A Bayesian clustering approach for detecting gene-gene interactions in high-dimensional genotype data

Sui-Pi Chen and Guan-Hua Huang

Institute of Statistics
National Chiao Tung University
Hsinchu, Taiwan

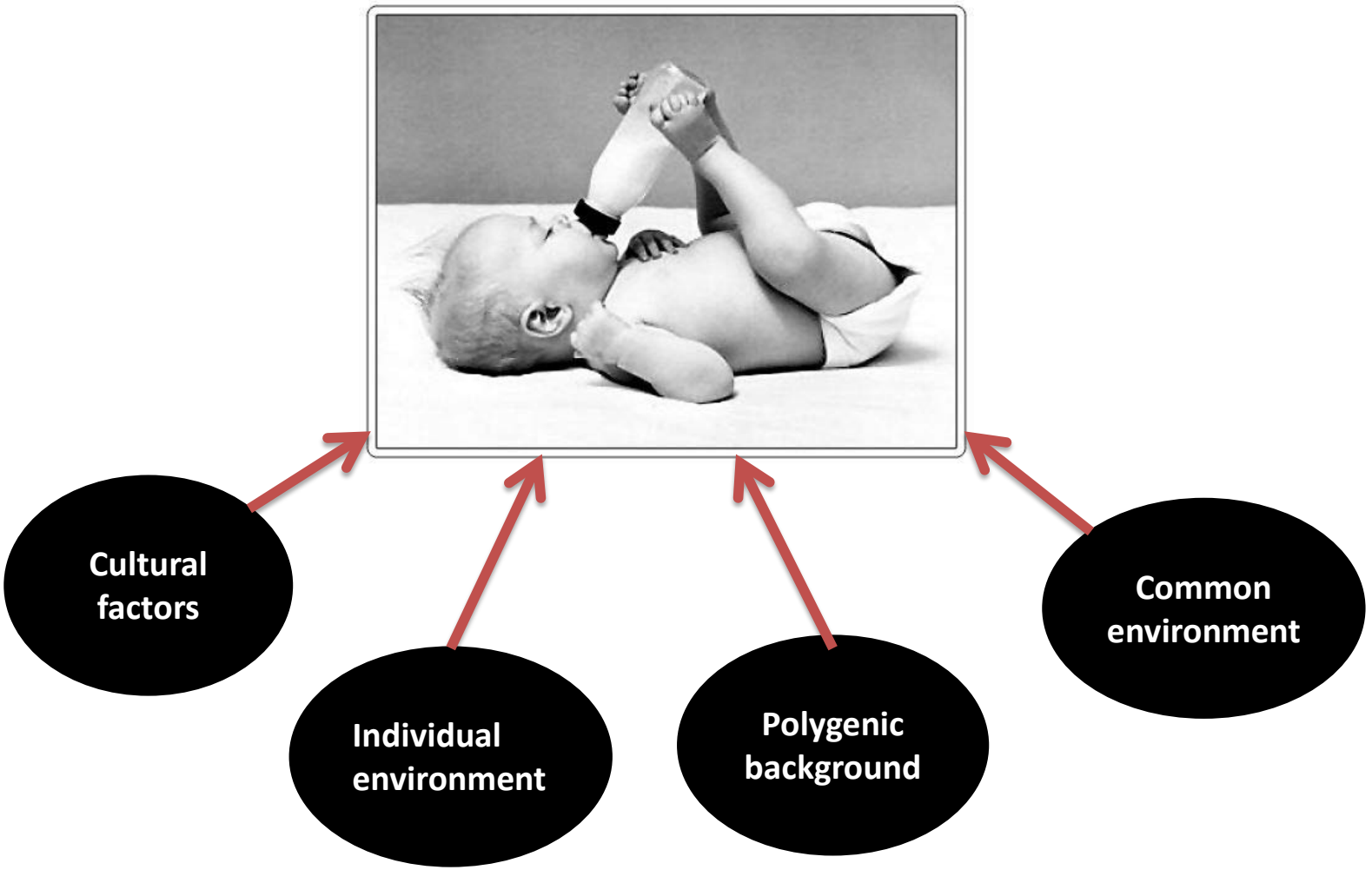
✉:ghuang@stat.nctu.edu.tw

2012.8.16

Outline

Outline

Motivation


Cultural factors
Individual environment
Polygenic background
Common environment

Single nucleotide polymorphism (SNP)

- A DNA sequence variation

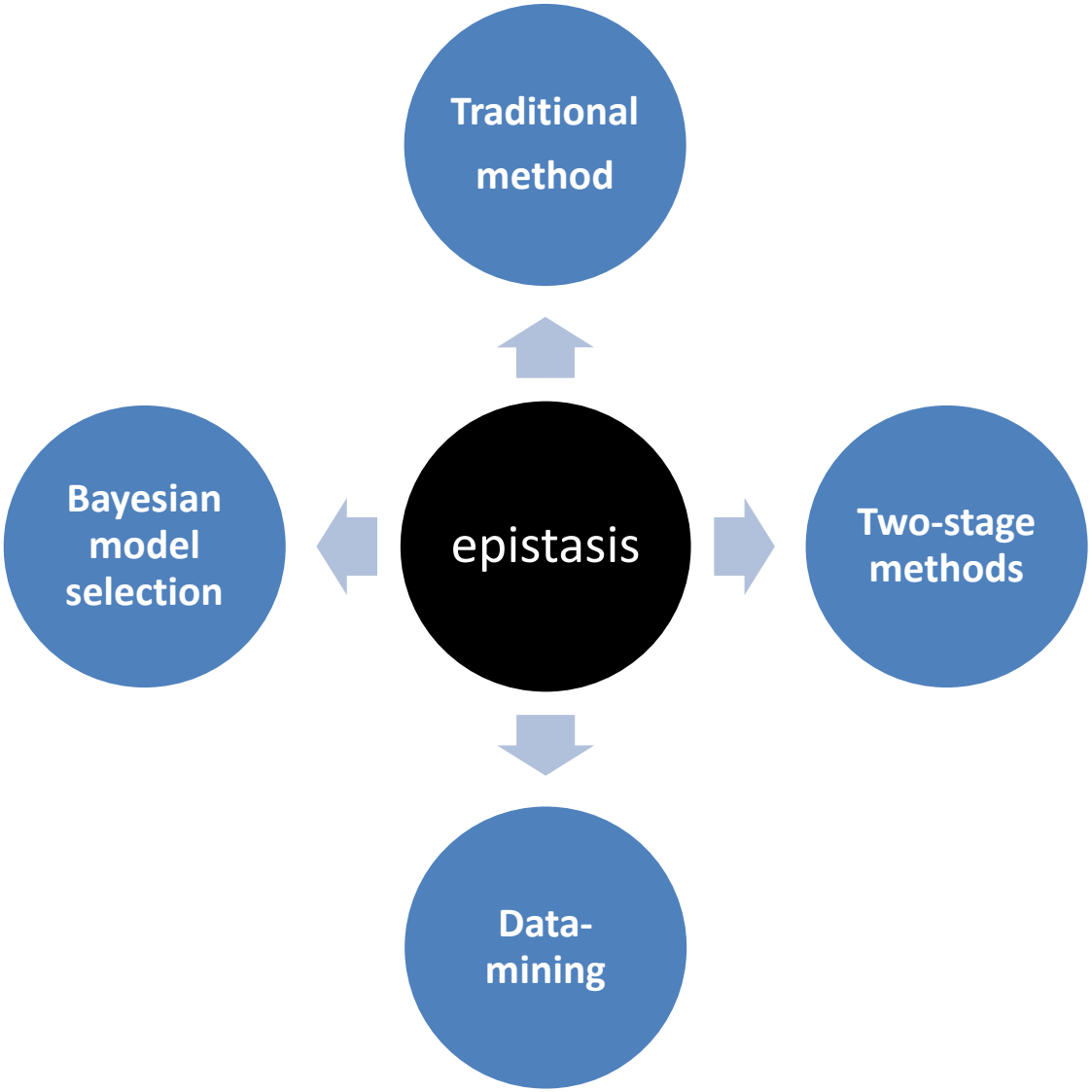
- Two alleles: A and a
- Treating SNPs as categorical features that have three possible values: AA, Aa, aa.
- Relabel AA (2),Aa (1),aa (0).

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.
- As increasing body of evidence has suggested that epistasis ploy an important role in susceptibility to human complex disease, such as Type 1 diabetes, breast cancer, obesity, and schizophrenia.
- More evidences have confirmed that display interaction effects without displaying marginal effect.

