Detecting gene-gene interactions in high-throughput genotype data through a Bayesian clustering procedure

Sui-Pi Chen and Guan-Hua Huang

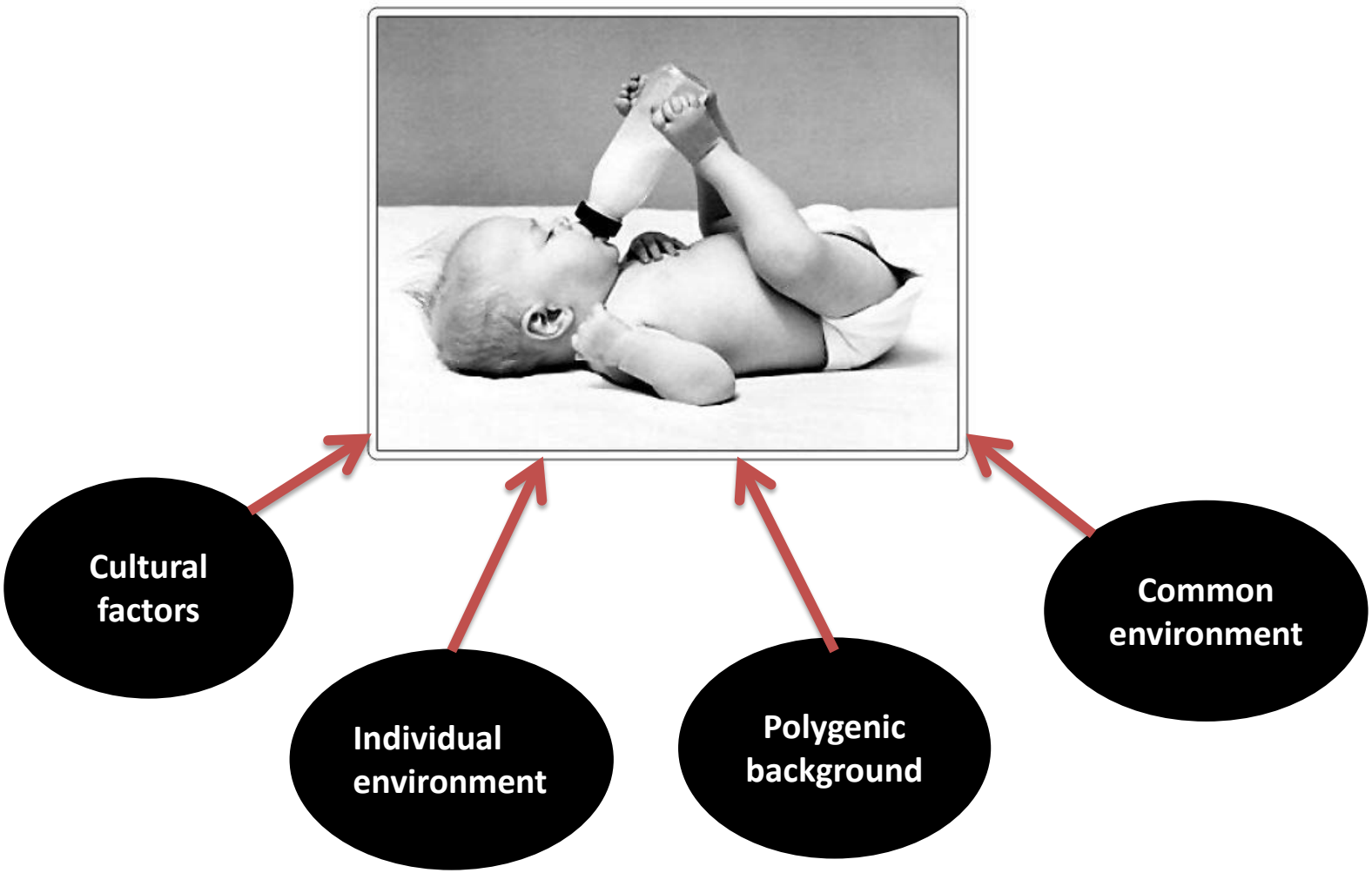
Institute of Statistics
National Chiao Tung University
Hsinchu, Taiwan

✉:ghuang@stat.nctu.edu.tw

2015.05.27

Statistical Applications in Genetics and Molecular Biology
13, 275-297, 2014

Motivation


Cultural factors
Individual environment
Polygenic background
Common environment

Single nucleotide polymorphism (SNP)

