Laboratory Note

Genetic Epistasis
V - Assessing Algorithm SNPRuler
LN-5-2014

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 SNPRuler is presented. SNPRuler is an epistatic detection algorithm written in Java that creates rules based on the epistatic interactions detected in data sets. Using many configurations of data sets, the results obtained show a correlation between Power and the number of sampled individuals and a correlation between Power and minor allele frequency. This shows that the algorithm has a very high accuracy in optimal conditions, but has very low accuracy in below optimal conditions. The algorithm is very scalable with different number of individuals, with only a slight increase in running time and memory usage. The Type I Error Rate is very low in all configurations.

1 Introduction

SNPRuler [WYY+10] is a rule based algorithm that, based on the relations between SNPs and the phenotype related to the expression of a disease, creates rules of association, between SNPs and the phenotype expression. The order or magnitude of these interactions can have any amount of SNPs. For each rule, a 3x3 table is generated, relating to the probability of each possible genotype combination and phenotype expression.
The way that rules are defined is described in the following steps:

1. **Literal** - *A literal s is an index-value pair (i,v) with i denoting an index and v a value in 1,2,3 representing the possible genotypes. A sample satisfies a literal(i,v) if and only if its i-th SNP has the value v.*

2. **Predictive Rule** - *A predictive rule* $(r, \zeta) : s_1 \cap s_2 \cap ... \cap s_n \rightarrow \zeta$*, is an association between a conjunction of n literals denoted as r and a class label* ζ*. A sample satisfies* (r, ζ) *if and only if it satisfies all literals in r and its class label is* ζ*.*

3. **Literal Relevance** - *Given a predictive rule* (r, ζ) *and a utility function* $U(r, \zeta)$ *for rule measurement, a literal* s_i *in the rule r is relevant if and only if* $U(r, \zeta) > U(r - s_i, \zeta)$*. Here,* $R - s_i$ *means removing* s_i *from r.*

4. **Closed Rule** - *A predictive rule* (r, ζ) *if closed if and only if there is not there is no literal* s_i *which satisfies* $U(r + s_i, \zeta) > U(r, \zeta)$*. Here,* $R + s_i$ *means adding* s_i *into r.*

The measurement rule of relevance is χ^2 statistic. Considering that most of the epistatic interactions involve many SNPs, before creating rules, an upper bound is used to determine if a new SNP will reveal to be significant to a rule. This decreases the amount of rules created immensely, compared to exhaustive searches. A branch-and-bound approach is used for this effect.

1.1 Input files

The algorithm is written in Java and receives a file containing the genotype and the phenotype expressed for each individual. The first row contains the number of each SNP and the final column corresponds to the label. Each subsequent row contains an individuals genotype 0,1,2 corresponding to homozygous dominant genotype (AA), heterozygous genotype (Aa), and homozygous recessive genotype (aa). The Label 0,1 corresponds to control and disease affected, respectively.

X1	X2	X3	X4	Label
1	1	0	2	0
1	0	1	1	1
1	2	2	0	1

Table 1: An example of the input file containing genotype and phenotype information with 4 SNPs and 3 individuals.

1.2 Output files

The output is a list of interactions ranked by their significance in the χ^2 test. A post-processing calculates the P-value of these interactions adjusting the χ^2 test with a Bonferroni correction with a significance threshold of 0.3.

1.3 Parameters

There are 3 configurable parameters:

- *listSize* - The expected number of interactions.
- *depth* - Order of interaction. Number of interacting SNPs.
- *updateRatio* - The step size of updating a rule. Takes a value between 0 and 1, 0 being not updated and 1 updating a rule at each step.

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 algorithms settings consist of a -Xmx7000M heap size, with a maximum number or rules set as 50 000. The length of the rules is 2, considering that the data sets used contain ground-truths of pairs of SNPs. The pruning threshold is 0, which means that all possible combinations will be tested.

