MTHSA-DHEI: Multitasking Harmony Search Algorithm for Detecting High-*Order* Epistatic Interactions

1. The harmony search algorithm 1
   1.1 The steps of the standard harmony search algorithm 1
   1.2 An Example for introducing the proposed algorithm 3
2. Disease models 4
   2.1 Eight EINME models 4
   2.2 Twelve EIME models 4
3. Experimental results 5
4. References 9

1. The harmony search algorithm

The harmony search algorithm (HSA) is a new swarm intelligence optimization algorithm inspired by the process of improvising a musical harmony in a music orchestra (Geem, et al., 2001,2008) [1][2]. To solve an optimization model with k decision variables, a harmony represents a vector that consists of values of the k decision variables. Harmony memory (HM), which is composed of harmonies, is similar to the population in genetic algorithm (GA). The number of harmonies in HM is called HMS. An HM is a matrix of order HMS×k or an augmented matrix of order HMS × (k + 1) [18], as follows:

$$\mathrm{HM}=\left[\begin{array}{c|c} X^1 & f(X^1) \\ X^2 & f(X^2) \\ \vdots & \vdots \\ X^{\mathrm{HMS}} & f(X^{\mathrm{HMS}}) \end{array}\right]=\left[\begin{array}{cccc|c} x_1^1 & x_2^1 & \cdots & x_k^1 & f(X^1) \\ x_1^2 & x_2^2 & \cdots & x_k^2 & f(X^2) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ x_1^{\mathrm{HMS}} & x_2^{\mathrm{HMS}} & \cdots & x_k^{\mathrm{HMS}} & f(X^{\mathrm{HMS}}) \end{array}\right]$$

where X^i (i=1, 2,…, HMS) denotes the i-th harmony in HM and $f(X^i)$ is the value of the objective function of X^i. HSA aims to optimize the harmony in HM by adjusting the pitch.

1.1 The steps of the standard harmony search algorithm

Step 1. Initializing the parameters.

The parameters include harmony memory consideration rate (HMCR), pitch adjustment rate (PAR), harmony memory size (HMS), pitch fret width (*fw*) and terminal condition (i.e., the maximum number of objective function evaluated (MaxFEs)). HMCR is primarily used to balance the exploration power and exploitation power of the algorithm. PAR and *fw* are used for local pitch adjustment.

Step 2. Initializing the harmony memory (HM) and calculating the fitness value of each harmony.

For i=1: HMS

$X^i = \varnothing$;

For j=1: ***k***

Do

$a = \lceil rand(0,1) \times N \rceil$;

While ($a \in X^i$)

$X^i \leftarrow a$;

End

// calculating the fitness value of harmony X^i

$Score(i) = f(X^i)$

End

$$\mathbf{HM} = \begin{bmatrix} \boldsymbol{X}^1 \\ \boldsymbol{X}^2 \\ \vdots \\ \boldsymbol{X}^{\mathrm{HMS}} \end{bmatrix} = \left[\begin{array}{cccc|c} x_1^1 & x_2^1 & \cdots & x_k^1 & Score(1) \\ x_1^2 & x_2^2 & \cdots & x_k^2 & Score(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{\mathrm{HMS}} & x_2^{\mathrm{HMS}} & \cdots & x_k^{\mathrm{HMS}} & Score(\mathrm{HMS}) \end{array}\right]$$

Step 3. Improvising a new harmony $X^{new} = (x_1^{new}, x_2^{new}, \cdots, x_k^{new})$.

For i=1→N

If rand(0,1) < HMCR

x_i^{new} ← randomly choose from $\{x_i^1, x_i^2, \cdots, x_i^{HMS}\}$ with probability HMCR.

If rand (0,1)<PAR

$x_i^{new} = x_i^{new} + rand(0,1) \times fw$.

EndIf

Else

x_i^{new} ← select a value from the feasible search space randomly with probability 1-HMCR

EndIf

EndFor

Step 4. Updating the HM.

If X^{new} is better than the worst harmony $X^{idworst}$ in the HM

$X^{idworst}$ ← X^{new}

end

Step 5. Checking the stopping condition.

If the stopping criterion is met, the computation is terminated. Otherwise, Steps (3) and (4) are repeated.