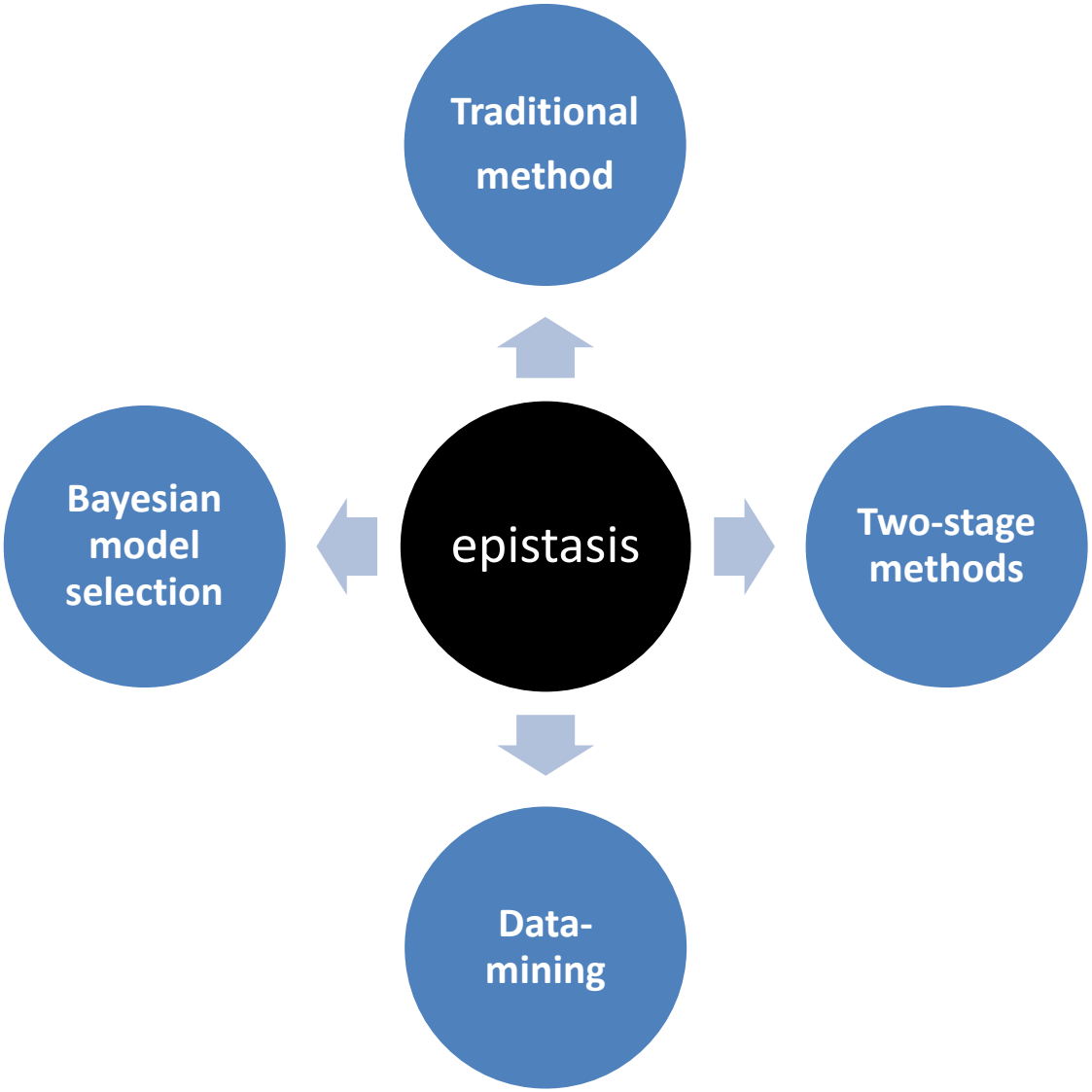
- A DNA sequence variation

- Two alleles: A and a
- Treating SNPs as categorical features that have three possible values: AA, Aa, aa.
- Relabel AA (2),Aa (1),aa (0).

What is the gene−gene interaction (epistasis)?

- The effects of a given gene on a biological trait are masked or enhanced by one or more genes.
- An increasing body of evidence has suggested that epistasis plays an important role in susceptibility to human complex disease, such as Type 1 diabetes, breast cancer, obesity, and schizophrenia.
- More evidences have confirmed that displaying interaction effects without displaying marginal effect.
- When analyzing thousands and thousands genes from high-throughput SNP arrays, this can further complicate the problem due to computational burden.