3 Results

SNPRuler is used exclusively for the interactive effect, therefore data sets with main effect and full effect wont be analyzed.
In the Figure 1 the Power obtained from each allele frequency with different

population sizes is displayed. For data sets with 500 individuals, the Power is nearly 0 for all allele frequencies. However, as the population size increases, the Power starts to rise in data sets with allele frequency higher than 0.1. This is also true for data sets with 2000 individuals, with a slightly higher Power than smaller data sets. The configuration with the most Power corresponds to the datasets with 2000 population and 0.5 minor allele frequency.

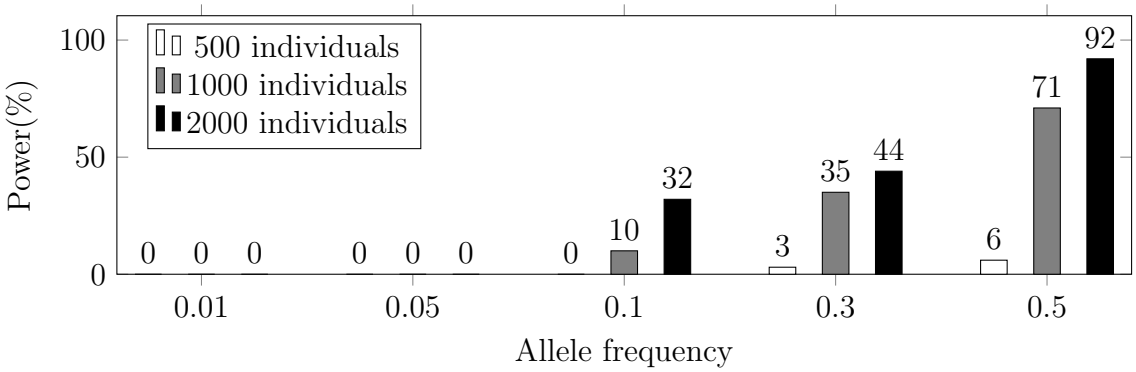
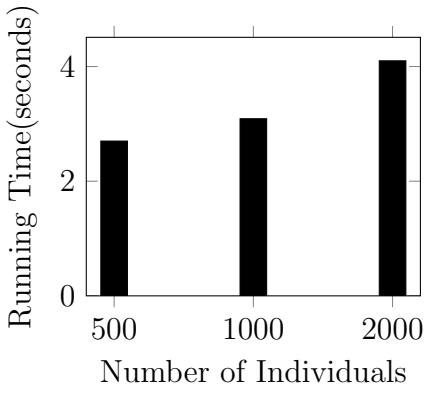


Figure 1: Power by allele frequency. For each frequency, three sizes of data sets were used to measure the Power, with an odds ratio of 2.0 and prevalence of 0.02. The Power is measured by the amount of data sets where the ground truth was amongst the most relevant results, out of all 100 data sets.

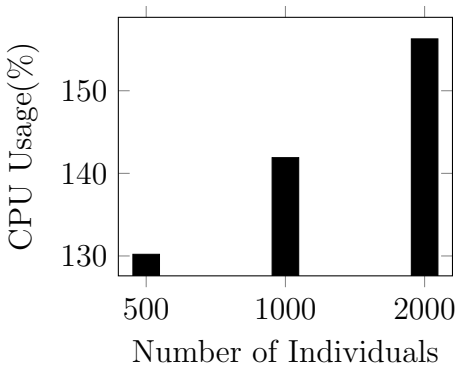
In Figure 2 the average running time, percentage of CPU usage, and memory usage are displayed by individuals in the data set, to evaluate the scalability of the algorithm. The results show that there is a slight increase in running time when applied to larger data sets. In this results, the increase in running time is not very significant. The CPU usage shows an increase with the data set size, with all the data sets having a CPU usage higher than 100%. This means that for each data set, more than one core was used. The memory usage results show that there is an increase of nearly 10 megabytes in memory usage. This increase may be significant in more complex data sets but is not as significant as the running time increase or the CPU usage.

