BMC Medical Genetics

BioMed Central

Research article

Genome-wide association reveals three SNPs associated with sporadic amyotrophic lateral sclerosis through a two-locus analysis

Qiuying Sha[1], Zhaogong Zhang[1,2], Jennifer C Schymick[3,4], Bryan J Traynor[3,5] and Shuanglin Zhang*[1,6]

Address: [1]Department of Mathematical Sciences, Michigan Technological University, Houghton, MI, USA, [2]School of Computer Science and Technology, Heilongjiang University, Harbin, PR China, [3]Laboratory of Neurogenetics, National Institute on Aging, NIH, Bethesda, MD, USA, [4]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK, [5]Department of Neurology, Johns Hopkins University, Baltimore, MD, USA and [6]Department of Mathematics, Heilongjiang University, Harbin, PR China

Email: Qiuying Sha - qsha@mtu.edu; Zhaogong Zhang - zhaogong@mtu.edu; Jennifer C Schymick - schymickj@mail.nih.gov; Bryan J Traynor - traynorb@mail.nih.gov; Shuanglin Zhang* - shuzhang@mtu.edu

* Corresponding author

Published: 9 September 2009

BMC Medical Genetics 2009, **10**:86 doi:10.1186/1471-2350-10-86

Received: 19 September 2008
Accepted: 9 September 2009

This article is available from: http://www.biomedcentral.com/1471-2350/10/86

Abstract

Background: Amyotrophic lateral sclerosis (ALS) is a fatal, degenerative neuromuscular disease characterized by a progressive loss of voluntary motor activity. About 95% of ALS patients are in "sporadic form"-meaning their disease is not associated with a family history of the disease. To date, the genetic factors of the sporadic form of ALS are poorly understood.

Methods: We proposed a two-stage approach based on seventeen biological plausible models to search for two-locus combinations that have significant joint effects to the disease in a genome-wide association study (GWAS). We used a two-stage strategy to reduce the computational burden associated with performing an exhaustive two-locus search across the genome. In the first stage, all SNPs were screened using a single-marker test. In the second stage, all pairs made from the 1000 SNPs with the lowest p-values from the first stage were evaluated under each of the 17 two-locus models.

Results: we performed the two-stage approach on a GWAS data set of sporadic ALS from the SNP Database at the NINDS Human Genetics Resource Center DNA and Cell Line Repository http://ccr.coriell.org/ninds/. Our two-locus analysis showed that two two-locus combinations--rs4363506 (SNP1) and rs3733242 (SNP2), and rs4363506 and rs16984239 (SNP3) -- were significantly associated with sporadic ALS. After adjusting for multiple tests and multiple models, the combination of SNP1 and SNP2 had a p-value of 0.032 under the Dom∩Dom epistatic model; SNP1 and SNP3 had a p-value of 0.042 under the Dom × Dom multiplicative model.

Conclusion: The proposed two-stage analytical method can be used to search for joint effects of genes in GWAS. The two-stage strategy decreased the computational time and the multiple testing burdens associated with GWAS. We have also observed that the loci identified by our two-stage strategy can not be detected by single-locus tests.

Background

Amyotrophic lateral sclerosis (ALS) is a fatal progressive neurodegenerative disease that attacks nerve cells in the brain and spinal cord resulting in muscle weakness and atrophy. Although ALS is listed as a rare disease with a prevalence of approximately 1 per 10,000, it is the most common adult onset form of motor neuron diseases [1,2]. Epidemiological studies have showed that 1.5-5.3% of cases are familial in nature [3-6]. The remaining 95% of cases are not associated with a family history of the disease and seem to occur sporadically throughout the community. Several genes that cause familial ALS have been identified [7-14], especially the SOD1 gene which is believed to be responsible for 20% of familial ALS.

The identification of susceptibility genes of sporadic ALS has been slow in arriving. The search for sporadic ALS genes has generated a large number of candidate-gene association studies [15-19]. To date, we do not have a functional SNP or haplotype that has made a credible contribution to our understanding of disease pathogenesis in the way that the *APOE*-e4 allele does in Alzheimer disease (AD) and the H1 MAPT haplotype does in parkinsonian syndromes [20]. There is an urgent need to understand the genetic architecture of sporadic ALS and ultimately to develop novel drugs for this fatal disease. Sporadic ALS is hypothesized to be a complex disorder in which the disease is modulated by variations in multiple genetic loci interacting with each other and environmental exposures [18]. The lack of major genes may be a reason for the unsuccessful candidate gene studies which investigated one gene at a time.

Recently, Schymick et al. made the first attempt to identify genetic factors that might be relevant in the pathogenesis of sporadic ALS by using a well-designed GWAS [1]. The first stage single-marker analysis performed by Schymick et al. showed that 34 SNPs had a p-value less than 0.0001 with the smallest one being 6.8×10^{-7}. After adjusted by permutation procedure, none of these SNPs reached the significance level of 0.05. This finding suggests that the ALS phenotype is not driven by a single powerful locus. By testing one marker at a time, the first stage analysis made the implicit assumption that susceptibility loci can be identified through their independent, marginal contributions to the trait variability. More recently, other GWAS in ALS have been conducted by different research groups [21-24]. However, all these GWAS used single-marker analysis. Recent human and animal studies of complex diseases have identified susceptibility genes that marginally contribute to a common trait, to a minor extent only or not at all, but that interact significantly in combined analyses [25-32]. Thus, methods that can account for joint effects of genes may be appropriate for analyzing genome-wide association data sets.

