Proceedings of IICMA 2009
Topic, pp. xx—xx.

SNP DATA ANALYSIS USING LOGISTIC REGRESSION

ADI SETIAWAN

Abstract. Whole-genome association studies have become overwhelmingly popular in the last few years to find association between genes and the disease of interest. In this paper we describe SNP (single nucleotide polymorphism) data analysis by using logistic regression. The *p*-value is determined based on statistical value of logistic regression model. The *p*-value is used to reject or accept the association between gene related to the SNP and the disease. Bonferroni multiple testing correction is used to adjust the *p*-values. The described method is then applied to the whole-genome SNP data to find the SNPs associated to the disease of interest.

Key words and Phrases : logistic regression, single nucleotide polymorphism, whole genome SNP data.

1. INTRODUCTION

Association studies have become overwhelmingly popular in the last few years as means to elucidate association between particular alleles in one's DNA and a predisposition to disease, using genetic data from unrelated individuals randomly sampled from a population (Balding, 2006). In this paper we focus on case-control association studies, which simply compare the genotypes of individuals who have a disease (cases) with the genotype of individuals without the disease (controls). The proportions in each group having a characteristic of interest (for instance the numbers of alleles of a given type) are then compared to determine whether there is an association between the disease and the characteristic of interest. Association studies must use markers (the observable characteristics of the genome) in the analysis. In this paper we study in particular methods based on Single Nucleotide Polymorphisms (SNPs). Such markers may be measured to investigate a particular target region of the genome, for fine-mapping a functional gene. Increasingly SNP data are also gathered all across the genome (e.g. 10000 SNPs). Whole-genome analysis is focused on scanning this complete set of markers. A simple case-control test applied to this multitude of markers, or set of markers, would of course lead to many false positives, whence a correction is necessary.

2000 Mathematics Subject Classification:

Genetic markers are tools that can be used in genetic association studies to identify genes responsible for disease susceptibility. Genetic markers that have played significant roles in genetic research are restriction fragment length polymorphisms (RFLPs), short tandem repeats (STRs), and SNPs. RFLP is a polymorphic difference in the size of allelic restriction fragments as a result of the polymorphic presence or absence of a particular restriction site. STR (also called microsatellite) is a small run (usually less than 0.1 kilo base pairs) of tandem repeats of a very simple DNA (Deoxyribonucleic Acid) sequence, usually 1 to 4 base pairs, for example $(CA)_n$ where nucleotide C and A are cytosine and adenine, respectively. A SNP (pronounced snip) is a DNA sequence variation, occurring when a single nucleotide : adenine (A), thymine (T), cytosine (C) or guanine (G) in the genome differs. For example, base pair sequences T**A**GATA and T**G**GATA differ in the second nucleotide. The human genome consists of 10^9 base pairs and around 10 million SNPs with frequency bigger than 1 %. SNPs play a central role for mapping complex diseases among different types of genetic markers because of their high frequency throughout the genome.

Traditional association studies begin with a candidate gene or genetic region that researchers already suspect is associated with the disease. This approach will depend on prior knowledge about the genes or regions and cannot identify new genes or genetic regions that might be associated.

Without prior knowledge about regions that may be associated to the disease, we can compare SNPs in the whole-genome of individuals from case and control samples. This approach (whole-genome association studies) will provide the opportunity to discover novel genes. Approximately 300K SNPs would be required to cover the whole-genome. The strategy to use partial measurement (300K SNPs) instead of the whole-genome, could be successful only if the population is in linkage disequilibrium or if the causal loci are among the measured markers. Linkage equilibrium by definition would imply that different loci are independent within the population, so that a marker would never be informative about any other locus than itself.

In order to find a causal locus we need linkage disequilibrium between the causal locus and an observed marker locus. Such linkage disequilibrium is likely to exist for markers that are close to the disease locus, but not for markers at some distance. If we measure markers close enough to the disease location, a two-sample (case and control) comparison approach might work.

