Laboratory Note

Genetic Epistasis
II - Assessing Algorithm BEAM 3.0
LN-2-2014

Ricardo Pinho and Rui Camacho
FEUP
Rua Dr Roberto Frias, s/n,
4200-465 PORTO
Portugal
Fax: (+351) 22 508 1440
e-mail: ei09045@fe.up.pt
www : http://www.fe.up.pt/∼ei09045
rcamacho@fe.up.pt
www : http://www.fe.up.pt/∼rcamacho

May 2014

Abstract

In this lab note, the algorithm BEAM 3.0 is presented and tested for main effect detection. This is a bayesian algorithm that creates a graph with SNPs and the relations between them and the disease expression. The results obtained reveal a high detection for data sets with higher allele frequencies. This is also true for the population size, however this increases Type I Error Rates, therefore Power values are nearly equal to the error rates. The algorithm seems very scalable with the data sets used, and may be scalable to large genome wide association studies.

1 Introduction

The Bayesian Epistasis association Mapping (BEAM) [ZL07] is a stochastic algorithm that uses a Markov chain Monte Carlo (MCMC) [ADH10] to create posterior probabilities that each marker is associated with the disease phenotype.

Instead of the standard epistatic detection using χ^2 statistic, BEAM uses a new B statistic. The B statistic is defined by:

$$B_M = ln\frac{P_A(D_M, U_M)}{P_0(D_M, U_M)} = ln\frac{P_{join}(D_M)[P_{ind}(U_M) + P_{join}(U_M)]}{P_{ind}(D_M, U_M) + P_{join}(D_M, U_M)} \quad (1)$$

where M represents each set of k markers, representing different complexities of interactions. D_M and U_M are genotype data from M cases and controls and $P_0(D_M, U_M)$ and $P_A(D_M, U_M)$ are the Bayes factors. P_{ind} is the distribution that assumes independence among markers in M and P_{join} is a saturated joint distribution of genotype combinations among all markers in M.
BEAM3 introduces multi-SNP associations and high-order interactions flexibility, using graphs, reducing the complexity and increasing the Power. BEAM3 [Zha12] produces cleaner results with improved mapping sensitivity and specificity.
Initially, the disease graph is built based on the probability that a given genotype configuration is related to the phenotype, considering the frequencies of that genotype in controls and cases. Cliques (non overlapping groups of SNPs) are then generated based on the disease related SNPs. A joint probability model and MCMC are used to update the disease graph and create undirected edges between dependent SNPs.

1.1 Input files

The input file contains the phenotypes of all the individuals in the first row and the genotypes of each SNP on the subsequent rows.

1.2 Output files

The algorithm outputs 3 files: posterior file; g.dot file; and chi.txt. The posterior file contains the posterior probabilities of marginal and interaction

ID	Chr	Pos	0	1	0	0	1
rs1	chr1	1	1	0	2	0	1
rs2	chr1	2	1	2	1	1	0
rs3	chr1	3	1	2	2	0	1

Table 1: An example of the input file containing the index of the SNPs, the chromosome that they belong to, the position of the SNP, the phenotype, corresponding to the first row and subsequent rows correspond to the genotype of each SNP for all individuals.

associations per SNP. The g.dot file contains the disease graph. The file requires a graph visualization software, such as GraphViz. The chi.txt contains the chi square results, together with allele counts.

1.3 Parameters

There are some options available to the user:

- "-filter k": Tells the program to filter SNPs with too many missing genotypes.
- "-sample burnin mcmc": Specifies the number of sampling interactions by the MCMC.The default value is 100.
- "-prior p": specifies how likely each SNP is associated with the disease. By default, p=5/L, where L is the number of SNPs.
- "-T t": Specifies the temperature which the MCMC starts running. With a high temperature, the program can jump out of local modes with few iterations. However, it can make the program very slow in the first iterations.

2 Experimental Settings

The datasets used in the experiments are characterized in Lab Note 1. The computer used for this experiments used the 64-bit Ubuntu 13.10 operating system, with an Intel(R) Core(TM)2 Quad CPU Q6600 2.40GHz processor and 8,00 GB of RAM memory.
The parameters used in this experiment are the default parameters, with the exception of "-prior p", which is p=2/L.