In this article, we used seventeen two-locus models to analyze the previously published genome-wide association data for ALS. We found that three SNPs were significantly associated with sporadic ALS. After we observed the significant two-locus combinations, we further estimated the impact (relative risk and odds ratio) of each of the two-locus combinations on sporadic ALS. It has been recognized that the traditional method will over estimate the odds ratio or relative risk in GWAS [32,33]. Recently, Zollner and Pritchard proposed a new method to estimate penetrance and then odds ratio and relative risk [32]. Through extensive simulation studies, Zollner and Pritchard showed that the estimations of odds ratio and relative risk by their method were not upward biased. By modifying Zollner and Pritchard's method, we proposed a new method to estimate two-locus penetrance, and then estimate the odds ratio, relative risk and sample size needed to replicate the findings for this rare disease.

Methods

In this section, we will give details of the data set and describe a new analytical method to analyze this data set.

The Data Set from GWAS for Sporadic ALS

Schymick et al. have made their data set publicly available through the website of the National Institute of Neurological Disorders and Stroke (NINDS) Human Genetics Resource Center at the Coriell Institute http://ccr.coriell.org/ninds[1]. The data set contained 555,352 unique SNPs across the genome in 276 patients with sporadic ALS and 271 neurologically normal controls. The 555,352 SNPs were carefully chosen tagging SNPs from phase I and II of the HapMap Project. The sampled individuals were all non-Hispanic white Americans. There were 102 females and 174 males in cases, and 142 females and 129 males in controls. All sampled individuals had a more than 95% genotype call rate. The average call rate across all samples was 99.6%. Of the 555,352 SNPs studied, the genotype call rate was greater than 99% for 514,088 (representing 92.6% of all SNPs assayed) and greater than 95% for 549,062 (98.9%) SNPs. The phenotype file of this data set contained the status of sporadic ALS, age of onset, site of onset (bulbar-onset, upper-limb-onset, and lower-limb-onset), gender, and smoking status among other information.

Statistical Analysis

Two-locus Analysis Based on Seventeen Two-locus Models

In this article, we used seventeen two-locus models to analyze the genome-wide association data. For each SNP, we called one allele a high-risk allele if its frequency in cases was larger than the frequency in controls. For SNP A with alleles A, a and SNP B with alleles B, b, Figure 1 and 2 give eight epistatic two-locus models and nine multiplicative two-locus models with high-risk alleles A and B, respec-

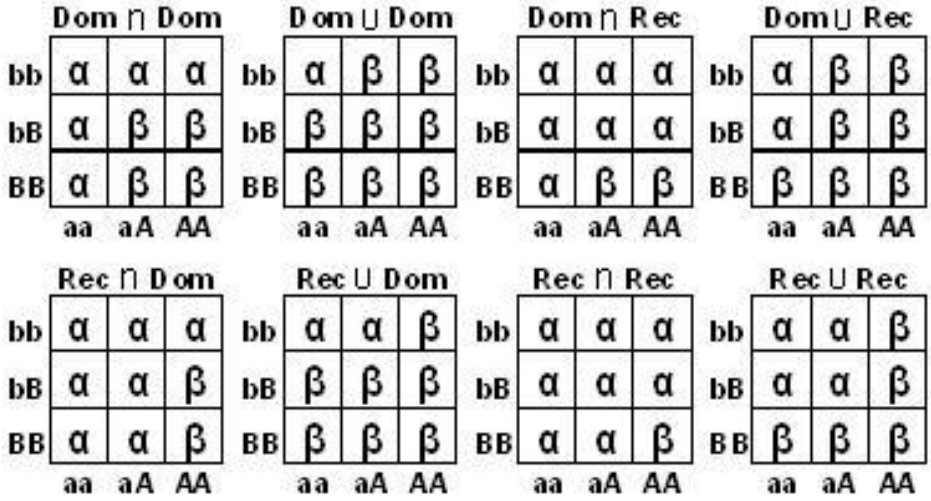

Figure 1
Eight two-locus epistatic models. A and B are the high-risk alleles in the two markers. α and β are the penetrance. ∩: two-locus genotypes with both high-risk genotypes at SNP A and SNP B are high-risk genotypes. ∪: two-locus genotypes with at least one high risk genotype at SNP A or SNP B are high-risk genotypes.

tively. Some of the eight epistatic two-locus models have been used and discussed by Xiong et al. and Zhao et al. [34,35]. The multiplicative models that are good approximations of additive models have been discussed by Hodge and Risch [36,37].

Under each of the epistatic models, the nine two-locus genotypes were divided into two groups: high-risk genotype group and low-risk genotype group. For example, under the model Dom∩Dom, the high-risk group was G_H = {*aAbB*, *AAbB*, *aABB*, *AABB*} and the low-risk group was G_L = {*aabb*, *aAbb*, *AAbb*, *aaBB*} For the eight epistatic models, we used one degree of freedom (df) χ^2 test statistic given by

$$T_{epi} = \frac{mn}{(m+n)} \bullet \frac{(p-q)^2}{P(1-P)}$$

to test for association of two-locus joint effects, where $\hat{p}$, $\hat{q}$, and $\hat{P}$ denote the frequencies of the high-risk genotype group in cases, controls and the pooled sample (cases and controls are pooled together).