For the Type I Error Rate test, Figure 3 shows that the Type I Error Rate is relatively small across all the data sets, having outliers with allele frequency of 0.1 and 2000 individuals. This is the only groups of configurations that yield a Type I Error Rate higher than 1%.

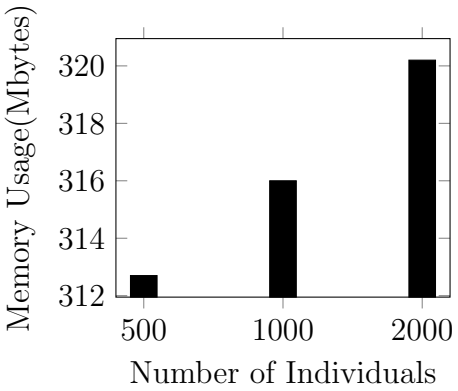
According to Figure 4 we can conclude that the number of individuals has a big influence in the Power of the algorithm. This is also true for the allele frequency. With very small number of individuals, the Power is nearly 0.The



(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.

Power also increases with the frequency of the alleles with the ground truth. On Figure 5 and 6, the influence of odds ratio, through the penetrance table of the disease, and the prevalence of the disease are undetermined. There is an increase in Power with the odds ratio of 1.5, but it decreases for 2.0 odds ratio. The difference in prevalence does not show a very significant difference in Power. Figure 7 shows the Power by frequency, independent of population size.

Overall, the Power of the algorithm shows very high accuracy in certain configurations with the optimal conditions, but also shows very low Power in many configurations.

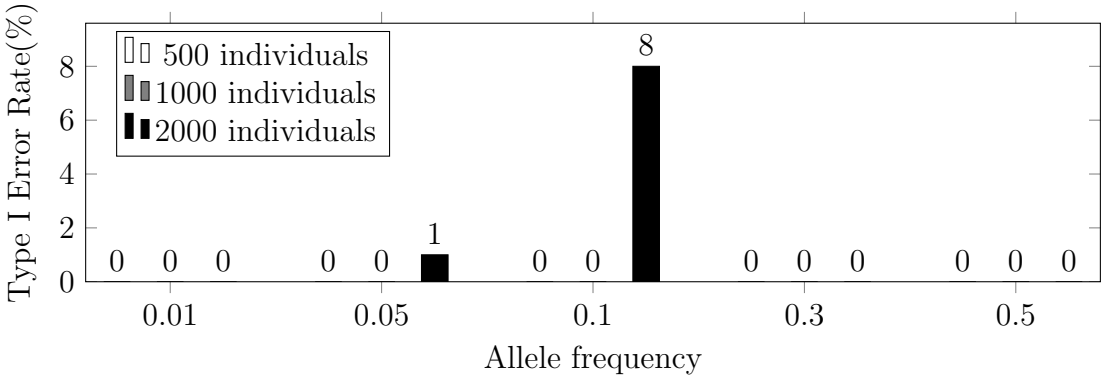


Figure 3: Type I Error Rate by allele frequency and population size, with an odds ratio of 2.0 and prevalence of 0.02. The Type I Error Rate is measured by the amount of data sets where the false positives were amongst the most relevant results, out of all 100 data sets.

4 Summary

In this lab note, the algorithm SNPRuler was presented and tested to detect epistasis interactions that manifest complex diseases using generated data sets. The results obtained showed that The number of individuals is important to epistasis detection. Diseases with ground truths in high frequency SNPs are easier to detect. The scalability test revealed a significant increase in the use of computer resources and running time with the increase in number of individuals, which may have a significant impact in datasets with a higher amount of SNPs. The Type I Error Rate results show very low error rate in all configurations.