Outline

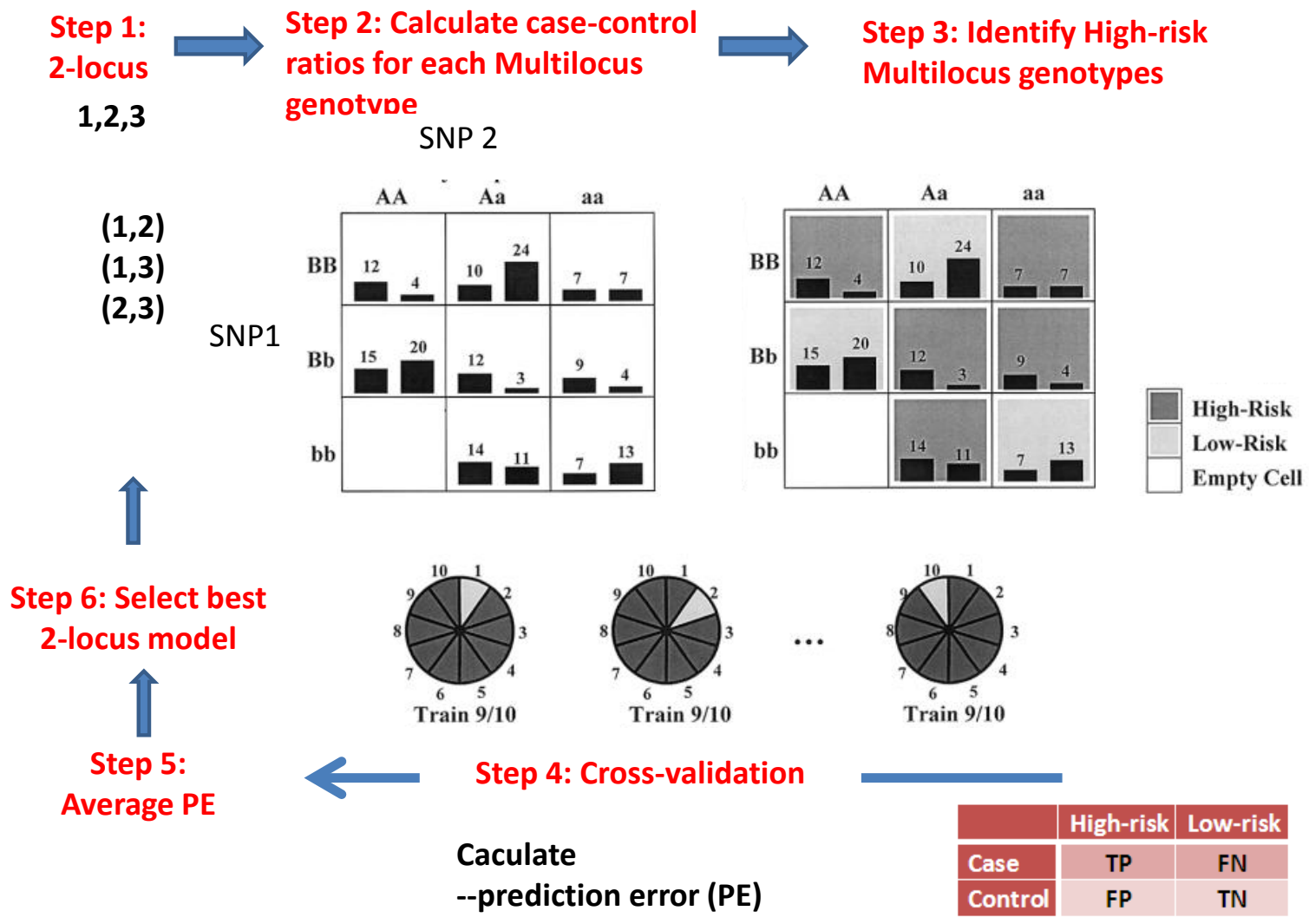
Methods for detecting gene-gene interaction



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 has 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 an 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)

Multifactor Dimensionality Reduction (MDR)



MDR

- From all best models, the model with minimal average prediction error is the final best model.
- MDR is the data reduction strategy which is the nonparametric model and genetic model-free.

Permutation test for the final best model.

- Applying MDR to 1000 permutation datasets, we use the PE of the 1000 final best models for the original data to create an empirical distribution for estimate of a p-value.

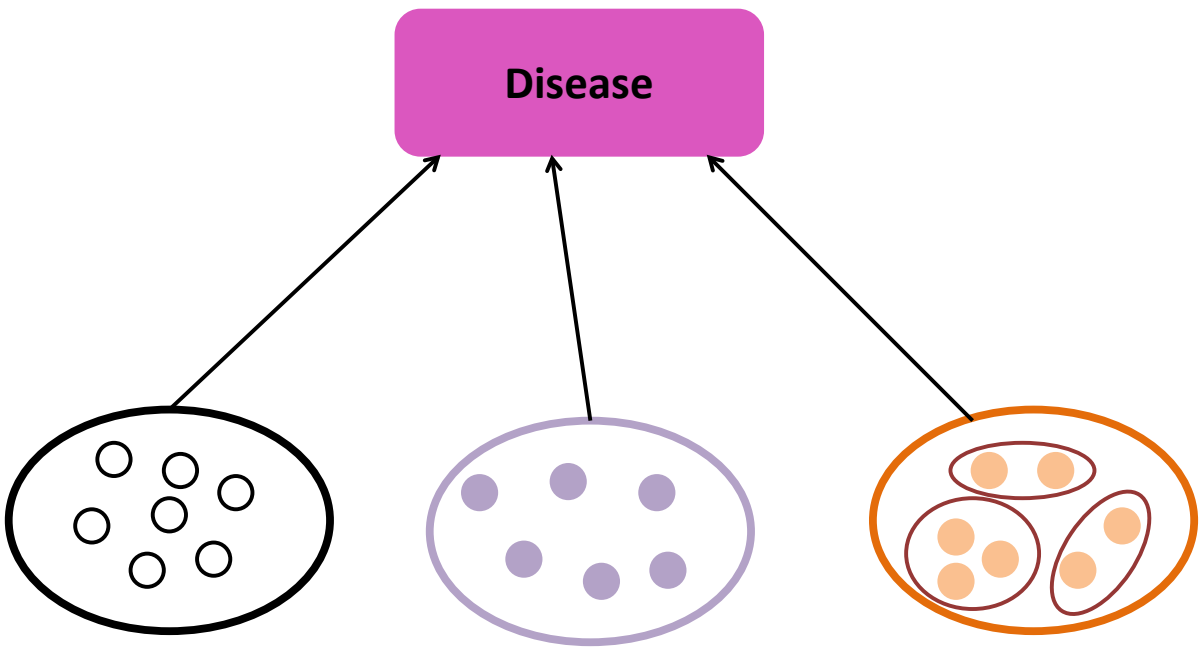
Note. This permutation test includes the variation of the search.

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

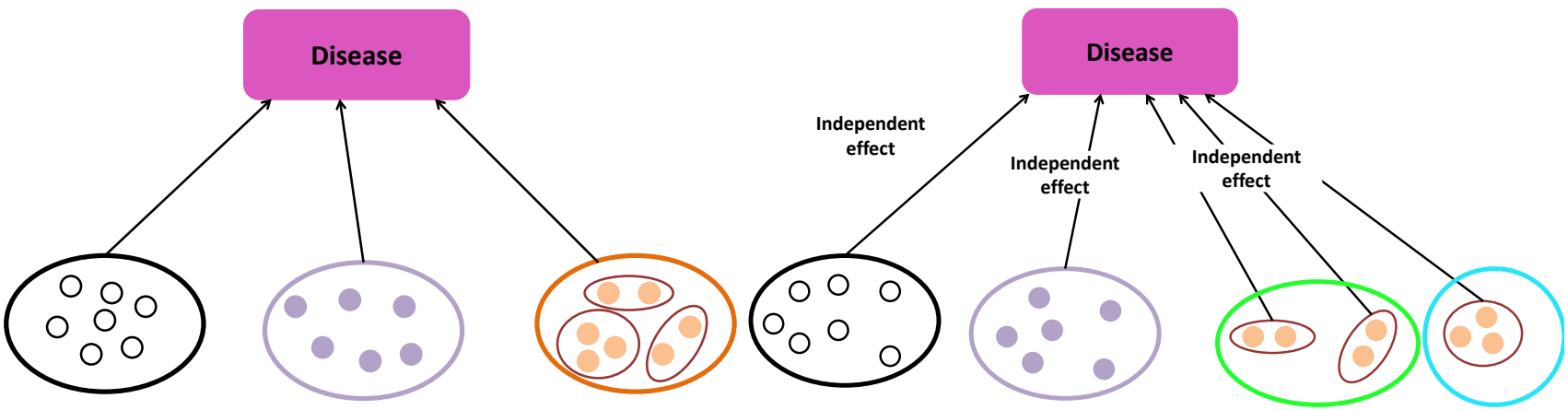
BEAM algorithm

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



Outline

Algorithm via Bayesian Clustering to Detect Epistasis (ABCDE)



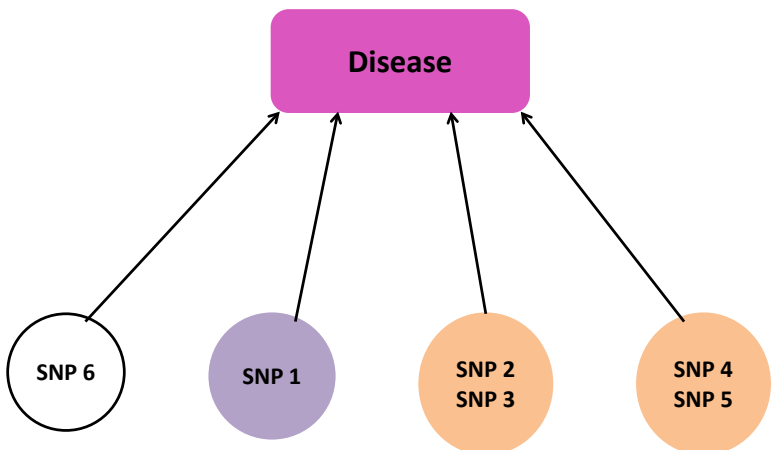
(a) BEAM (b) ABCDE

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.

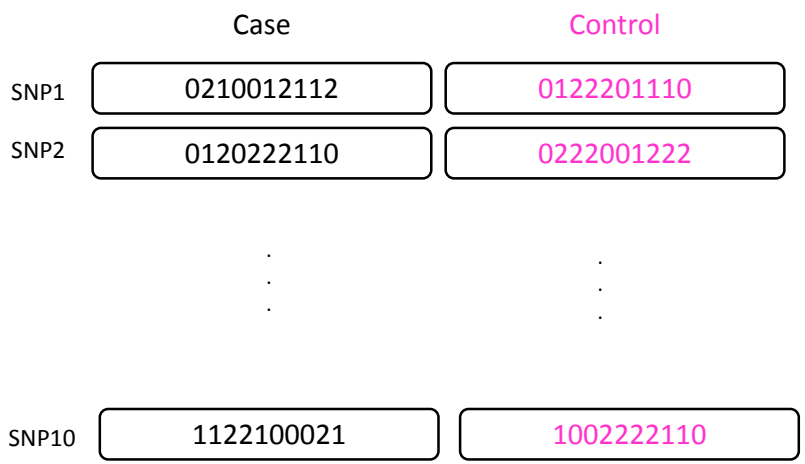
Example

- $\mathbf{c} = (C_1, \cdots, C_{n(\mathbf{c})})$.
- $\mathbf{c} = (\{1\}, \{2,3\}, \{4,5\}, \{6\})$.
- Add the group indicator $\mathbf{a} = (a_1, a_2, \cdots, a_{n(\mathbf{c})})$.
- Group membership of subset C_j: $a_j \in \{0, 1, 2, \cdots, g(\mathbf{c})\}$.
- The partition of interest is $\mathbf{h} = (H_1, \cdots, H_{n(\mathbf{h})})$, where $H_j = (C_j, a_j)$.
- $\mathbf{h} = (\{1\}, \{2,3\}, \{4,5\}, \{6\}), (0, 2, 2, 1))$.



