likely that only 17 of the top 40 models will be true discoveries. Then we may decide to further investigate the top 20 models but not the next 20 models.

To measure the quality of their discoveries concerning microRNA target predictions[9,10], researchers used a signal-to-noise ratio measure, which is estimated by performing random shuffling of the data set. To our knowledge this strategy has not be applied to analyzing GWAS data sets. We developed a variant of this strategy that addresses the matter just discussed; i.e. it concerns determining the likely number of

true models in the first *k* models. We applied the strategy to the analysis of a GWAS data set on Alzheimer's disease. A Bayesian network scoring criterion was used to measure the association of each SNP with the disease.

Background

In what follows the indicators could be any variables and the target could be any phenotype. However, for the sake of focus we assume the indicators are SNPs and the phenotype is a disease.

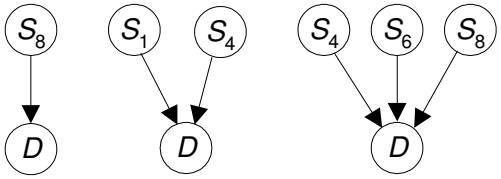

Figure 1. DAGs representing the relationships between SNPs and a disease *D*.

In previous studies[11,12] we represented the relationships between SNPs and a disease using *directed acyclic graph* (*DAG*) models like those shown in Figure 1. The first DAG indicates that SNP S_8 is associated with disease D by itself, the second indicates that S_1 and S_4 together are associated with D, and the last indicates that S_4, S_6, and S_8 together are associated with D. We computed the likelihood of such models using the *Bayesian Dirichlet equivalent uniform* (*BDeu*) score[13] for Bayesian network DAG models. This score gives the $P(Data \mid G)$, where G is the DAG being scored and *Data* is data about *D* and its predictors. In general, a Bayesian network can contain a much more complex DAG than those shown in Figure 1. Neapolitan[14] provides an introduction to Bayesian Networks. For specialized DAG models like those shown in Figure 1 the simplified formula for the BDeu score is as follows:

$$P(Data \mid G) = \prod_{j=1}^{q} \frac{\Gamma(\alpha/q)}{\Gamma(\alpha/q + s_j + t_j)} \frac{\Gamma(\alpha/2q + s_j)}{\Gamma(\alpha/2q)} \frac{\Gamma(\alpha/2q + t_j)}{\Gamma(\alpha/2q)}$$

where q is the number of different states the parents of D can assume, s_j is the number of cases that have the parents of D in their jth state, t_j is the number of controls that have the parents of D in their jth state, and α is a hyperparameter such that smaller values of α have larger DAG penalties. This score has exhibited good discovery performance in previous analysis of GWAS and simulated data sets[11,12]. The best results were obtained using moderate values of α; in particular, $\alpha = 54$ exhibited the top performance. In these previous studies traditional statistical methods were used to analyze results; we had not yet developed the technique presented here.

Method

We scored DAG models using the BDeu score, set thresholds for model discovery, and then used the signal-to-noise ratio to measure the quality of our discoveries. The details are provided next. First we discuss shuffling a data set.

Shuffling a Case-Control Data Set: A *Case-control data set* is a data set consisting of records concerning individuals with a disease (cases) and individuals without the disease (controls). The measured attributes (features) for each subject are possible predictors of the disease. The top table in Figure 2 shows a possible case-control data set where there are 8 subjects, 3 of them are cases ($D = 1$), and 5 of them are controls ($D = 0$). We have three binary attributes S_1, S_2, and S_3. A *random shuffle* of the data set is a new data set in which the subjects are assigned random values of the disease attribute while keeping the total number of cases the same. The bottom table in Figure 2 shows a random shuffle of the data set in the top table. Note that the number of cases (D = 1) in both tables is 3, and that values of the non-disease attributes are the same in both tables.