References

[WYY+10] Xiang Wan, Can Yang, Qiang Yang, Hong Xue, Nelson L S Tang, and Weichuan Yu. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics (Oxford, England)*, 26:30–37, 2010.

A Bar graphs

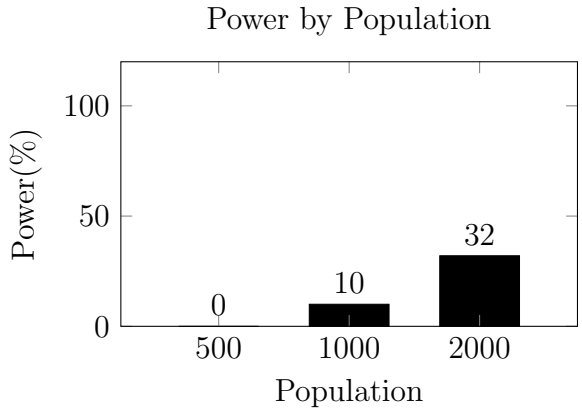


Figure 4: Distribution of the Power by population. The allele frequency is 0.1, the odds ratio is 2.0, and the prevalence is 0.02.

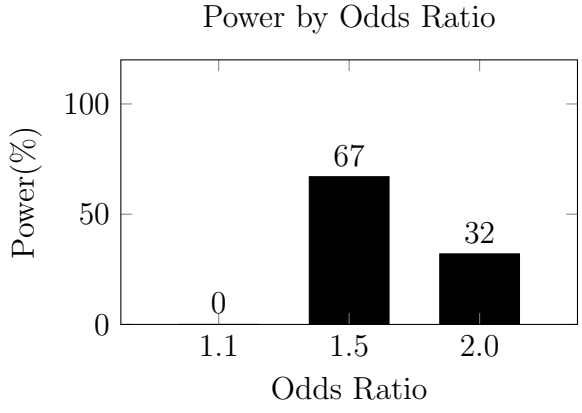


Figure 5: Distribution of the Power by odds ratios. The allele frequency is 0.1, the number of individuals is 2000, and the prevalence is 0.02.

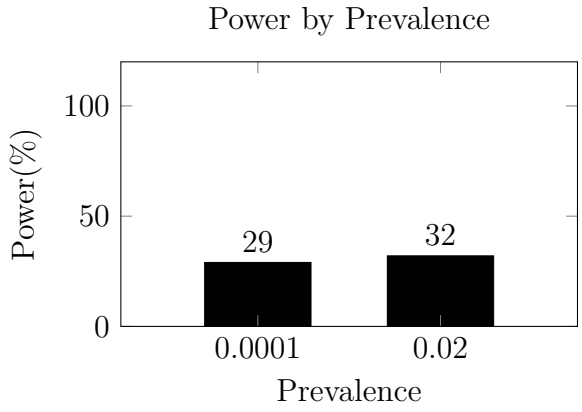


Figure 6: Distribution of the Power by prevalence. The allele frequency is 0.1, the number of individuals is 2000, and the odds ratio is 2.0.

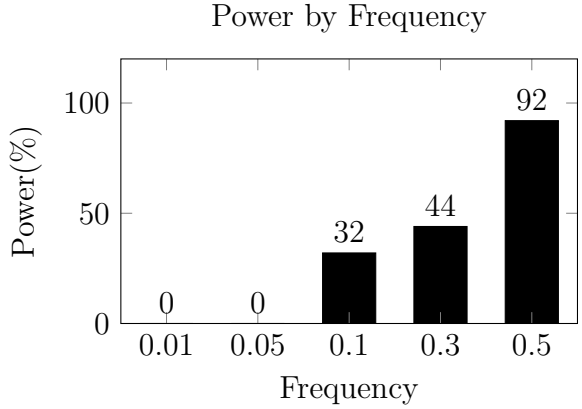


Figure 7: Distribution of the Power by allele frequency. The number of individuals is 2000, the odds ratio is 2.0, and the prevalence is 0.02.