1.2 An Example for introducing the proposed algorithm

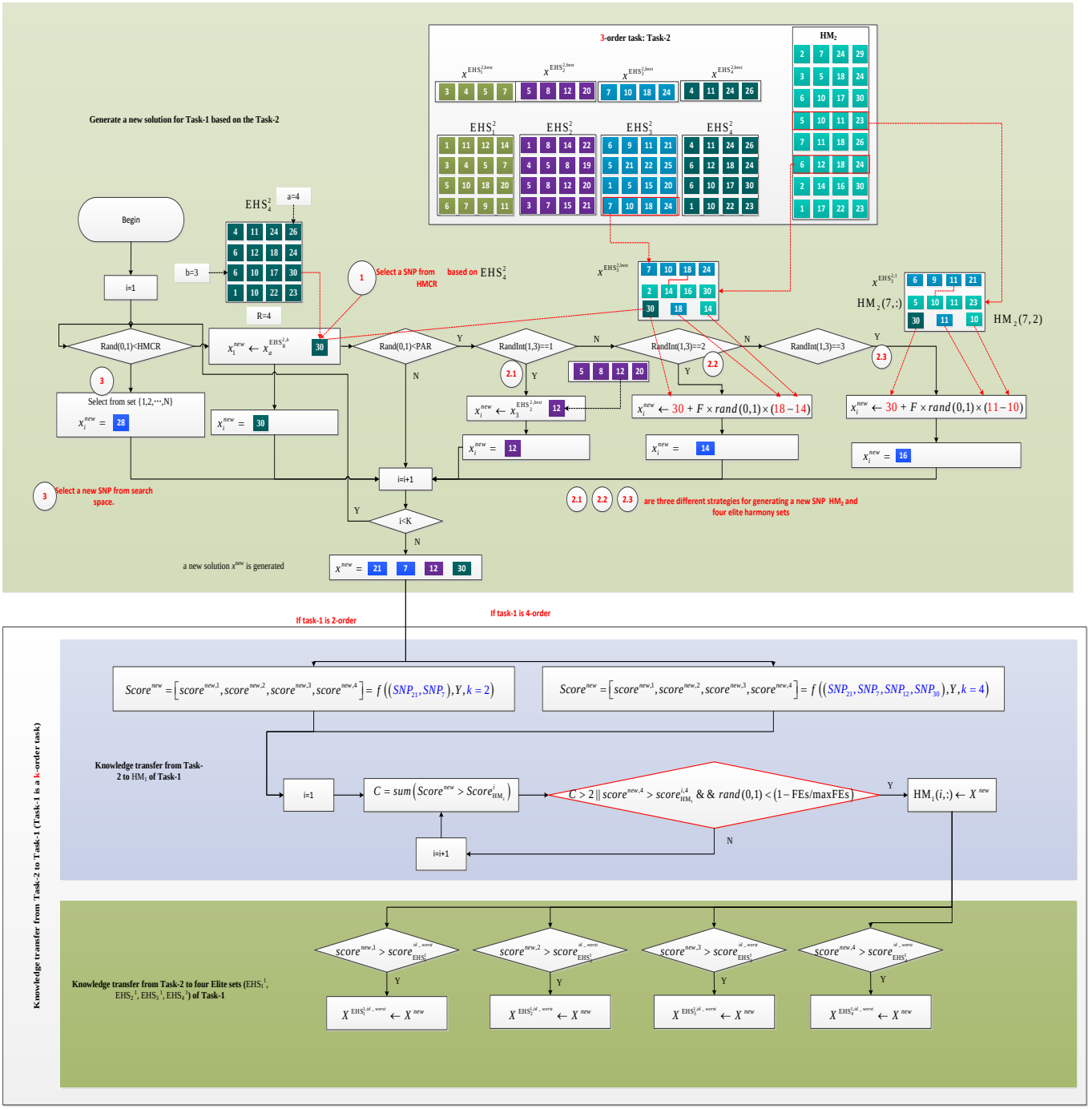

Figure S1. An example that a new solution (SNP combination) generated based on the harmony HM2 and four elite harmony sets of task-2 is transferred to task-1, where task-2 aims for detecting 3-order epistatic interactions and task-1 aims to detect 2-order or 4-order epistatic interactions.

2. Disease models

2.1 Eight EINME models

Table S1. Eight EINME models.

The 3rd column denotes whether the model satisfies the Hardy-Weinberg equilibrium (HWE). In column 4–column 8, the values represent the prediction accuracy from k-order (k=1, 2,…,5) epistatic interaction.