Subject	S_1	S_2	S_3	D
1	1	1	0	1
2	0	1	1	1
3	0	1	0	1
4	1	0	0	0
5	1	0	1	0
6	0	1	1	0
7	1	0	1	0
8	1	1	0	0

Subject	S_1	S_2	S_3	D
1	1	1	0	0
2	0	1	1	1
3	0	1	0	0
4	1	0	0	0
5	1	0	1	1
6	0	1	1	0
7	1	0	1	0
8	1	1	0	1

Figure 2. The original data set is on the top and a shuffled data set is on the bottom.

Signal-to-Noise Ratio: The signal-to-noise ratio is a measure of how much a signal has been corrupted by noise. It is the ratio of the signal power to the noise power. When the ratio exceeds one, it indicates that there is more signal than noise. Depending on the context, there are different precise mathematical definitions of this term. We use the signal-to-noise ratio to measure whether it is likely that we have some true discoveries, and therefore define it as follows. Suppose we have a score representing the plausibility of a model M. For the sake of focus we take this score to be the probability of some $Data$ given that M is correct. We set a threshold T and call the model a discovery if $P(Data \mid M) > T$. The number *m* of models whose scores exceed T is the *signal*. However, the set of models whose scores exceed T includes false models. The number *n* of such false models is the *noise*. We then have that the *signal-to-noise ratio* (*SNR*) is given by

$$SNR = \frac{m}{n}.$$

Figure 3 depicts possible values of the signal, noise, and SNR at a possible threshold. The figure indicates that at this threshold we would obtain 5 findings of which 2 are noise.

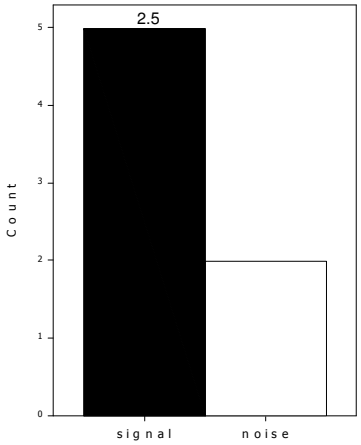

Figure 3. The signal is on the left; the noise is on right; the signal-to-noise ratio is 2.5.

We cannot measure the noise n directly because we do not know whether a model is true or false. However, we can estimate it by randomly shuffling the data set, and counting the number of models whose scores exceed T given the shuffled data set. The idea is that a random data set should only contain noise. We obtain this count for many random shuffles and then take the average to obtain our final estimate of n. This is the approach taken by researchers in microRNA target predictions[9,10]. We extend this analysis as follows. We treat the noise n in our actual data set as a random variable, and our probability distribution (belief) concerning this random variable is the distribution of the noise for the randomly shuffled data sets. Using this distribution, we can not only determine the expected value of n, but also probability intervals for n. For example, we can determine a value n' such that the probability is 0.99 that $n < n'$. Using this expected value of the noise and a probability interval for the noise, we can compute the expected value of the SNR and a probability interval for the SNR.

For a given threshold T, Figure 4 shows a possible distribution of the noise for shuffled data sets and a value of n'. This distribution was developed using the data set discussed in the Experiments section. There were 861 cases, 550 controls, and 312,317 SNPs. We considered all 1-SNP models, and developed 1088 shuffled data sets. The threshold T is the log likelihood of the data given the SNP model, and in this example T = -947.18. The value plotted on the x-axis is the number of models whose log likelihoods exceeded T for a given shuffled data set, and the value plotted on the y-axis is the number of shuffled data sets having a given value of x. The expected value of the noise $E(n)$ is the average noise over all 1088 data sets. In this example $E(n)$ = 18143.4.