2. STATISTICAL METHODS

The methods are based on a case-control design and try to find marker loci associated to the disease by comparing genotype frequencies between random samples of cases (diseased individuals) and controls. The methods can be classified as single marker, double marker or multiple markers according to whether they take into account frequencies of markers at one locus or combinations of markers at two or more loci. Under assumptions of infinite population size, discrete generations, random mating, no selection, no migration, no mutation and equal initial genotype frequencies in the two sexes, Hardy-Weinberg equilibrium arises after one generation and thereafter the genotype frequencies in the population are constant from generation to generation.

Let (p_1, p_2, p_3) and (q_1, q_2, q_3) be the genotype frequencies (*AA*, *Aa*, *aa*) in the populations of controls and cases, respectively. We take random samples of *n* controls and *m* cases, respectively. The layout of the data is given in Table 1. Testing association between the genotype of the marker locus and the disease is equivalent to testing the null hypothesis $H_0 : \boldsymbol{p} = \boldsymbol{q}$ versus the alternative hypothesis $H_1 : \boldsymbol{p} \neq \boldsymbol{q}$ where $\boldsymbol{p} = (p_1, p_2, p_3)$ and $\boldsymbol{q} = (q_1, q_2, q_3)$.

Table 1. Table of the number of genotypes *AA*, *Aa* and *aa* in control and case samples.

	AA	*Aa*	*aa*	Total
Controls	X_1	X_2	X_3	*n*
Cases	Y_1	Y_2	Y_3	*m*
Pooled	$X_1 + Y_1$	$X_2 + Y_2$	$X_3 + Y_3$	*n* + *m*

The idea is to code for each individual *i*, the genotype information in regression variables X_{i1}, X_{i2}, ….., X_{iL} and apply logistic regression of the affection status Y_i on X_{i1}, X_{i2}, ….., X_{iL}, where L is the number of markers that are used in the analysis. The observed marker data (unordered genotypes at one or more marker loci) are mapped in the regression variables in a simple and direct manner.

Formulation of the testing problem within the context of logistic regression give great flexibility, both allowing the use of smaller and bigger models to describe the genotypic effect. Smaller models are for instance obtained by modeling the effect of different loci additively. On the other hand, by using enough dummy variables the regression can also be carried out on a complete classification of the set genotypes.

Suppose we summarize the genotype or haplotype information of individual *i* for *i* = 1, 2, …,*N*, where *N* is the total number of cases and controls, in *L* numerical scores X_{i1}, X_{i2}, ….., X_{iL}. Let $Y_i = 1$ denote that the *i*-th individual is affected, and let $Y_i = 0$ otherwise. Let

$$P(Y_i = 1 \mid X_{i1}, X_{i2}, ..., X_{iL})$$

be the probability that an individual is affected given the genetic information. We adopt the logistic regression model

$$p_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_L X_{iL})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_L X_{iL})}$$

where $X_i = (1, X_{i1}, X_{i2}, ..., X_{iL})^t$ dan $\beta = (\beta_0, \beta_1, ..., \beta_L)^t$. The log-likelihood function of the data can be expressed as

$$l(\beta) = \ln(L(\beta)) = \ln\left(\prod_{i=1}^{N} p_i^{Yi} (1 - p_i)^{1-Yi} \right) = \sum_{i=1}^{N} [Y_i \ln(p_i) + (1 - Y_i) \ln(1 - p_i)]$$

$$= \sum_{i=1}^{N} \left[Y_i \ln\left(\frac{\exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_L X_{iL})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_L X_{iL})} \right) \right]$$

$$- \sum_{i=1}^{N} (1 - Y_i) \ln(1 + \exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_L X_{iL}))$$

$$= \sum_{i=1}^{N} Y_i (\beta_0 + \beta_1 X_{i1} + ... + \beta_L X_{iL}) - \sum_{i=1}^{N} \ln(1 + \exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_L X_{iL}))$$

$$= \sum_{i=1}^{N} Y_i X_i^t \beta - \sum_{i=1}^{N} \ln(1 + \exp(X_i^t \beta)).$$

We want to test the association between the markers and the disease, i.e. to test the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_L$.