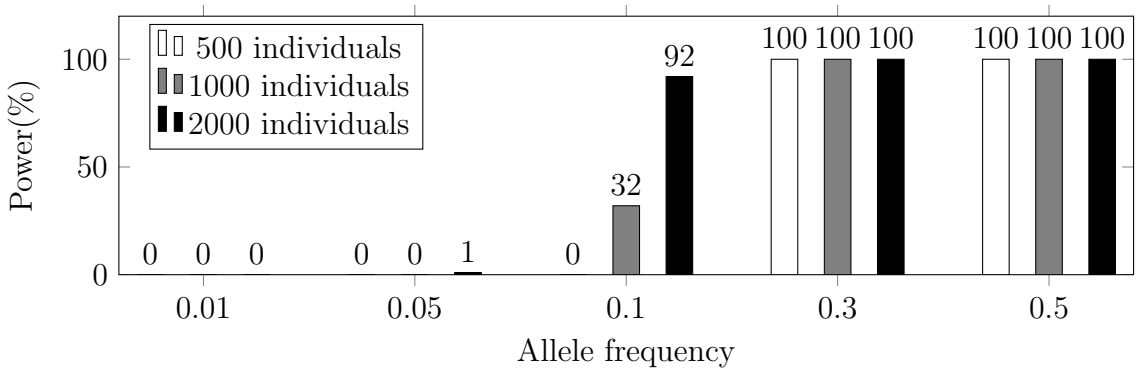


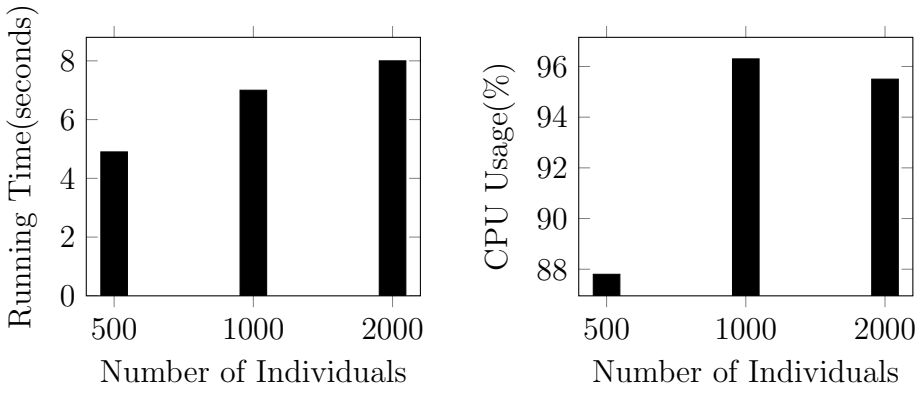
Figure 1: Power by allele frequency. For each frequency, three sizes of data sets were used to measure the Power, with odds ratio of 2.0 and prevalence of 0.02. The Power is measured by the amount of data sets where the ground truth was amongst the most relevant results, out of all 100 data sets.

3 Results

The results of epistasis detection of the algorithm consist of posterior probabilities. This is not comparable with χ^2 tests, therefore only main effect detections will be considered for this experiment.
Figure 1 shows near 0% Power for allele frequencies lower than 0.1, but increases greatly reaching 100% Power for frequencies of 0.3 and 0.5. There is also a clear growth with population size, especially in data sets with 0.1 minor allele frequency.

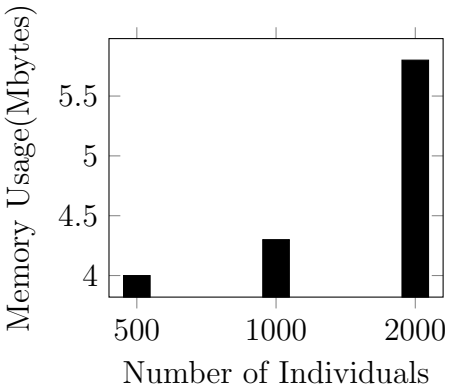
The running time (a) of these experiments show a steady increase, with a difference of nearly 3 seconds between data sets with 500 individuals and data sets with 2000 individuals. The increase in running time is not very significant, which may translate to larger data sets. This is also true for memory usage (c), with only 1.5 MB increase from 500 to 2000 individuals in a data set. The CPU usage (b increased has an increase of nearly 10% from 500 individuals to 1000, lowering slightly for 2000 individuals.