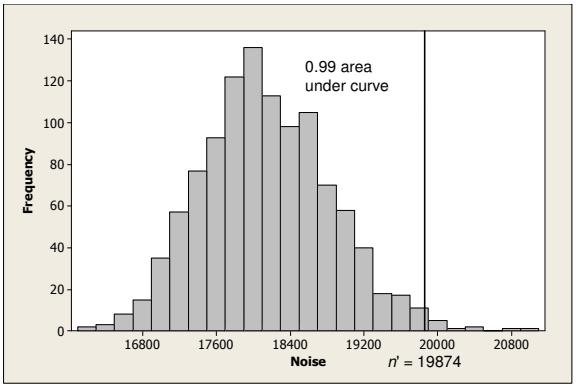

Figure 4. For a given threshold T, this histogram shows a possible probability distribution of the noise for randomly shuffled data sets and the value of n' such that the probability is 0.99 that the noise $n < n'$.

Experiments: We investigated a *late onset Alzheimer's disease* (*LOAD*) data set that was developed by Reiman et al.[15]. The indicators in this data set include 312,317 SNPs and *APOE* status. It is well-known that the APOE gene is linked with LOAD where increased risk is associated with the $\varepsilon 4$ allele. The data set consists of three cohorts containing a total of 1411 participants. Of the 1411 participants, 861 had LOAD and 550 did not. In addition, 644 participants were *APOE* $\varepsilon 4$ carriers, who carry at least one copy of the *APOE* $\varepsilon 4$ allele, and 767 were APOE $\varepsilon 4$ non-carriers.

The following procedure was followed. Using this LOAD data set and $\alpha = 54$ in the BDeu score we scored each 1-SNP model, which is a model containing one SNP as depicted on the far left in Figure 1. We then determined a number of thresholds. For each threshold we computed the signal by counting how many models had log BDeu scores exceeding the threshold. Next, based on computational considerations, we performed 1088 random shuffles of the data set, and for each of these data sets we computed its noise by counting how many models had log BDeu scores exceeding the threshold. As discussed in the Methods section, we used the distribution of the noise for the randomly shuffled data sets to determine the expected value of the noise n in the actual data set, and a value n' such the probability is 0.99 that $n < n'$. We did this by sorting the noise values, and then determining the value of the noise at the 99th percentile. Using the expected value of the noise and 99th percentile value, we compute the expected value of the SNR and its 99th percentile value. Note that in the case of the SNR

this is the value S' such that the probability is 0.99 that SNR > S' (since the SNR is the signal divided by the noise).

All experiments were run using a Mac Pro Server A1289, which had 8 cores consisting of Two 2.93 GHz Quad-Core Intel Xeon 5500 series processors. The server had 32 GB 1066MHz DDR3 ECC SDRAM and 8TB of internal storage. The operating system was Mac OS X v10.6 Snow Leopard. The data shuffling and scoring programs were developed in Java using eclipse 3.1. It took approximately 184 seconds to process each shuffled data set including the time it took to shuffle the data.

Results

Table 1 shows the results at various thresholds for the LOAD data set. We discuss these results next. Note that all probabilities stated in what follows are based on the assumption that our belief (probability distribution) concerning the noise is obtained from the noise distribution for our 1088 randomly shuffled data sets.

The signal at the highest threshold (bottom row in Table 1) is 2 and the *E*(noise) is 0.This means that no randomly shuffled data set resulted in any SNP models having a log BDeu score exceeding this threshold. So we can be very confident that the first two loci are true discoveries. The signal at the 2nd highest threshold is 3, the *E*(noise) is 0.08455, and the 99th noise percentile is 2. This means that the expected value of the number of true discoveries among the first 3 loci is 3-0.08455 ≈ 3, and the probability is 0.99 that at least 3-2 = 1 is a true discovery. The signal at the 3rd highest threshold is 6, the *E*(noise) is 0.284926, and the 99th noise percentile is 3. This means that the expected value of the number of true discoveries among the first 6 loci is about 6, and the probability is 0.99 that at least 3 are true discoveries. The signal at the 4th highest threshold is 19, the *E*(noise) is 0.956801, and the 99th noise percentile is 6. This means that the expected value of the number of true discoveries among the first 19 loci is 18, and the probability is 0.99 that at least 13 are true discoveries.

