Statistical Analysis of the TSAD Interactome in Multiple Sclerosis: Multiple Testing, High Dimensional Regression and Interactions

Azimach Ginjo Girmma Woldeselassie Azige Alito
Department of Statistics, College of Natural and Computational Sciences, Wolaita Sodo University, P.O. Box 138, Wolaita Sodo Ethiopia.

Abstract
Multiple Sclerosis (MS) is a chronic inflammatory disease of the Central Nervous System (CNS). The presence of Oligoclonal Band (OCB) in Cerebrospinal Fluid (CSF) is an important diagnostic tool in MS. The main aim of this study was to determine SNPs and SNPs-SNPs interactions in the genomic TSAD (T-Cell Specific Adaptor Proteins) region which explain the difference between two MS conditions: OCB positive vs. OCB negative in sampled Norwegian patients. The data to this study was obtained from the MS Registry Ulleval University Hospital, Oslo, Norway. Of 899 patients, 802 were OCB positive and 97 OCB negative, each has 923 SNPs at specific position in their chromosome measure. The study incorporated two different statistical methods to our data analysis. First, we apply variable selection based on Lasso method, here we discuss Lasso for Logistic Regression Analysis and see interaction effect for the Lasso selected SNPs. In the second section, we analyze variable selection based on test of association. Here, we used Chi-Square and Fisher's exact test of association to see association between the statuses of OCB to each SNP. We found out that the Chi-Square test of association selected 34 significant SNPs and the association test for Fisher's exact test selected 38 significant SNPs at significance level of 0.05. Then we used Boferroni and False Discovery Rate to statistically significant SNPs for the multiple testing corrections. Finally, we looked for interaction effects to some selected SNPs from test of association and we have determined SNPs and SNP-SNP interactions which appear to have significant associated to the OCB subpopulation of MS patients, based on our study of the Norwegian cohort. These are selected SNPs which have been selected by any of the methods that we used (various hypothesis testing and regressions). This is the SNPs that we think should be studied further and validated on a new dataset. Of particular importance are the seven SNP-SNP interactions which we found. It is the first time a SNP-SNP study has been performed on these data, and the finding will be communicated to the Norwegian molecular biologists to be followed up.

Keywords: Multiple Sclerosis, OCB, TSAD, Multiple Testing, High Dimensional Regressions, LASSO
DOI: 10.7176/JMPB/61-01
Publication date: November 30th 2019

1. Introduction

1.1 Background

Multiple sclerosis (MS), also known as disseminated sclerosis or encephalomyelitis disseminate, is an inflammatory disease in which the fatty myelin sheaths around the axons of the brain and spinal cord are damaged, leading to demyelination and scarring as well as a broad spectrum of signs and symptoms (Coles et al, 2008)[1]. In patients with multiple sclerosis, myelin, the fatty tissue which protects nerve fibers in the CNS, deteriorates leading to neurological problems and disorders of the CNS.

Patients with multiple sclerosis suffer from a variety of symptoms, including visual problems, muscle weakness, difficulties with coordination and speech, fatigue, cognitive impairment, problems with balance, and pain. Multiple sclerosis is a debilitating disease which leads to impaired mobility and disability in most cases. It appears that multiple sclerosis is a modern day disease, but there has been mention of symptoms associated with the disease as far back as the 14th century. In 1947, the skeleton of St. Lidwina of Schiedam, which is a suburb of Rotterdam, Holland, was discovered during a reconstruction project.

It appeared that Lidwina, who lived from 1380 to 1433 had a case of MS. The skeleton showed signs of muscular atrophy and paralysis of both legs. It was recorded that Lidwina suffered from difficulty walking and a "hanging lip" which could possibly be indicative of facial paralysis. It was also recorded that in the last 15 years of her life, Lidwina had large wounds on her body, difficulty swallowing, blindness, and other pains (Murray, 2005)[2]. Disease onset usually occurs in young adults, and it is more common in women (Coles et al, 2008) [1]. MS was first described in 1868 by Jean-Martin Charcot (Clanet, 2008)[3]. MS affects the ability of nerve cells in the brain and spinal cord to communicate with each other effectively. Nerve cells communicate by sending electrical signals called action potentials down long fibers called axons, which are contained within an insulating substance called myelin. In MS, the body own immune system attacks and damages the myelin. When myelin is lost, the axons can no longer effectively conduct signals (Coles et al, 2002)[4]. The name multiple sclerosis refers to scars (sclerae-better known as plaques or lesions) particularly in the white matter of the brain and spinal cord, which is mainly composed of myelin (Clanet, 2008)[3]. Although much is known about the mechanisms involved in the disease process, the cause remains unknown. Theories include genetics or infections.

Different environmental risk factors have also been found (Coles et al, 2002)[4] and Munger et al, 2007)[5]. Almost any neurological symptom can appear with the disease, and the disease often progresses to physical and cognitive disability (Coles et al, 2002)[4]. MS takes several forms, with new symptoms occurring either in discrete attacks (relapsing forms) or accumulating over time (progressive forms) (Reingold et al, 1996)[6]. Between attacks, symptoms may go away completely, but permanent neurological deficits often occur, especially as the disease advances (Reingold et al, 1996)[6]. There is no known cure for multiple sclerosis. Treatments attempt to return function after an attack, prevent new attacks, and prevent disability (Coles et al, 2002)[4].

MS medications can have adverse effects or be poorly tolerated, and many people pursue alternative treatments, despite the lack of supporting scientific study. The prognosis is difficult to predict; it depends on the subtype of the disease, the individual's disease characteristics, the initial symptoms and the degree of disability the person experiences as time advances (Weinshenker, 1994)[7]. Life expectancy of people with MS is 5 to 10 years lower than that of the unaffected population (Coles et al, 2008)[1]. The clinical manifestations typically first develop in young adults as acute relapses, and then evolve into a gradually progressive course with permanent disability after 10-15 years. It is a complex disease influenced by many factors rather than driven by a single cause. Research into the genetics of MS therefore involves the search for genes that contribute to susceptibility and/or to the severity and other aspects of the disease. More recently, genetic research has extended into the study of inherited variations in response to treatment (pharmacogenetics). For many years it has been evident that close family members of a person with MS have a higher risk of having the disease, and the closer they are genetically, the higher the risk. Unrelated family members (such as husband or wife) show no increased risk but the children of marriages where both parents have MS have a particularly high risk. A large study of people with MS who were adopted under the age of one year clearly showed that risk is largely due to genetic factors rather than the environment MSIF (2007)[8]. Relatives of patients with MS are at increased risk for the disease and several lines of evidence indicate that MS has a genetic component. However, the genetic basis of MS is complex (i.e., multiple genes contribute cumulatively to the risk of MS and disease behavior) and is heterogeneous (i.e., the genes and alleles involved probably differ from patient-to-patient). The prevalence of MS is approximately 1/1000, affecting approximately 400, 000 people in the US and 2 million worldwide (Marrie, 2004)[9]. It is the most common non-traumatic cause of neurologic disability in young adults. Onset typically is between the ages of 20 and 40 years, and women are affected more frequently than men (or approximately 3:2). The presence of OCB in CSF is an important diagnostic tool in MS and is thought to reflect a local B-cell response of unknown specificity and significance (Poser et al, 1983, McDonald et al, 2001 and Polman et al, 2010)[10].