The error rate results in Figure 3 contain high values of false positives. The percentage of Type I Error Rate is bigger than the Power for smaller allele frequencies. In frequencies higher than 0.1 the Type I Error Rate is lower than the Power but the difference of both percentages decrease as the number of individuals increases. This means that for a bigger sized data sets, it is more likely to find the ground truth but it is also more likely to be accompanied by false positives.



(a) Average running time

(b) Average CPU usage



(c) Average memory usage

Figure 2: Comparison of scalability measures between different sized data sets. This figures shows the average running time, CPU usage, and memory usage by each data set. The data sets have a minor allele frequency is 0.5, 2.0 odds ratio, 0.02 prevalence.

The distribution of Power by odds ratio reveals a big increase in Power with the increase of odds ratio in Figure 5. This is similar to the Power by population size in Figure 4. Data sets with low allele frequencies have a near 0% Power. With 0.1 minor allele frequency, there is a significant increase, having 92% of Power, and reaching 100% for higher allele frequencies in Figure 7. There is no clear difference in Power with prevalence changes on Figure 6.

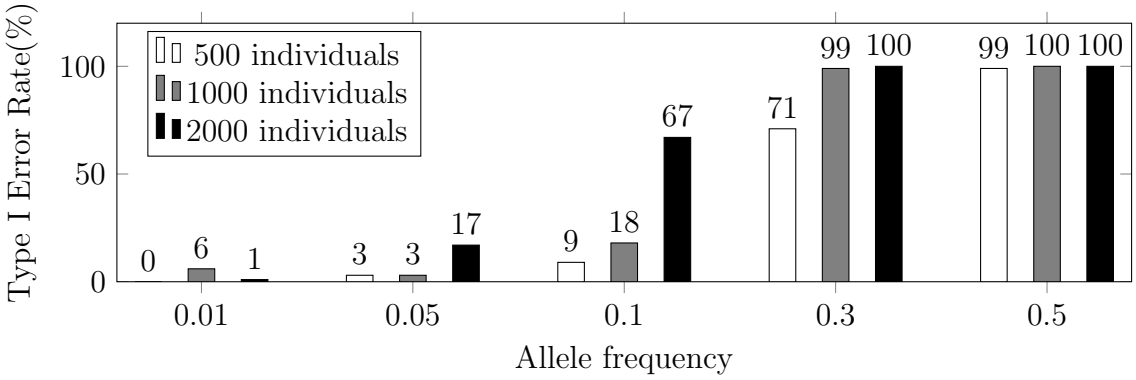


Figure 3: Type I Error Rate by allele frequency and population size, with odds ratio of 2.0 and prevalence of 0.02. The Type I Error Rate is measured by the amount of data sets where the false positives were amongst the most relevant results, out of all 100 data sets.

4 Summary

BEAM3 is the third iteration of a bayesian algorithm that uses posterior probabilities to detect epistasis. BEAM3 generates a disease graph representing multi-SNP associations that have a high probability of being related to the disease phenotype expression. This graph is updated using MCMC. This version of BEAM also outputs χ^2 values of single SNPs, which are comparable with other algorithms. Due to this the results consist of main effect detection only. The Power obtained reveals similar values for Power and Type I Error Rate, increasing with allele frequency and population size, but type 1 errors are lower in relation to Power in data sets with high allele frequency and low population size. The scalability of the algorithm is promising.

References

[ADH10] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:269–342, 2010.

[Zha12] Yu Zhang. A novel bayesian graphical model for genome-wide multi-SNP association mapping. *Genetic Epidemiology*, 36:36–47, 2012.

[ZL07] Yu Zhang and Jun S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, 39:1167–1173, 2007.