Notations in ABCDE

- Treating SNPs as categorical features that have three possible values: AA(2), Aa(1), aa(0).
- N_d cases and N_u controls are genotyped at L SNPs.
- $\mathbf{G} = (\mathbf{D}, \mathbf{U})$
  $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_{N_d})$ be the case genotype ;
  $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_{N_u})$ be the control genotype.
- Genotypes of patient i at L SNPs: $\mathbf{d_i} = (d_{i1}, \cdots, d_{iL})$.
  Genotypes of control i at L SNPs: $\mathbf{u_i} = (u_{i1}, \cdots, u_{iL})$.



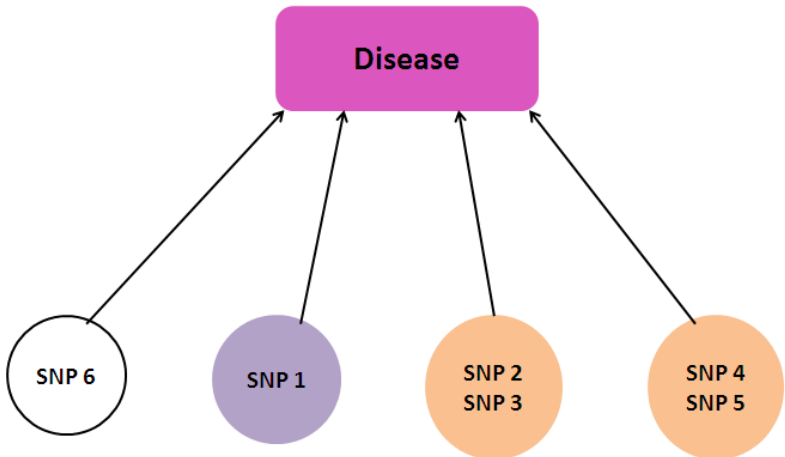
Product partition model

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

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

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

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



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

The data model- Group 0

Case genotype frequencies at unlinked SNPs are the same as control frequencies.

	Case			Control		
Genotype	AA	Aa	aa	AA	Aa	aa
Count	m_{0j1}	m_{0j2}	m_{0j3}	n_{0j1}	n_{0j2}	n_{0j3}

	Case+Control		
Genotype	AA	Aa	aa
Frequencies	θ_{0j1}	θ_{0j2}	θ_{0j3}
Count	$m_{0j1}+n_{0j1}$	$m_{0j2}+n_{0j2}$	$m_{0j3}+n_{0j3}$

The data model- Group 0

- Conditional distribution of $\mathbf{G}_{C_j}$ given $\mathbf{h}$ and $\boldsymbol{\theta}_{0j}$ as

$$f_0(\mathbf{G}_{C_j}|\boldsymbol{\theta}_{0j}) = \prod_{i=1}^{3} \theta_{0ji}{}^{(m_{0ji}+n_{0ji})},$$

- Specify a Dirichlet($\boldsymbol{\alpha}_0$) prior for $\boldsymbol{\theta}_{0j} = (\theta_{0j1}, \theta_{0j2}, \theta_{0j3})$, where $\boldsymbol{\alpha}_0 = (\alpha_{01}, \alpha_{02}, \alpha_{03})$.
- We integrate out $\boldsymbol{\theta}_{0j}$ and get the marginal distribution given $\mathbf{h}$ as

$$f_0(\mathbf{G}_{C_j}) = \frac{\Gamma(|\boldsymbol{\alpha}_0|)}{\Gamma(|\boldsymbol{\alpha}_0| + N_d + N_u)} \prod_{i=1}^{3} \frac{\Gamma(\alpha_{0i} + m_{0ji} + n_{0ji})}{\Gamma(\alpha_{0i})},$$

- $|\boldsymbol{\alpha}_0|$: the sum of all elements in $\boldsymbol{\alpha}_0$.

The data model- Group k

- SNP subset C_j associated with the disease should show different genotype frequencies between cases and controls.
- 3^k possible genotype combinations.

	Case				Control			
Genotype	AABB...	AABB...	$\cdots$	aabb...	AABB...	AABB...	$\cdots$	aabb...
Count	m_{kj1}	m_{kj2}	$\cdots$	m_{kj3^k}	n_{kj1}	n_{kj2}	$\cdots$	n_{kj3^k}

	Case				Control			
	AABB...	AABB...	$\cdots$	aabb...	AABB...	AABB...	$\cdots$	aabb...
Frequencies	θ_{kj1}	θ_{kj2}	$\cdots$	θ_{kj3^k}	γ_{kj1}	γ_{kj2}	$\cdots$	γ_{kj3^k}

The data model- Group k

- Conditional likelihood given $\mathbf{h}$, $\boldsymbol{\theta}_{kj}$ and $\boldsymbol{\gamma}_{kj}$

$$f_k(\mathbf{G}_{C_j}|\boldsymbol{\theta}_{kj}, \boldsymbol{\gamma}_{kj}) = \prod_{i=1}^{3^k} \theta_{kji}^{m_{kji}} \gamma_{kji}^{n_{kji}},$$

- We Specify a Dirichlet($\boldsymbol{\alpha}_k$) prior for $\boldsymbol{\theta}_{kj} = (\theta_{kj1}, \cdots, \theta_{kj3^k})$ and a Dirichlet($\boldsymbol{\beta}_k$) prior for $\boldsymbol{\gamma}_{kj} = (\gamma_{kj1}, \cdots, \gamma_{kj3^k})$.
  - $\boldsymbol{\alpha}_k = (\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{k3^k})$.
  - $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \cdots, \beta_{k3^k})$.
- Integrating out $\boldsymbol{\gamma}_{kj}$ and $\boldsymbol{\theta}_{kj}$, we obtain the marginal distribution $\mathbf{h}$

$$f_k(\mathbf{G}_{C_j}) = \frac{\Gamma(|\boldsymbol{\alpha}_k|)}{\Gamma(|\boldsymbol{\alpha}_k| + N_d)} \frac{\Gamma(|\boldsymbol{\beta}_k|)}{\Gamma(|\boldsymbol{\beta}_k| + N_u)} \prod_{i=1}^{3^k} \frac{\Gamma(\alpha_{ki} + m_{kji})}{\Gamma(\alpha_{ki})} \frac{\Gamma(\beta_{ki} + n_{kji})}{\Gamma(\beta_{ki})}.$$

The prior part

- A conjugate prior distribution of partition for the product partition model is the Dirichlet process.
- To distinguish subsets from group 0 and group 1, we assign a single SNP to be either group 0 or group 1 with equal probability.

$$p(\mathbf{h}) = p(\mathbf{c}, \mathbf{a}) \propto \frac{\delta^{n(\mathbf{h})} \prod_{j=1}^{n(\mathbf{h})} \Gamma(\#(C_j))}{2^B} = \prod_{j=1}^{n(\mathbf{h})} g(C_j),$$

$$\mathrm{E}(n(\mathbf{h})) = \delta \sum_{i=1}^{L-1} \frac{1}{\delta + i}.$$

- δ approaches 0 and ∞, the expected number has limiting values 1 and L, respectively.

MCMC sampling

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

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

Posterior

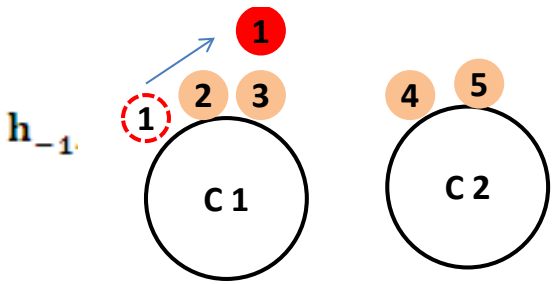
$$p(\mathbf{h}|\mathbf{G}) \propto \prod_{j=1}^{n(\mathbf{h})} g^*(C_j) \text{ with } g^* = g(C_j) \times f_{a_j}(G_{C_j})$$

$\Rightarrow$ Need a procedure to simulate from a distribution proportional to $\prod_{j=1}^{n(\mathbf{h})} g^*(C_j)$.

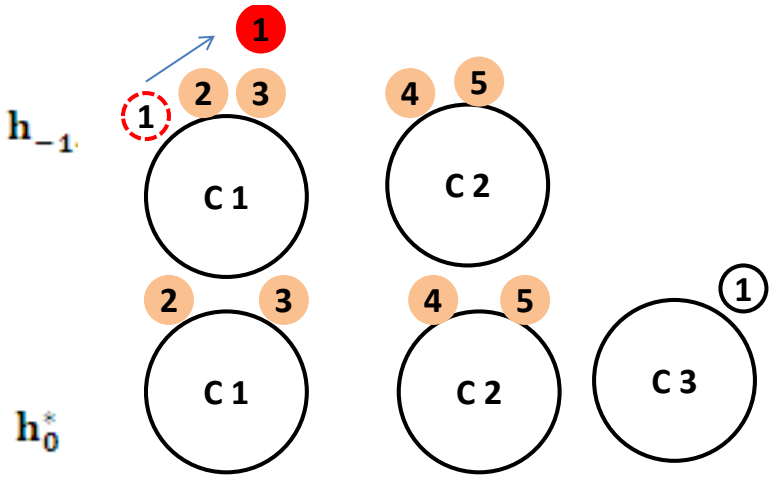
Gibbs weighted Chinese restaurant (GWCR) procedure

- Choose an initial partition $\mathbf{h}_0$
- The following Gibbs cycle, for $i = 1, \cdots, L$, do
  1. Remove $\{i\}$, from $\mathbf{h}_{-i}$
  2. Reseat $\{i\}$ according to the seating probabilities $p(\mathbf{h}^*|\mathbf{G})/p(\mathbf{h}_{-i}|\mathbf{G})$,where $\mathbf{h}^*$ is the resulting partition after the reassignment of marker t
- To get a new partition of $1, \cdots, n$.