The deviance D is defined as twice the log-likelihood ratio statistic, given by

$$D = 2 \ln\left(\frac{L(\hat{\beta})}{L(\hat{\beta}_{Ho})} \right) = 2\left[\ln(L(\hat{\beta})) - \ln\left(L(\hat{\beta}_{Ho}) \right) \right] = 2\left[l(\hat{\beta}) - l(\hat{\beta}_{Ho}) \right],$$

where $\hat{\beta}_{Ho}$ is the MLE of β under H_0 and $\hat{\beta}$ is the MLE of β under the model in which (β_0, β_1, ….., β_L) vary freely over R^{L+1}. Under H_0, the statistic D has a chi-squared distribution with L degrees of freedom, asymptotically as $N \to \infty$ (provided that the matrix $X^t X$ defined below converges to a matrix of full rank).

We illustrate the techniques on data collected by the Department of Medical Genetics at the Vrije Universiteit Medical Center Amsterdam. The genotype of data came from a genetically isolated population in Turkey with current population size around 6000 people. Ninety percent of these people are supposed to be descendants of 23 families that originally inhabited the region approximately 400 years ago. Genotyping was done using the Affimetrix 10K SNP chip to 27 controls and 31 cases We summarized characteristics of the 11229 SNPs, such as the identity of the SNP in the chromosome and the genotype of every individual in the control and case samples The genotype of individuals are defined as *AA*, *AB* or *BB*, a missing genotype is coded as "NoCall", meaning that the marker did not pass the discrimination filter (Setiawan, 2007).

A case-control study with a biallelic marker was conducted from SNPs analysis with identity 1513978 in chromosome 2 and the results are given in Table 2. Using Table 2 and the logistic regression models for single marker genotype-based methods, we find the deviance is 20.9886 and the related *p*-value is 4.6204×10^{-6}. Thus, we reject the null hypothesis and thus there is an association between the marker and the disease.

Table 2. Table of the number of genotypes in control and case samples.

	AA	*Aa*	*aa*	Total
Controls	11	14	2	27
Cases	29	2	0	31
Pooled	40	16	2	58

Current technology allows to measure 10K SNPs per individual, a number that is likely to increase in the comming years. The case-control methods can be applied to each SNP separately or to sets of SNPs, adjacent or not, but then require adjustment for multiple testing. Carrying out such tests at the usual level of significance (e.g. 5 %) would result in many false positives, i.e. markers that are found to be statistically significant but in reality do not differ between the conditions. In this section, we review Bonferroni correction method to adjust *p*-values.

We assume that *M* tests are performed, yielding *p*-values $p_1, p_2, \ldots, p_M$. We let denote the ordered *p*-values. The adjustment for multiple testing is expressed through replacing each of the *p*-values by a different (larger) *p*-values. The Bonferroni correction is the simplest possible correction. It simply multiplies every *p*-value by the number of tests

$$\tilde{p}_{r_i} = \min\{ M\ p_{r_i}, 1\}.$$

3. SIMULATION STUDY, APPLICATIONS AND DISCUSSION

In this section, we describe a simulation study of SNP data analysis and apply the described method in the whole-genome SNP data.

Simulation Study

Case control association simulated data can be generated as follows. We use 50 controls and 50 cases to describe the method. Genotype *AA*, *Aa* and *aa* are generated for 50 individuals in controls sample by using Multinomial distribution with parameter 50 and (p_{AA}, p_{Aa}, p_{aa}) = (0.1, 0.1, 0.8). Similarly, 50 individuals in cases sample can be generated. The result of 10 simulated data in controls sample, cases sample and the related p-values are given in Table 3. By using 5 % level of significance and based on the table, we tend to conclude that there is no association between the marker and the disease. Table 4 presents simulated data and their *p*-values when we use Multinomial distribution with parameter 30 and (p_{AA}, p_{Aa}, p_{aa}) = (0.1, 0.1, 0.8) for controls sample and for (q_{AA}, q_{Aa}, q_{aa}) = (0.1, 0.7, 0.2) cases sample. By using 5 % level of significance and based on the table, we conclude that the *p*-values tend to conclude that there is an association between the marker and the disease.