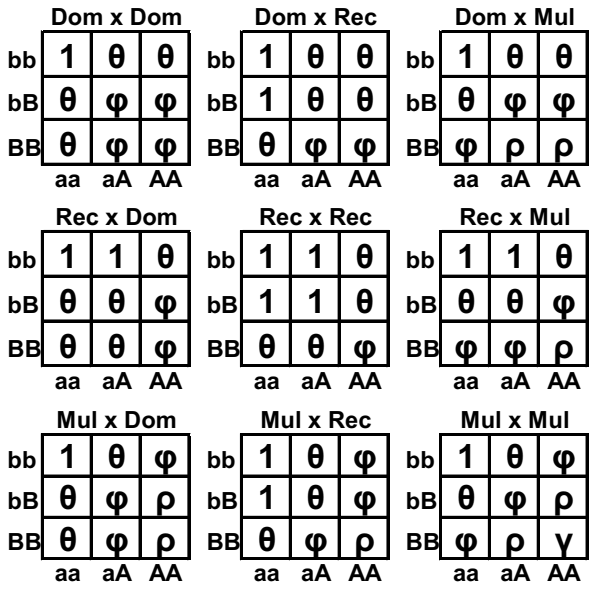

Figure 2
Nine two-locus multiplicative models. A and B are the high-risk alleles in the two markers. The symbol in each cell denotes the relative risk of this cell. $\phi = \theta^2$, $\rho = \theta^3$ and $\gamma = \theta^4$.

For the nine multiplicative models, we constructed a two-locus association test as follows. Let P(*Disease*|g) denote the penetrance of two-locus genotype combination $g = (g_1, g_2)$, where g_1 and g_2 are the genotypes in the first and second markers, respectively. Let β_0 denote the logarithm of the penetrance of genotypes with a relative risk of 1 in the models (see Figure 2) and $\beta_1 = \log\theta$, where θ is the relative risk given in Figure 2. Then, the nine multiplicative models can be described by the following log linear model $\log P(Disease|g) = \beta_0 + \beta_1 X$, where $X = x_1 + x_2$, x_1 is the numerical code of g_1 and is given by

$$x_1 = \begin{cases} 0, & \text{if } g_1 = aa \\ 1, & \text{if } g_1 = aA \text{ or } AA \end{cases}, \quad x_1 = \begin{cases} 0, & \text{if } g_1 = aa \text{ or } aA \\ 1, & \text{if } g_1 = AA \end{cases}, \quad \text{or } x_1 = \begin{cases} 0, & \text{if } g_1 = aa \\ 0.5, & \text{if } g_1 = aA \\ 1, & \text{if } g_1 = AA \end{cases}$$

for a dominant, recessive or multiplicative model, respectively; x_2 is similarly defined as the numerical code of g_2. Under the log linear model $\log P(Disease|g) = \beta_0 + \beta_1 X$, $\beta_1 = 0$ means that all the genotypes have the same penetrance which implies that $\theta = 1$. So a test of the association between the disease and the two loci under the nine multiplicative models is equivalent to a test of the null hypothesis H_0: $\beta_1 = 0$. For the i^{th} individual, let γ_i denote the trait value (1 for diseased individual and 0 for normal individual) and X_i denote the numerical code of the genotype (X in the log linear model). The score test statistic is given by

$$T_{score} = \frac{(\sum_{i=1}^{N}(X_i-\bar{X})(\gamma_i-\bar{\gamma}))^2}{\bar{\gamma}(1-\bar{\gamma})\sum_{i=1}^{N}(X_i-\bar{X})^2},$$

where N is the sample size, $\bar{X}$ is the average of $X_1,..., X_N$, and $\bar{\gamma}$ is the average of $\gamma_1,..., \gamma_N$. Under the null hypothesis, T_{score} follows a χ^2 distribution with 1 df. Note that under each of the two-locus epistatic models, if we code $X = 1$ for a high-risk genotype group and $X = 0$ for a low-risk genotype group, then $T_{epi} = T_{score}$.

The method to search for significant two-locus combinations for each of the seventeen models has the following two steps:

Step 1: For each SNP, let n and m denote the number of individuals in cases and controls (different SNPs may have a different number of cases and controls due to missing genotypes). Let n_1, n_2, n_3 and m_1, m_2, m_3 denote the number of three genotypes in cases and controls, respectively. The 2 df genotypic test statistic is given by

$$T = \sum_{i=1}^{3}\frac{(p_i-q_i)^2}{\frac{p_i}{m}+\frac{q_i}{n}},$$

where $\hat{p}_i = \frac{n_i}{n}$ and $\hat{q}_i = \frac{m_i}{m}$. We applied this test statistic to each SNP, calculated the corresponding p-value, and returned M SNPs with the smallest p-values (M = 1,000 was used in this article).

***Step 2*:** Under each of the seventeen two-locus models, we applied a two-locus association test to each of the L two-locus combinations among the M retained SNPs, where $L = M(M\text{-}1)/2$. For a two-locus epistatic model given in figure 1, we used the two-locus test T_{epi}. For a multiplicative model given in figure 2, we used the score test T_{score}. In this step, we got a p-value (called raw p-value) for each of the L two-locus combinations and each of the seventeen two-locus models.

A permutation procedure was used to adjust for multiple tests and multiple models. In each permutation, we randomly shuffled the cases and controls and repeated step 1 and step 2 based on the permuted data. We performed the permutation procedure B times (B = 1,000 was used in this article). For the i^{th} model and l^{th} two-locus combination ($i = 1,...,17$; $l = 1,..., L$), let p_{il} and p_{il}^b denote the raw p-values of the two-locus tests in step 2 based on the original data and on the b^{th} permutated data, respectively. Let

$$q_b = \min\{p_{il}^b, 1 \le i \le 17, 1 \le l \le L\}.$$

Then, for the i^{th} model and l^{th} two-locus combination, P_{il}, the p-value adjusted for multiple tests and multiple models, was given by $P_{il} = \frac{\#\{b:p_{il}>q_b\}}{B}$.