The single most consistent laboratory Abnormality in patients with MS exclusive of magnetic resonance imaging is increased oligoclonal immunoglobuns in cerebrospinal fluid (Paty, 1988)[11]. In patients with a single demyelinating episode, detection of intrathecal immunoglobunin synthesis may predict progression to MS (Paolino, 1996)[12], and oligoclonal band (OCBs) in CSF during the early phase of disease are associated with a worse outcome (Amato, 2000)[13]. Up to 95% of MS patients in Northern Europe have OCB in the CSF, but this frequency varies de- pending on laboratory routines, study populations and was recently also related to latitude (Link and Huang, 2006 and Andersson et al, 1994)[14]. (Joseph, 2008)[15] in his result to case control study on 100 MS patients sampled from southwest England and south Wales, indicated that an approximate minimum 3% of patients with MS were OCB negative. They were significantly more likely to exhibit neurological or systemic clinical features a typical of MS (headaches, neuropsychiatric features and skin changes). Non-specific MRI (Magneti Resonance Imaging), blood and (other) CSF abnormalities were also more common, emphasizing the need for continued diagnostic vigilance, although the incautious application of McDonald diagnostic criteria in OCB negative cases renders categorization as definite, MS more likely.

1.2 Statement of the Problem

Multiple sclerosis is an autoimmune disease which has a high and growing incidence in many parts of the world. It is a chronic or relapsing disease, and incurable. Two and half million people worldwide are affected by MS (Kerstin, 2009)[16].

Biologists believe that inheritable diseases like cancer or multiple sclerosis can be recognized or predicted by looking to SNPs (or Single nucleotide polymorphisms) of people which is single base positions in the human DNA which are mutated (polymorphic), that is different in some individuals with respect to the whole population. Ideally, a set of SNPs can be the cause of the disease. There is, indeed, a genetic susceptibility to multiple sclerosis which originated in Scandinavia (high prevalence occurred), and the susceptibility alleles have been transmitted to other races by gene flow (Gunderson et al, 2007)[17]. So, in this study we are aimed to look for Statistical analysis of the Tsad interactome in multiple sclerosis:

Research Questions

- Can we use some of the SNPs which are associated to OCB conditions in genomic TSAD region, so that they will allow a classification of future patients in the OCB conditions of MS?
- Do SNPs in the genomic TSAD region have significant association with the OCB condition?
- Can we find interactions of SNPs in the genomic TSAD region which are associated to the OCB condition, and therefore describe first elements of the biological mechanisms of the OCB condition in MS?

1.3 Objectives of the Study

The general objective of this study is to identify SNPs and SNPs-SNPs interactions in the genomic TSAD region which explain the difference between two MS conditions: OCB positive vs. OCB negative.

Specific Objectives the Study

- To apply logistic regression and fit the model with lasso penalty to determine SNPs in the genomic TSAD region which are associated to the status of OCB so that they will allow a classification of future patients in the OCB conditions of MS.
- To investigate genetic factors of the genomic TSAD region which important for the disease condition of OCB in Multiple Sclerosis with different kind of statistical hypothesis (Fisher Exact and Chi-Square test of association).
- To assess which SNPs in the TSAD region have significant association with the OCB condition, using multiple testing correction method: Bonferroni and False Discovery Rate.
- To identify SNP-SNP interactions in the genomic TSAD region which are associated to the OCB condition, and therefore describe first elements of the biological mechanisms of the OCB condition in MS.

2. Methodology

2.1 Description of the Data

The data for this study was secondary data collected from two Norwegian samples: The Oslo MS DNA Bio Bank and The Norwegian MS Registry and Bio Bank held in Bergen. In the Oslo MS DNA Bio Bank the majority of patients are recruited by the neurologists at Oslo University Hospital, Ulleval with the remainder coming from local MS societies and other neurological departments are serving the suburban Oslo areas. Samples in the Norwegian MS Registry and Bio bank were recruited from all other parts of Norway. This collection started in 2007, and currently includes approximately 1/5 of the prevalent MS patients in Norway.

2.2 Description of Variables Considered under the Study

In this study, we categorized the status of a total of 899 MS patient into two groups. Of these 802 persons are OCB positive and 97 persons OCB negative. OCB are proteins found in the spinal fluid of MS patients. They are called oligoclonal bands, because they appear in the part of the electrophoresis of the spinal fluid where the gamma globulins are found. It is thought to be produced by single clones of B-cells located within the CNS of the MS patients. The origin of the OCB is not understood, but they are viewed as quite specific for MS, although not all MS patients do have these bands. It is important to try to find a genetic signature of the OCB positive versus OCB negative MS patients. This will allow a better understanding of the diseases and also a possibility to predict the long term prognoses. For each person, we have used 923 SNPs (or Single Nucleotide Polymorphism) in 923 specific positions of their DNA. These 923 SNPs have been chosen because they are located in, or around, genes on the DNA who have to do with an important protein, called T cell specific adapter protein (TSAd), which is believed to have a very important role in development of MS. This means that these SNPs have some good chances to be relevant in the distinction between OCB and not OCB, as they reside on genes which biologists believe are important in MS. T cells play a crucial role in our defense against infection and cancer, but these cells are also dangerous sometimes, as they can start to react against the body they are supposed to defend, leading to autoimmune disease, like MS. This probably happens because these cells are triggered in some way. We do not know exactly how and why. The last 10 years many research groups worldwide have therefore focused primarily on elucidating molecular mechanisms for control of T cell activation, and in particular the role of T-cell specific adapter protein (TSAd) in this process. One group is led by Professor Anne Spurkland, and she has selected for us these 923 SNPs.

The importance of TSAd is described in many papers. We cite here the ones of the group of Professor Spurkland. They have found that TSAd modulates early signaling events in T cells (Kolltveit et al, 2008)[18]. Spurkland has found that a certain gene participating to the production of TSAd, is associated with increased susceptibility to MS and juvenile arthritis (Lorentzen et al, 2008)[19]. Genetically determined variation in the expression level of TSAd may provide a mechanism for how TSAd contribute to genetic susceptibility to autoimmune disease. Each person has 923 SNP (or Single Nucleotide Polymorphism) at specific position in their chromosome measure. Each patient has age at onset 33.7 years, old age of diagnosis 48.5 years clinical course and patients with primary progressive symptoms: 13%.

2.2.1 The Response Variable of the Study

The outcome of interest in this study is the status of OCB, where OCB is categorized into OCB positive and OCB negative i.e., the outcome is categorical response with two categories. Hence, the response variable for the ith person is represented by the random variable Yi with two possible values coded as 1 for the success OCB ”positive” and ”0” otherwise Mathematically, we write this as:

$$\textbf{Yi}= \begin{cases} \mathbf{1}, \boldsymbol{if} \ \ \textbf{the ith person is OCB positive} \\ \mathbf{0}, \qquad\qquad\qquad\quad \textbf{otherwise}, \end{cases} \qquad \textbf{(1)}$$

2.2.2 Independent Variables of the Study

Each person in this study has **923 SNPs** (or Single Nucleotide Polymorphism) at specific position in their chromosome measure. The layout for these **923 SNPs** (explanatory variables) and their coding is given below. The unavailable data value is coded as ”**NA**” and each **SNP** has three allele sequence in their **DNA**. That is **AA TA** or **AT** and **TT**. For the sake of analysis it was coded as 0 for **AA** 1 for **AT** or **TA** and 2 for **TT**. Therefore, the summary in the following table 1 shows the coding of this predictor variable.