Simulation can be extended to different sample size such as 100 and 200 and simulation can be done for big number *B*=10000 times. Table 5 presents the result of mean and standard deviation of *p*-values by using sample size 50, 100 and 200. We conclude that the larger sample size will have the smaller *p*-value.

Tabel 3. The result of 10 simulated data using 50 individuals in controls sample and 50 individuals in cases sample and the related *p*-value.

No	Controls Sample			Cases Sample			*p*-value
	n_{AA}	n_{Aa}	n_{aa}	n_{AA}	n_{Aa}	n_{aa}	
1	3	7	40	0	8	42	0,2895
2	6	6	38	1	6	43	0,0788
3	5	3	42	7	7	36	0,2380
4	5	6	39	4	5	41	0,6288
5	4	4	42	8	4	38	0,2343
6	5	2	43	8	9	33	0,0610
7	7	2	41	6	6	38	0,7738
8	6	4	40	7	4	39	0,7738
9	5	5	40	8	2	40	0,6551
10	7	8	35	6	6	38	0,5712

Tabel 4. The result of 10 simulated data using 50 individuals in controls sample, 50 individuals in cases sample and the related *p*-value.

No	**Controls Sample**			**Cases Sample**			***p*-value**
	n_{AA}	n_{Aa}	n_{aa}	n_{AA}	n_{Aa}	n_{aa}	
1	2	5	43	2	35	13	3.0787×10^{-8}
2	4	5	41	7	30	13	1.8801×10^{-6}
3	3	7	40	2	37	11	6.7918×10^{-7}
4	5	9	36	4	38	8	1.1329×10^{-5}
5	4	6	40	7	34	9	1.0094×10^{-7}
6	4	5	41	6	33	11	4.1417×10^{-7}
7	4	4	42	4	26	2	1.9404×10^{-6}
8	5	6	39	2	23	5	1.3189×10^{-5}
9	2	7	41	14	29	7	7.8119×10^{-12}
10	3	5	42	5	37	8	1.1125×10^{-9}

Application

We applied the single marker allele-based method discussed in section 2 to find markers associated to the disease. The method is applied to each SNP in the whole-genome 10 K SNPs data. The 10 smallest unadjusted *p*-values together with the corresponding SNP identies and the adjusted *p*-value for Bonferroni multiple testing correction method are given in Table 3. We find only one significant marker at the level of significance 5 %, i.e. marker with identity 1513978 in chromosome 2.

Tabel 5. Relation between parameter, sample size and *p*-value that is used to generate controls sample and cases sample.

No.	Parameter of Controls Sample	Parameter of Cases Sample	Mean and Standard deviation of *p*-value for n=50	Mean and Standard deviation of *p*-value for n=100	Mean and Standard deviation of *p*-value for n=200
1.	(0.1, 0.1, 0.8)	(0.1, 0.1, 0.8)	0.4959 (0.2943)	0.4973 (0.2920)	0.4977 (0.2890)
2.	(0.1, 0.1, 0.8)	(0.1, 0.3, 0.6)	0.2408 (0.2731)	0.1221 (0.2000)	0.0317 (0.0909)
3.	(0.1, 0.1, 0.8)	(0.1, 0.5, 0.4)	0.0319 (0.0910)	2.5897×10^{-3} (0.0163)	3.1156×10^{-5} (0.0012)
4.	(0.1, 0.1, 0.8)	(0.1, 0.7, 0.2)	9.7392×10^{-4} (8.0676×10^{-3})	1.0812×10^{-5} (6.0991×10^{-4})	1.1876×10^{-11} (8.3589×10^{-10})
5.	(0.1, 0.3, 0.6)	(0.1, 0.1, 0.8)	0.2376 (0.2726)	0.1210 (0.1976)	0.0319 (0.0853)
6.	(0.1, 0.3, 0.6)	(0.1, 0.3, 0.6)	0.4932 (0.2923)	0.4981 (0.2898)	0.4958 (0.2892)
7.	(0.1, 0.3, 0.6)	(0.1, 0.5, 0.4)	0.2377 (0.2713)	0.1207 (0.1991)	0.0300 (0.0856)
8.	(0.1, 0.3, 0.6)	(0.1, 0.7, 0.2)	0.0238 (0.0774)	1.3366×10^{-5} (0.0101)	4.6073×10^{-6} (8.4603×10^{-5})
9.	(0.1, 0.5, 0.4)	(0.1, 0.1, 0.8)	0.0306 (0.0867)	0.0028 (0.0183)	1.9387×10^{-5} (3.9089×10^{-4})
10.	(0.1, 0.5, 0.4)	(0.1, 0.3, 0.6)	0.2380 (0.2702)	0.1229 (0.1985)	0.0320 (0.0904)
11.	(0.1, 0.5, 0.4)	(0.1, 0.5, 0.4)	0.5021 (0.2924)	0.4988 (0.2905)	0.4946 (0.2900)
12.	(0.1, 0.5, 0.4)	(0.1, 0.7, 0.2)	0.2043 (0.2542)	0.0893 (0.1694)	0.0169 (0.0609)
13.	(0.1, 0.7, 0.2)	(0.1, 0.1, 0.8)	0.0010 (0.0106)	3.2862×10^{-6} (7.8937×10^{-5})	8.7585×10^{-12} (5.7496×10^{-10})
14.	(0.1, 0.7, 0.2)	(0.1, 0.3, 0.6)	0.0230 (0.0728)	0.0016 (0.0137)	5.7751×10^{-6} (1.4118×10^{-4})
15.	(0.1, 0.7, 0.2)	(0.1, 0.5, 0.4)	0.2064 (0.2583)	0.0851 (0.1608)	0.0176 (0.0623)
16.	(0.1, 0.7, 0.2)	(0.1, 0.7, 0.2)	0.4982 (0.2935)	0.4982 (0.2905)	0.4997 (0.2885)