A New Method to Estimate Penetrance

When a study identifies a locus or locus-combination that shows evidence of association with a disease, it is common to estimate the impact of this locus or locus-combination on the phenotype of interest. This impact is often expressed as an odds ratio. Estimation of the odds ratio is also helpful for planning successful replication studies.

It is recognized that the traditional estimate of odds ratio is up-biased because it is typically estimated for the locus which was significant for association [32,33]. Recently, Zollner and Pritchard proposed a new method to estimate penetrance (odds ratio can be calculated based on the penetrance) [32]. This new method was based on the likelihood of observed genotypes given that the locus was significant for association. We modified Zollner and Pritchard's method to estimate the penetrance and odds ratio for two-locus combinations under each of the seventeen models given in Figure 1 and Figure 2. We use the Dom∩Dom model given in Figure 1 as an example to describe our method.

We use the following notation:

n, m: the number of cases and controls

the data $D = \{n_1,..., n_9; m_1,..., m_9\}$: the counts of nine two-locus genotypes in cases and controls that constitute the significant signal for association

$(q_1,..., q_9)$: the population frequencies of the genotypes

R: the relative risk of high-risk genotype combination to low-risk genotype combination, $R = \beta/\alpha$.

F: the population prevalence of the disease which is assumed to be known.

Because ALS is a rare disease with $F = 0.0001$, we can estimate q_i from the sampled controls. Thus, we assume that q_i = (number of i^{th} genotype in controls)/m is known in the following discussion. In the Dom∩Dom model, the 5^{th}, 6^{th}, 8^{th} and 9^{th} genotype combination {(aA, bB), (AA, bB), (aA, BB), (AA, BB)} is the high-risk genotype combination, and the combination of the other genotypes is the low-risk genotype combination. Let $q_H = q_5 + q_6 + q_8 + q_9$ denote the population frequency of the high-risk genotype combination. Then, the penetrance α and β (see Figure 1) can be calculated by

$$\alpha = \frac{F}{(R-1)q_H+1} \text{ and } \beta = \frac{F \bullet R}{(R-1)q_H+1} \qquad (1)$$

Thus, we have only one unknown parameter *R* Let *S* indicate that the two-locus combination of interest shows significant association. As described in the previous section, we use a two-step approach for the two-locus analysis. A significant association of the two-locus combination from our two-step method means that each of the two loci shows significant marginal association at level α_1 in step 1 and significant joint association at level α_2 in step 2. We calculate the likelihood $L(R)$ using the equation

$$L(R) = \Pr(D \mid S, R) = \frac{\Pr(S|D,R)\Pr(D|R)}{\Pr(S|R)} = \frac{\Pr(D|R)}{\Pr(S|R)},$$

where the data $D = \{n_1,..., n_9; m_1,..., m_9\}$. Since the data D constitutes, by definition, a significant result, so D implies S; hence $\Pr(S|D,R) = 1$. If the value of $L(R)$ can be calculated for each given *R*, we can obtain the MLE of *R* by using a numerical optimization method (grid search was used in this article). For each *R*, the numerator can be calculated by the product of two multinomial distributions

$$\Pr(D \mid R) = n! \prod_{k=1}^{9} (p_k^{n_k} / n_k!) \bullet m! \prod_{k=1}^{9} (q_k^{m_k} / m_k!),$$

where $p_k = \frac{\alpha}{F} q_k$ if the k^{th} genotype is a low-risk genotype; $p_k = \frac{\beta}{F} q_k$ otherwise. The traditional method to estimate the relative risk is to maximize $\Pr(D|R)$, the numerator in the likelihood function $L(R)$, without considering the fact that the loci were significant for association. There is no simple method to calculate the denominator $\Pr(S|R)$, the power of our two-step test. We propose to use a simulation method as described below. For a given *R*, the values of α and β can be calculated by equation (1). When α, β, and q_i are known, we can generate the two-locus genotypes for *n* cases and *m* controls. Next, we will perform the single-marker test and the two-locus test on the data set. If the p-values of the two single-marker tests are less than α_1 and the p-value of the two-locus test is less than α_2, the data set is said to be significant for association. We repeat the process to generate the data sets many times (1 million was used in this article). The proportion of significant data sets is the estimate of $\Pr(S|R)$.

When the relative risk *R* has been estimated, the corresponding estimates of α and β can be obtained from equation (1). The estimate of odds ratio of the high-risk genotype group is given by $\frac{\beta}{1-\beta} \cdot \frac{1-\alpha}{\alpha}$.

Following Zollner and Prichard, when there are more than two genotype groups in the models such as these in Figure 2, we define the odds ratio of one group to be the odds of this group divided by the odds of the combination of the others. For example, there are three genotype groups in the Dom × Dom model: low risk genotype group G_L = {aabb}, middle risk genotype group G_M = {aabB, aaBB, aAbb, AAbb}, and high risk genotype group G_H = {aAbB, aABB, AAbB, AABB}. The odd ratio of the high risk group OR^H is the odds of G_H divided by the odds of $G_M \cup G_L$ = {aabb, aabB, aaBB, aAbb, AAbb}. The odd ratio of the low risk genotype group OR^L is the odds of G_L divided by the odds of $G_M \cup G_H$ = {aabB, aaBB, aAbb, AAbb, aAbB, aABB, AAbB, AABB}. The odds ratio estimation method will be the same as the case of two genotype groups.

We used this new proposed method to estimate the odds ratio for each of the two-locus combinations that showed significant association with ALS in our two-locus analysis. Based on the estimated penetrance, we used a simulation method to estimate the sample size required to replicate the findings with 80% power.