Gibbs weighted Chinese restaurant (GWCR) procedure

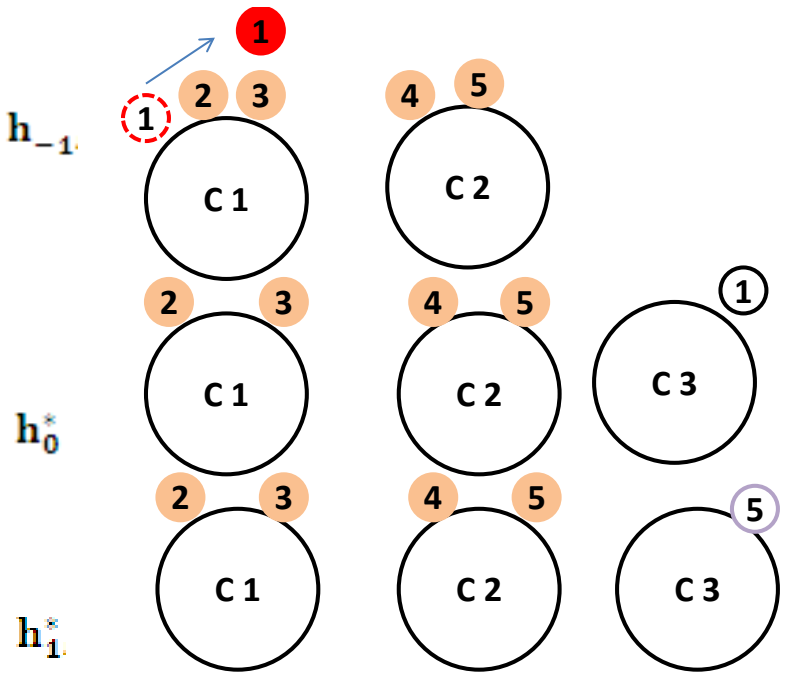


Gibbs weighted Chinese restaurant (GWCR) procedure



$$q_0 \propto \frac{p(\mathbf{h}_0^*|\mathbf{G})}{p(\mathbf{h}_{-1}|\mathbf{G})}$$

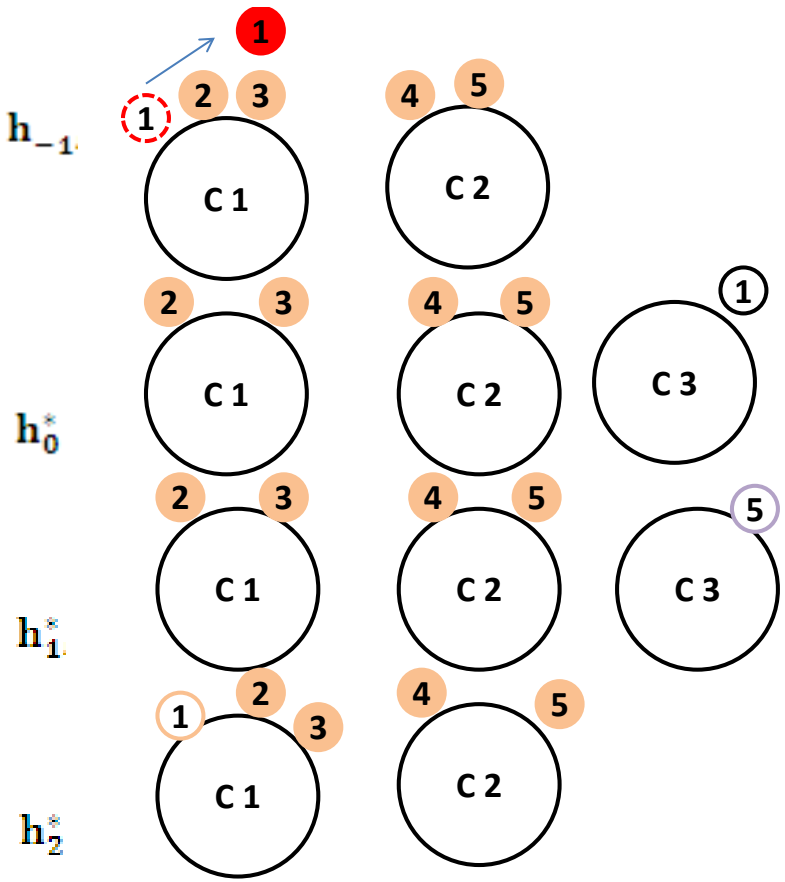
Gibbs weighted Chinese restaurant (GWCR) procedure



$$q_0 \propto \frac{p(\mathbf{h}_0^*|\mathbf{G})}{p(\mathbf{h}_{-1}|\mathbf{G})}$$

$$q_1 \propto \frac{p(\mathbf{h}_1^*|\mathbf{G})}{p(\mathbf{h}_{-1}|\mathbf{G})}$$

Gibbs weighted Chinese restaurant (GWCR) procedure

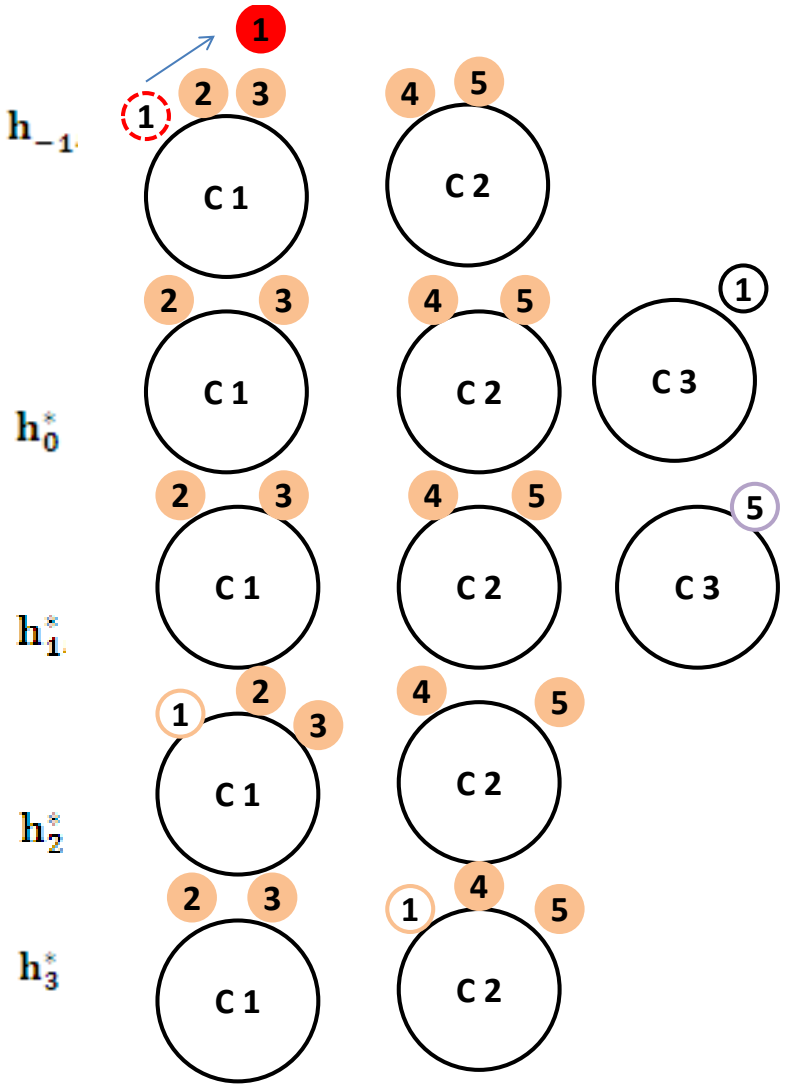


$$q_0 \propto \frac{p(\mathbf{h}_0^*|\mathbf{G})}{p(\mathbf{h}_{-1}|\mathbf{G})}$$

$$q_1 \propto \frac{p(\mathbf{h}_1^*|\mathbf{G})}{p(\mathbf{h}_{-1}|\mathbf{G})}$$

$$q_2 \propto \frac{p(\mathbf{h}_2^*|\mathbf{G})}{p(\mathbf{h}_{-1}|\mathbf{G})}$$

Gibbs weighted Chinese restaurant (GWCR) procedure



$$q_0 \propto \frac{p(\mathbf{h}_0^*|\mathbf{G})}{p(\mathbf{h}_{-1}|\mathbf{G})}$$

$$q_1 \propto \frac{p(\mathbf{h}_1^*|\mathbf{G})}{p(\mathbf{h}_{-1}|\mathbf{G})}$$

$$q_2 \propto \frac{p(\mathbf{h}_2^*|\mathbf{G})}{p(\mathbf{h}_{-1}|\mathbf{G})}$$

$$q_3 \propto \frac{p(\mathbf{h}_3^*|\mathbf{G})}{p(\mathbf{h}_{-1}|\mathbf{G})}$$

Gibbs weighted Chinese restaurant (GWCR) procedure

- Output:
  1. Posterior Mode: $\mathbf{h}^* = \max_{\mathbf{h}} p(\mathbf{h}|\mathbf{G})$
  2. The posterior distribution of single SNPs and subset of SNPs association with the disease.