Table 6. Table of the 10 smallest and adjusted *p*-values in single marker method.

No.	Chromosome	SNP Identity	Logistic Regression Statistical Value	*p*-value	Adjusted *p*-value
1.	2	1513978	20.9886	4.6204×10^{-6}	0.0462
2.	9	1510558	20.6612	5.4815×10^{-6}	0.0548
3.	8	15451758	15.5096	8.2115×10^{-5}	0.8211
4.	X	1511631	14.5918	1.3350×10^{-4}	1.0000
5.	9	157485	13.9938	1.8342×10^{-4}	1.0000
6.	3	1516091	13.9762	1.8514×10^{-4}	1.0000
7.	9	1518902	13.5946	2.2683×10^{-4}	1.0000
8.	11	1508808	13.5792	2.2871×10^{-4}	1.0000
9.	7	1510253	13.4534	2.4456×10^{-4}	1.0000
10.	11	1508850	13.3465	2.5891×10^{-4}	1.0000

Disscussion

In this paper, we have described how to apply multiple testing correction methods in the whole-genome SNP data. Bonferroni multiple testing correction method gives only a small number of significant SNPs. However, the disease of interest is a complex disease so that many genes involved to the disease of individuals. Thus the described method may not a good method to find genes related to the disease.

Using recent technology, it is allowed to have highly dense SNP data in the whole-genome which use hundreds of thousand SNPs (see for examples, Nishida et al. (2008), Pengyuan et al. (2008), Zheng et al. (2009), and the multiple testing correction methods become very conservative and finding an excellent method still become a challenge research.

REFERENCES

1. D. Balding, A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7, 781-791, 2006.
2. N. Nishida, Evaluating the performance of affimetrix SNP 6.0 platform with 400 Japanese individuals. *BMC Genomics* 9(1):431.2008.
3. L. Pengyuan, Familial Aggregation of Common Sequence Variants on 15q24-25.1 in Lung Cancer. Journal of the National Cancer Institute 100:1326-1330. 2008.
4. A. Setiawan, *Statistical Data Analysis of Genetic Data in Twin Studies and Association Studies*, Vrije Universiteit, Amsterdam, Ph.D Thesis, 2007.
5. W. Zheng, Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1 *Nature Genetics* 41:324-328. 2009.

Adi Setiawan : Department of Mathematics, Satya Wacana Christian University, Jl. Diponegoro 52-60 Salatiga 50711, Indonesia. E-mail: adi_setia_03@yahoo.com