Table 1: Independent Variables and their coding

Variables	**Coding**
SNP1	**AA = 0, AT,TA = 1, TT = 2 AA =**
SNP2	**0, AT,TA = 1, TT = 2 AA = 0,**
SNP3	**AT,TA = 1, TT = 2**
⋮	⋮
SPk **(k≤923)**	**AA = 0, AT,TA = 1, TT = 2**

2.3 Method of Data Analysis

2.3.1 Fitting the Logistic Regression Model to Data

As in multiple regression analysis, there are two important stages in the analysis of data. First, estimation of the parameters in the model must be obtained. Second, some determination must be made of how well the model actually fits the observed data. In multiple regression analysis, the parameter estimates are obtained using the least-squares principle and assessment of fit is based on significance tests for the regression coefficients as well as on interpreting the multiple correlation coefficients. But in LRA, the parameters that must be estimated from the available data are the constant,α, and the logistic regression coefficients, βj. Because of the nature of the model, estimation is based on the maximum likelihood principle rather than on the least-squares principle. Maximum likelihood estimation (MLE) is the standard method of estimating the unknown parameters in a logistic regression model. This method yields values for the unknown parameters which maximize the probability of obtaining the observed response values. The likelihood function expresses the probability of the observed response values as a function of the unknown parameters. In the context of logistic regression analysis, maximum likelihood estimation (MLE) involves the following. First, we define the likelihood, L(parameter|data), of the sample data as the product, across all sampled cases, of the probabilities for success or for failure:

$$\textbf{L(parameter|data)} \ = \prod_{i=1}^{n} \boldsymbol{P(Y_i, |X_{i1}, \cdots X_{ip})}$$

$$= \prod_{i=1}^{n} \left[\left(\frac{e^{\alpha + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\alpha + \sum_{j=1}^{p} \beta_j X_j}} \right)^{Y_i} \left(\frac{1}{1 + e^{\alpha + \sum_{j=1}^{p} \beta_j X_j}} \right)^{1-Y_i} \right] \qquad \textbf{(1)}$$

Note that Y is the 0/1 outcome for the ith case and, Xi1,⋯,Xip are the values of the predictor variables for the ith case based on a sample of n observations. The use of Yi and 1-Yi as exponents in the equation above includes in the likelihood the appropriate probability term dependent upon whether Yi =1 or Yi =0. Using the methods of calculus, a set of values for and the can be calculated that maximize L(parameter|data) and these resulting values are known as maximum likelihood estimates (MLE’s). This maximization process is somewhat more complicated than the corresponding minimization procedure in multiple regression analysis for finding least-square estimates. However, the general approach involves establishing initial guesses for the unknown parameters and then continuously adjusting these estimates until the maximum value of L(parameter|data) is found. This iterative solution procedure is available in statistical packages. The usefulness of the model as a whole can be assessed by testing the hypothesis that, simultaneously, all of the partial logistic regression coefficients are 0; i.e., $\mathbf{H}\colon \boldsymbol{\beta}\mathbf{j} = \mathbf{0}$ for all $\mathbf{j}$. In effect, we can compare the general model given above with the restricted model ln($\frac{\pi}{1-\pi}$) =$\boldsymbol{\alpha}$. This test, that is equivalent to testing the significance of the multiple R in multiple regression analysis, is based on a chi-squared statistic (R software calculates the value of ”Chi-Square”). Finally, different logistic regression coefficients models fitted to the same set of data can be compared statistically in a simple manner if the models are hierarchical. The hierarchy principle requires that the model with the larger number of predictors include among its predictors all of the predictors from the simpler model. Given this condition, the difference in model chi-squared values is (approximately) distributed as chi-squared with degrees of freedom equal to the difference in degrees of freedom for the models. In effect, this procedure tests a conditional null hypothesis. If the models are specified, R software calculates Chi-square value for each model and this can be used to test whether or not the additional predictors result in significantly better fit of the model to the data.

2.3.2 Variable Selection for Logistic Regression: Lasso (Least absolute shrinkage and selection operator)

A statistical model is a simplification of reality (Agresti, 2007)[20]. At the initial stage of modeling, a large number of candidate predictors are considered to minimize possible modeling biases (Fan and Li, 2006)[21]. However, in most cases, not all the predictors have significant effects on the response variable. In statistics, a result from certain hypothesis testing is called statistically significant if it is unlikely to have occurred by chance. A simpler model that contains only the important predictors is preferred because it is easy to explain. Parsimony is especially important for high dimension data. The parsimony means that the simplest plausible model with the fewest possible number of predictors is desired. Variable selection plays an important role in regression analysis and is intended to select the best subset of predictors. There are typically two competing goals in statistical modeling: The model should be complex enough to fit the data well, and also should be simple to interpret (Agresti, 2007)[20]. In linear regression, parameter estimation by the ordinary least square (**OLS**) method is unbiased. However the estimates may have large variance in some cases, the occurrence of multi-co linearity for instance is one case. With slight sacrifice of bias, ridge regression tends to improve the prediction accuracy by shrinking some coefficients. But ridge regression will not shrink values of any coefficients to exact 0, and the fitted model might be too complex to interpret. In 1996, Tibshirani introduced a different shrinkage method, called the Lasso (least absolute shrinkage and selection operator). This method shrinks values of some coefficients to 0 by a constraint on the sum of absolute values of regression coefficients, so Lasso can serve as a tool for variable selection. The Lasso is a shrinkage method like ridge regression, with subtle but important differences. Like ridge regression, penalizing the absolute values of the coefficients introduces shrinkage towards zero. However, unlike ridge regression, some of the coefficients are shrunken all the way to zero; such solutions, with multiple values that are identically zero, are said to be sparse. The penalty thereby performs a sort of continuous variable selection. The resulting estimator was thus named the Lasso, for ”Least Absolute Shrinkage and Selection Operator” and defined by:

$$\widehat{\boldsymbol{\beta}}^{Lasso} = \underset{\boldsymbol{\beta}}{argmin} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}_0 - \sum_{j=1}^{P} x_{ij}\boldsymbol{\beta}_j \right)^2 \right\} \qquad \textbf{(2)}$$

Subject to $\sum_{j=1}^{P} |\boldsymbol{\beta}_j| \le \boldsymbol{t}$ Where, $\mathbf{t} \ge \mathbf{0}$ is a tuning parameter which controls the amount of shrinkage that is applied to the estimates and for all **t** $\boldsymbol{\beta}_0 = \overline{\boldsymbol{y}}$. Just as in ridge regression, we can re-parameterize the constant $\boldsymbol{\beta}\mathbf{0}$ by standardizing the predictors; the solution for $\widehat{\boldsymbol{\beta}}\mathbf{0}$ is $\overline{\boldsymbol{y}}$, and thereafter we fit a model without an intercept by assuming $\overline{\boldsymbol{y}} = \mathbf{0}$ and omitting $\boldsymbol{\beta}\mathbf{0}$ without loss of generality. Computing the Lasso solution is a quadratic programming problem, although efficient algorithms are available for computing the entire path of solutions as $\boldsymbol{\lambda}$ is varied, with the same computational cost as for ridge regression. Because of the nature of the constraint, making **t** sufficiently small will cause some of the coefficients to be exactly zero. Thus the Lasso does a kind of continuous subset selection. If t is chosen larger than

$\mathbf{t0} = \sum_{j=1}^{P} |\boldsymbol{\beta}_j| \le \boldsymbol{t}$ (where $\widehat{\boldsymbol{\beta}}\mathbf{j} = \widehat{\boldsymbol{\beta}}\mathbf{jls}$**,** the least square estimates), then the Lasso estimates are the $\widehat{\boldsymbol{\beta}}\mathbf{jls}$. (Hui et al, 2007)[22] develop versions of **AIC** and **BIC** for Lasso that can be used to find an ”optimal” value or or equivalently **t**. They suggested using **BIC** to find the ”optimal” Lasso model when sparsity of the model is of primary concern. Lars, least angle regression (Efron et al, 2004)[23] provides a clever and hence very efficient way of computing the complete Lasso sequence of solutions as s is varied from 0 to infinity. In fact, (Hui et al, 2007)[22] show that it is possible to find the optimal Lasso fit with the computational effort equivalent to obtaining a single least squares fit. Thus, the lasso has the potential to revolutionize variable selection. It employs an **L1**-type penalty on the regression coefficients which tends to produce sparse models, and thus is often used as a variable selection tool as in (Tibshirani, 1997 and Osborne, et al, 2000)[24]. (Knight and Fu, 2000)[25] studied the asymptotic properties of Lasso-type estimators. They showed that under appropriate conditions, the Lasso estimators are consistent for estimating the regression coefficients, and the limit distribution of the Lasso estimators can have positive probability mass at **0** when the true value of the parameter is **0**. It has been demonstrated in (Tibshirani, 1996)[26] that the Lasso is more stable and accurate than traditional variable selection methods such as best subset selection. For Multiple Regression Analysis, lasso penalty give as:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{^{argmin}} \left\{ \sum_{i=1}^{N} (y_i - X^T)^2 \right\} + \boldsymbol{\lambda} \sum_{j=1}^{P} |\boldsymbol{\beta}_j| \qquad \textbf{(3)}$$