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 (%)
0.5,2000,I,2.0,0.02	92	0
0.3,2000,I,1.5,0.02	89	7
0.3,2000,I,1.5,0.0001	84	3
0.5,2000,I,1.5,0.02	81	1
0.5,1000,I,2.0,0.02	71	0
0.1,2000,I,1.5,0.02	67	2
0.5,2000,I,2.0,0.0001	52	8
0.5,1000,I,2.0,0.0001	50	1
0.05,2000,I,2.0,0.0001	50	19
0.3,2000,I,2.0,0.02	44	0
0.3,1000,I,1.5,0.02	41	1
0.5,1000,I,1.5,0.02	40	0
0.3,1000,I,2.0,0.0001	36	0
0.3,1000,I,2.0,0.02	35	0
0.05,2000,I,1.5,0.0001	35	12
0.5,2000,I,1.5,0.0001	34	1
0.1,2000,I,2.0,0.02	32	8
0.3,1000,I,1.5,0.0001	29	1
0.1,2000,I,2.0,0.0001	29	0
0.1,2000,I,1.5,0.0001	29	1
0.05,2000,I,1.5,0.02	23	14
0.3,2000,I,2.0,0.0001	12	1
0.1,1000,I,2.0,0.02	10	0
0.5,500,I,2.0,0.02	6	0
0.3,500,I,2.0,0.0001	6	0
0.5,2000,I,1.1,0.02	5	1
0.5,500,I,2.0,0.0001	4	0
0.5,1000,I,1.5,0.0001	3	0
0.3,500,I,2.0,0.02	3	0
0.5,2000,I,1.1,0.0001	2	0
0.3,500,I,1.5,0.0001	2	0

0.3,2000,I,1.1,0.0001	2	0
0.1,1000,I,1.5,0.02	2	0
0.05,1000,I,2.0,0.0001	2	0
0.5,500,I,1.5,0.02	1	0
0.3,2000,I,1.1,0.02	1	0
0.1,2000,I,1.1,0.0001	1	1
0.1,1000,I,2.0,0.0001	1	0
0.5,500,I,1.5,0.0001	0	0
0.5,500,I,1.1,0.02	0	0
0.5,500,I,1.1,0.0001	0	0
0.5,1000,I,1.1,0.02	0	0
0.5,1000,I,1.1,0.0001	0	0
0.3,500,I,1.5,0.02	0	0
0.3,500,I,1.1,0.02	0	0
0.3,500,I,1.1,0.0001	0	0
0.3,1000,I,1.1,0.02	0	0
0.3,1000,I,1.1,0.0001	0	0
0.1,500,I,2.0,0.02	0	0
0.1,500,I,2.0,0.0001	0	0
0.1,500,I,1.5,0.02	0	0
0.1,500,I,1.5,0.0001	0	0
0.1,500,I,1.1,0.02	0	0
0.1,500,I,1.1,0.0001	0	0
0.1,2000,I,1.1,0.02	0	0
0.1,1000,I,1.5,0.0001	0	0
0.1,1000,I,1.1,0.02	0	0
0.1,1000,I,1.1,0.0001	0	0
0.05,500,I,2.0,0.02	0	0
0.05,500,I,2.0,0.0001	0	0
0.05,500,I,1.5,0.02	0	0
0.05,500,I,1.5,0.0001	0	0
0.05,500,I,1.1,0.02	0	0
0.05,500,I,1.1,0.0001	0	0
0.05,2000,I,2.0,0.02	0	1
0.05,2000,I,1.1,0.02	0	0
0.05,2000,I,1.1,0.0001	0	1
0.05,1000,I,2.0,0.02	0	0
0.05,1000,I,1.5,0.02	0	0