Model	k-order	HWE	1-order(sd)	2-order(sd)	3-order(sd)	4-order(sd)	5-order(sd)	tar.gz link
EINME-1	3	No	.502(.001)	.511(.007)	.886(.023)			threewayBests
EINME-2	3	Yes	.504(.002)	.509(.003)	.680(.024)			HWthreewayBests
EINME-3	4	No	.502(.001)	.510(.003)	-	.897(.018)		fourwayBests
EINME-4	4	Yes	.507(.003)	.513(.003)	-	.673(.009)		HWfourwayBests
EINME-5	4	No	.501(.000)	.504(.001)	.518(.003)	.567(.010)		fourwayNoLowBests
EINME-6	5	No	.502(.001)	.510(.002)	-	-	.895(.009)	fivewayBests
EINME-7	5	Yes	.511(.003)	.518(.003)	-	-	.693(.008)	HWfivewayBests
EINME-8	5	No	.503(.001)	.508(.001)	.518(.002)	.543(.004)	.690(.008)	fivewayNoLowBests

The eight datasets are generated by Himmelstein et al, 2011[3], which disables the discovery of disease-causing models for certain existing heuristic methods due to the lack of clues of causative SNP markers.

2.2 Twelve EIME models

Table S2. The parameters and the values of penetrance of 12 EIME models.

Model type	EIME	order	Heritability(H^2)	MAF	Heterogeneity proportion
Additive model	EIME -1	5	0.1	0.1	1.0
	EIME -2	5	0.1	0.25	1.0
	EIME -3	5	0.1	0.5	1.0
	EIME -4	5	0.1	0.2	1.0
Threshold model	EIME -5	5	0.1	0.1	1.0
	EIME -6	5	0.25	0.1	1.0
	EIME -7	5	0.5	0.1	1.0
	EIME -8	5	0.1	0.2	1.0
Multiplicative model	EIME -9	4	0.005	0.1	1.0
	EIME -10	4	0.005	0.2	1.0
	EIME -11	4	0.005	0.4	1.0
	EIME -12	4	0.004	0.05	1.0

H^2 denotes the genetic heritability. **MAF** represents the minor allele frequencies.

The datasets are generated using GAMETES software [4].

3. Experimental results

Table S3. Experimental results of proposed algorithm for 8 EINME datasets with 100SNPs, 1k SNPs, 10k SNPs and 50k SNPs

Models	Number of SNPs	1st Power	2nd Power	3nd Power	FEs	runtime (s)	Precision	Recall	F1-score
EINME-1	100	100	100	100	1002	1.2	100%	100%	100%
EINME-2	100	100	100	100	995	1.2	100%	100%	100%
EINME-3	100	100	100	100	2601	4.0	100%	100%	100%
EINME-4	100	100	100	100	3315	5.3	100%	100%	100%
EINME-5	100	97	97	97	5100	10.6	100%	97%	98%
EINME-6	100	100	100	100	3800	6.9	100%	100%	100%
EINME-7	100	95	95	95	5838	12.9	100%	95%	97%
EINME-8	100	53	53	53	18362	32.7	100%	53%	69%
EINME-1	1000	100	100	100	8105	9.8	100%	100%	100%
EINME-2	1000	100	100	100	8693	10.2	100%	100%	100%
EINME-3	1000	100	100	100	13958	18.1	100%	100%	100%
EINME-4	1000	100	100	100	19292	26.8	100%	100%	100%
EINME-5	1000	93	93	93	83365	203.6	100%	93%	96%
EINME-6	1000	100	100	100	14229	20.2	100%	100%	100%
EINME-7	1000	95	95	95	32701	269.3	100%	95%	97%
EINME-8	1000	37	37	37	282899	526.6	100%	37%	54%
EINME-1	10000	100	100	100	12722	13.5	100%	100%	100%
EINME-2	10000	100	100	100	15051	18.3	100%	100%	100%
EINME-3	10000	100	100	100	23639	30.6	100%	100%	100%
EINME-4	10000	100	100	100	28969	42.1	100%	100%	100%
EINME-5	10000	87	87	87	141556	271.2	100%	87%	93%
EINME-6	10000	100	100	100	27876	36.4	100%	100%	100%
EINME-7	10000	71	71	71	142607	310.9	100%	71%	83%
EINME-8	10000	11	11	11	797659	3237.3	100%	11%	19.8%