Results

We applied the two-locus analysis with two steps to the genome-wide association data set for sporadic ALS. The analysis was done for all genotypes with a call rate greater than or equal to 95% (549,062 SNPs left). SNPs on the sex chromosome were excluded in the analysis. In the first step, we returned 1,000 SNPs with the smallest p-values which corresponded to use a p-value cut-off $\alpha_1 = 0.0023$. Then we tested all of the $L = 499{,}500$ two-locus combinations under each of the seventeen models and used 1,000 permutations to evaluate the adjusted p-value for each of

Table 1: Information of the three SNPs. HRA: high-risk allele.

SNP	dbSNP ID	Chromosome Location	Gene	Two alleles	Allele frequency		HRA
					Controls	Cases	
				T	0.656	0.505	
SNP1	rs4363506	10q26.13	Intergenic	C	0.344	0.495	C
				A	0.467	0.341	
SNP2	rs3733242	4q21.1	SHROOM3	G	0.533	0.659	G
				C	0.887	0.786	
SNP3	rs16984239	2p24	Intergenic	A	0.113	0.214	A

the two-locus combinations. After adjusting for multiple tests and multiple SNPs, we found two two-locus combinations with p-values less than 0.05. There were three SNPs involved in the two two-locus combinations. The details of the three SNPs are given in Table 1. The combination of SNP1 and SNP2 followed the Dom∩Dom model with a p-value of 0.032 and SNP1 and SNP3 followed the Dom × Dom model with a p-value of 0.042. Table 2 gives the number of cases and controls in each of the nine genotypes for the two two-locus combinations. This table shows that the two two-locus combinations fit the two models, Dom∩Dom and Dom × Dom. For example, for SNP1 and SNP2, there were more cases than controls for genotypes with at least one C allele at SNP1 and at least one G allele at SNP2 and there were more controls than cases for the other genotypes, which indicated that SNP1 and SNP2 followed the Dom∩Dom model. In Schymick et al.'s 2 df single-gene analysis [1], SNP1 was ranked 1st with a p-value of 6.8×10^{-7}, SNP 2 was ranked 10th with a p-value of 2.2×10^{-5}, and SNP 3 was ranked 2nd with a p-value of 1.7×10^{-6}.

To estimate the impact of the two two-locus combinations on sporadic ALS, we first estimated the penetrance of the two-locus genotypes for each of the two two-locus combinations under the corresponding model. Based on the estimated penetrance, we estimated the relative risk, odds ratio and sample size required to replicate the significant findings with 80% power. We followed what is in Zollner and Pritchard to obtain the 95% CI of the estimates [32], that is, we generated 95% CI by comparing the likelihood of all initial parameter points with the likelihood of the point estimate. We included all points for which twice the difference of log-likelihoods was < 95th percentile of a χ^2 distribution with 1 df. The estimations using both the proposed method (adjusted estimates) and the traditional method (unadjusted estimates) are summarized in Table 3. From this table, we can see that the unadjusted relative risk, odds ratio were higher than the adjusted ones, and the unadjusted sample size was smaller than the adjusted one. These results were consistent with the finding of others that the traditional estimates of relative risk and odds ratio are up-biased [33,34].

Discussion

In this study we proposed a new analytical method that considered joint effects of genes to analyze a data set from the GWAS in sporadic ALS previously performed by Schymick et al. [1]. Our analysis showed that the combination of SNP1 and SNP2 and the combination of SNP1 and SNP3 had significant effects on sporadic ALS.

Population stratification may lead to false-positive results. We had also checked the population stratification problem in this data set using the following method. We randomly chose 5,000 SNPs and got their p-values by a single marker test. If population stratification did exist in this

Table 2: (number of cases)/(number of controls) in each of the two-locus genotypes.

SNP	Genotype	SNP1		
		TT	TC	CC
SNP2	AA	11/23	14/37	3/7
	AG	29/50	73/56	29/11
	GG	23/45	65/24	28/16
SNP3	CC	33/95	95/89	37/30
	CA	29/20	52/25	22/4
	AA	1/3	5/3	1/0

Table 3: Penetrence, relative risk and odds ratio of the two-locus combinations.

Two-locus combination		SNP1 and SNP2	SNP1 and SNP3
Penetrance	Unadjusted	*Pen*(G_L^1) = 0.48*F*, *Pen*(G_H^1) = 1.78*F*.	*Pen*(G_L^2) = 0.40*F*; *Pen*(G_M^2) = 1.02*F*; *Pen*(G_H^2) = 2.60*F*.
	Adjusted	*Pen*(G_L^1) = 0.51*F*; *Pen*(G_H^1) = 1.73*F*.	*Pen*(G_L^2) = 0.44*F*; *Pen*(G_M^2) = 1.03*F*; *Pen*(G_H^2) = 2.43*F*.
R and 95% CI	Unadjusted	3.70, (2.85, 4.85)	2.55, (2.10, 3.15)
	Adjusted	3.40, (2.40, 4.60)	2.35, (1.85, 2.95)
OR^H and 95% CI	Unadjusted	3.70, (2.85, 4.85)	3.37, (2.66, 4.34)
	Adjusted	3.40, (2.40, 4.60)	3.05, (2.27, 4.01)
OR^L and 95% CI	Unadjusted	0.27, (0.21,0.35)	0.31, (0.23, 0.40)
	Adjusted	0.29, (0.22, 0.42)	0.34, (0.25, 0.47)
SS and 95% CI	Unadjusted	680, (480, 1040)	680, (460, 1040)
	Adjusted	800, (500, 1500)	810, (520, 1520)