0.05,1000,I,1.5,0.0001	0	0
0.05,1000,I,1.1,0.02	0	0
0.05,1000,I,1.1,0.0001	0	0
0.01,500,I,2.0,0.02	0	0
0.01,500,I,2.0,0.0001	0	0
0.01,500,I,1.5,0.02	0	0
0.01,500,I,1.5,0.0001	0	1
0.01,500,I,1.1,0.02	0	0
0.01,500,I,1.1,0.0001	0	0
0.01,2000,I,2.0,0.02	0	0
0.01,2000,I,2.0,0.0001	0	1
0.01,2000,I,1.5,0.02	0	0
0.01,2000,I,1.5,0.0001	0	0
0.01,2000,I,1.1,0.02	0	0
0.01,2000,I,1.1,0.0001	0	1
0.01,1000,I,2.0,0.02	0	0
0.01,1000,I,2.0,0.0001	0	0
0.01,1000,I,1.5,0.02	0	0
0.01,1000,I,1.5,0.0001	0	0
0.01,1000,I,1.1,0.02	0	0
0.01,1000,I,1.1,0.0001	0	1

*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,I,2.0,0.02	2.70	130.19	320211.28
0.5,500,I,2.0,0.0001	2.69	136.88	319311.36
0.5,500,I,1.5,0.02	2.68	140.78	319508.72
0.5,500,I,1.5,0.0001	2.69	141.46	320285.24
0.5,500,I,1.1,0.02	2.73	136.88	320504.08
0.5,500,I,1.1,0.0001	2.70	136.47	319897.04
0.5,2000,I,2.0,0.02	4.10	156.28	327876.12
0.5,2000,I,2.0,0.0001	4.16	143.03	330393.48

0.5,2000,I,1.5,0.02	4.10	140.41	329206.28
0.5,2000,I,1.5,0.0001	4.01	136.85	327414.84
0.5,2000,I,1.1,0.02	3.96	125.00	325492.92
0.5,2000,I,1.1,0.0001	3.97	126.28	325792.92
0.5,1000,I,2.0,0.02	3.09	141.88	323600.36
0.5,1000,I,2.0,0.0001	3.12	139.30	324334.68
0.5,1000,I,1.5,0.02	3.08	141.47	323865.08
0.5,1000,I,1.5,0.0001	3.11	140.43	323880.44
0.5,1000,I,1.1,0.02	3.09	142.06	323780.88
0.5,1000,I,1.1,0.0001	3.12	141.69	323507.80
0.3,500,I,2.0,0.02	2.75	148.18	321318.64
0.3,500,I,2.0,0.0001	2.73	149.82	319605.00
0.3,500,I,1.5,0.02	2.73	149.43	321487.72
0.3,500,I,1.5,0.0001	2.75	150.12	320878.40
0.3,500,I,1.1,0.02	2.74	150.35	320952.24
0.3,500,I,1.1,0.0001	2.74	150.21	319914.16
0.3,2000,I,2.0,0.02	4.05	124.62	325950.12
0.3,2000,I,2.0,0.0001	4.04	119.74	325417.16
0.3,2000,I,1.5,0.02	4.04	122.47	325669.04
0.3,2000,I,1.5,0.0001	4.07	126.54	326147.32
0.3,2000,I,1.1,0.02	4.12	125.71	325679.80
0.3,2000,I,1.1,0.0001	4.11	123.02	325735.24
0.3,1000,I,2.0,0.02	3.07	118.96	322399.76
0.3,1000,I,2.0,0.0001	3.10	127.03	323056.56
0.3,1000,I,1.5,0.02	3.07	124.95	322673.52
0.3,1000,I,1.5,0.0001	3.11	131.41	323709.60
0.3,1000,I,1.1,0.02	3.09	134.61	323485.68
0.3,1000,I,1.1,0.0001	3.09	138.13	323444.76
0.1,500,I,2.0,0.02	2.75	119.13	320066.32
0.1,500,I,2.0,0.0001	2.74	119.29	319312.12
0.1,500,I,1.5,0.02	2.73	118.35	320222.28
0.1,500,I,1.5,0.0001	2.77	119.58	319002.32
0.1,500,I,1.1,0.02	2.77	118.50	320626.68
0.1,500,I,1.1,0.0001	2.76	121.01	320034.20
0.1,2000,I,2.0,0.02	4.01	119.18	325869.52
0.1,2000,I,2.0,0.0001	4.05	122.05	325484.96
0.1,2000,I,1.5,0.02	4.07	127.11	326038.04
0.1,2000,I,1.5,0.0001	4.09	126.69	326636.80