Table S4. Experimental results of proposed algorithm for 12 EIME datasets with 100SNPs, 1k SNPs, 10k SNPs and 50k SNPs

Models	Number of SNPs	1st Power	2nd Power	3nd Power	FEs	runtime (s)	Precision	Recall	F1-score
EIME-1	100	100	100	**100**	2195	3.3	100.0%	100.0%	100.0%
EIME-2	100	100	100	**100**	2331	3.7	100.0%	100.0%	100.0%
EIME-3	100	100	100	**100**	2567	4.2	100.0%	100.0%	100.0%
EIME-4	100	100	100	**100**	2261	3.4	100.0%	100.0%	100.0%
EIME-5	100	100	100	**0**	2700	4.3	/	0.0%	/
EIME-6	100	100	100	**100**	2391	3.7	100.0%	100.0%	100.0%
EIME-7	100	100	100	**100**	2173	3.2	100.0%	100.0%	100.0%
EIME-8	100	100	100	**1**	3831	6.7	100.0%	1.0%	2.0%
EIME-9	100	100	100	**0**	1379	1.7	/	0.0%	/
EIME-10	100	100	100	**0**	1286	1.5	/	0.0%	/
EIME-11	100	100	100	**100**	1259	1.5	100.0%	100.0%	100.0%
EIME-12	100	100	100	**100**	1242	1.4	100.0%	100.0%	100.0%
EIME-1	1000	100	100	**100**	13106	17.0	100.0%	100.0%	100.0%
EIME-2	1000	100	100	**100**	12931	17.7	100.0%	100.0%	100.0%
EIME-3	1000	100	100	**100**	14619	22.5	100.0%	100.0%	100.0%
EIME-4	1000	100	100	**100**	13142	18.3	100.0%	100.0%	100.0%
EIME-5	1000	100	100	**0**	16264	26.0	/	0.0%	/
EIME-6	1000	100	100	**100**	12906	17.9	100.0%	100.0%	100.0%
EIME-7	1000	100	100	**100**	12506	16.4	100.0%	100.0%	100.0%
EIME-8	1000	100	100	**1**	27796	56.3	100.0%	1.0%	2.0%
EIME-9	1000	100	100	**0**	2823	5.5	/	0.0%	/
EIME-10	1000	100	100	**0**	2694	5.3	/	0.0%	/
EIME-11	1000	100	100	**100**	2690	5.3	100.0%	100.0%	100.0%
EIME-12	1000	100	100	**100**	2559	4.8	100.0%	100.0%	100.0%
EIME-1	10000	100	100	100	18378	27.3	100.0%	100.0%	100.0%
EIME-2	10000	100	100	100	18080	26.9	100.0%	100.0%	100.0%
EIME-3	10000	100	100	100	20787	33.8	100.0%	100.0%	100.0%
EIME-4	10000	100	100	100	18436	27.3	100.0%	100.0%	100.0%
EIME-5	10000	100	100	0	24061	40.7	/	0.0%	/
EIME-6	10000	100	100	100	20166	31.1	100.0%	100.0%	100.0%
EIME-7	10000	100	100	100	18189	25.6	100.0%	100.0%	100.0%
EIME-8	10000	100	100	1	40976	89.8	100.0%	1.0%	2.0%
EIME-9	10000	99	99	0	14488	23.2	/	0.0%	/
EIME-10	10000	95	98	0	13711	21.4	/	0.0%	/
EIME-11	10000	100	100	100	12372	17.2	100.0%	100.0%	100.0%
EIME-12	10000	100	100	100	13254	18.9	100.0%	100.0%	100.0%

