



EPISNPmpi

**A supercomputer parallel computing program for
epistasis testing in genome-wide association studies**

USER MANUAL VERSION 2.0

LI MA¹, H. BIRALI RUNESHA² AND YANG DA¹

Department of Animal Science¹ and Supercomputer Institute²
University of Minnesota

March 27, 2008

AUTHOR CONTRIBUTIONS

Li Ma is the author of the PC Windows version of the EPISNP program was used for developing EPISNPmpi, is the main author of EPISNPmpi and is the author of the EPISNP and CPUHD programs.

H. Birali Runesha directed the development of the parallel computing coding, and did a portion of the coding of the EPISNPmpi program.

Yang Da is the project leader and designed most functionalities of the EPISNPmpi program.

RELEASE HISTORY

- EPISNPmpi version 1.0: First release by L. Ma, H. Birali Runesha and Y. Da, January 10, 2008. This version was tested to run on up to 128 processors/cores.
- EPISNPmpi version 2.0: Second release by L. Ma, H. Birali Runesha and Y. Da, March 27, 2008. This version used a new data distribution implementation among processors/cores and added estimates of individual epistasis effects to the output file of test results for identifying the allele and/or genotype combinations with the most desirable and undesirable effects. This version was tested for 500,000 SNPs on 2000 individuals using 528 cores. The computing time for this test was about 20 hours on SGI Altix XE 1300 Linux cluster system with 2.66 GHz Intel Clovertown processor (Calhoun).

Table of Contents

1. Functionality and applicability of EPISNPmpi	4
2. Input files.....	4
2.1 Parameter file (parameter.dat)	4
2.2 SNP genotype files	6
2.3 Phenotype (quantitative trait) file	7
3. Running EPISNPmpi parallel program	8
3.1 How to download EPISNPmpi	8
3.2 How to run	8
4. Output files.....	10
4.1 Screen output	10
4.2 Output files of test results	10
5. References.....	12

1. **Functionality and Applicability of EPISNPmpi**

EPISNPmpi is a parallel computing program designed for large scale epistasis testing of SNP markers on quantitative traits in genome-wide association (GWA) analysis. This program produces statistical test results for:

- Five epistasis effects for each pairs of SNPs
- Three single locus effects for each SNP

The five epistasis effects for each pairs of SNPs are the interaction between the two SNPs, additive \times additive, additive \times dominance, dominance \times additive, and dominance \times dominance epistasis effects, and the three single locus effects for each SNP are the single-locus marker effect, additive effect, and dominance effect. The method of statistical test for pairwise epistasis effects is based on an extended Kempthorne model (Mao et al., 2006).

The EPISNPmpi computer package is applicable to all bi-allelic loci of diploid species. The X chromosome loci can be analyzed but only females can be used. Similarly, the Z chromosome loci in birds can be analyzed but only male individuals can be used. The program assumes male to be the heterogametic sex, which is the case for mammals. For birds, the user can flip the gender definitions in the parameter file described in section **2.1 Parameter file (parameter.dat)**, i.e., in the parameter.dat file, male birds are defined as females and female birds are defined males. SNPs in the pseudoautosomal region of the Y chromosome in mammals (or W in birds) are analyzed as autosomal markers.

Through the control of the parameter file, EPISNPmpi has the flexibility to consider a number of factors in the statistical testing, including:

- an arbitrary number of quantitative traits;
- an arbitrary classification non-genetic factors such as smoking status and gender of the individual;
- an arbitrary number of continuous covariables such as age and body weight;
- a user specified number of significant effects to be stored in the output file for epistasis tests or single-locus tests;

2. **Input Files**

2.1 **Parameter file (parameter.dat)**

A parameter file with the name parameter.dat is required to run EPISNPmpi. The parameter file provides various user-specified controls and must have the name 'parameter.dat'. The following is an example of the parameter.dat file:

```

3 # number of traits
4 # starting position of traits in the trait file
3 1 2 # number of column for gender information and
codes for male and female
-1 # code for missing traits, non-genetic factors and
covariables
295 # number of individuals
2 # number of non-genetic factors
2 3 # positions of the non-genetic factors
0 # number of covariables
4 5 # positions of covariables in the trait file
3 # number of chromosomes
3 # sex chromosome number
1 # 1: 0=A1A1, 1=A1A2, 2=A2A2 and others=missing used for
coding; 2: A1/A1, A1/A2 and A2/A2 used for coding, 0/0
used for missing
4 # starting position of SNPs in the SNP data file
6 chr14.dat # number of SNPs and file name of the first
Chromosome data
10 chr19.dat # number of SNPs and file name of the second
Chromosome data
3 chr22.dat # number of SNPs and file name of the third
Chromosome data
sample_data.txt # file name of the phenotype data
10 # number of most significant results for single SNP
tests to be printed in the output file
15 # number of most significant results for pairwise
tests to be printed in the output file

```

Table 1: Example of parameter.dat

WARNINGS:

a) Adding or deleting any line except adding chromosome files creates errors and is not allowed. The user may change the parameter values but may not add or delete any line in the parameter file provided by the program. The number of lines for chromosome files must not exceed the number of chromosomes.

b) The recommended largest number of ‘most significant results’ to be saved in the output file of significant results is 1000 for single locus analysis and 10,000 for pairwise analysis. Larger numbers will result in increased computing time.