A Bar graphs

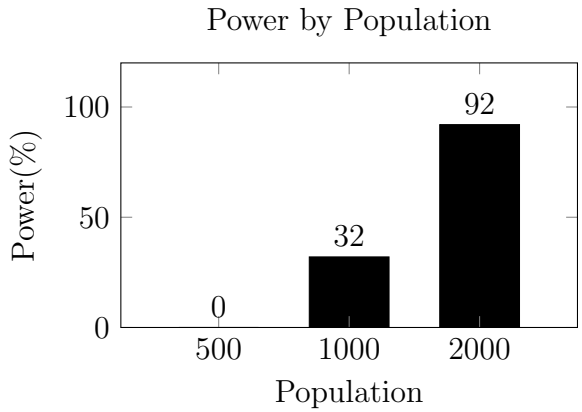


Figure 4: Distribution of the Power by population. The allele frequency is 0.1, the odds ratio is 2.0, and the prevalence is 0.02.

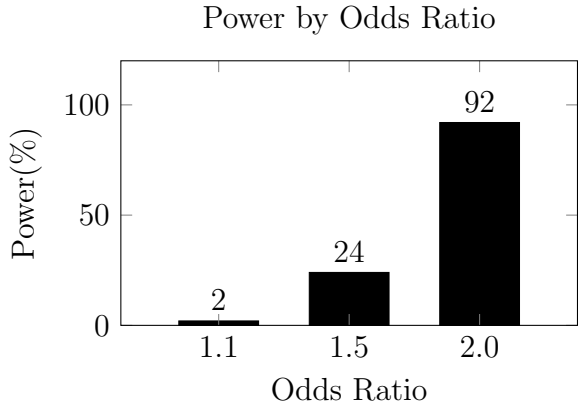


Figure 5: Distribution of the Power by odds ratios. The allele frequency is 0.1, the number of individuals is 2000, and the prevalence is 0.02.

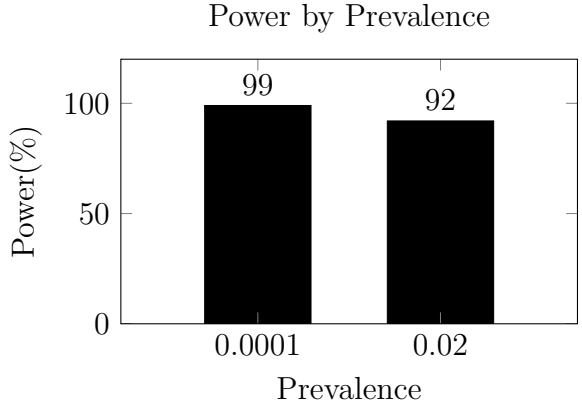


Figure 6: Distribution of the Power by prevalence. The allele frequency is 0.1, the number of individuals is 2000, and the odds ratio is 2.0.

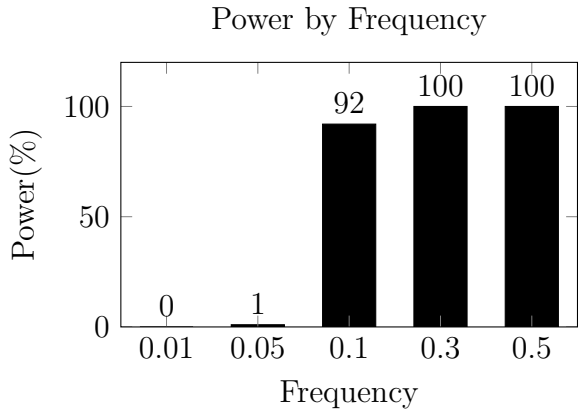


Figure 7: Distribution of the Power by allele frequency. The number of individuals is 2000, the odds ratio is 2.0, and the prevalence is 0.02.

B Table of Results

Table 2: A table containing the percentage of true positives and false positives in each configuration. The first column contains the description of the configuration. The second and third columns contain the number of datasets with true positives and false positives respectively, out of all 100 data sets per configuration.