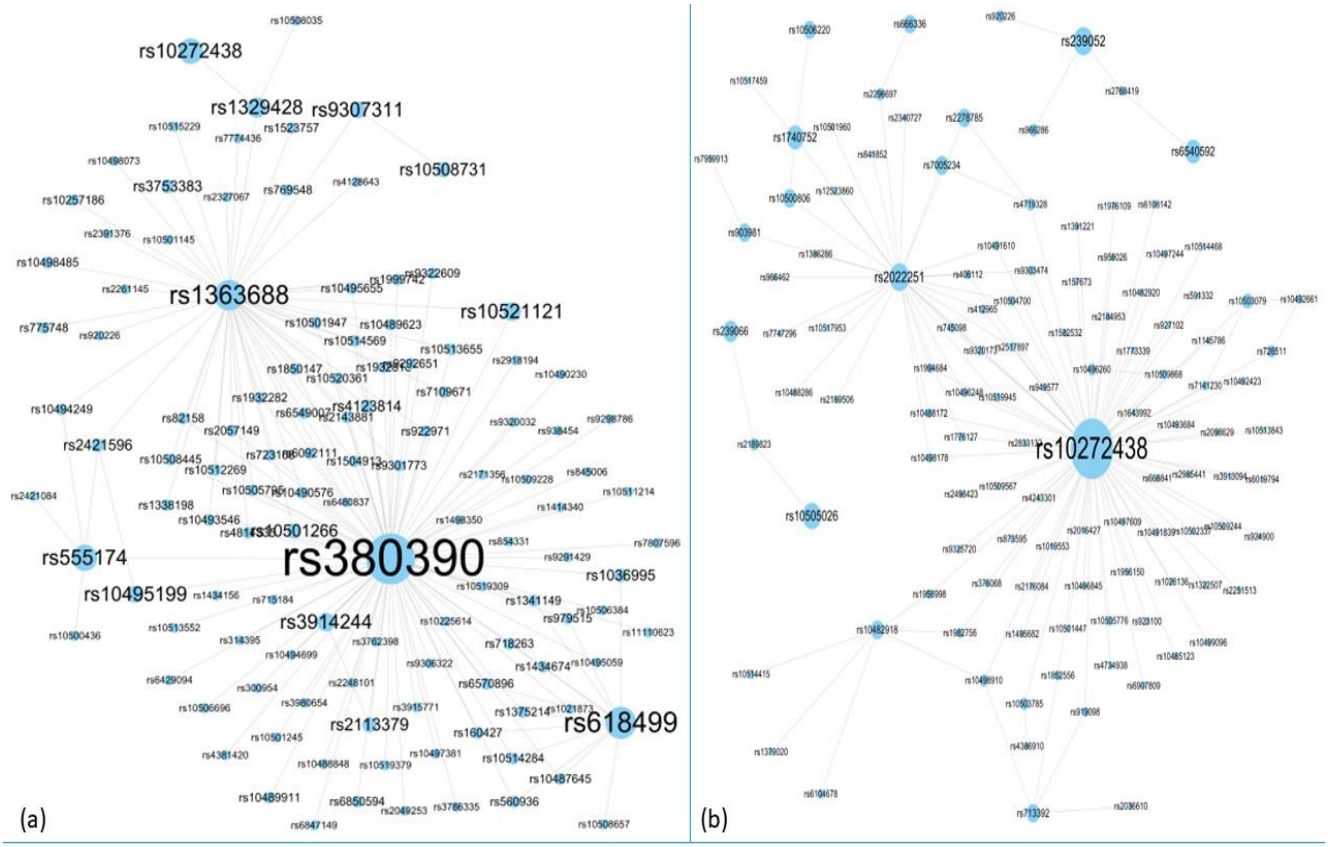

Figure S2. A 3rd-*order* SNP interaction network. A node denotes an SNP locus, and an edge indicates that the 2 SNPs connected by that edge are a subset of a 3rd-*order* SNP combination that is strongly associated with AMD (its significance level for the G-test is $< 1 \times 10^{-9}$, and the classification accuracy of the 3rd-*order* SNP combination is >80%). The larger the node is, the greater the number of nodes connected to it.