2.2 SNP genotype files (chrxxx files)

The SNP genotypes are in the chrxxx.dat files. Each chrxxx.dat file is for one chromosome. Each chromosome file must start with 'chr', and 'xxx' is a user specified chromosome number and will be used in the output files as the chromosome number. Alternatively, the SNP genotypes can be placed in one file with a maximum of 25 characters. However, the output files will not have the true chromosome numbers but rather, the file name except the first three characters is used as the chromosome name.

The coding for SNP genotypes has two options. The first option, with '1' in the parameter file for the SNP genotype definitions, uses the following coding, 0 = AA, 1 = Aa, 2 = aa, i.e., '0' and '2' denote the two homozygous genotypes and '1' denotes the heterozygous genotype. A number of '3' or greater denotes a missing SNP genotype.

The general format of the SNP genotype file for the first option is as follows:

```
{name of each column}
{family ID} {individual ID} {sex: 2 if female, 1 if male,
others if unknown} {locus1}{locus2}{locus3}{locus4}...
```

The second option for coding SNP genotype is that of Problem #1 of the Genetic Analysis Workshop 15 held in Tampa, Florida, November 12-15, 2006, where the three SNP genotypes of each locus are coded as A1/A1, A1/A2, and A2/A2, where A1 and A2 are nonzero integer numbers. A missing genotype is coded as 0/0.

fam_ID	ind_ID	Sex	M1	M2	M3	M4	M5	M6
1	31	1	0	2	0	1	0	0
1	32	1	1	1	0	2	0	0
1	33	1	1	2	0	2	1	0
1	34	1	2	2	0	1	0	0
1	35	1	3	2	0	1	0	0
1	36	2	0	2	0	2	0	0
1	37	2	1	0	0	2	0	0
1	38	2	0	2	0	2	0	0
1	39	2	1	1	2	1	1	0
2	40	1	1	2	0	2	0	0
2	41	1	0	2	0	2	0	0
2	42	1	0	2	2	2	0	0
2	43	1	1	2	0	1	0	0
2	44	1	0	2	0	2	2	2
2	45	1	3	1	0	1	0	0
2	46	1	0	2	2	2	0	0
2	47	1	0	2	0	2	2	2
2	48	2	0	1	0	2	0	0

Table 2: Example of a SNP genotype input file

2.3 Phenotype (quantitative trait) file

The phenotype file contains quantitative traits, fixed non-genetic effects such as gender, herd, year, season, blocks, treatment and living conditions, and covariables such as body weight and age. The name of the phenotype file is determined by the user in the parameter file.

The general format of the SNP genotype file is as follows:

```
{name of each column}
{family ID} {individual ID} {fixed effects, one fixed effect
per colum} {covariables, one covariable per column}{trait
1}{trait 2}{trait 3} . . .
```

Table 3 gives an example of a phenotype input file

The family IDs and individual IDs must match those in each chrxxx.dat file. Note that a trait can be used as a covariable. For example, birth weight can be a trait and can also be a covariable for analyzing the weight at six month old. Table 3 is an example of the phenotype file where 'sex' is a non-genetic fixed effect:

ind_ID	fam_ID	Sex	trait1	trait2	trait3
29	1	1	1.29	4.15	4.61
30	1	1	1.22	4.28	4.7
31	1	1	1.21	3.99	4.42
32	1	1	1.12	4.29	4.78
33	1	1	1.13	4.24	4.73
34	1	1	1.06	3.82	4.21
35	1	1	0.95	3.71	4.15
36	1	2	1.27	4.35	4.83
37	1	2	1.06	3.75	4.07
38	1	2	1.34	4.98	5.36
39	1	2	1.17	4.18	4.57
40	2	1	1.31	4.12	6.62
41	2	1	1.17	3.37	5.4
42	2	1	1.23	4.27	6.4
43	2	1	1.17	3.78	5.5
44	2	1	1.05	3.4	5.51
45	2	1	0.93	3.36	5.51
46	2	1	0.6	1.6	2.87
47	2	1	0.59	2.05	3.25
48	2	2	1.35	3.87	5.84
49	2	2	1.25	3.81	5.52