Note: There were two genotype combinations for SNP1 and SNP2, G_L^1 = {TTAA, TCAA, CCAA, TTAG, TTGG} and G_H^1 = {TCAG, CCAG, TCGG, CCGG}, three genotype combinations for SNP1 and SNP3, G_L^2 = {TTCC}, G_M^2 = {TCCC, CCCC, TTCA, TTAA} and G_H^2 = {TCCA, CCCA, TCAA, CCAA}. *Pen*(*G*) denotes the penetrance of G. R: relative risk. For SNP 1 and SNP2, $R = pen(G_H^1)/pen(G_L^1) = \alpha/\beta$; for SNP1 and SNP3, $R = pen(G_M^2)/pen(G_L^2) = pen(G_H^2)/pen(G_M^2) = \theta$. OR^H (OR^L): the odds ratio of the high-risk (low-risk) genotype group. SS: the sample size required to reach 80% power. Adjusted (Unadjusted): based on the penetrance estimated using the method proposed in this article (the traditional method). *F* is the prevalence and $F = 10^{-4}$.

data set, among the 5,000 p-values, there should be more small p-values than expected under the uniform distribution. We used the one-side Kolmorgorov test statistic to test if the 5,000 p-values followed a uniform distribution. We repeated the procedure 10 times. The Kolmorgorov test results showed that the p-values followed a uniform distribution for all 10 replications, which indicated that there was no population stratification in this data set. The lack of population stratification in the data set was consistent with the results of Schymick et al. [1]. Schymick et al. studied the potential population structure in this data by using STRUCTURE program [38]. The analysis with STRUCTURE showed that there was no discernible difference in the population substructure between cases and controls.

Significant associations claimed by association studies often fail to be replicated. One possible reason is the overestimation of the effect in terms of the odds ratio or relative risk of the claimed variants. The overestimation of the effect leads to the underestimation of the sample size required to replicate the finding. In this article, we proposed a new method to estimate the effect of claimed variants. Based on the study of Zollner and Pritchard [32], we expected that the estimates of odds ratio and relative risk based on our proposed method would be nearly unbiased. Thus we provided a useful tool to estimate the sample size for the follow up studies. For example, in order to replicate the finding of SNP1 and SNP2 (the adjusted p-value less than 0.05 under the Dom $\cap$ Dom model) with 80% power, the sample size required is 800 estimated

using our proposed method instead of 680 estimated using the traditional method.

Currently, several methods are available to test associations by taking joint effects of genes into account, such as combinatorial searching method (CSM) and the multifactor dimensionality reduction (MDR) method [39,40]. We used the two-step CSM and MDR, replacing the two-locus analysis test in step 2 by the CSM or MDR, to perform the two-locus analysis. For the two-step MDR, we returned 50 SNPs instead of 1, 000 SNPs in the first step due to the computational intensity. Both of the two-step CSM and MDR found rs4363506 (SNP1) and rs12680546 (on chromosome 8) as the best two-locus combination. However, the adjusted p-values of the two-step CSM and MDR were 0.2 and 0.156. This means that the two-step CSM and MDR did not find any two-locus combinations that had significant association with sporadic ALS. The possible reasons are as follows: The genotypes of the two-locus combinations we found (such as those given in Table 3) are ordered. For example, penetrance of $H_1H_2 \geq$ penetrance of $H_1h_2 \geq$ penetrance of h_1h_2, where $H_1(h_1)$ and $H_2(h_2)$ are the high-risk (low-risk) genotypes in the first and second marker, respectively. The CSM and MDR ignore the order of genotypes and therefore can group any two genotypes together-in essence searching for the "best" one among 21,146 different partitions of the two-locus genotypes. By searching for irrelevant two-locus genotype combinations, the CSM and MDR did not gain more information but increased the noise level, and thus lost power.

Conclusion

The proposed two-stage analytical method can be used to search for two-locus joint effects of genes in GWAS. The two-stage strategy significantly decreased the computational time and the multiple testing burdens associated with GWAS. We have also observed that the three SNPs identified by our two-stage strategy can not be detected by single-locus tests.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

QS and SZ designed the study. ZZ contributed the two-locus data analysis under the direction of SZ. SZ performed the penetrance estimation. JCS & BJT assisted in data interpretation and approved the final manuscript. QS and SZ contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the National Institute of Health (NIH) grants R01 GM069940 and the Overseas-Returned Scholars Foundation of Department of Education of Heilongjiang Province (1152 HZ01). This work was supported in part by the Intramural Research Program of the National Institute on Aging (project Z01 AG000949-02).