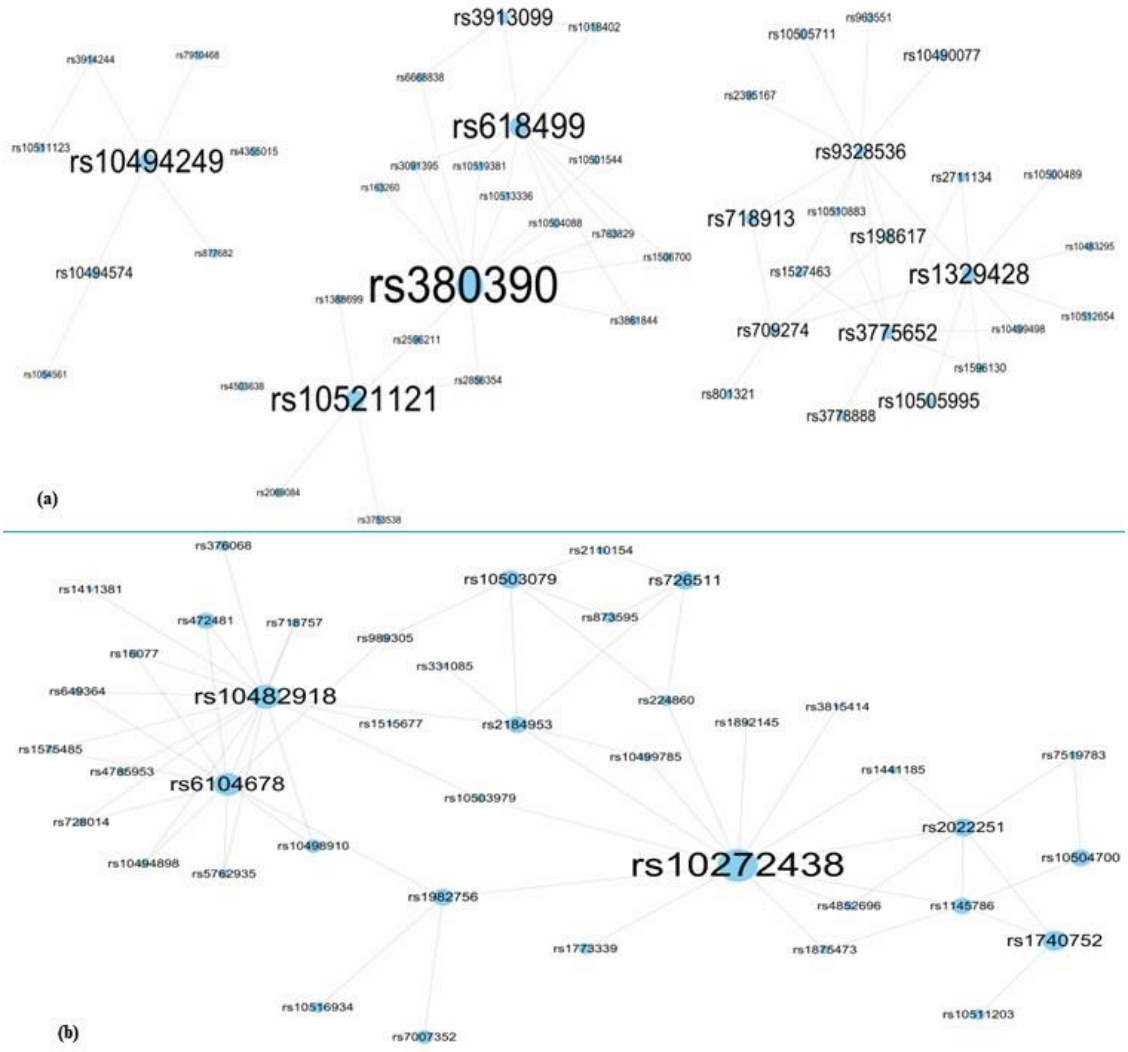

Figure S3. A 4th-*order* SNP interaction network, in which a node denotes an SNP locus, and an edge indicates that the 2 SNPs connected by that edge are a subset of a 4th-*order* SNP combination that is strongly associated with AMD (its significance level for the G-test is $<1 \times 10^{-11}$, and the classification accuracy of the 4th-*order* SNP combination is >80%). The larger the node is, the greater the number of nodes connected to it.