Table 3: Example of a phenotype input file

3. Running EPISNPmpi parallel program

3.1. How to download EPISNPmpi:

EPISNPmpi program has been implemented using MPI and has been tested on multiple parallel systems and architectures. The following are the currently supported processors type, MPI libraries, compilers and corresponding binaries:

<i>MPI library</i>	<i>Compiler</i>	<i>Processor</i>	<i>Binary</i>
Voltaire MPI	Intel	Intel	EPISNPmpi_2.0_Voltaire_intel_intel.tar.gz
Voltaire MPI	Intel	AMD	EPISNPmpi_2.0_Voltaire_intel_AMD.tar.gz
Voltaire MPI	Intel	Intel (EM64T)	EPISNPmpi_2.0_Voltaire_suse_EM64T.tar.gz
PathMPI	Pathscale	AMD	EPISNPmpi_2.0_Pathscale_suse_AMD.tar.gz
IntelMPI	Intel	AMD	EPISNPmpi_2.0_intelMPI.suse_AMD.tar.gz
OpenMPI	Intel	Intel (EM64T)	EPISNPmpi_2.0_OpenMPI_suse_EM64T.tar.gz
IBM MPI	Intel	Power4	EPISNPmpi_2.0_IBM_AIX_pwr.tar.gz
MPT	Intel	Itanium	EPISNPmpi_2.0_SGI-Altix_SUSE_itanium.tar.gz

The computer systems and processors for the above executables are described below:

<i>System</i>	<i>Processors</i>	<i>Operating System</i>	<i>MPI Library</i>
IBM Regatta	Power4	AIX	IBM MPI
SGI Altix BX2	Itanium	Linux	MPT
SGI Altix XE	Intel Clovertown	Linux	Voltaire, Intel, OpenMPI
IBM BladeCenter	AMD Opteron	Linux	Voltaire, Intel

Note: To request a build of a parallel executable on a different parallel system architecture, please send an email to Dr. Yang Da at yda@umn.edu and Dr. H. Birali Runesha at runesha@msi.umn.edu.

All the above executables of EPISNPmpi are distributed in one file: EPISNPmpi_2.0.tar, which can be downloaded from

<http://animalgene.umn.edu/EPISNPmpi/index.html>

After downloading a binary, you can untar the file using the following commands:

```
Gunzip EPISNPmpi.tar.gz
tar xvf EPISNPmpi.tar
```

3.2 How to run EPISNPmpi

To run EPISNPmpi, parallel program, you need the parameter.dat, phenotype and SNP genotype (chromosome) files to be in the same directory as the executable. Below are selected examples on how to run EPISNPmpi on a SGI Altix XE 1300 and IBM BladeCenter.

A) SGI Altix XE 1300

The SGI Altix XE 1300 has 256 compute nodes and a total of 2048 cores. Each node has two quad-core 2.66 GHz Intel Xeon processors sharing 16 GB of memory.

1. To run EPISNP with Voltaire MPI built with intel compiler
 - gunzip EPISNPmpi_intel.calhoun.tar.gz
 - tar xvf EPISNPmpi_intel.calhoun.tar

```
module load intel  
module load vmpi  
mpirun -np 2 login2 login2 ./EPISNPmpi
```

where np = number of cores, login2 = the node you are logged on to.

2. To run EPISNP with Intel MPI libraries built with Intel compilers
 - gunzip EPISNPmpi_intel.calhoun.tar.gz
 - tar xvf EPISNPmpi_intel.calhoun.tar

```
module load intel  
module load impi  
mpirun -np 2 login2 login2 ./EPISNPmpi
```

3. To run EPISNP with OpenMPI libraries built with Intel compilers
 - gunzip EPISNPmpi_openmpi.calhoun.tar.gz
 - tar xvf EPISNPmpi_openmpi.calhoun.tar

```
module load intel  
module load omp  
mpirun -np 2 login2 login2 ./EPISNPmpi
```

B) IBM BladeCenter

The IBM BladeCenter H* with 309 LS 21 nodes. Each node has two dual-core 2.6 GHz AMD Opteron processors sharing 8 GB of memory. This gives a total of 1236 cores on the system.

1. To run EPISNP with MPI libraries built with Intel MPI
 - gunzip EPISNPmpi_intel.tar.gz
 - tar xvf EPISNPmpi_intel.tar

```
module load intelmpi  
mpirun -np 2 blade286 blade286 ./EPISNPmpi
```

The above command assumes the user is logged onto a node called blade286.