Methods for detecting gene-gene interaction



Methods for detecting gene-gene interaction

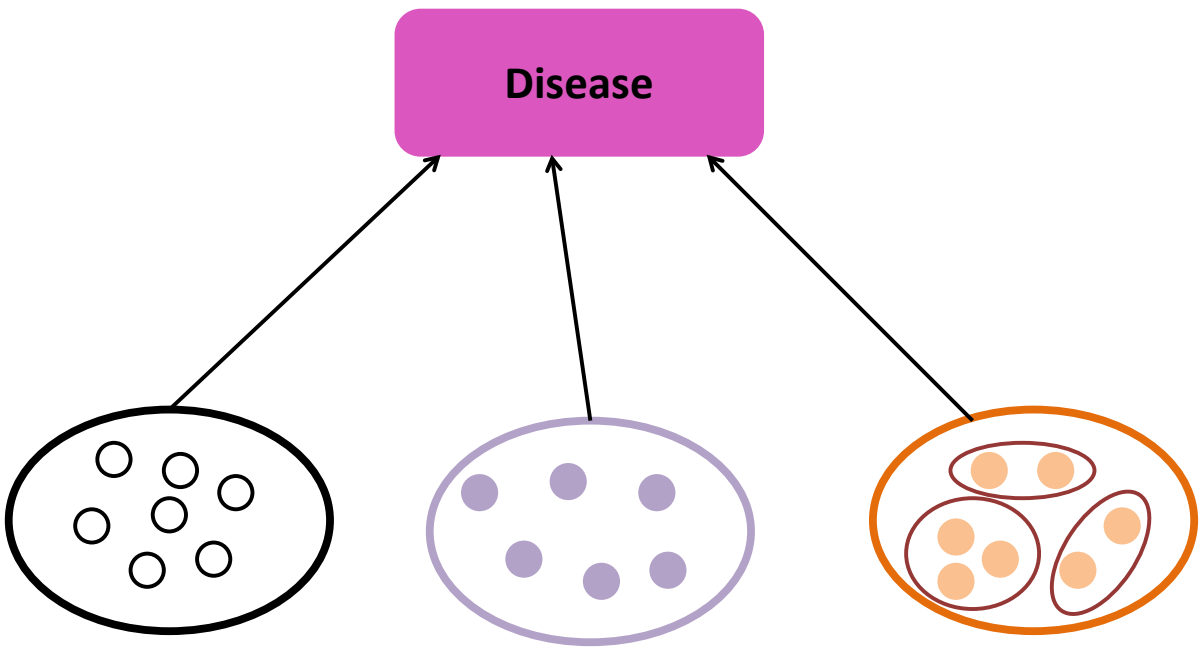
Traditional method	–Logistic regression, contingency table χ^2 test – It dose not include the interaction terms without main effect. – High-dimensional data that have high-order interactions, the contingency table have many empty cells.
Two-stage method	– A subset of loci that pass some single-locus significance threshold is chosen as the "filtered" subset. – An exhaustive search of all two-locus or higher-order interactions is carried out at the "filtered" subset.
Data-mining method	–Nonparametic –Not doing an exhaustive search –Multifactor Dimensionality Reduction (MDR)
Bayesian model selection	–Bayesian epistasis association mapping (BEAM) –Algorithm via Bayesian Clustering to Detect Epistasis (ABCDE)

BEAM algorithm

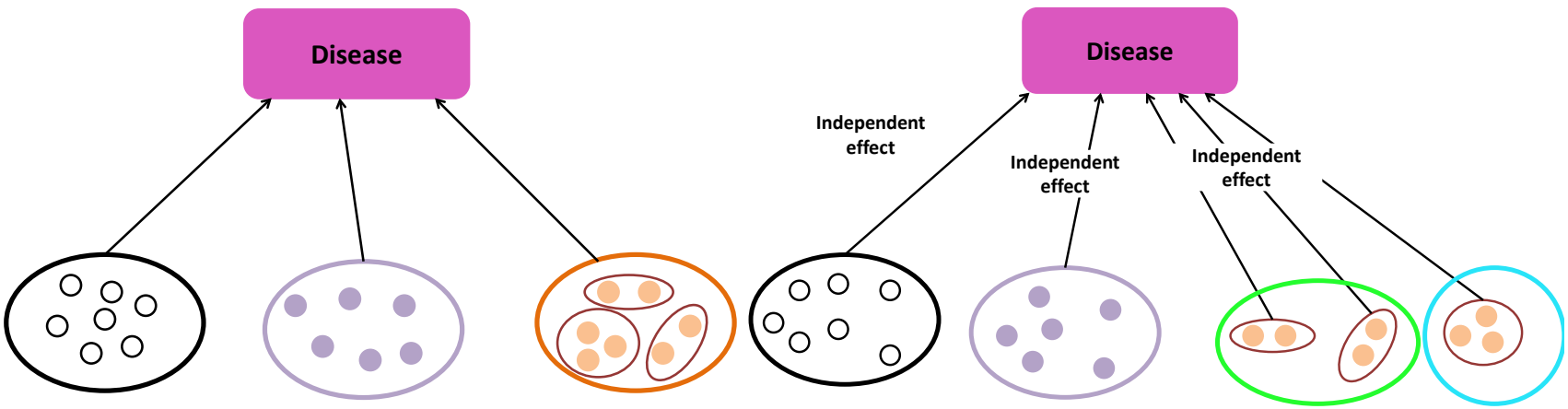
- BEAM (Zhang and Liu, 2007) algorithm
  - case-control study
  - Metropolis-Hasting algorithm
  - posterior probabilities
    - each SNP not associated with the disease
    - each SNP associated with the disease
    - each SNP involved with other SNPs in epistasis
- B statistic
  - each SNP or set of SNPs for significant association
  - asymptotically distributed as a shifted χ^2 with $3^k - 1$ degrees of freedom

BEAM algorithm

- $\mathbf{I} = (I_1, \cdots, I_L)$ indicator the membership of the SNPs with $I_j = 0, 1, 2$.
- BEAM found no significant interactions associated in the AMD data.



Algorithm via Bayesian Clustering to Detect Epistasis (ABCDE)



(a) BEAM (b) ABCDE

ABCDE algorithm

- ABCDE algorithm
  - bayesian clustering approach
  - case-control study
  - Gibbs weighted Chinese restaurant (GWCR) procedure
  - posterior probabilities
    - each SNPs is associated with the disease
    - clustered SNPs is associated with the disease.
- Permutation test for candidate disease subset selected by ABCDE
  - 10-fold cross validation
  - the heart of MDR approach: dimensional reduction.