Table 1. Signal, noise, *E*(noise), *E*(SNR) and 99th percentile values at various thresholds for the LOAD data set. The column labeled "99th noise perc." contains the value such that the probability is 0.99 that the noise in the actual data set is less than this value. The column labeled "99th SNR perc." contains the value such that the probability is 0.99 that the SNR for the actual data set is greater than this value.

threshold	signal	*E*(noise)	99th noise perc.	*E*(SNR)	99th SNR perc.
-950.533	296109	295189.6	295519	1.003	1.002
-949.416	175840	171481.6	174658	1.025	1.007
-948.298	69081	64054.55	67036	1.078	1.031
-947.18	21769	18143.4	19874	1.200	1.095
-946.062	6884	5160.163	5885	1.334	1.170
-944.945	2250	1481.453	1776	1.519	1.267
-943.827	776	427.4752	555	1.815	1.398
-942.709	261	124.4072	180	2.099	1.450
-941.592	106	36.43658	62	2.909	1.710
-940.474	46	10.8943	26	4.222	1.769
-939.356	26	3.191176	12	8.147	2.167
-938.238	19	0.956801	6	19.858	3.167
-937.121	6	0.284926	3	21.058	2.000
-936.003	3	0.084559	2	35.478	1.500
-931.532	2	0	0	infinity	Infinity

The results just discussed are consistent with our previously knowledge of locus-LOAD association. The first two loci were *APOE* status and SNP rs41377151. As mentioned previously, it is well-known that the *APOE* gene is associated with LOAD. SNP rs41377151 is on the *APOC1* gene, which is in strong linkage disequilibrium with *APOE* and previous studies have indicated that they predict LOAD equally

well[16]. So, we would expect that we can be very confident that the first 2 loci are true discoveries. The 3rd discovered locus is rs1082430, which is on the *PRKG1* gene. There are a number of previous studies associating this gene with LOAD[17,18]. The 4th, 5th, and 6th SNPs are rs4356530, rs17330779, and rs6784615. There is evidence associating these latter two SNPs with LOAD[19]. So it is not surprising that it is very probable that 3 of the first 6 loci are true discoveries. Of the remaining top 19 loci we found previous studies linking four of them to LOAD. So previous studies have linked a total of 9 of the first 19 loci with LOAD. This number is somewhat less than 13 (the highly probable number of true discoveries in the first 19 loci). However, it is reasonably close, and indeed the reason we are conducting GWAS is to discovery loci not previously known to be associated with diseases.

It is perhaps surprising that, for example, the probability is 0.99 that the noise is less than 67036 when the signal is 69081 (3rd row in Table 1). This means that it is probable that there are 69081-67036 = 2045 true discoveries among the first 69081 loci. This number seems high. However, these are SNPs and not genes, and many of the SNPs are located on the same gene. For example, the study included 14 *GAB*2 SNPs. Regardless, the number of loci that are likely unaccounted for by noise is more than we expected. This investigation does not even include epistasis. These loci by themselves are generating the signal. Perhaps there are a multitude of loci that have a relatively small effect on LOAD. On the other hand, perhaps there is an extreme amount of noise in the LOAD data set.

To investigate this matter further, we looked at the noise at the 100th percentile. This value was obtained by looking at the value of the noise in the shuffled data set that had the most noise. Table 2 shows the results. Interestingly, even at this extreme probability, for many thresholds there is quite a bit of signal not accounted for by noise.