2. To run EPISNP with MPI libraries built with Pathscale compilers

- gunzip EPISNPmpi_pathscale.tar.gz
- untar the file: tar xvf EPISNPmpi_pathscale.tar

```
module load pathmpi
mpirun -np 2 blade286 blade286 ./EPISNPmpi_pathscale
```

C) EPISNPmpi commodity cluster-based processing

EPISNPmpi has been developed and tested on many modern high-performance computers and supercomputer systems. When evaluating high performance computing systems it is important to keep in mind the price-to-performance ratio of the system. It is with that in mind that EPISNPmpi has also been implemented to also run on commodity cluster or on an inexpensive network of workstations using MPICH message passing libraries. MPICH is a freely available, portable implementation of MPI, a standard for message-passing for distributed-memory applications. MPICH is available for download at www.mcs.anl.gov/mpi/mpich1/download.html.

4. Output files

4.1 Screen output:

Screen output is designed to monitor the status of the execution of EPISNPmpi. Version 1.0 of EPISNPmpi tested for 128 processors is shown in Table 4.

```
Pairwise analysis ...
Single locus analysis ...
Pairwise analysis ...
Number of tests =          171
Number of tests =          171
          1 Number of tests per processor =          86
          0 Number of tests per processor =          85
The CPU time for running is 3.2002000000000002E-002
seconds
Cleaning up all processes ...
done.
```

Table 4: Screen output of EPISNPmpi Version 1.0

4.2 Output files of test results

The results of statistical tests for single-locus effects are saved in the output file named “single_locus_sig.out”. Single-locus effects include the marker genotypic effect, and additive and dominant effects. Table 5 shows an example of the single-locus test results in “single_locus_sig.out”.

7 MOST SIGNIFICANT RESULTS OF SINGLE LOCUS ANALYSIS				

m = overall marker effect				
a = additive effect, d = dominance effect				

Chr	Locus	Trait	Test	P_value

01	rs2017	trait1	d	0.113E+00
	D	12 21	11	
	Estimate	4.132 -3.487	-5.273	
01	rs2017	trait1	m	0.243E+00
01	rs2840	trait1	a	0.291E+00
	A	2 1		
	Estimate	0.604 -4.655		
01	rs2477	trait1	a	0.121E+00
	A	2 1		
	Estimate	4.725 -1.968		
01	rs7703	trait1	m	0.202E+00
01	rs7703	trait1	d	0.371E+00
	D	12 11	21	
	Estimate	2.585 -0.559	-6.230	
01	rs4999	trait1	a	0.243E+00
	A	1 2		
	Estimate	2.439 -1.912		

Table 5: Example of single_locus_sig.out

The results of statistical tests for pairwise epistasis effects are saved in the output file named “pairwise_sig.out”. Pairwise epistasis effects include the two-locus interaction effects, additive \times additive, additive \times dominance, dominance \times additive, and dominance \times dominance effects. Table 6 shows an example of the pairwise test results in pairwise_sig.out.

5 MOST SIGNIFICANT RESULTS OF PAIRWISE EPISTASIS ANALYSIS										
Chr1	SNP1	Chr2	SNP2				Trait	Effect	p-value	
5	M7	19	M12				Trait1	I	0.345E-06	
5	M7	19	M12				Trait1	AA ^a	0.793E-07	
	i_k ^a	2_1	1_2	1_1	2_2					
	Estimate	0.388	0.314	-0.305	-0.365					
8	M8	22	M93				Trait1	AD ^b	0.158E-07	
	i_kl ^b	1_11	2_12	1_22	2_22	1_12	2_11			
	Estimate	2.335	1.049	0.238	-0.303	-0.868	-3.937			
9	M9	14	M612				Trait2	DA ^c	0.383E-08	
	ij_k ^c	11_1	13_1	12_2	22_2	11_2	12_1			
	Estimate	4.892	3.344	0.874	-0.259	-1.871	-2.074			
19	M112	23	M182				Trait3	DD ^d	0.793E-07	
	ij_kl ^d	22_11	11_11	11_22	12_12	22_22	12_22	22_12	11_12	12_11
	Estimate	2.969	1.839	1.719	1.543	0.616	-0.374	-0.691	-1.446	-2.118

Table 6: Example of pairwise_sig.out

The second output file of single-locus tests named “single_locus_fig.out” is used as the input file of the EPISNPLOT program to draw figures of chromosome view of significant results, and the second output file of pairwise epistasis tests named “pairwise_net.out” is used as the input file of the EPINET program to draw figures of epistasis network. The EPISNPLOT and EPINET are compiled for *Windows* and are distributed as the epiSNP package, which is freely available at <http://animalgene.umn.edu/episnp/index.html>.

5. References

Yongcai Mao, Nicole R. London, Li Ma, Daniel Dvorkin and Yang Da (2006) Detection of SNP epistasis effects of quantitative traits using an extended Kempthorne model. *Physiological Genomics* 28: 46-52.