Where, λ is the Lasso penalizing parameter. For logistic regression, Lasso modifies the traditional parameter estimation method, maximum log likelihood, by adding the **L1** norm of the parameters to the negative log likelihood function, so it turns a maximization problem into a minimization one. To solve this problem, we first need to give the value for the parameter of the **L1** norm, called tuning parameter. Since the tuning parameter affects the coefficients estimation and variable selection, we want to find the optimal value for the tuning parameter to get the most accurate coefficient estimation and best subset of predictors in the **L1** regularized regression model. There are two popular methods to select the optimal value of the tuning parameter that results in a best subset of predictors, Bayesian information criterion (**BIC**) and cross validation (**CV**). Therefore, best subsets of predictors are selected after standardizing the predictor variable by applying **BIC** or **k**-fold cross-validation (**CV**) (Tibshirani1, 996)[26]. Then, the package glmnet gives an optimum tuning parameter or **λ** for **CV**. In case of logistic regression with lasso, the model can be expressed as:

$$\hat{\boldsymbol{\beta}} = {}^{\boldsymbol{argmax}}_{\boldsymbol{\beta}} \left\{ \boldsymbol{l}(\boldsymbol{\beta}) - \boldsymbol{\lambda} \textstyle\sum_{j=1}^{P} |\boldsymbol{\beta}_j| \right\} \qquad \textbf{(4)}$$

$$\hat{\boldsymbol{\beta}} = {}^{\boldsymbol{argmax}}_{\boldsymbol{\beta_0\beta}} \left\{ \left[\boldsymbol{y}_i(\boldsymbol{\beta}_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) - \boldsymbol{log}(\mathbf{1} + \boldsymbol{e}^{\boldsymbol{\beta}_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i}) \right] - \boldsymbol{\lambda} \textstyle\sum_{j=1}^{P} |\boldsymbol{\beta}_j| \right\} \qquad \textbf{(5)}$$

Where λ is positive integer that determines the amount of shrinkage. As with the Lasso, we typically do not penalize the intercept term, and standardize the predictors for the penalty to be meaningful. (Efron, 2004)[23] proposed the Least Angle Regression (the **Lars**), and showed that there is a close connection between the lars, the Lasso, and another model selection procedure called the Forward Stage wise regression. Each of these procedures involves a tuning parameter that is chosen to minimize the prediction error.

2.3.3 Selection of Tuning Parameter: Cross Validation

Cross validation is a popular method for estimating the prediction error and comparing different models. Typically, the dataset partition into two parts: the training data and the testing data. In **k-fold** cross validation, the dataset will be randomly split into **k** mutually exclusive subsets of approximately equal size. Among the **k** subsets, one subset is retained as validation data for testing the model, and the remaining **k-1** subsets are used as training data to fit the model. The cross validation process will be repeated **k** times, and each of the subsets is used exactly once as validation data. Different values of the tuning parameter could result in different fitted model using the same training data. The optimal model is the one that has the minimum cross-validated errors, and the corresponding value of the tuning parameter for the optimal model is preferred (Jerome et al, 2008)[27].

2.3.4 Lasso for Logistic Regression to See Interaction Effect

Testing for interactions after identifying main effects or marginal predictors is the next step. This strategy is prompted by the number of interactions possible. With p predictors, we have: $\binom{P}{2}$ Two-way interactions but, with hundreds of thousands of SNPs, it is impossible even to examine all two-way interactions (Tong, 2009)[28]. To evaluate the performance of Lasso penalized regression in association testing, we focus on underdetermined problems where the number of predictors' p far exceeds the number of observations n. Here for two-way case, the model is: Which involve both marginal and two way interactions.

$$\textbf{log}\left(\frac{\boldsymbol{P}_i}{\mathbf{1}-\boldsymbol{P}_i}\right) = \boldsymbol{\mu} + \textstyle\sum_{j=1}^{P} \boldsymbol{x}_{ij}\boldsymbol{\beta}_j + \sum_{k=1}^{P} \sum_{l=1}^{P} \boldsymbol{x}_{ik}\boldsymbol{x}_{il}\boldsymbol{\eta}_{kl} \qquad \textbf{(6)}$$

2.4 Statistical Hypotheses Testing

Hypothesis testing is concerned with using observed data to make decisions regarding properties of (i.e., hypotheses for) the unknown data generating distribution. In any testing problem, two types of errors can be committed. A Type I error, or false positive, is committed by rejecting a true null hypothesis. A Type II error, or false negative, is committed by failing to reject a false null hypothesis. Ideally, one would like to simultaneously minimize both the number of Type I errors and the number of Type II errors. Unfortunately, this is not feasible and one seeks a trade-off between the two types of errors. This trade-off typically involves the minimization of Type II errors, i.e., the maximization of power, subject to a Type I error constraint. As in the case of single hypothesis testing, one can report the results of a multiple testing procedure in terms of the following quantities: rejection regions for the test statistics, confidence regions for the parameters of interest, and adjusted p-values. Adjusted p-values, for the test of multiple hypotheses, are defined as straightforward extensions of unadjusted p-values, for the test of individual hypotheses: the adjusted p-value for a particular null hypothesis is the smallest nominal Type I error level (for the multiple test of all hypotheses) at which one would reject this null hypothesis. The smaller the adjusted p-values indicates, the stronger the evidence against the corresponding null hypothesis (Sandrine et al, 2008)[29].

2.4.1 Chi-Square Test of Independence

The Chi-Square test may be used both as a test of goodness-of-fit (comparing frequencies of one categorical variable to theoretical expectations) and as a test of independence. The underlying arithmetic of the test is the same to test of goodness-of-fit but the only difference is the way the expected values are calculated. However, goodness-of-fit tests and tests of independence are used for quite different experimental designs and test different null hypotheses. The Chi-Squared test of independence is used when we have two categorical variables, each with two or more possible values.

2.4.2 Fisher's Exact Test

This test is used when we have two nominal variables. A data set like this is often called an (R x C table), where R is the number of rows and C is the number of columns. Fisher's exact test is more accurate than the Chi-Squared test of independence when the expected numbers are small. The most common use of Fisher's exact test is for (2 x 2 tables). But for our case we use (2 x 3 tables) for each SNPs. Fisher's Exact test assumes that the row and column totals are fixed. In the much more common design, the row totals and/or column totals are free to vary. In this case, the Fisher's Exact test is not, strictly speaking, exact. It is still considered to be more accurate than the chi-square and we should feel comfortable using it for any test of independence with small numbers.

2.4.3 The Problem of Multiple Testing Correction

Any time you reject the null hypothesis because a p-value is less than your critical value. It is possible that you are wrong; the null hypothesis might really be true, and your significant result might be due to chance. A P-value of 0.05 means that there is a 5% chance of getting your observed result, if the null hypothesis were true. This problem, that when you do multiple statistical tests, some fraction will be false positives, has received increasing attention in the last few years. This is important for such techniques as the use of microarrays, which make it possible to measure RNA quantities for tens of thousands of genes at once (Mcdonald, 2001)[30]. Controlling for multiple testing to accurately estimate significance thresholds is a very important aspect of studies involving many genetic markers, particularly GWA studies. The type I error, also called the significance level or false-positive rate, is the probability of rejecting the null hypothesis when it is true. The significance level indicates the proportion of false positives that an investigator is willing to tolerate in his or her study. The family-wise error rate (FWER) is the probability of making one or more type I errors in a set of tests. Lower FWERs restrict the proportion of false positives at the expense of reducing the power to detect association when it truly exists. It is then important to keep track of the number of statistical comparisons performed and correct the individual SNP-based significance thresholds for multiple testing to maintain the overall FWER (Zondervan et al, 2007)[31]. In order to choose an appropriate multiple testing methods, it is critical to select the definition of correct decisions. The following sub-sections introduce the common multiple testing correction methods.