Permutation test

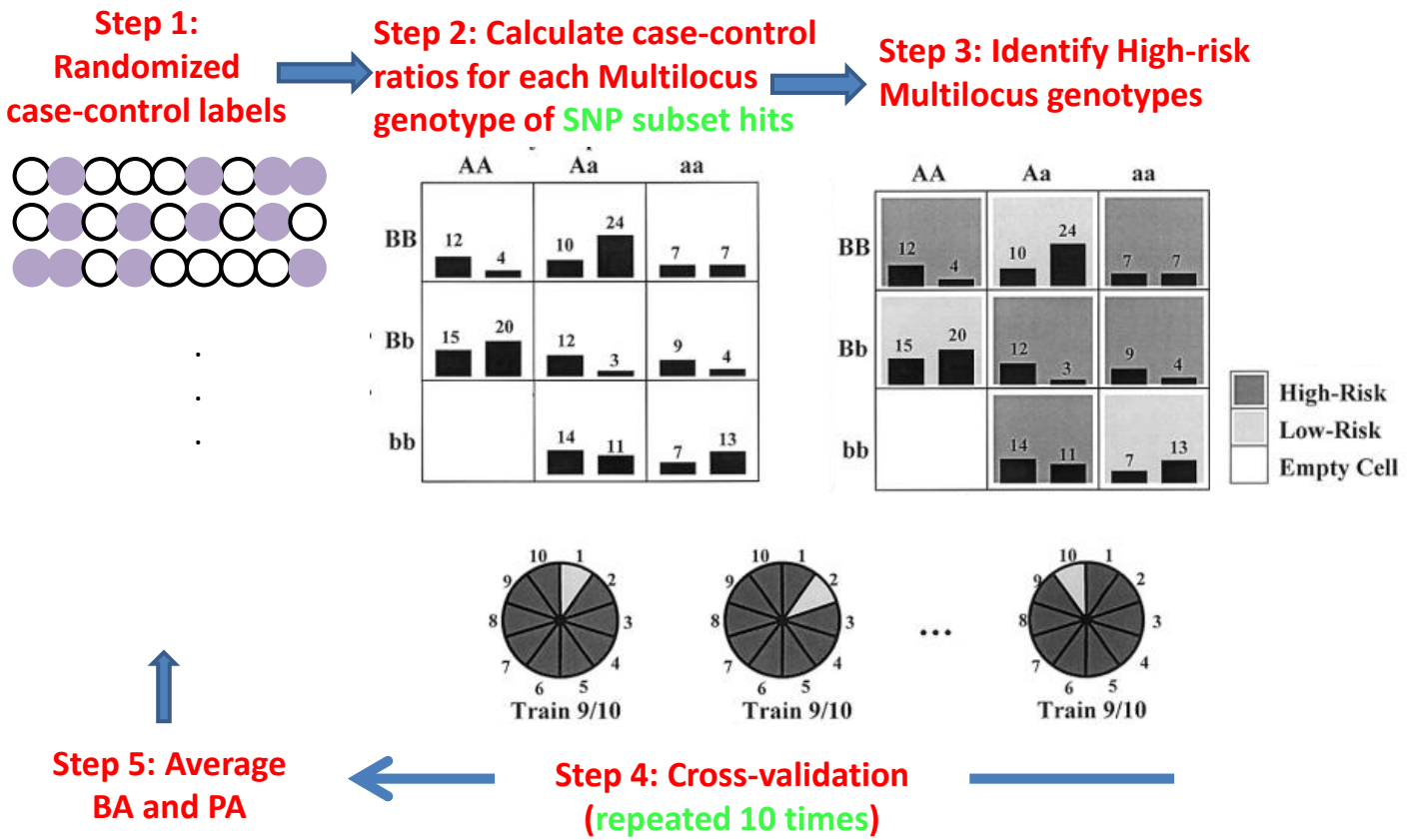
- 10-fold cross-validation and the heart of MDR.
- disease association for SNP subsets selected by ABCDE.
- validation test.
- Don't take the variation of SNP subset selection into count.
- Balance accuracy (BA) and prediction accuracy (PA).

$$\text{BA} = \frac{\text{sensitivity} + \text{specificity}}{2} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right),$$

$$\text{PA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}},$$

- The BA function (Velez et al.,2007) is preferable to PA when there is an imbalanced dataset.

Permutation test



Calculate
--Balance accuracy (BA)
--Prediction accuracy (PA)

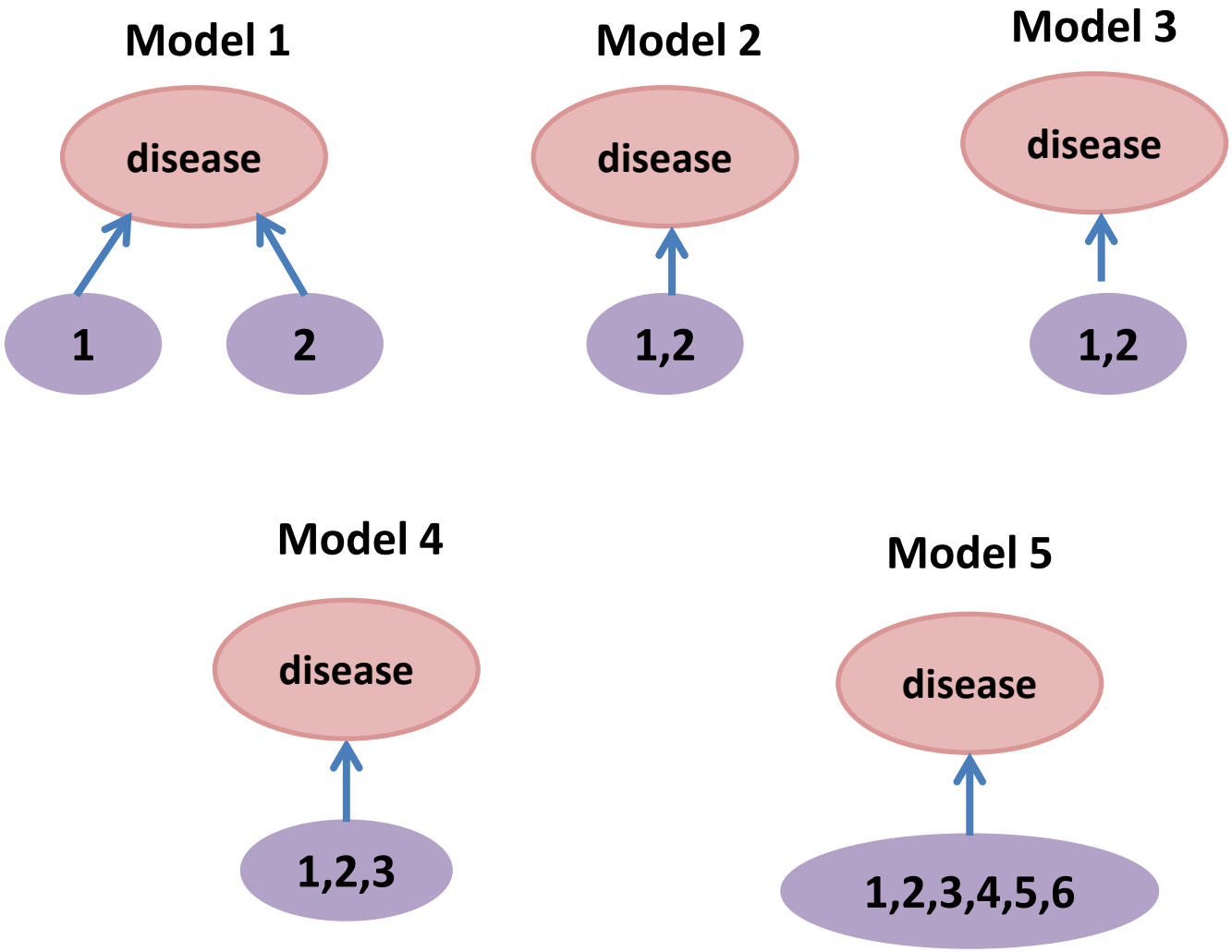
	High-risk	Low-risk
Case	TP	FN
Control	FP	TN

Outline

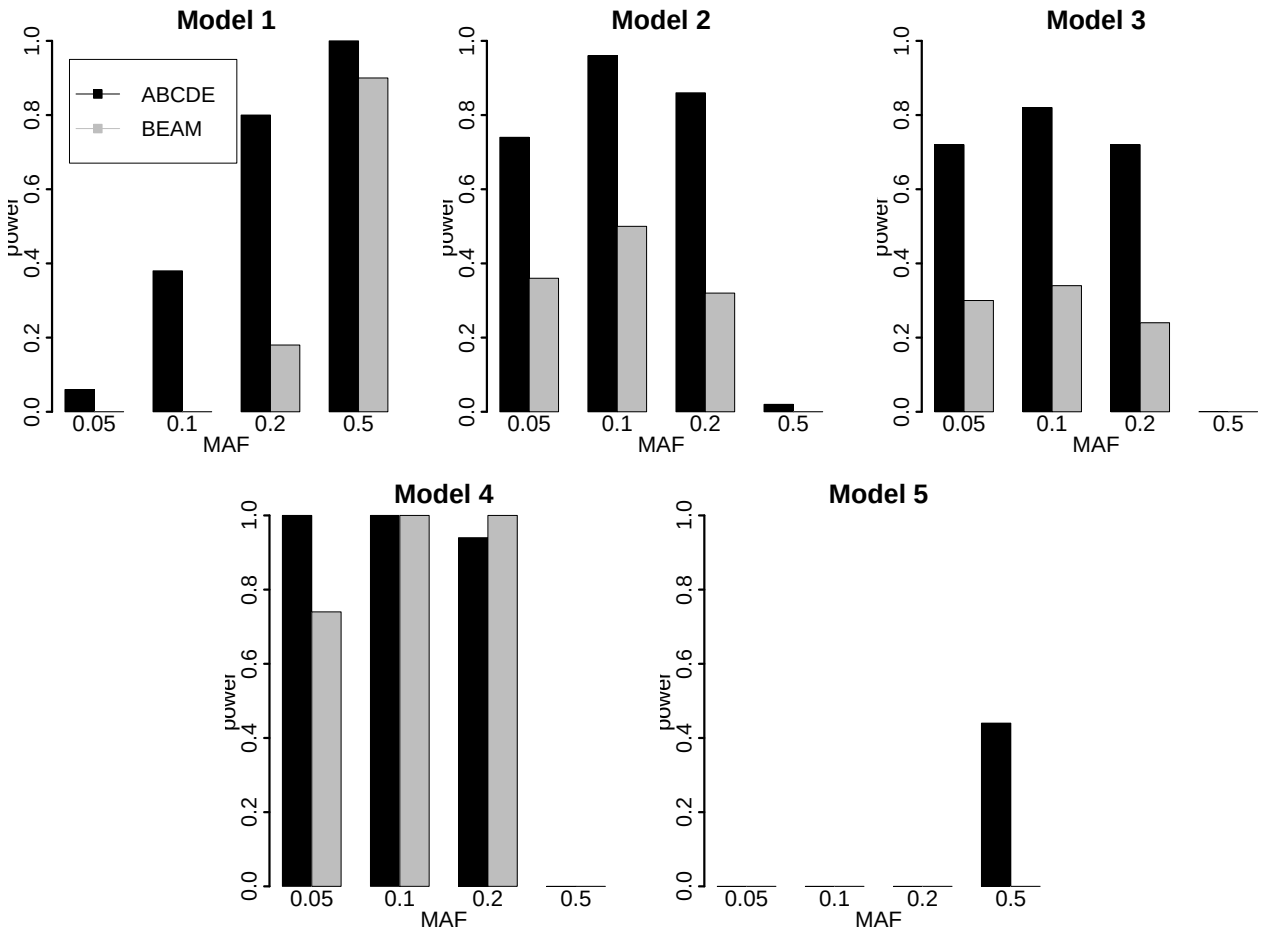
Simulation

- 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.