Table2. Signal, noise, 100th noise perc., and 100th signal-noise perc. The 100th noise percentile is obtained by taking the value of the noise in the shuffled data set that had the most noise. The 4th column shows the amount of signal that is unaccounted for by noise.

threshold	signal	100th noise perc.	100th signal - noise perc.
-950.533	296109	295613	496
-949.416	175840	176540	0
-948.298	69081	69513	0
-947.18	21769	20995	774
-946.062	6884	6321	563
-944.945	2250	1932	318
-943.827	776	606	170
-942.709	261	202	59
-941.592	106	76	30
-940.474	46	31	15
-939.356	26	20	6
-938.238	19	17	2
-937.121	6	12	0
-936.003	3	5	0
-931.532	2	0	2

Figure 5 shows the noise distributions for a representative subset of the rows in Table 1. For small expected values the distributions are skewed to the left with many zero values, but as the expected values get larger, they look increasingly like normal distributions.

Conclusion

We developed a method based on the signal-to-noise ratio that enables us to determine the likely number of true models in the first *k* models. We argued that this information would be very useful to researchers when deciding how many models to investigate further. The method was applied to a real GWAS

Alzheimer's data set, and for relatively small values of k the results were consistent with known associations with Alzheimer's disease. However, for large values of k, the likely number of true discoveries was more than we expected. Furthermore, even at the most extreme noise levels, there was a substantial amount of signal that was unaccounted for by noise. Galvin[20] notes that there seems to still be a great deal of dark matter concerning the genetic basis of disease. Perhaps there are many undiscovered loci that have a modest affect on Alzheimer's disease. This matter bears further investigation. We plan to obtain additional GWAS Alzheimer's data sets and repeat the signal-to-noise ratio study using them.

The signal-to-noise ratio technique worked well using an Alzheimer's data set. Future research should investigate its application to the investigation of other GWAS data sets such as those concerning breast and other cancers.

We developed 1088 shuffled data sets based on computational considerations. A further avenue for future research would be to investigate how sensitive the results are to the number of shuffled data sets.

Another application of the technique described here would be in the comparison of two discovery measures. For example, *multifactor dimensionality reduction* (*MDR*)[21] is another score for measuring the prospect of a model based on data. Its performance has been compared to that of the BDeu score using simulated data sets[12]. When using simulated data sets the comparison is straightforward since we know the correct model (the one generating the data), and we can investigate which score better identifies the correct model. However, if we want to compare MDR and the BDeu score using a real data set, we don't know which models are correct. As an alternative we can investigate the signal-to-noise ratios for the scores and see which method yields higher signal-to-noise ratios.

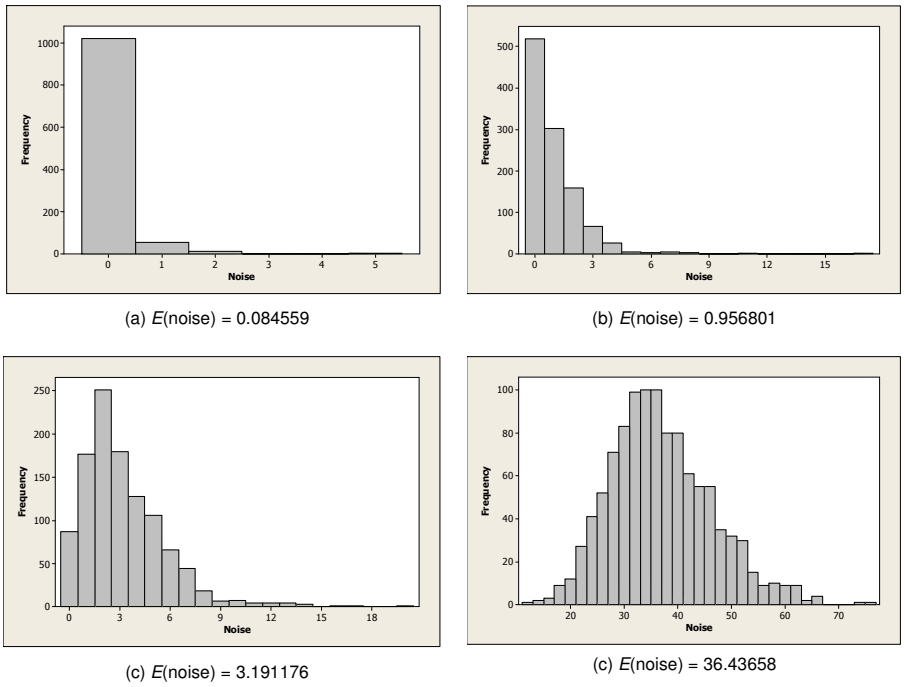

Figure 5. Histograms representing several of the noise probability distributions obtained from the randomly shuffled data sets.