0.1,2000,I,1.1,0.02	4.10	127.66	326390.36
0.1,2000,I,1.1,0.0001	4.12	126.83	326720.76
0.1,1000,I,2.0,0.02	3.13	128.79	323402.72
0.1,1000,I,2.0,0.0001	3.13	128.00	323800.64
0.1,1000,I,1.5,0.02	3.12	126.40	323558.52
0.1,1000,I,1.5,0.0001	3.14	125.43	323584.04
0.1,1000,I,1.1,0.02	3.14	126.95	323569.56
0.1,1000,I,1.1,0.0001	3.14	126.27	323193.08
0.05,500,I,2.0,0.02	2.73	135.34	319177.48
0.05,500,I,2.0,0.0001	2.76	139.71	320980.88
0.05,500,I,1.5,0.02	2.73	131.66	320560.40
0.05,500,I,1.5,0.0001	2.76	139.02	320381.20
0.05,500,I,1.1,0.02	2.75	137.41	320737.96
0.05,500,I,1.1,0.0001	2.77	132.74	320620.16
0.05,2000,I,2.0,0.02	3.85	128.39	325633.16
0.05,2000,I,2.0,0.0001	3.93	135.36	324273.96
0.05,2000,I,1.5,0.02	3.87	144.42	326558.92
0.05,2000,I,1.5,0.0001	3.88	137.91	325713.84
0.05,2000,I,1.1,0.02	3.99	131.54	325690.40
0.05,2000,I,1.1,0.0001	3.94	131.49	324629.08
0.05,1000,I,2.0,0.02	2.94	147.28	323110.24
0.05,1000,I,2.0,0.0001	3.00	149.84	323443.36
0.05,1000,I,1.5,0.02	3.00	146.13	323144.92
0.05,1000,I,1.5,0.0001	3.02	143.14	323136.72
0.05,1000,I,1.1,0.02	3.00	143.31	323410.08
0.05,1000,I,1.1,0.0001	3.02	146.23	323356.00
0.01,500,I,2.0,0.02	2.63	154.11	320784.96
0.01,500,I,2.0,0.0001	2.65	150.07	320432.16
0.01,500,I,1.5,0.02	2.64	126.83	320529.56
0.01,500,I,1.5,0.0001	2.75	129.40	319814.80
0.01,500,I,1.1,0.02	2.76	129.15	320633.56
0.01,500,I,1.1,0.0001	2.72	182.19	321332.20
0.01,2000,I,2.0,0.02	3.99	130.97	325901.32
0.01,2000,I,2.0,0.0001	4.03	129.72	325971.00
0.01,2000,I,1.5,0.02	4.06	126.38	325816.40
0.01,2000,I,1.5,0.0001	4.02	110.41	324423.52
0.01,2000,I,1.1,0.02	4.00	121.24	325429.32
0.01,2000,I,1.1,0.0001	4.04	128.06	326333.80

0.01,1000,I,2.0,0.02	3.06	127.62	323421.92
0.01,1000,I,2.0,0.0001	3.07	127.73	323639.96
0.01,1000,I,1.5,0.02	3.07	126.69	323483.56
0.01,1000,I,1.5,0.0001	3.05	156.55	325006.56
0.01,1000,I,1.1,0.02	3.03	163.41	324945.28
0.01,1000,I,1.1,0.0001	3.01	156.46	320749.28

*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.