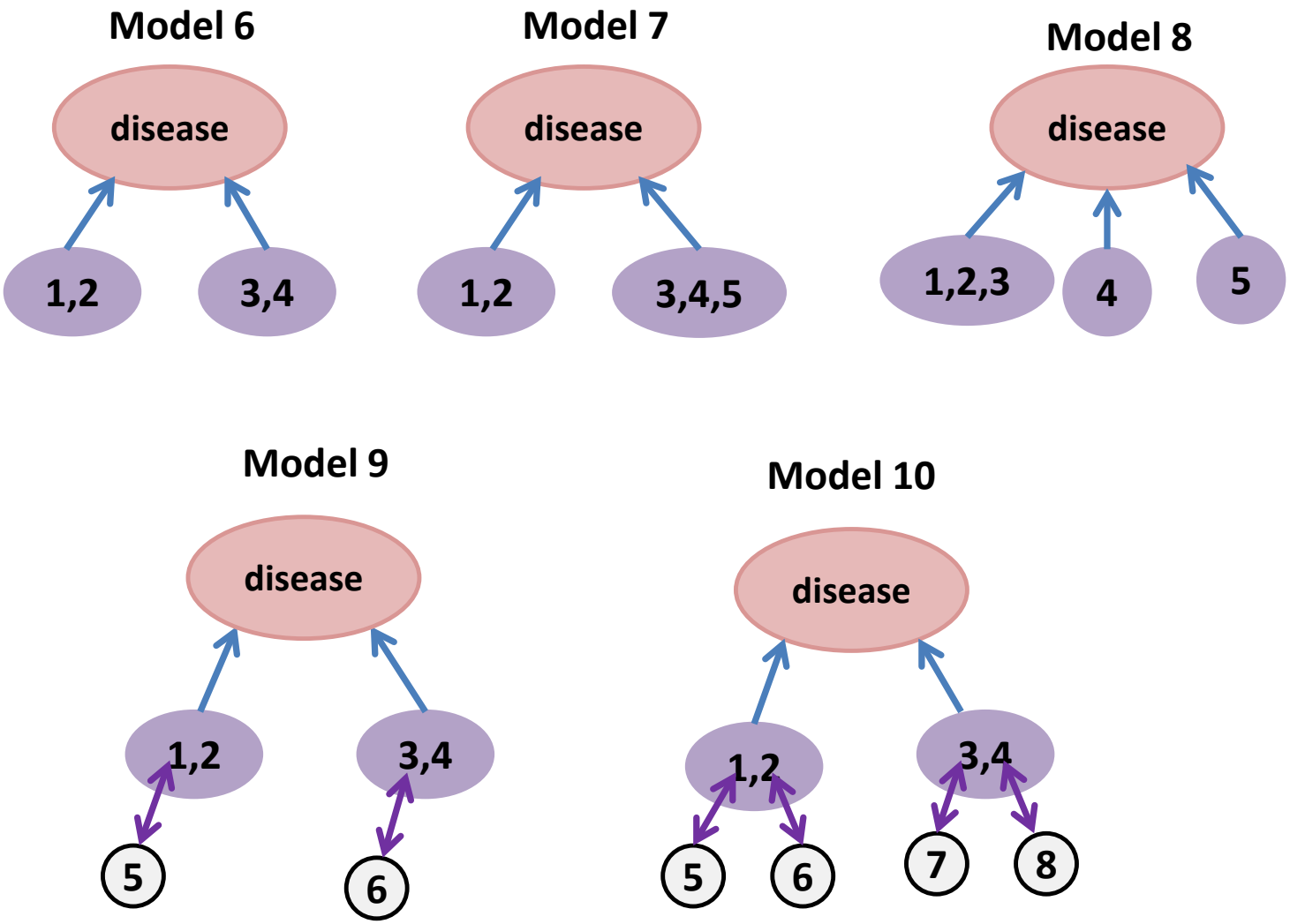
Single-set models



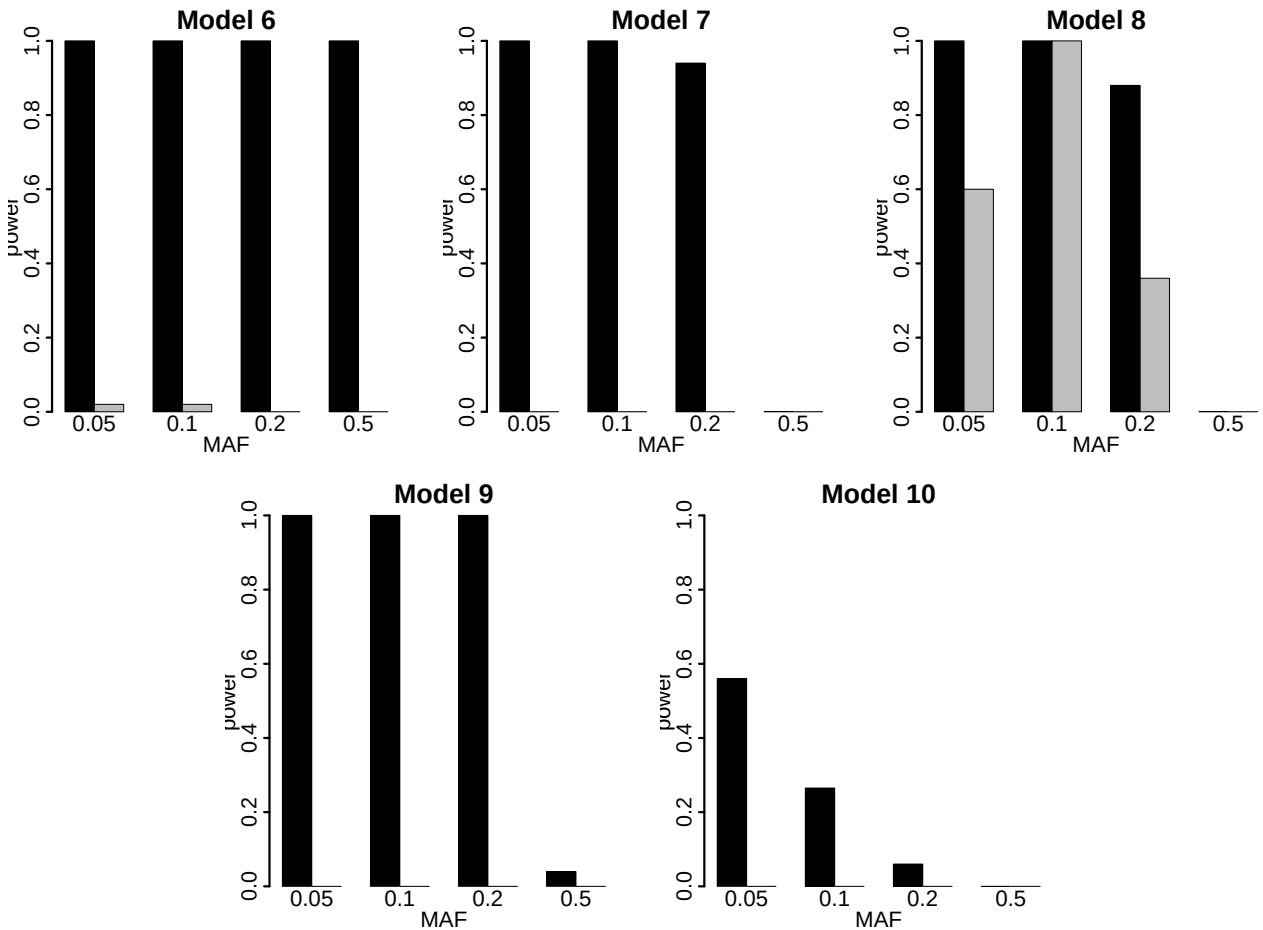
Result for Single-set models



Multiple-set models and LD-extend models



Result for Multiple-set models and LD-extend models



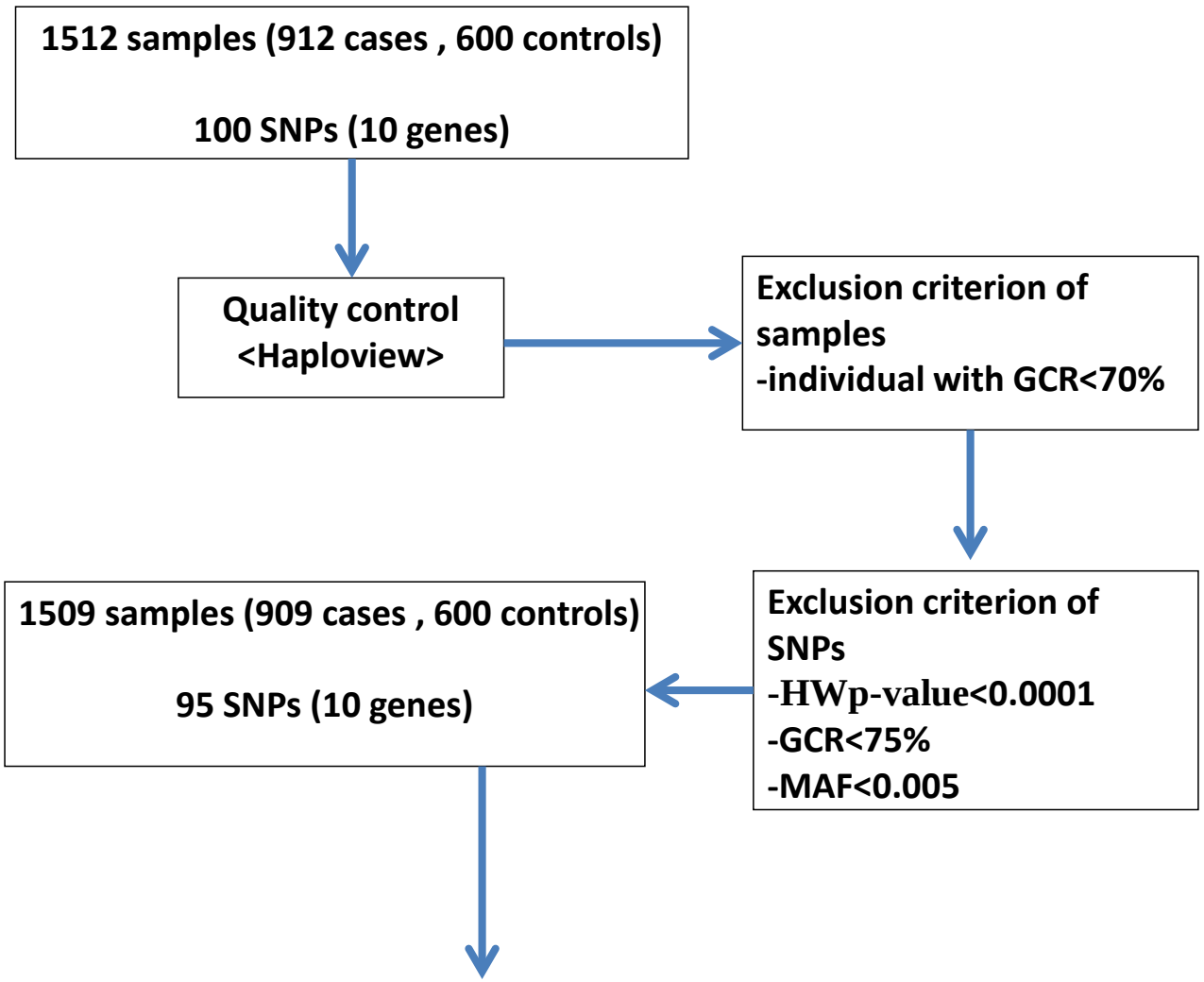
Outline

Real data

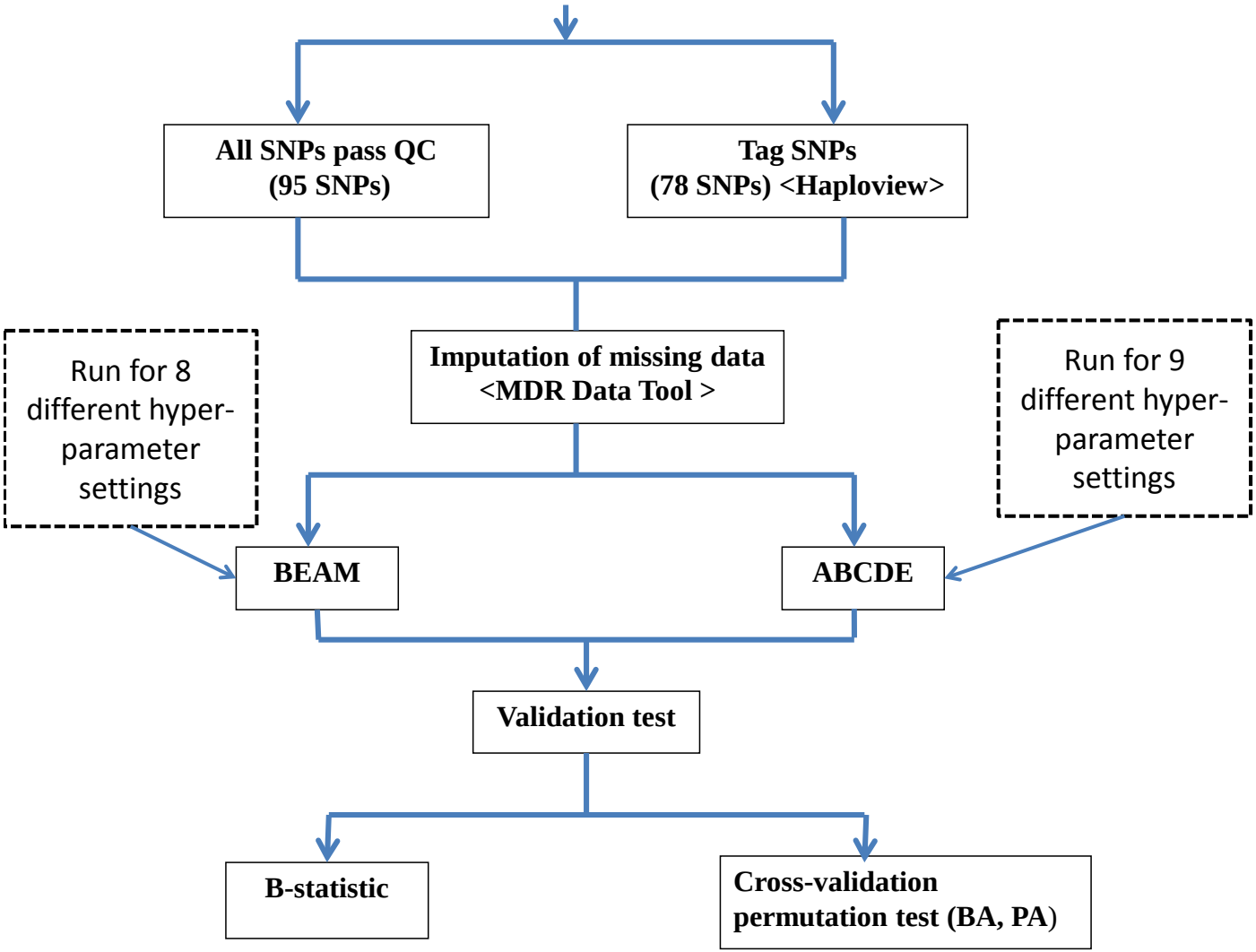
- Detect pairwise and/or higher-order SNP interactions and understand the genetic architecture of schizophrenia through ABCDE and BEAM.
- 1512 individuals, including 912 schizophrenia cases and 600 controls.

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



Flow chart


All SNPs pass QC
(95 SNPs)
Tag SNPs
(78 SNPs) <Haploview>
Imputation of missing data
<MDR Data Tool >
Run for 8 different hyper-parameter settings
Run for 9 different hyper-parameter settings
BEAM
ABCDE
Validation test
B-statistic
Cross-validation permutation test (BA, PA)

Detection of gene-gene interaction

To obtain robust results, we adopted the two-stage approach.

- Candidate SNP or subset SNPs hit by ABCDE (BEAM): In at least 3 out of different settings, candidate SNP subset hit with the posterior probability higher than a predefined cut-off, 0.3.
- Susceptibility SNPs: permutation test (p-value< 0.001) or B-statistic (p-value< 0.1).

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 BEAM using 78 selected tag SNPs.

SNP	Chr.	Gene	B-statistic(p-value)	BA(p-value)	PA(p-value)
rsDISC1-23	1q	DISC1	31.31(1.24×10^{-5})	0.5705(0)	0.5434(0.179)
rsDPYSL-4	8p	DPYSL	21.26(0.0018)	0.5561(0)	0.5176(0.415)
rsDPYSL-15	8p	DPYSL	13.59(0.087)	0.5328(0)	0.4606(0.574)