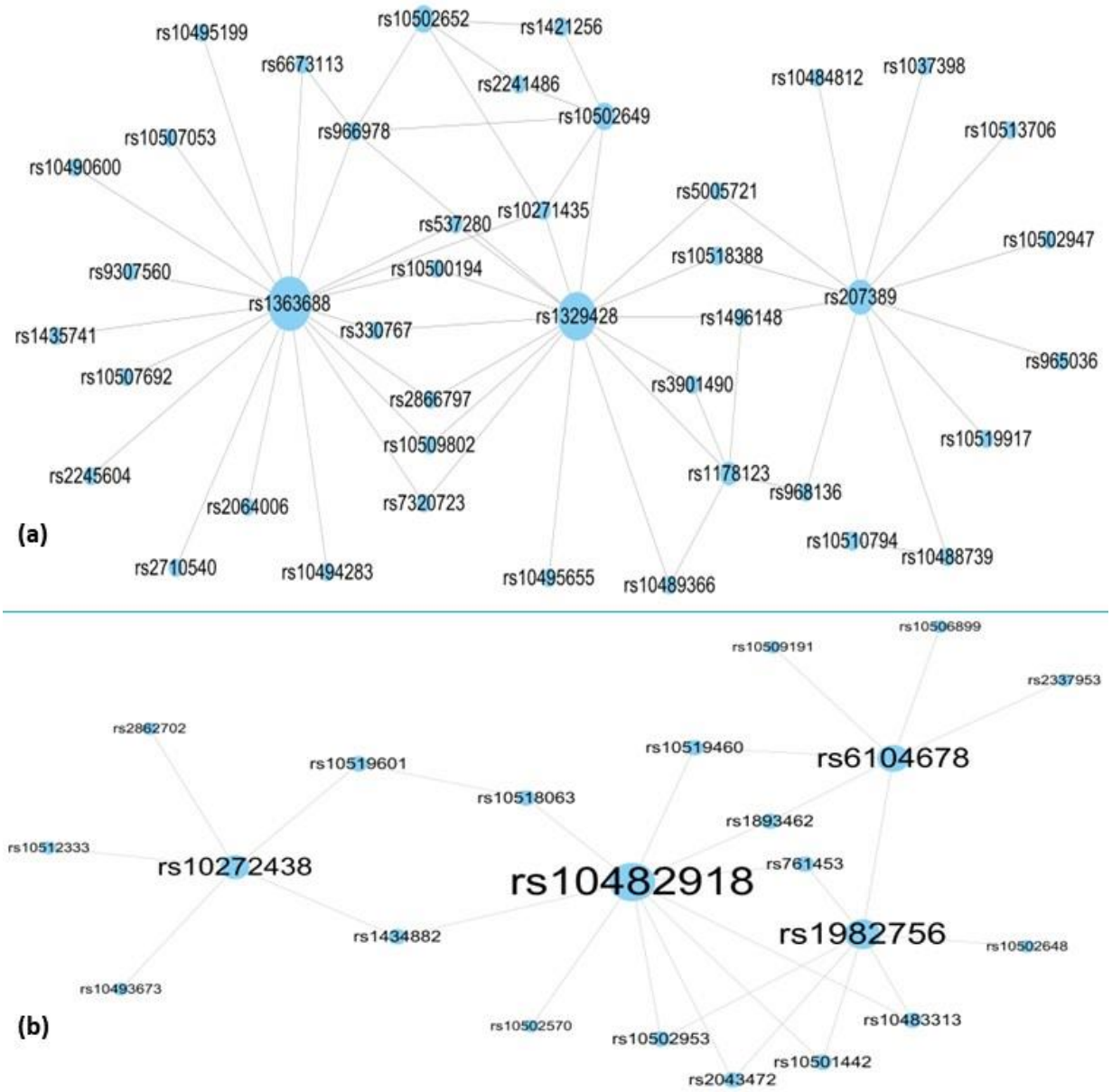

Figure S4. A 5th-*order* SNP interaction network, in which a node denotes an SNP locus, and an edge indicates that the 2 SNPs connected by that edge are a subset of a 5th-*order* SNP combination that is strongly associated with AMD (its significance level for G-test is $<1 \times 10^{-13}$, and the classification accuracy of the 5th-*order* SNP combination using MDR is >85%). The larger the node is, the greater the number of nodes connected to it.

4. References

[1]. Z.W. Geem, J. Kim, G. Loganathan, Music-inspired optimization algorithm harmony search, Simulation 76, 60-68(2001).

[2]. Z, W. Geem. Novel derivative of harmony search algorithm for discrete design variables. Applied Mathematics & Computation 199(1), 223-230(2008).

[3] Himmelstein et al. Evolving hard problems: Generating human genetics datasets with a complex etiology. BioData Mining 4, 21(2011). doi:10.1186/1756-0381-4-21. http://discovery.dartmouth.edu/model_free_data/.

[4] Urbanowicz, R. J., Kiralis, J., Sinnott-Armstrong, N. A., Heberling, T., Fisher, J. M. & Moore, J. H. GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. ***BioData mining*** 2012; 5:1–14.