2.4.4 Bonferroni Correction

The classical approach to the multiple comparison problem is to control family- wise error rate (Bland et al, 1995)[32]. Instead of setting p-value for significance, or, $\boldsymbol{\alpha}$ to 0.05, a lower $\boldsymbol{\alpha}$ is used. If the hypothesis is true for all tests, the probability of getting one result that is significant at this new lower, level is 0.05. In other words, if the null hypotheses are true, the probability that the family of tests includes one or more false positive due to chance is 0.05. The most common way and simplest of the p-value-based procedures to control the family-wise error rate is the well-known Bonferroni procedure. The basic procedure is: The significance level ($\boldsymbol{\alpha}$) for an individual test is found by dividing the family-wise error rate (usually 0.05) by the number of tests. If we are doing 100 statistical tests, the $\boldsymbol{\alpha}$ level for an individual test would be $\frac{0.05}{100}$ = 0.0005, and only individual tests with p-value < 0.0005 would be considered significant. The Boferroni correction assumes that the tests are independent of each other, and the method has good job of controlling family-wise error rate for multiple, independent comparisons; but important issue with Bonferroni correction is deciding what a ”family” of statistical test is. However there is no firm rule on this; we have to use our judgement, based on just how bad a false positive would be.

2.4.5 False Discovery Rate

A different approach to multiple testing does not try to control the family-wise error rate, (the probability can be computed under the assumption that all hypotheses are simultaneously true), but focuses instead on the proportion of falsely significant genes. As we will see, this approach has a strong practical appeal and it is an alternative approach to control family wise error rate. It is the proportion of ”discoveries” (significant results) that are actually false positive. For the example, suppose we are using microarray to compare expression levels of 100, 000 genes between liver tumors and normal liver cells. We are going to do additional experiments on any genes that show a significant difference between the normal and tumor cells, and we are willing to accept up to 10% of the genes with significant results being false positive; we find out they are false positives when we do the follow-ups experiment. In this case, we would set our false discovery rate to 10%. A good technique for controlling false discovery rate was briefly mentioned by (Simes, 1986)[29] and developed in detail by (Benjamini and Hochberg, 1995)[33].

Table 2: Possible outcomes of individual tests

	H0Accepted	H0Rejected	Total
H0 True	nT ;A	nT ;R	nT
H0 False	nF ;A	nF ;R	n-nT ornF
Total	n-nR	nR	N

In the above table:

- n is the total number hypotheses tested
- n_T is the number of true null hypotheses
- n - n_T or n_F is the number of false null hypotheses or true alternative hypotheses
- n_T ; R is the number of false-positive tests (Type I error) (also called ”false discoveries”)
- n_F ;A is the number of false negatives (Type II error)
- n_T ;A is the number of true negatives
- n_T ; R is the number of true positives (also called ”true discoveries”)
- n_R is the number of rejected null hypotheses (also called ”discoveries”)

In n hypothesis tests of which n_T are true null hypotheses, n_R is an observable random variable and n_F; R, n_F ;A, n_T;A and n_T;R are unobservable random variables. The type I error rate is E[n_T;R]/n_T , the type-II error rate is E[n_F ;A]/ n_F, family-wise error rate is Pr(n_T;R ≥ 1) and the power is 1 - E[n_T ;A]/ n_F . Here our focus is on the false discovery rate which is defined as: FDR = E(n_T;R |n_R) that is, the expected proportion of genes that are incorrectly called significant, among the R genes that are called significant. The expectation is taken over the population from which the data are generated. If the hypotheses are independent, (Benjamini and Hochberg, 1995)[33] show that regardless of how many null hypotheses are true and regardless of the distribution of the p-values when the null hypothesis is false, this procedure has the property:

$$\mathbf{FDR} \leq \frac{n_T}{n}\boldsymbol{\alpha} \leq \boldsymbol{\alpha} \qquad (7)$$

Algorithm 1 Benjamini-Hochberg (BH) Method:

1. Fix the false discovery rate α and $P_1 \leq P_2 \leq, \cdots \leq, P_n$ denote the ordered p-values
2. Define, L = maxj : pj <$\alpha \cdot \frac{j}{n}$
3. Reject all hypotheses H0j for which $P_j \leq P_l$; BH rejection threshold.

That is put the individual p-values in order, from smallest to largest. The smallest p-value has a rank of i=1, the next has i=2, etc. Then compare each individual p-value to ($\frac{i}{n}$)Q, where, m is the total number of tests and Q is the chosen false discovery rate. The largest p-value that has p < ($\frac{i}{n}$)Q, is significant, and all p-values smaller than it are also significant.

3. Results and Discussion

3.1 Results

The main objective of our work is to determine SNPs and SNPs-SNPs interactions in the genomic TSAD region which explain the difference between two MS conditions: OCB positive vs. OCB negative using the data from the Oslo MS DNA Bio Bank and The Norwegian MS Registry and Bio Bank. Accordingly, the analysis is carried out in two different approaches or variable selection methods: Variable Selection Based on Lasso applied to Binary Logistic Regression analysis and Variable Selection Based on Test of Association.

3.2 Analysis I: Variable Selection Based on Lasso Method

Variable selection plays an important role in regression analysis and is intended to select the best subset of predictors. Therefore, here we use the Lasso (Least absolute shrinkage and selection operator) which is a recent method of variable selection which applied when the number of samples is relatively smaller than the number parameters (variables). The Lasso method shrinks values of some coefficients to 0 by a constraint on the sum of absolute values of regression coefficients (penalty), so the Lasso can serve as a tool for variable selection. Such solutions, with multiple values that are identically zero, are said to be sparse. The penalty thereby performs a sort of continuous variable selection and the resulting estimator was thus named the Lasso, for Least Absolute Shrinkage and Selection Operator. In the following section, we apply the Lasso for Binary Logistic Regression analysis and we want to see interaction effects to some selected SNPs.

3.3 Lasso for Logistic Regression Analysis

Here we apply the Lasso variable selection to binary logistic regression on 923 SNPs. We run Lasso for 120 times, because it selects different random folds, each at different times. Then, Lasso method selects four SNPs after the optimum lambda

(i.e., lambda is obtained by subtracting one standard deviation from lambda.min) with ten and five-fold cross validation with glmnet package in R. These are most frequently selected SNPs which are the one which we selected in the end (See the Lasso Output for 923 SNPs at Appendix Section). Finally, we fitted the model to our dataset which contains those selected SNPs by our method (see Equation 11). The Table 3 shows the four Lasso selected SNPs with the corresponding regression coefficients and chromosome positions. Then we plot the corresponding cross validation curve to our dataset (See Figure 3).

$$\mathbf{Log(\frac{\hat{\pi}}{1-\hat{\pi}}) = 0.00098{*}NA + 0.08{*}NA\text{-}\ 0.0261{*}TEK\text{-}0.0894{*}EGFR} \qquad (8)$$

where, $\hat{\boldsymbol{\pi}}$ is a conditional probability of the form **P(Y=OCB|SNP1;···;SNP923).** That is, it is assumed that success or Y=OCB is more or less likely depending on combinations of values of the SNPs or predictor variables.

3.3.1 Interaction Effect for the Lasso Selected SNPs

In this section, we study interaction effects of the ten variables. Four SNPs which are selected by Lasso (See Table 3) and six interactions obtained by combination of these four SNPs which we selected by Lasso method with Logistic Regression Analysis on 923 SNPs. Then we have run Lasso to Logistic Regression Analysis for these 10 variables. So, we found out that none of the interaction variables are selected (See the Lasso Output for 10 Variables at Appendix Section).