Configuration*	TP (%)	FP (%)
0.5,500,ME,2.0,0.02	100	99
0.5,500,ME,2.0,0.0001	100	95

0.5,500,ME,1.5,0.02	100	53
0.5,500,ME,1.5,0.0001	100	57
0.5,500,ME,1.1,0.02	80	20
0.5,500,ME,1.1,0.0001	79	22
0.5,2000,ME,2.0,0.02	100	100
0.5,2000,ME,2.0,0.0001	100	100
0.5,2000,ME,1.5,0.02	100	100
0.5,2000,ME,1.5,0.0001	100	100
0.5,2000,ME,1.1,0.02	100	100
0.5,2000,ME,1.1,0.0001	100	98
0.5,1000,ME,2.0,0.02	100	100
0.5,1000,ME,2.0,0.0001	100	100
0.5,1000,ME,1.5,0.02	100	100
0.5,1000,ME,1.5,0.0001	100	97
0.5,1000,ME,1.1,0.02	100	57
0.5,1000,ME,1.1,0.0001	100	60
0.3,500,ME,2.0,0.02	100	71
0.3,500,ME,2.0,0.0001	100	79
0.3,500,ME,1.5,0.02	88	24
0.3,500,ME,1.5,0.0001	89	30
0.3,500,ME,1.1,0.02	21	11
0.3,500,ME,1.1,0.0001	23	6
0.3,2000,ME,2.0,0.02	100	100
0.3,2000,ME,2.0,0.0001	100	100
0.3,2000,ME,1.5,0.02	100	99
0.3,2000,ME,1.5,0.0001	100	100
0.3,2000,ME,1.1,0.02	100	54
0.3,2000,ME,1.1,0.0001	100	50
0.3,1000,ME,2.0,0.02	100	99
0.3,1000,ME,2.0,0.0001	100	100
0.3,1000,ME,1.5,0.02	100	68
0.3,1000,ME,1.5,0.0001	100	63
0.3,1000,ME,1.1,0.02	90	25
0.3,1000,ME,1.1,0.0001	81	25
0.1,500,ME,2.0,0.02	0	9
0.1,500,ME,2.0,0.0001	12	17
0.1,500,ME,1.5,0.02	0	5
0.1,500,ME,1.5,0.0001	0	6

0.1,500,ME,1.1,0.02	0	6
0.1,500,ME,1.1,0.0001	0	5
0.1,2000,ME,2.0,0.02	92	67
0.1,2000,ME,2.0,0.0001	99	76
0.1,2000,ME,1.5,0.02	24	16
0.1,2000,ME,1.5,0.0001	44	29
0.1,2000,ME,1.1,0.02	2	8
0.1,2000,ME,1.1,0.0001	1	7
0.1,1000,ME,2.0,0.02	32	18
0.1,1000,ME,2.0,0.0001	59	38
0.1,1000,ME,1.5,0.02	1	6
0.1,1000,ME,1.5,0.0001	6	10
0.1,1000,ME,1.1,0.02	0	7
0.1,1000,ME,1.1,0.0001	0	5
0.05,500,ME,2.0,0.02	0	3
0.05,500,ME,2.0,0.0001	0	6
0.05,500,ME,1.5,0.02	0	4
0.05,500,ME,1.5,0.0001	0	4
0.05,500,ME,1.1,0.02	0	5
0.05,500,ME,1.1,0.0001	0	1
0.05,2000,ME,2.0,0.02	1	17
0.05,2000,ME,2.0,0.0001	7	25
0.05,2000,ME,1.5,0.02	0	3
0.05,2000,ME,1.5,0.0001	0	13
0.05,2000,ME,1.1,0.02	0	5
0.05,2000,ME,1.1,0.0001	0	6
0.05,1000,ME,2.0,0.02	0	3
0.05,1000,ME,2.0,0.0001	1	18
0.05,1000,ME,1.5,0.02	0	2
0.05,1000,ME,1.5,0.0001	0	5
0.05,1000,ME,1.1,0.02	0	7
0.05,1000,ME,1.1,0.0001	0	3
0.01,500,ME,2.0,0.02	0	0
0.01,500,ME,2.0,0.0001	0	6
0.01,500,ME,1.5,0.02	0	0
0.01,500,ME,1.5,0.0001	0	6
0.01,500,ME,1.1,0.02	0	0
0.01,500,ME,1.1,0.0001	0	6