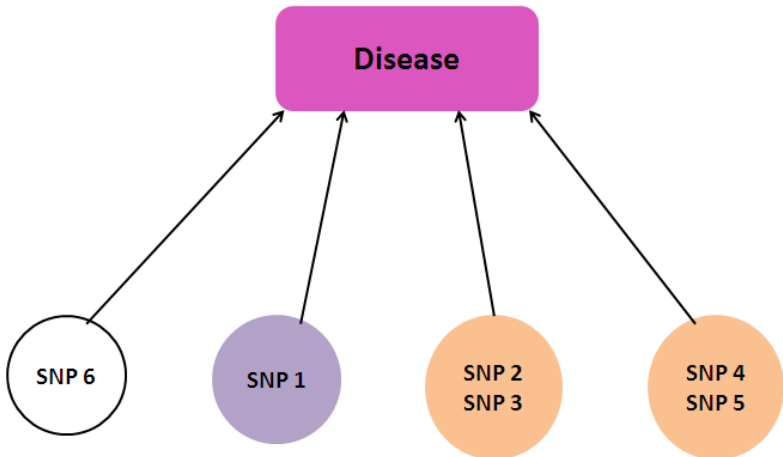
Product partition model

$$p(\mathbf{h}|\mathbf{G})$$

$$\propto p(\mathbf{h}) \times p(\mathbf{G}|\mathbf{h})$$

$$\propto p(\mathbf{h}) \prod_{j=1}^{n(\mathbf{h})} f_{a_j}(G_{C_j})$$

$$\propto p(\mathbf{h}) \times \prod_{A \in \mathbf{S}_0} f_0(\mathbf{G}_A) \times \prod_{A \in \mathbf{S}_1} f_1(\mathbf{G}_A) \times \cdots \times \prod_{A \in \mathbf{S}_{g(\mathbf{h})}} f_{g(\mathbf{h})}(\mathbf{G}_A),$$

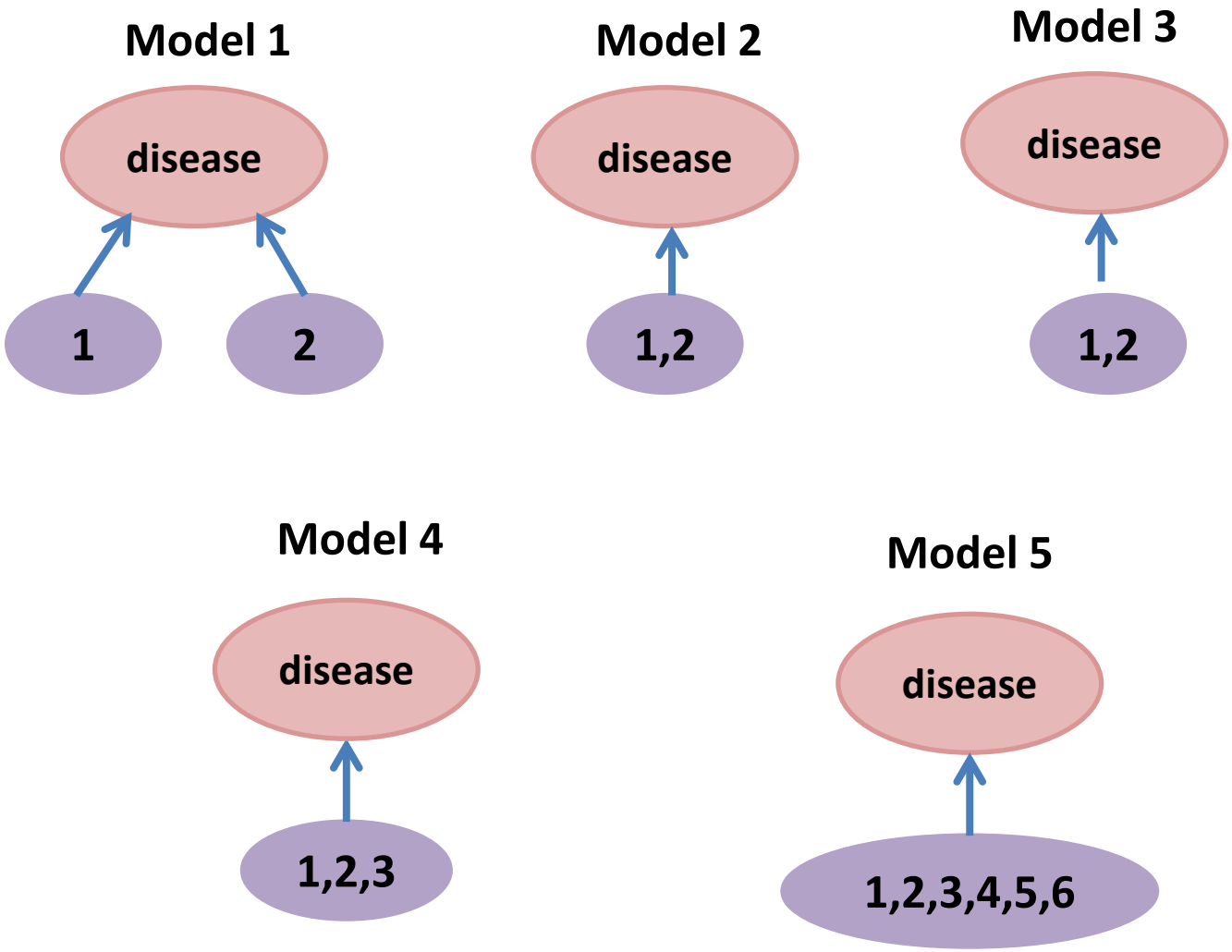


- $\mathbf{S}_k = \{C_j : a_j = k, j = 1, \cdots, n(\mathbf{h})\}$, for $k = 0, 1, \cdots, g(\mathbf{h})$.
- Note that some $\mathbf{S}_k$ may be empty.