Table 3: The Lasso Selected SNPs

S.no	SNPs	Chr	Coeff
1	rs706862A	6	0.0296
2	rs4245566A	7	-0.1004
3	rs2002680A	11	-0.0294
4	rs11230645G	11	-0.0982

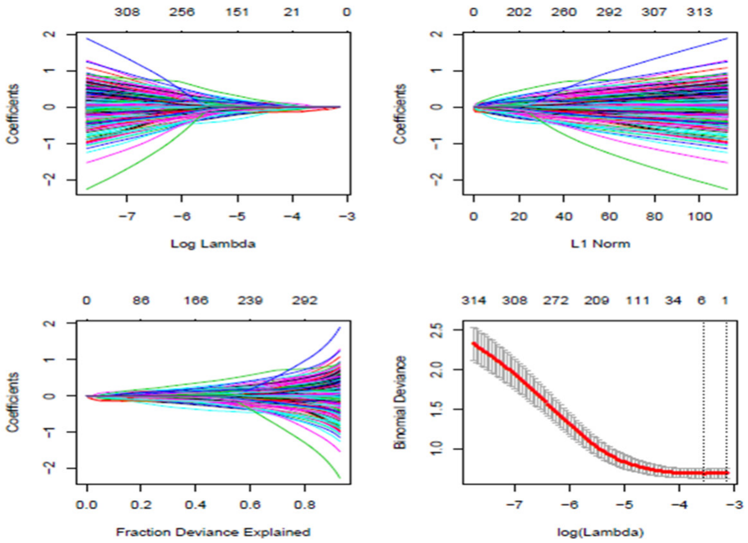

Figure 3: CV for Lasso Logistic Regression Analysis

3.4 Analysis II: Variable Selection Based on Test of Association

In this section, we use different approach to select variables. First section identifies an important variable by using test of associations. Here, we used Chi-Square and Fisher’s exact test of association. In the second section, we perform multiple independent hypotheses, that is we need to do some multiple testing corrections to control family wise error rate or Type I error rate for multiple hypotheses testing to multiple hypotheses from Fisher’s exact test of association. Here we use FDR and Bonferroni adjustment method for the 38 P-value (raw or unadjusted P-value) obtained by Fisher’s exact test. Finally, we identified the interaction effects to SNPs selected by Fisher’s Exact test.

3.4.1 Fisher’s Exact and Chi-Square Test of Association

We conducted 923 independent tests one for each SNP. We found out that the Chi-Square test of association selected 34 significant SNPs and the association test for Fisher’s Exact test selected 38 significant SNPs at significance level of 0.05 (See Table 4 and Table 5 respectively).

Table 4: SNPs Selected by Chi-Square Test

no.	SNPs	Pv.chi	no.	SNPs	Pv.chi	no.	SNPs	Pv.chi
1	rs4234103A	0.0045	13	rs692946A	0.0106	25	rs17092209G	0.0059
2	rs13412634C	0.0411	14	rs1360773C	0.0062	26	rs633903C	0.0082
3	rs11686987A	0.0494	15	rs1923332A	0.0062	27	rs6088659A	0.0003
4	rs13419955A	0.0290	16	rs638203G	0.0042	28	rs491892G	0.0042
5	rs824097G	0.0455	17	rs489451G	0.0038	29	rs6088646A	0.0489
6	rs16843013A	0.0161	18	rs2756900G	0.0036	30	rs2145926G	0.0371
7	rs17704348G	0.0353	19	rs1111782G	0.0417	31	rs6120757G	0.0489
8	rs11238349A	0.0200	20	rs28599952A	0.0027	32	rs669102A	0.0134
9	rs10280515G	0.0398	21	rs11616506A	0.0408	33	rs6088640G	0.0195
10	rs4733037A	0.0260	22	rs16970646G	0.0169	34	rs3117638A	0.0152
11	rs2292979G	0.0400	23	rs6088635A	0.0489			
12	rs11789885A	0.0314	24	rs4911163G	0.0489			

Table 5: SNPs Selected by Fisher's Exact Test

no.	SNPs	Pv.fish	no.	SNPs	Pv.fish	no.	SNPs	Pv.fish
1	rs4234103A	0.0009	14	rs692946A	0.0102	27	rs17092209G	0.0059
2	rs13412634C	0.0308	15	rs1360773C	0.0055	28	rs633903C	0.0158
3	rs11686987A	0.0420	16	rs1923332A	0.0065	29	rs6088659A	0.0004
4	rs17261971A	0.0472	17	rs638203G	0.0047	30	rs491892G	0.0059
5	rs13419955A	0.0237	18	rs489451G	0.0074	31	rs6088655A	0.0462
6	rs824097G	0.0167	19	rs2756900G	0.0069	32	rs2145926G	0.0385
7	rs16843013A	0.0469	20	rs1111782G	0.0384	33	rs6088646A	0.0440
8	rs17704348G	0.0442	21	rs2002680A	0.0443	34	rs669102A	0.0152
9	rs17586344G	0.0346	22	rs28599952A	0.0054	35	rs6120757G	0.0440
10	rs11238349A	0.0270	23	rs11616506A	0.0359	36	rs3117638A	0.0248
11	rs10280515G	0.0338	24	rs16970646G	0.0178	37	rs6088640G	0.0160
12	rs4733037A	0.0288	25	rs6088635A	0.0440	38	rs11789885A	0.0391
13	rs2292979G	0.0385	26	rs4911163G	0.0440			

3.4.2 Bonferroni and FDR for Multiple Testing Correction

At FDR of 0.05, Benjamini and Hockberg (BH) adjustment for FDR shows all 38 SNPs are significant that is, all 38 raw P-value is below adjusted FDR. In similar manner, at family-wise-error rate of 0.05, the Bonferroni adjustment shows none of SNPs are significant that is, none of the adjusted P-value is below unadjusted P-value (See Table 6).

Table 6: Identified SNPs by Bonferroni and FDR

S.no	SNPs	Pv.unadj	Pv.adj.Bon	Pv.adj.FDR
1	rs6088659A	0.00040	0.01396	0.01396
2	rs4234103A	0.00089	0.03141	0.17944
3	rs638203G	0.00470	0.03141	0.20621
4	rs28599952A	0.00540	0.03141	0.21023
5	rs1360773C	0.00550	0.03141	0.22413
6	rs491892G	0.00590	0.03141	0.22420
7	rs17092209G	0.00590	0.03141	0.24613
8	rs1923332A	0.00650	0.03141	0.26377
9	rs2756900G	0.00690	0.03141	0.28271
10	rs489451G	0.00740	0.03341	0.33406
11	rs692946A	0.01020	0.03526	0.38788
12	rs669102A	0.01520	0.04217	0.57860
13	rs633903C	0.01580	0.04217	0.60217
14	rs6088640G	0.01580	0.04217	0.60718
15	rs16843013A	0.01670	0.04217	0.63467
16	rs16970646G	0.01780	0.04217	0.67469
17	rs13419955A	0.02370	0.04724	0.90147
18	rs3117638A	0.02480	0.04724	0.94265
19	rs11238349A	0.02700	0.04724	1.00000
20	rs4733037A	0.02880	0.04724	1.00000
21	rs13412634C	0.03080	0.04724	1.00000
22	rs10280515G	0.03380	0.04724	1.00000
23	rs17586344G	0.03460	0.04724	1.00000
24	rs11616506A	0.03590	0.04724	1.00000
25	rs1111782G	0.03840	0.04724	1.00000
26	rs2292979G	0.03850	0.04724	1.00000
27	rs2145926G	0.03850	0.04724	1.00000
28	rs11789885A	0.03910	0.04724	1.00000
29	rs11686987A	0.04200	0.04724	1.00000
30	rs6088635A	0.04410	0.04724	1.00000
31	rs4911163G	0.04410	0.04724	1.00000
32	rs6120757G	0.04410	0.04724	1.00000
33	rs6088646A	0.04410	0.04724	1.00000
34	rs17704348G	0.04420	0.04724	1.00000
35	rs2002680A	0.04430	0.04724	1.00000
36	rs6088655A	0.04610	0.04724	1.00000