0.01,2000,ME,2.0,0.02	0	1
0.01,2000,ME,2.0,0.0001	0	3
0.01,2000,ME,1.5,0.02	0	3
0.01,2000,ME,1.5,0.0001	0	2
0.01,2000,ME,1.1,0.02	0	3
0.01,2000,ME,1.1,0.0001	0	2
0.01,1000,ME,2.0,0.02	0	6
0.01,1000,ME,2.0,0.0001	0	3
0.01,1000,ME,1.5,0.02	0	7
0.01,1000,ME,1.5,0.0001	0	3
0.01,1000,ME,1.1,0.02	0	3
0.01,1000,ME,1.1,0.0001	0	4

*MAF,POP,MOD,OR,PREV where MAF represents the minor allele frequency, POP is the number of individuals, MOD is the used model (with or without main effect and with or without epistasis effect), OR is the odds ratio and PREV is the prevalence of the disease.

Table 3: A table containing the running time, cpu usage and memory usage in each configuration.

Configuration*	Running Time (s)	CPU Usage (%)	Memory Usage (KB)
0.5,500,ME,2.0,0.02	04.90	87.81	4152.80
0.5,500,ME,2.0,0.0001	03.30	87.16	3446.24
0.5,500,ME,1.5,0.02	02.16	86.74	2723.76
0.5,500,ME,1.5,0.0001	02.15	82.36	2757.20
0.5,500,ME,1.1,0.02	01.82	80.97	2566.12
0.5,500,ME,1.1,0.0001	01.73	83.54	2556.08
0.5,2000,ME,2.0,0.02	08.02	95.53	5986.72
0.5,2000,ME,2.0,0.0001	05.17	94.16	4108.72
0.5,2000,ME,1.5,0.02	02.78	92.74	3512.88
0.5,2000,ME,1.5,0.0001	02.59	93.39	3508.48
0.5,2000,ME,1.1,0.02	02.34	93.38	3493.44
0.5,2000,ME,1.1,0.0001	02.30	93.32	3492.60
0.5,1000,ME,2.0,0.02	06.96	96.31	4437.08
0.5,1000,ME,2.0,0.0001	03.79	95.00	3240.00
0.5,1000,ME,1.5,0.02	02.38	93.54	2771.80
0.5,1000,ME,1.5,0.0001	02.25	93.99	2729.16
0.5,1000,ME,1.1,0.02	02.10	93.08	2686.12

0.5,1000,ME,1.1,0.0001	02.02	93.41	2665.64
0.3,500,ME,2.0,0.02	02.60	94.60	2970.00
0.3,500,ME,2.0,0.0001	02.32	93.51	2917.44
0.3,500,ME,1.5,0.02	01.93	93.41	2615.88
0.3,500,ME,1.5,0.0001	01.83	92.49	2607.24
0.3,500,ME,1.1,0.02	01.17	89.70	2483.28
0.3,500,ME,1.1,0.0001	01.09	88.25	2476.68
0.3,2000,ME,2.0,0.02	02.77	94.79	3534.72
0.3,2000,ME,2.0,0.0001	02.95	95.25	3563.44
0.3,2000,ME,1.5,0.02	02.38	94.49	3493.60
0.3,2000,ME,1.5,0.0001	02.32	94.27	3492.92
0.3,2000,ME,1.1,0.02	02.30	94.73	3491.44
0.3,2000,ME,1.1,0.0001	02.28	94.56	3490.44
0.3,1000,ME,2.0,0.02	02.42	94.03	2886.64
0.3,1000,ME,2.0,0.0001	02.45	94.19	2831.72
0.3,1000,ME,1.5,0.02	02.04	93.96	2675.80
0.3,1000,ME,1.5,0.0001	02.04	93.88	2671.00
0.3,1000,ME,1.1,0.02	01.82	93.43	2665.28
0.3,1000,ME,1.1,0.0001	01.76	92.86	2662.68
0.1,500,ME,2.0,0.02	0.80	85.95	2471.00
0.1,500,ME,2.0,0.0001	0.95	88.33	2520.12
0.1,500,ME,1.5,0.02	0.61	82.27	2383.04
0.1,500,ME,1.5,0.0001	0.64	84.21	2432.96
0.1,500,ME,1.1,0.02	0.57	82.88	2367.56
0.1,500,ME,1.1,0.0001	0.58	81.66	2408.72
0.1,2000,ME,2.0,0.02	02.24	93.47	3493.84
0.1,2000,ME,2.0,0.0001	02.26	94.12	3492.40
0.1,2000,ME,1.5,0.02	01.37	90.66	3489.68
0.1,2000,ME,1.5,0.0001	01.45	91.55	3484.24
0.1,2000,ME,1.1,0.02	01.02	90.22	3482.16
0.1,2000,ME,1.1,0.0001	0.99	90.46	3483.44
0.1,1000,ME,2.0,0.02	01.38	89.81	2681.04
0.1,1000,ME,2.0,0.0001	01.50	91.44	2696.48
0.1,1000,ME,1.5,0.02	0.78	88.49	2655.24
0.1,1000,ME,1.5,0.0001	0.83	88.49	2653.08
0.1,1000,ME,1.1,0.02	0.69	83.77	2652.16
0.1,1000,ME,1.1,0.0001	0.68	89.10	2648.20
0.05,500,ME,2.0,0.02	0.59	81.11	2380.88