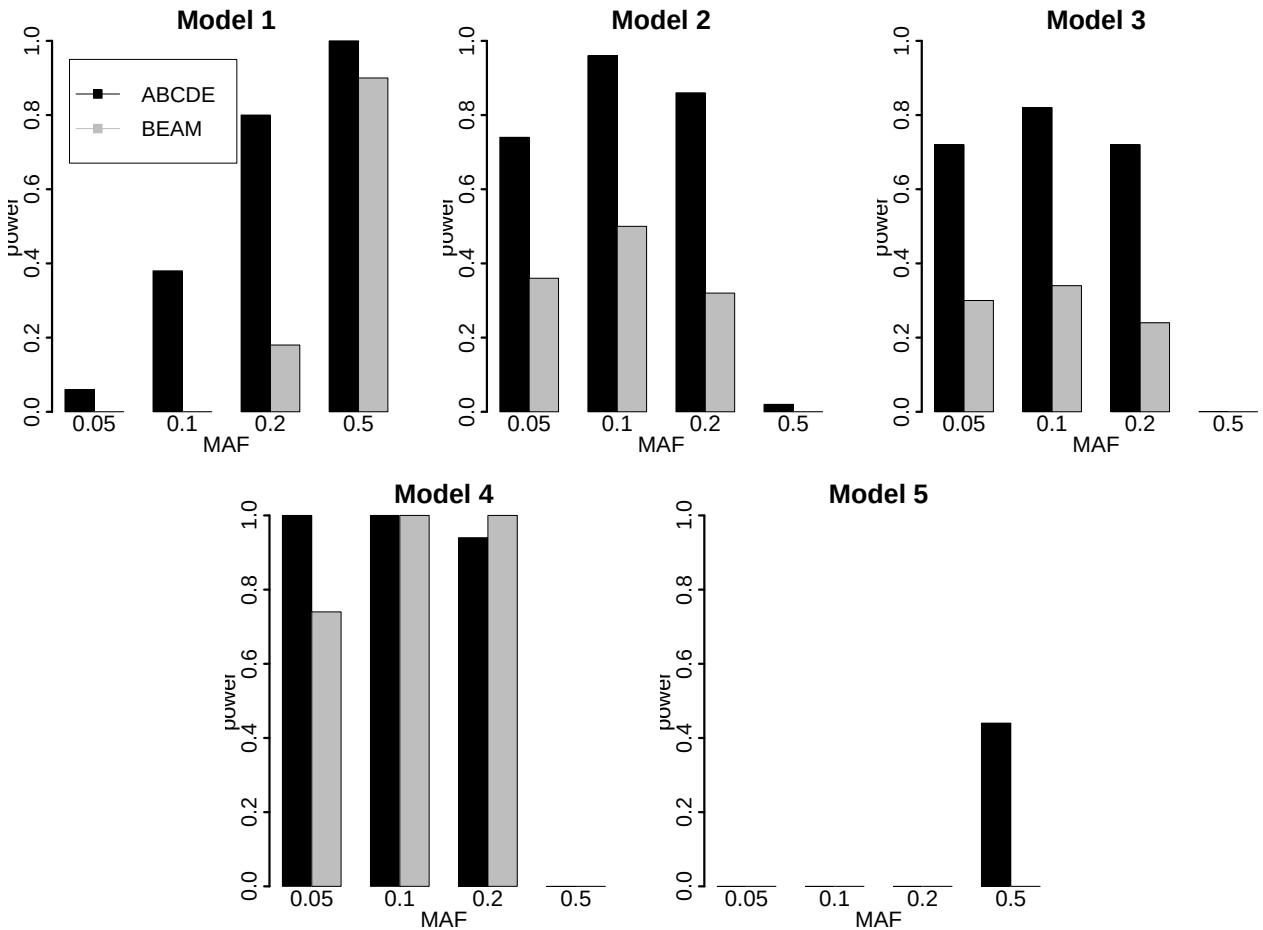
Simulation

- To evaluate the performance of ABCDE, we simulated data from 10 different models.
  - Single-set models (models 1-5)
  - Multiple-set models (models 6-8)
  - LD-extend models (models 9-10)
- Comparison between ABCDE and BEAM.