Acknowledgements

This research was funded by grant 1K99LM010822-01 from the National Library of Medicine at the National Institutes of Health.

References

1. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature; 2010;467:1061-1073.
2. The International Cancer Genome Consortium. International network of cancer genome projects. Nature; 2010;464:993-998.
3. Jiang X, Neapolitan RE, Barmada MM, Visweswaran S, Cooper GF. A fast algorithm for learning epistatic genomic relationships. AMIA Annu Symp Proc; 2010: 341–345.
4. Wacholder S, Chanock S, Garcia-Closas M, et al. Assessing the probability that a positive report is false; an approach for molecular epidemiology studies. J Nat Can Inst; 2004;96:434-432.
5. Wakefield J. Reporting and interpreting in genome-wide association studies. International Journal of Epidemiology; 2008;37(3): 641-653.
6. Neapolitan RE. A polemic for Bayesian statistics. In: Holmes D, Jain L, editors. Innovations in Bayesian Networks, New York: Springer Verlag; 2008.
7. Hoggart C, Clark T, De Iorio M, Whittaker J, Balding D. Genome-wide significance for dense SNP and resequencing data. Genetic Epidemiology; 2008; 32:179-185.
8. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science; 1996;273:1516-1517.
9. Lewis et al. Prediction of mammalian microRNA targets. Cell; 115:787-798.
10. Krek et al. Combinatorial microRNA target predictions. Nature Genetics; 37(5):495-499.
11. Jiang X, Barmada MM, Visweswaran S. Identifying genetic interactions in genome-wide data using Bayesian networks. Genetic Epidemiology; 2010;34:575-581.
12. Jiang X, Neapolitan RE, Barmada MM, Visweswaran S. Learning genetic epistasis using Bayesian network scoring criteria. To appear in BMC Bioinformatics; 2011.
13. Heckerman D, Geiger D, Chickering D. Learning Bayesian networks: the combination of knowledge and statistical data. Technical Report MSR-TR-94-09. Redmond, Washington; Microsoft Research; 1995.
14. Neapolitan RE. Learning Bayesian Networks. Upper Saddle River, NJ; Prentice Hall; 2004.
15. Reiman EM, et al. GAB2 alleles modify Alzheimer's risk in APOE carriers. Neuron; 2007;54:713-720.
16. Benjamin T, et al. *APOE* and APOC1 promoter polymorphisms and the risk of Alzheimer disease in African American and Caribbean Hispanic individuals. Neurology; 2004;61(9):1434-9.
17. Fallin MD, Szymanski M, Wang R, Gherman A, Bassett SS, Avramopoulos D. Fine mapping of the chromosome 10q11-q21 linkage region in Alzheimer's disease cases and controls. Neurogenetics; 2010; 11(3):335-348.
18. Liang X, et al. Genomic convergence to identify candidate genes for Alzheimer disease on chromosome 10. Human Mutation; 2009;30(3):463-471.
19. Shi H, Medway C, Bullock J, Brown K, Kalsheker N, Morgan K. Analysis of Genome-Wide Association Study (GWAS) data looking for replicating signals in Alzheimer's disease (AD). Int J Mol Epidemiol Genet. 2010;1(1):53-66.
20. Galvin A, Ioannidis JPA, Dragani TA. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. Trends in Genetics; 2010;26(3):132-41.
21. Hahn LW, Ritchie MD, Moore, JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics; 2003;19(3);376-382.