S.no	SNPs	Pv.unadj	Pv.adj.Bon	Pv.adj.FDR
37	rs17054409G	0.04690	0.04724	1.00000
38	rs17261971A	0.04720	0.04724	1.00000

3.4.3 Interaction Effects of SNPs which found to have Significant by Test of Association

Here we study interaction effects of the 38 SNPs which are found to have significant association. These are 703 interactions obtained by combination and 38 SNPs which we found to have significant association by Fisher's Exact test. So, we have made variable selection among the 741 variables using Lasso method again with Binary Logistic Regression Analysis. We found out that seven SNP-SNP interactions are significant by our method and none of main effects are significant (See Lasso output for 741 variables at Appendix Section). These seven interactions variables are which listed below in Table 7. Then we plot the corresponding CV which is shown in Figure 4. Finally, we fitted the model to our data which contains only seven interaction effects (See Equation 12 below)

$$\mathbf{Log(\hat{\pi}/(1-\hat{\pi})) = -0.020 * (a*b) - 0.054\ (c*d) - 0.001\ (f*g) - 0.06\ (f*b) - 0.018\ (c*e) - 0.051\ (c*e) - 0.0067\ (e*e)} \qquad \mathbf{(9)}$$

Where, a=MYO7B, b=EGFR, c=NA, d=ITK), f=FYN, g=NA and e=TEK and is a conditional probability of the form P(Y = OCB|SNP1,⋯,SNP923) or P(Y = OCB|SNP1,2,⋯,SNP922,923). That is, it is assumed that success or Y=OCB is more or less likely depending on combinations of values of the SNPs or combination of SNPs.

Table 7: Interaction Effects of Selected SNPs by Lasso

S.no	SNPs(Symbol)	Coeff
1	rs4662738A(MY O7B)rs7780270C(EGFR)	-0.0195
2	rs17776723G(NA) rs13190673A(ITK)	-0.0544
3	rs3752545C(FY N) rs12678502A(NA)	-0.0007
4	rs1558542A(FY N) rs7780270C(EGFR)	-0.0600
5	rs12678502A(NA) rs7040866A(TEK)	-0.0167
6	rs12678502A(NA) rs1923332A(TEK)	-0.0514
7	rs3739542C(TEK) rs2208637A(TEK)	-0.0057

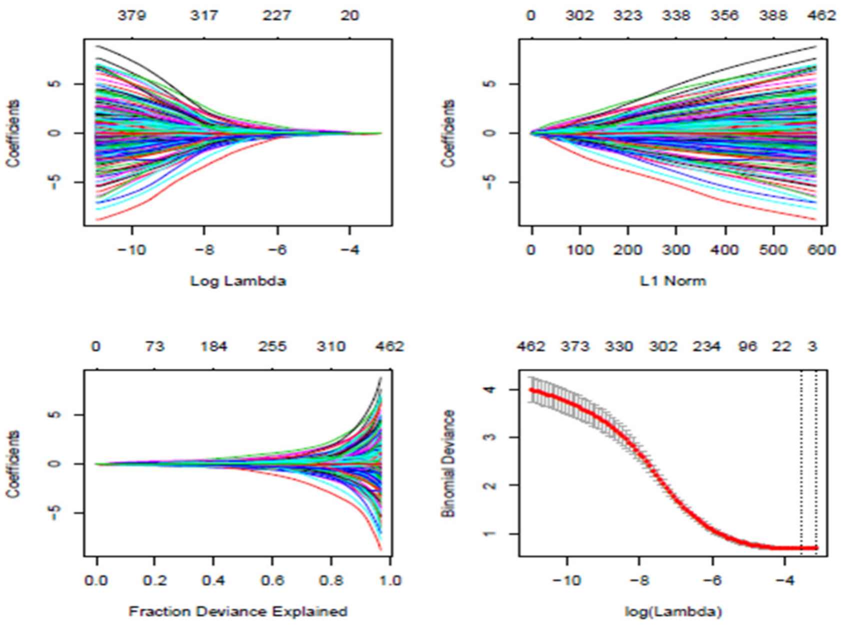

Figure 4: CV for Interaction Effects

3.5 Discussion

The results from analysis I with our proposed method (Variable Selection Based on Lasso Method) as a framework to determine important SNPs in the genomic TSAD region presented in this study identified four SNPs which appear to be associated to the Oligoclonal Bands(OCBs) subpopulation of multiple sclerosis patients (See Table 3). We also tested interaction effect by combining these four SNPs and finally the analysis showed none of the combinations (interactions) has an effects to Oligoclonal Bands (OCBs) subpopulation of multiple sclerosis patients.

This findings suggests that these four SNPs independently determine the status of Oligoclonal Bands (OCBs) subpopulation of multiple sclerosis patients based on our study of the Norwegian cohort. Then, the finally model contains four important SNPs which the method selected at the end (**See Equation 8**). The results from analysis II, shows that at 0.05 level of significance, the Fisher Exact test identified 38 important SNPs from 923 SNPs in the genomic TSAD region that are associated with Oligoclonal Bands (OCBs) subpopulation of multiple sclerosis patients (See Table 5). The results from analysis II for Chi-Square test showed 34 important SNPs at 0.05 significances level (See Table 4). The results from analysis II for FDR at 0.05 significant level of Benjamini and Hockberg (BH) procedure showed 38 SNPs are all important to determine Oligoclonal Bands (OCBs) subpopulation of multiple sclerosis patients (See Table 6). The similar result for Bonferroni correction or adjustment at family wise error rate of 0.05 showed none of SNPs are significant. The final results from analysis II for interaction effect to the variables obtained by combination among 38 SNPs selected previously from Fisher Exact test showed seven important SNP-SNP interactions which determine Oligoclonal Bands (OCBs) subpopulation of multiple sclerosis patients (See Table 7). Our study will be confirmed with similar studies which identified potentially interesting logic expressions that represent SNP interactions and measures for quantifying the importance of these features for classification in case control studies (Holger, 2008)[34]. Similar study indicated that many common diseases are influenced by interaction of certain genes and quadratic penalization not only correctly characterizes the influential genes along with their interaction structures but also yields additional benefits in handling high dimensional, discrete factors with a binary response (Young, 2008)[35]. Other similar analysis indicated that penalizing the size of the coefficients is a common strategy for robust modeling in regression classification with high dimensional data and examined the properties of the Lasso constraints applied to the coefficients in generalized linear models (GLM) to the specific application of modeling gene interactions (Young, 2006)[36]. Results from (Yoav, 2002)[27] in their experiments aims to identify genes with altered expression in the livers of mice with very low cholesterol levels compared to inbred control mice. They examined the p-values obtained directly from the raw t-statistics with 14 degrees of freedom. Then, Bonferroni adjustment points to eight rejections. Also applying the FDR controlling BH procedure on the raw p-values, they came up with the same eight genes identified as differentially expressed in the original analysis.

4. Conclusions and Recommendations

4.1 Conclusions

Lasso method selects four important SNPs which appear to be associated to the Oligoclonal Bands subpopulation of multiple sclerosis patients. These are most frequently selected SNPs which are the one which we selected in the end.