rsTRIM35-5	8p	TRIM	32.23(7.82×10^{-6})	0.5693(0)	0.5315(0.343)
rsNRG1P-7	8p	NRG1	59.88(7.76×10^{-12})	0.5996(0)	0.5832(0.013)
rsG72-E-2	13q	G72	43.16(3.31×10^{-8})	0.5839(0)	0.5712(0.022)
rsSDISC1-1,rsDISC1-23	1q	DISC1	50.89(8.29×10^{-5})	0.5672(0)	0.5838(0.004)
rsDISC1-27,rsDISC1-23	1q	DISC1	55.85(9.05×10^{-6})	0.5632(0)	0.5885(0.001)
rsDISC1-23,rsDISC1-4	1q	DISC1	35.71(0.059)	0.5765(0)	0.5765(0.002)
rsSDISC1-1,rsDISC1-23,rsDISC1-27	1q	DISC1	74.51(0.109)	0.5692(0)	0.5792(0.001)
rsSDISC1-1,rsDISC1-23,rsDISC1-4	1q	DISC1	63.09(1)	0.5678(0)	0.5885(0)
rsDISC1-23,rsDISC1-27,rsDISC1-4	1q	DISC1	70.62(0.41)	0.5588(0)	0.5779(0.002)
rsSDISC1-1,rsDISC1-23, rsDISC1-27,rsDISC1-4	1q	DISC1	87.56(1)	0.5708(0)	0.5905(0.001)

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)

Result

Table: Identified significant epistatic sets by ABCDE using 78 selected tag SNPs.

SNPs	Chr.	Gene	B-statistic(p-value)	BA(p-value)	PA(p-value)
rsDPYSL-15,rsSDPYSL2-11	8p	DPYSL	58.48(2.78×10^{-6})	0.5304(0.007)	0.5933(0.006)
rsSDPYSL2-1,rsDPYSL-3,rsDPYSL-4	8p	DPYSL	81.63(0.0089)	0.5678(0)	0.6619(0)
rsTRIM35-4,rsTRIM35-5,rsTRIM35-6	8p	TRIM35	157.49(0)	0.5651(0)	0.5256(0.38)
rsNRG1-1,rsNRG1P-6,rsNRG1P-7	8p	NRG1	75.64(0.074)	0.5888(0)	0.5736(0.006)
rsG72-E-1,rsG72-E-2,rsG72-13	13q	G72	91.00(2.92×10^{-4})	0.5866(0)	0.575(0.006)
rsDPYSL2-1,rsDPYSL-3, rsDPYSL-4,rsDPYSL-21	8p	DPYSL	197.15(1.01×10^{-5})	0.5656(0)	0.6223(0)
rsDAO-6,rsDAO-8, rsG72-E-2,rsG72-13	(12q, 13q)	(DAO,G72)	181.52(0.0011)	0.6289(0)	0.6769(0)
rsSDISC1-1,rsDISC1-23,rsDISC1-27, rsDISC1-2,rsDISC1-35	1q	DISC1	25.62(1)	0.5919(0)	0.5969(0)