References

1. Schymick JC, Scholz SW, Fung HC, Britton A, Arepalli S, Gibbs JR, Lombardo F, Matarin M, Kasperaviciute D, Hernandez DG, Crews C, Bruijn L, Rothstein J, Mora G, Restagno G, Chiò A, Singleton A, Hardy J, Traynor BJ: **Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data.** *Lancet Neurol* 2007, **6:**322-28.
2. Monk PN, Shaw PJ: **ALS: life and death in a bad neighborhood.** *Nat Med* 2006, **12:**885-87.
3. The Scottish Motor Neuron Disease Register: **A prospective study of adult onset motor neuron disease in Scotland: methodology, demography and clinical features of incident cases in 1989.** *J Neurol Neurosurg Psychiatry* 1992, **55:**536-41.
4. Traynor BJ, Codd MB, Corr B, Forde C, Frost E, Hardiman O: **Incidence and prevalence of ALS in Ireland, 1995-1997: a population-based study.** *Neurology* 1999, **52:**504-09.
5. Logroscino G, Beghi E, Zoccolella S, Palagano R, Fraddosio A, Simone IL, Lamberti P, Lepore V, Serlenga L, SLAP Registry: **Incidence of amyotrophic lateral sclerosis in southern Italy: a population based study.** *J Neurol Neurosurg Psychiatry* 2005, **76:**1094-98.
6. Incidence of ALS in Italy: **evidence for a uniform frequency in Western countries.** *Neurology* 2001, **56:**239-44.
7. Rosen DR, Siddique T, Patterson D, Figlewicz DA, Sapp P, Hentati A, Donaldson D, Goto J, O'Regan JP, Deng H, Rahmani Z, Krizus A, McKenna-Yasek D, Cayabyab A, Gaston SM, Berger R, Tanzi RE, Halperin JJ, Herzfeldt B, Bergh RV, Hung W, Bird T, Deng G, Mulder DW, Smyth C, Laing NG, Soriano E, Pericak-Vance MA, Haines J, Rouleau GA, Gusella JS, Horvitz HR, Brown RH Jr: **Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis.** *Nature* 1993, **362:**59-62.
8. Puls I, Jonnakuty C, La Monte BH, Holzbaur EL, Tokito M, Mann E, Floeter MK, Bidus K, Drayna D, Oh SJ, Brown RH Jr, Ludlow CL, Fischbeck KH: **Mutant dynactin in motor neuron disease.** *Nat Genet* 2003, **33:**455-56.
9. Münch C, Sedlmeier R, Meyer T, Homberg V, Sperfeld AD, Kurt A, Prudlo J, Peraus G, Hanemann CO, Stumm G, Ludolph AC: **Point mutations of the p150 subunit of dynactin (DCTN1) gene in ALS.** *Neurology* 2004, **63:**724-26.
10. Hadano S, Hand CK, Osuga H, Yanagisawa Y, Otomo A, Devon RS, Miyamoto N, Showguchi-Miyata J, Okada Y, Singaraja R, Figlewicz DA, Kwiatkowski T, Hosler BA, Sagie T, Skaug J, Nasir J, Brown RH Jr, Scherer SW, Rouleau GA, Hayden MR, Ikeda JE: **A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2.** *Nat Genet* 2001, **29:**166-73.
11. Chen YZ, Bennett CL, Huynh HM, Blair IP, Puls I, Irobi J, Dierick I, Abel A, Kennerson ML, Rabin BA, Nicholson GA, Auer-Grumbach M, Wagner K, De Jonghe P, Griffin JW, Fischbeck KH, Timmerman V, Cornblath DR, Chance PF: **DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS).** *Am J Hum Genet* 2004, **74:**1128-35.
12. Nishimura AL, Mitne-Neto M, Silva HC, Richieri-Costa A, Middleton S, Cascio D, Kok F, Oliveira JR, Gillingwater T, Webb J, Skehel P, Zatz M: **A mutation in the vesicle-traffi cking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis.** *Am J Hum Genet* 2004, **75:**822-31.
13. Greenway MJ, Alexander MD, Ennis S, Traynor BJ, Corr B, Frost E, Green A, Hardiman O: **A novel candidate region for ALS on chromosome 14q11.2.** *Neurology* 2004, **63:**1936-38.
14. Greenway MJ, Andersen PM, Russ C, Ennis S, Cashman S, Donaghy C, Patterson V, Swingler R, Kieran D, Prehn J, Morrison KE, Green A, Acharya KR, Brown RH Jr, Hardiman O: **ANG mutations segregate with familial and sporadic amyotrophic lateral sclerosis.** *Nat Genet* 2006, **38:**411-13.
15. Veldink JH, Kalmijn S, Hout AH Van der, *et al.*: **SMN genotypes producing less SMN protein increase susceptibility to and severity of sporadic ALS.** *Neurology* 2005, **65:**820-825.
16. Al-Chalabi A, Andersen PM, Nilsson P, Chioza B, Andersson JL, Russ C, Shaw CE, Powell JF, Leigh PN: **Deletions of the heavy neurofilament subunit tail in amyotrophic lateral sclerosis.** *Hum Mol Genet* 1999, **8:**157-164.
17. Lambrechts D, Storkebaum E, Morimoto M, Del-Favero J, Desmet F, Marklund SL, Wyns S, Thijs V, Andersson J, van Marion I, Al-Chalabi

A, Bornes S, Musson R, Hansen V, Beckman L, Adolfsson R, Pall HS, Prats H, Vermeire S, Rutgeerts P, Katayama S, Awata T, Leigh N, Lang-Lazdunski L, Dewerchin M, Shaw C, Moons L, Vlietinck R, Morrison KE, Robberecht W, Van Broeckhoven C, Collen D, Andersen PM, Carmeliet P: **VEGF is a modifier of amyotrophic lateral sclerosis in mice and humans and protects motoneurons against ischemic death.** *Nat Genet* 2003, **34:**383-394.