We also fitted the model to our dataset which contains those selected SNPs by our method (See Equation 9). From results of analysis I, we found out that none of the interaction variables are selected and the findings further suggest four SNPs in- dependently determine the status of Oligoclonal Bands subpopulation of multiple sclerosis patients. The results from analysis II, showed that 38 important SNPs from 923 SNPs in the genomic TSAD region that are associated with Oligoclonal Bands (OCBs) subpopulation of multiple sclerosis patients. Result for Bonferroni correction showed none of SNPs are significant at 0.05 levels. The results from analysis for FDR at 0.05 significant level showed 38 SNPs are all important to determine Oligoclonal Bands (OCBs) subpopulation of multiple sclerosis patients. Finally, results from analysis II showed seven important SNP-SNP interactions which determine Oligoclonal Bands (OCBs) subpopulation of multiple sclerosis patients, based on our study of the Norwegian cohort. Then, we collected all SNPs (See Table 8) (with the SNPs where they are located) which have been selected by any of the methods that we used (various hypothesis testing and regressions). These are the list of SNPs that we think should be studied further and validated on a new data set. Of particular importance are the seven SNP-SNP interactions which we found. It is the first time a SNP-SNP study has been performed on these data, and the finding will be communicated to the Norwegian molecular biologists to be followed up (See Table7).

4.2 Recommendations

The findings of this study have important implications to help biologists, healthy organizations, researchers and scientists to deal on disease prevalence and progression such that, genes are important factors for cause of MS. As the new study in this area, this may motivate interested groups and professionals to be aware of the disease, and it perhaps initiate them to their own contributions in the same area of research. Since Lasso selects SNPs which are best to perform classification (outcome) of a new patient. These are SNPs which each carry additional independent information on the classification. This means that two covariates which are both very strongly associated with the outcome and are also highly correlated with each other, will not be both selected by the Lasso, because only one covariate contributes to the best classification, the other carries the same information. The Bonferroni and FDR correction do not look to the correlation of the SNPs, but only on how many tests we performed, and reduce the significance level so that we make less false positives mistakes. Therefore, we have generated a series of biological hypothesis, supported by our stringent data analysis, which now need to be confirmed on a new population. If this will be the case, then it is possible to imagine that finding these SNPs in MS patients, will allow a better therapy. Therefore, our results need to be further validated.

References

1. Compston A. and Coles A. (2008). Multiple sclerosis. Lancet 372 (9648), 1502-17.
2. Murray T. (2005). Multiple Sclerosis: The History of a Disease. JAMA 294, 376-377.
3. Clanet M. (2008). International Multiple Sclerosis. Int J 15 (2), 59-61.
4. Compston A. and Coles A. (2002). Multiple sclerosis. Lancet 359 (9313),1221-31.
5. Ascherio A. and Munger K. (2007). Environmental Risk Factors for Multiple Sclerosis Part I: The Role of Infection. Ann Neurol 61 (4), 288-299.
6. Lublin F. and Reingold S. (1996). Defining the Clinical Course of Multiple Sclerosis: National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Neurology 46 (4), 907-11.
7. Weinshenker, B. (1994). Natural History of Multiple Sclerosis. Ann Neurol 36, 6-11.
8. Genetics and Hereditary Aspects of Multiple Sclerosis: MS in Focus. htp://www.worldmsday.org/.../tag/MSIF [Accessed on July 2007].
9. Marrie R. (2004). Environmental Risk Factors in Multiple Sclerosis Etiology. Lancet Neurol 3, 709-718.
10. Poser C., Paty D., Scheinberg L., McDonald W. and Davis F. (1983). New Diagnostic Criteria for Multiple Sclerosis: Guidelines for Research Protocols. Ann Neurol 13, 227-231.
11. Polman C., Reingold S., Banwell B., Clanet M. and Cohen J. (2010). Diagnostic Criteria for Multiple Sclerosis. Ann Neurol 69, 292-302.
12. Paolino E. and Fainardi P. (1996). A Prospective Study on the Predictive Value of CSF, Oligoclonal Bands (OCBs) and MRI in Acute Isolated Neurological Syndroms for Subsequent Progression to Multiple Sclerosis. J Neurol Neurosurg. Psychiatry 60, 572-575.
13. Amato M. and Ponziani G. (2000). A Prospective Study of Multiple Sclerosis. Neorol Sci 21, 831-838.
14. Andersson M., Alvarez J., Bernardi G., Cogato I. and Fredman P. (1994). Cerebrospinal Fluid in the Diagnosis of Multiple Sclerosis: A Consensus Report. Journal of Neurol Neurosurg Psychiatry 57, 897-902.
15. Joseph F. (2008). CSF Oligoclonal Band Status Informs Prognosis in Multiple Sclerosis: A Case Control Study, 89-85.
16. Kerstin I. (2009). Conquering Complexity: Successfull Strategies for Finding Disease Genes in Multiple Sclerosis Patients.
17. Kristin M. and Gunderson R. (2007). The Etiology of Multiple Sclerosis and Correlation of the Distribution of the Disease with Migration and Settlement History of Northern Europeans. Neuroimmunol 5, 67-80
18. Kolltveit K., Granum S., Aasheim H., Kristiansen M., Sundvold V., Dai K., Molberg O., Schjetne K., Bogen B., Shapiro V., Johansen F., Schenck K. and Spurkland A. (2008). Expression of SH2D2A in T Cells is Regulated Both at the Transcriptional and Translational Level. Mol Immunol 5, 2380-2390
19. Lorentzen A., Smestad C., Lie B., Oturai A., Akesson E., Saarela J., Myhr K., Vartdal F., Celius E., Sorensen P., Hillert J., Spurkland A. and Harbo H. (2008). The SH2D2A gene and susceptibility to multiple sclerosis. J Neuroimmunol 197, 152-60.
20. Agresti A. (2007). An Introduction to Categorical Data Analysis. New York:Wiley.
21. Fan J. and Li R. (2006). Statistical Challenges with High Dimensionality Feature Selection in Knowledge Discovery: Proceedings of the International Congress of Mathematicians, European Mathematical Society, zurich, 595-622.
22. Hui Z., Trevor H. and Robert T. (2007). On the Degrees of Freedom of the Lasso. The Annals of Statistics 5, 21732192
23. Efron B., Hastie T., Johnstone I. and Tibshirani R. (2004). Least Angle Regression with Discussion. Ann Statist 32, 407-451.
24. Osborne M., Presnell B. and Turlach, B. (2000). A New Approach to Variable Selection in Least Squares Problems. IMA J Numer Anal 20, 389-404.
25. Knight K. and Fu W. (2000). Asymptotics for Lasso-Type Estimators. Ann Statist 28, 1356-1378.
26. Tibshirani R. (1996). Regression Shrinkage and Selection via the Lasso. J Roy Statist Soc Ser 58, 267-288.
27. Trevor H., Robert T. and Jerome F. (2008). Elements of Statistical Learning, 241-245.
28. Tong W., Yi F., Trevor H., Eric S. and Kenneth L. (2009). Genome-Wide Association Analysis by Lasso Penalized Logistic Regression: Genome Analysis. Bioinformatics 6, 714-721.
29. Sandrine D. and Mark J. (2008). Multiple Testing Procedures with Applications to Genomics: Springer Series in Statistics, 9-10. Simes R. (1986). An Improved Bonferroni Procedure for Multiple Tests of Significance. Biometrika 731, 751-754.
30. McDonald W., Compston A., Edan G., Goodkin D. and Hartung H. (2001). Recommended Diagnostic Criteria for Multiple Sclerosis: Guidelines from the International Panel on the Diagnosis of Multiple Sclerosis. Ann Neurol 7, 121-127.
31. Zondervan K., and Cardon L. (2007). Designing Candidate Gene and Genome Wide Case-Control Association Studies. Nat Protoc 2, 24922501.

32. Bland J. and Altman D. (1995). Multiple significance tests: The Bonferroni method. BMJ 310, 170-182.
33. Benjamini Y. and Hochberg Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc 57, 289-300.
34. Holger S. (2008). Identification of SNP Interactions Using Logic Regression. Biostatistics 1, 187-198.
35. Young P. (2008). Penalized logistic regression for detecting gene interactions. Biostatistics, 9, (1), 3050.
36. Young P. (2006). Generalized Linear Modeles with Regularization: Autoimmune Disease-Regulator in Mouse and Man. Immunol Lett 97, 165-170.
37. Anat R., Daniel Y. and Yoav B. (2002). Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 45, 368-375.