0.05,500,ME,2.0,0.0001	0.93	84.09	2439.40
0.05,500,ME,1.5,0.02	0.57	81.72	2361.20
0.05,500,ME,1.5,0.0001	0.60	81.99	2390.04
0.05,500,ME,1.1,0.02	0.59	79.48	2361.20
0.05,500,ME,1.1,0.0001	0.57	81.46	2381.48
0.05,2000,ME,2.0,0.02	01.18	89.59	3485.56
0.05,2000,ME,2.0,0.0001	01.19	89.80	3484.76
0.05,2000,ME,1.5,0.02	0.98	89.07	3480.08
0.05,2000,ME,1.5,0.0001	0.98	89.80	3480.16
0.05,2000,ME,1.1,0.02	0.94	89.82	3479.28
0.05,2000,ME,1.1,0.0001	0.94	90.33	3481.12
0.05,1000,ME,2.0,0.02	0.70	85.95	2651.56
0.05,1000,ME,2.0,0.0001	0.81	86.89	2653.84
0.05,1000,ME,1.5,0.02	0.67	81.01	2647.16
0.05,1000,ME,1.5,0.0001	0.70	82.83	2648.96
0.05,1000,ME,1.1,0.02	0.66	84.68	2648.20
0.05,1000,ME,1.1,0.0001	0.69	80.38	2647.76
0.01,500,ME,2.0,0.02	0.55	77.93	2340.40
0.01,500,ME,2.0,0.0001	0.59	79.62	2391.20
0.01,500,ME,1.5,0.02	0.54	81.51	2345.64
0.01,500,ME,1.5,0.0001	0.58	79.48	2387.76
0.01,500,ME,1.1,0.02	0.55	78.36	2349.92
0.01,500,ME,1.1,0.0001	0.59	79.67	2393.76
0.01,2000,ME,2.0,0.02	0.91	85.28	3476.88
0.01,2000,ME,2.0,0.0001	0.93	91.10	3479.40
0.01,2000,ME,1.5,0.02	0.91	91.18	3478.80
0.01,2000,ME,1.5,0.0001	0.92	91.62	3480.64
0.01,2000,ME,1.1,0.02	0.91	91.07	3477.44
0.01,2000,ME,1.1,0.0001	0.93	91.07	3478.96
0.01,1000,ME,2.0,0.02	0.66	86.84	2645.76
0.01,1000,ME,2.0,0.0001	0.67	89.19	2649.60
0.01,1000,ME,1.5,0.02	06.55	88.46	6100.36
0.01,1000,ME,1.5,0.0001	0.67	80.52	2646.28
0.01,1000,ME,1.1,0.02	0.66	84.18	2644.68
0.01,1000,ME,1.1,0.0001	0.66	81.46	2645.48

*MAF,POP,MOD,OR,PREV where MAF represents the minor allele frequency, POP is the number of individuals, MOD is the used model (with or without main effect and with or without epistasis effect), OR is the odds ratio and PREV is the prevalence of the disease.