18. Saeed M, Siddique N, Hung WY, Usacheva E, Liu E, Sufit RL, Heller SL, Haines JL, Pericak-Vance M, Siddique T: **Paraoxonase cluster polymorphisms are associated with sporadic ALS.** *Neurology* 2006, **67:**771-776.
19. Slowik A, Tomik B, Wolkow PP, Partyka D, Turaj W, Malecki MT, Pera J, Dziedzic T, Szczudlik A, Figlewicz DA: **Paraoxonase gene polymorphisms and sporadic ALS.** *Neurology* 2006, **67:**766-770.
20. Shaw CE, Al-Chalabi A: **Susceptibility genes in sporadic ALS Separating the wheat from the chaff by international collaboration.** *Neurology* 2006, **67:**738-739.
21. Dunckley T, Huentelman MJ, Craig DW, Pearson JV, Szelinger S, Joshipura K, Halperin RF, Stamper C, Jensen KR, Letizia D, Hesterlee SE, Pestronk A, Levine T, Bertorini T, Graves MC, Mozaffar T, Jackson CE, Bosch P, Mc Vey A, Dick A, Barohn R, Lomen-Hoerth C, Rosenfeld J, O'connor DT, Zhang K, Crook R, Ryberg H, Hutton M, Katz J, Simpson EP, Mitsumoto H, Bowser R, Miller RG, Appel SH, Stephan DA: **Whole-genome analysis of sporadic amyotrophic lateral sclerosis.** *N Engl J Med* 2007, **357:**775-788.
22. van Es MA, Van Vught PW, Blauw HM, Franke L, Saris CG, Andersen PM, Bosch L Van Den, de Jong SW, van 't Slot R, Birve A, Lemmens R, de Jong V, Baas F, Schelhaas HJ, Sleegers K, Van Broeckhoven C, Wokke JH, Wijmenga C, Robberecht W, Veldink JH, Ophoff RA, Berg LH van den: **ITPR2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: a genome-wide association study.** *Lancet Neurol* 2007, **6:**869-877.
23. Cronin S, Berger S, Ding J, Schymick JC, Washecka N, Hernandez DG, Greenway MJ, Bradley DG, Traynor BJ, Hardiman O: **A genome-wide association study of sporadic ALS in a homogenous Irish population.** *Hum Mol Genet* 2008, **17:**768-774.
24. van Es MA, van Vught PW, Blauw HM, Franke L, Saris CG, Bosch L Van den, de Jong SW, de Jong V, Baas F, van't Slot R, Lemmens R, Schelhaas HJ, Birve A, Sleegers K, Van Broeckhoven C, Schymick JC, Traynor BJ, Wokke JH, Wijmenga C, Robberecht W, Andersen PM, Veldink JH, Ophoff RA, Berg LH van den: **Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis.** *Nat Genet* 2008, **40:**29-31.
25. De Miglio MR, Pascale RM, Simile MM, Muroni MR, Virdis P, Kwong KM, Wong LK, Bosinco GM, Pulina FR, Calvisi DF, Frau M, Wood GA, Archer MC, Feo F: **Polygenic control of hepatocarcinogenesis in Copenhagen # F344 rats.** *Int J Cancer* 2004, **111:**9-16.
26. Yanchina ED, Ivchik TV, Shvarts EI, Kokosov AN, Khodzhayantz NE: **Gene-gene interactions between glutathione-s transferaseM1 and matrix metalloproteinase 9 in the formation of hereditary predisposition to chronic obstructive pulmonary disease.** *Bull Exp Biol Med* 2004, **137:**64-66.
27. Yang P, Bamlet WR, Ebbert JO, Taylor WR, de Andrade M: **Glutathione pathway genes and lung cancer risk in young and old populations.** *Carcinogenesis* 2004, **25:**1935-1944.
28. Aston CE, Ralph DA, Lalo DP, Manjeshwar S, Gramling BA, DeFreese DC, West AD, Branam DE, Thompson LF, Craft MA, Mitchell DS, Shimasaki CD, Mulvihill JJ, Jupe ER: **Oligogenic combinations associated with breast cancer risk in women under 53 years of age.** *Hum Genet* 2005, **116:**208-221.
29. Dong C, Li WD, Li D, Price RA: **Interaction between obesity-susceptibility loci in chromosome regions 2p25-p24 and 13q13-q21.** *Eur J Hum Genet* 2005, **13:**102-108.
30. Roldan V, Gonzalez-Conejero R, Marin F, Pineda J, Vicente V, Corral J: **Five prothrombotic polymorphisms and the prevalence of premature myocardial infarction.** *Haematologica* 2005, **90:**421-423.
31. Millstein J, Conti DV, Gilliland FD, W James Gauderman WJ: **A testing framework for identifying susceptibility genes in the presence of epistasis.** *Am J Hum Genet* 2006, **78:**15-27.
32. Zollner S, Pritchard J: **Overcoming the winner's curse: estimating penetrance parameters from case-control data.** *Am J Hum Genet* 2007, **80:**605-615.
33. Garner C: **Upward bias in odds ratio estimates from genome-wide association studies.** *Genetic Epi* 2007, **31:**288-295.
34. Xiong M, Zhao J, Boerwinkle E: **Generalized T² test for genome association studies.** *Am J Hum Genet* 2003, **70:**1257-1268.
35. Zhao J, Jin L, Xiong M: **Test for Interaction.** *Am J Hum Genet* 2006, **79:**831-845.
36. Hodge SE: **Some epistatic two-locus models of disease. I. Relative risks and identity-by-descent distributions in affected sib pairs.** *Am J Hum Genet* 1981, **33:**381-395.
37. Risch N: **Linkage strategies for genetically complex traits. I. Multilocus models.** *Am J Hum Genet* 1990, **46:**222-228.
38. Falush D, Stephens M, Pritchard JK: **Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.** *Genetics* 2003, **164:**1567-87.
39. Sha Q, Zhu X, Zuo Y, Cooper R, Zhang S: **A combinatorial searching method for detecting a set of interacting loci associated with complex traits.** *Ann Hum Genet* 2006, **70:**677-692.
40. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: **Multifactor-Dimensionality reduction reveals high-order interactions among Estrogen-Metabolism genes in sporadic breast cancer.** *Am J Hum Genet* 2001, **69:**138-147.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2350/10/86/prepub