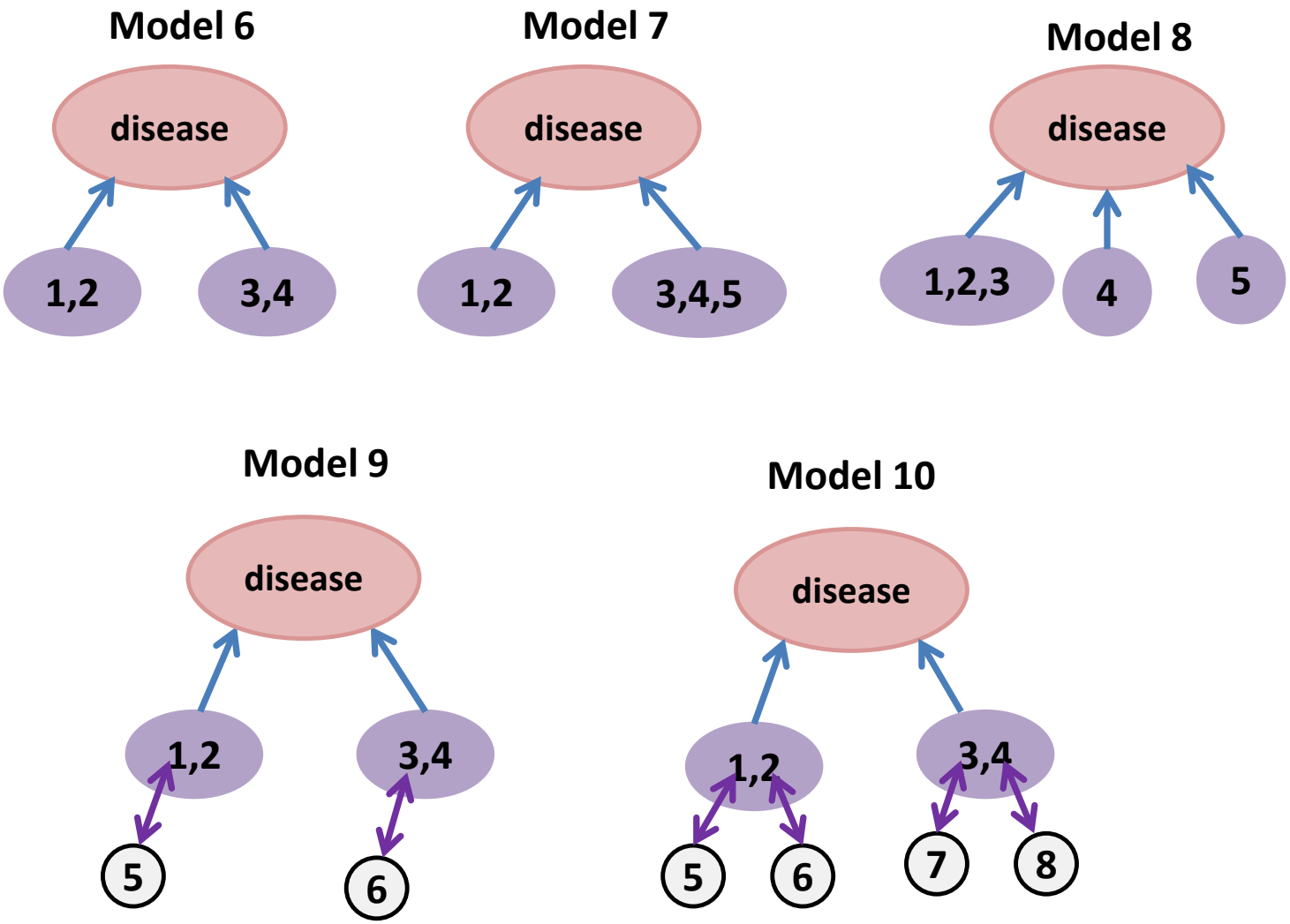
Single-set models



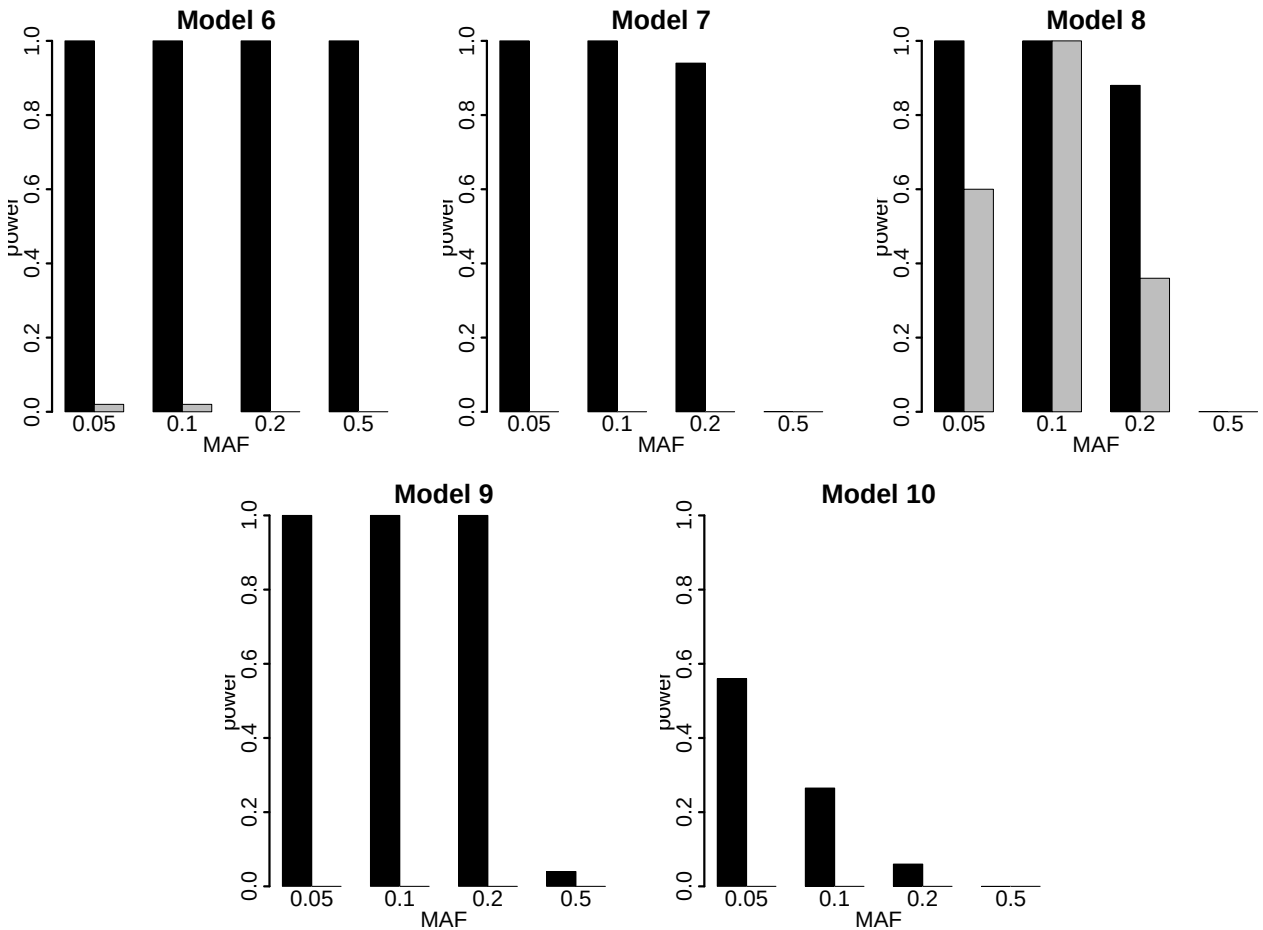
Result for Single-set models



Multiple-set models and LD-extend models



Result for Multiple-set models and LD-extend models

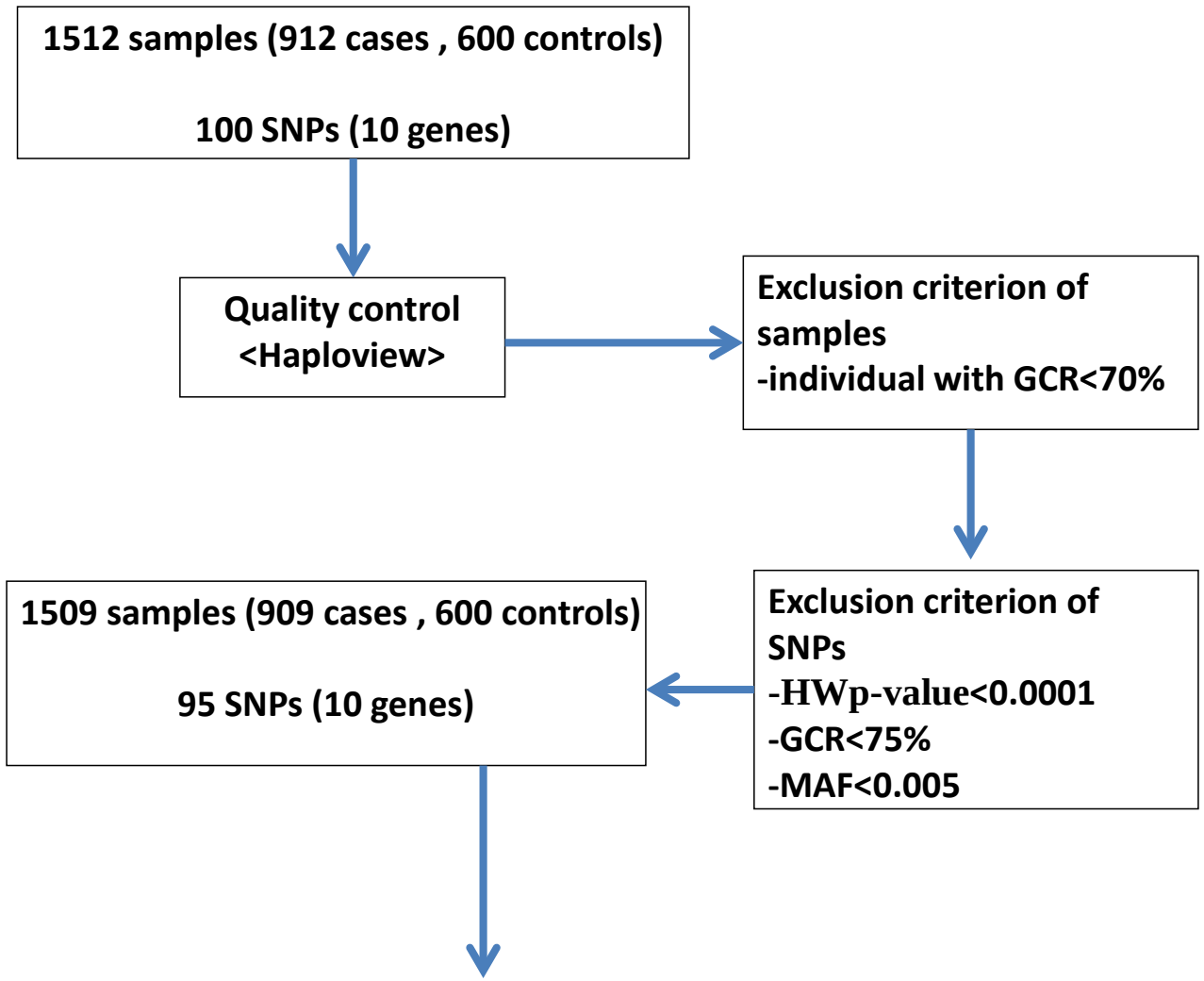


Real data

- Detect pairwise and/or higher-order SNP interactions and understand the genetic architecture of schizophrenia through ABCDE and BEAM.
- 1512 individuals, including 912 schizophrenia cases and 600 controls.

Gene	Chr	number
DISC1	1q	16
LMBRD1	6q	11
DPYSL2	8p	14
TRIM35	8p	10
PTK2B	8p	19
NRG1	8p	10
DAO	12q	5
G72	13q	5
RASD2	22q	4
CACNG2	22q	6

Flow chart-Quality Control



Result

Table: Identified significant epistatic sets by BEAM using all 95 SNPs.

SNP	Chr.	Gene	B-statistic(p-value)	BA(p-value)	PA(p-value)
rsDISC1P-3	1q	DISC1	55.19(9.89×10^{-11})	0.5944(0)	0.5557(0.018)
rsDISC1-23	1q	DISC1	31.31(1.51×10^{-5})	0.5705(0)	0.5416(0.224)
rsDPYSL-4	8p	DPYSL	21.26(0.002)	0.5561(0)	0.5156(0.399)
rsTRIM35-5	8p	TRIM	32.23(9.52×10^{-6})	0.5693(0)	0.5296(0.386)
rsNRG1P-7	8p	NRG1	59.88(9.44×10^{-12})	0.5996(0)	0.5815(0.024)
rsG72-E-2	13q	G72	43.16(4.03×10^{-8})	0.5839(0)	0.5695(0.029)

Result

Table: Identified significant epistatic sets by ABCDE using all 95 SNPs.

SNPs	Chr.	Gene	B-statistic(p-value)	BA(p-value)	PA(p-value)
rsDPYSL-15,rsSDPYSL2-11	8p	DPYSL	58.48(4×10^{-6})	0.5304(0.01)	0.5933(0.005)
rsSTRIM35-1,rsTRIM35-2,rsTRIM35-5	8p	TRIM35	127.97(0)	0.5647(0)	0.5146(0.412)
rsSDPYSL2-1,rsDPYSL-3,rsDPYSL-4	8p	DPYSL2	81.63(0.016)	0.5678(0)	0.6619(0)
rsDAO-6,rsDAO-7,rsDAO-8	12q	DAO	216.99(0)	0.582(0)	0.6531(0)
rsG72-E-1,rsG72-E-2,rsG72-13	13q	G72	91.00(5.32×10^{-4})	0.5866(0)	0.575(0.006)
rsSDISC1-1,rsDISC1P-3, rsDISC1-23,rsDISC1-27	1q	DISC1	251.41(0)	0.6325(0)	0.6178(0)
rsSDPYSL2-1,rsDPYSL-3, rsDPYSL-4,rsSDPYSL2-5	8p	DPYSL2	197.15(2.3×10^{-5})	0.5686(0)	0.6185(0)
rsNRG1P-6,rsNRG1P-7, rsCACNG2-16,rsCACNG2-15	(8p, 22q)	NRG1, CACNG2	86.96(1)	0.5962(0)	0.5642(0.05)
rsSTRIM35-1,rsTRIM35-2,rsTRIM35-4, rsTRIM35-5,rsTRIM35-6	8p	TRIM35	354.85(1)	0.572(0)	0.5255(0.403)
rsDAO-6,rsDAO-7,rsDAO-8 rsCACNG2-2,rsCACNG2P-1, rsCNCNG2-18	(12q,22q)	DAO, CACNG2	171.62(1)	0.5737(0)	0.6137(0)

Conclusion

- We propose the ABCDE algorithm which can character all explicit (interaction) effects, regardless of the number of groups.
- We further develop permutation tests to validate the disease association of SNP subsets selected by ABCDE.
- Applying ABCDE to the real data, we identify several known and novel schizophrenia-associated SNPs and sets of SNPs.
- We may develop a parallel implementation of the ABCDE, which is the algorithm for large scale epistatic interaction mapping, including genome-wide studies with hundreds of thousands of markers.