Outline

Efficient Stochastic Search

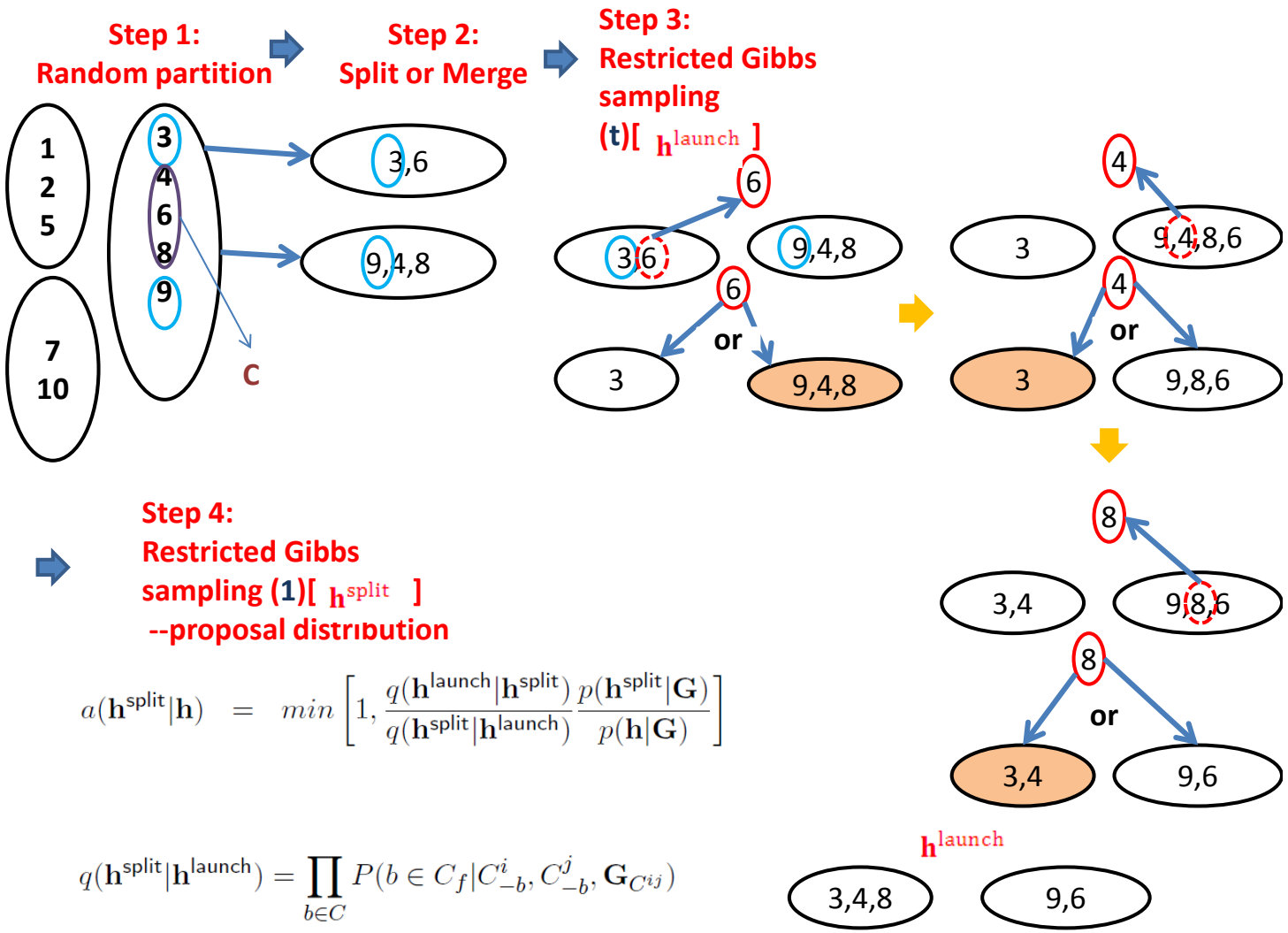
Although the GWCR algorithm works well high-dimensional data (simulation data with 1000 SNPs from 2000 cases and 2000 controls), genome-scale gene-gene interaction analysis is still infeasible.

- To improve the mixing of chains: Restricted Gibbs split merge procedure (RGSM) (Jain and Neal, 2004).
- Be easy to move between local modes: equi-energy (EE) sampler (Kou, Zhou and Wong, 2006)

Restricted Gibbs split merge procedure (RGSM)

- Simple random split-merge procedure:
  - The split proposals are unlikely to be appropriate, and hence are unlikely to be accepted.
- Restricted Gibbs split merge procedure (RGSM):
  - To employs a more complex proposal distribution obtained by using a Gibbs sampling on subset of data.
  - The split proposals with reference to the observed data is will likely be accepted.

Outline of Restricted Gibbs split merge procedure



$$a(\mathbf{h}^{\text{split}}|\mathbf{h}) = min\left[1, \frac{q(\mathbf{h}^{\text{launch}}|\mathbf{h}^{\text{split}})}{q(\mathbf{h}^{\text{split}}|\mathbf{h}^{\text{launch}})} \frac{p(\mathbf{h}^{\text{split}}|\mathbf{G})}{p(\mathbf{h}|\mathbf{G})}\right]$$

$$q(\mathbf{h}^{\text{split}}|\mathbf{h}^{\text{launch}}) = \prod_{b \in C} P(b \in C_f | C^i_{-b}, C^j_{-b}, \mathbf{G}_{C^{ij}})$$

Equi-Energy (EE) Sampler

The distribution of the system is thermal equilibrium at temperature T is described by the Boltzmann distribution,

$$p(\mathbf{h}) = \frac{1}{Z(T)} exp(\frac{-q(\mathbf{h})}{T})$$

where $Z(T) = \sum_{\mathbf{h}} exp(\frac{-q(\mathbf{h})}{T})$.

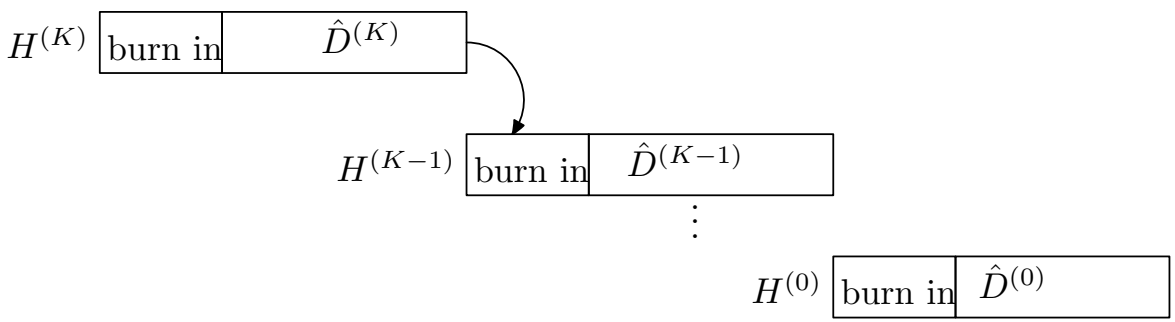
- $p(\mathbf{h})$: posterior distribution.
- $q(\mathbf{h})$: $-log(p(\mathbf{h}))$

Equi-Energy (EE) Sampler

$$1 = T_0 < T_1 < \cdots < T_K$$

$$p_i(h) = \frac{1}{Z(T_i)} exp(\frac{-q(\mathbf{h})}{T_i})$$

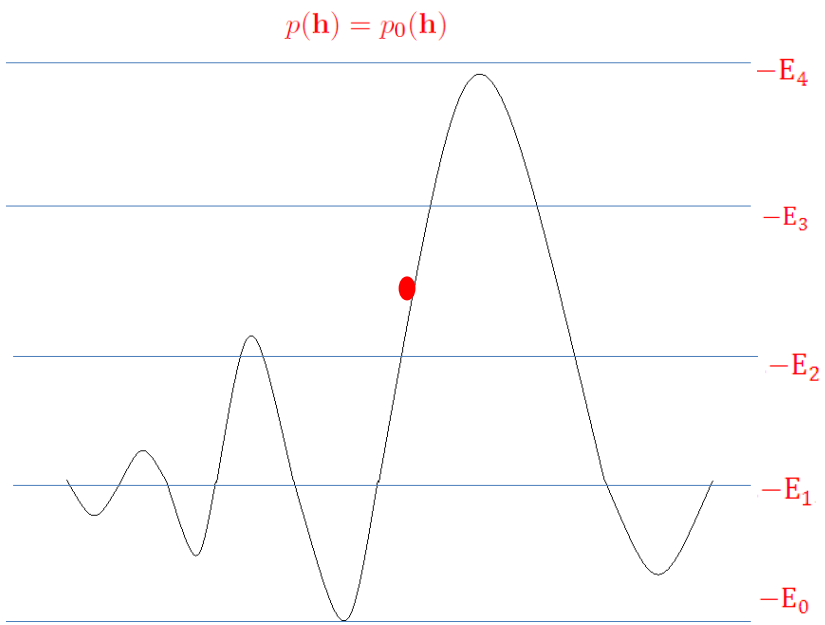
- The ideal is the perform sampling at different temperatures which make the distribution flat.



Equi-Energy (EE) Sampler

$$q(\mathbf{h}) = -log(p(\mathbf{h})) \in [E_k, E_{k+1})$$

$$E_0 < E_1 < E_2 < \cdots < E_K < E_{K+1} = \infty,$$

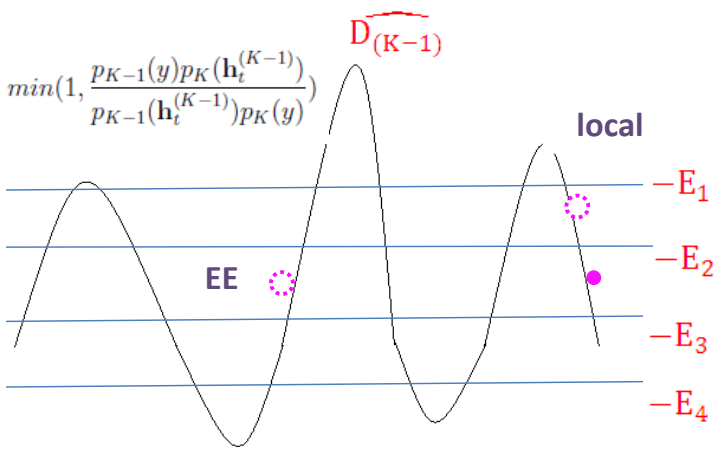


Hybird-GRE Sampler

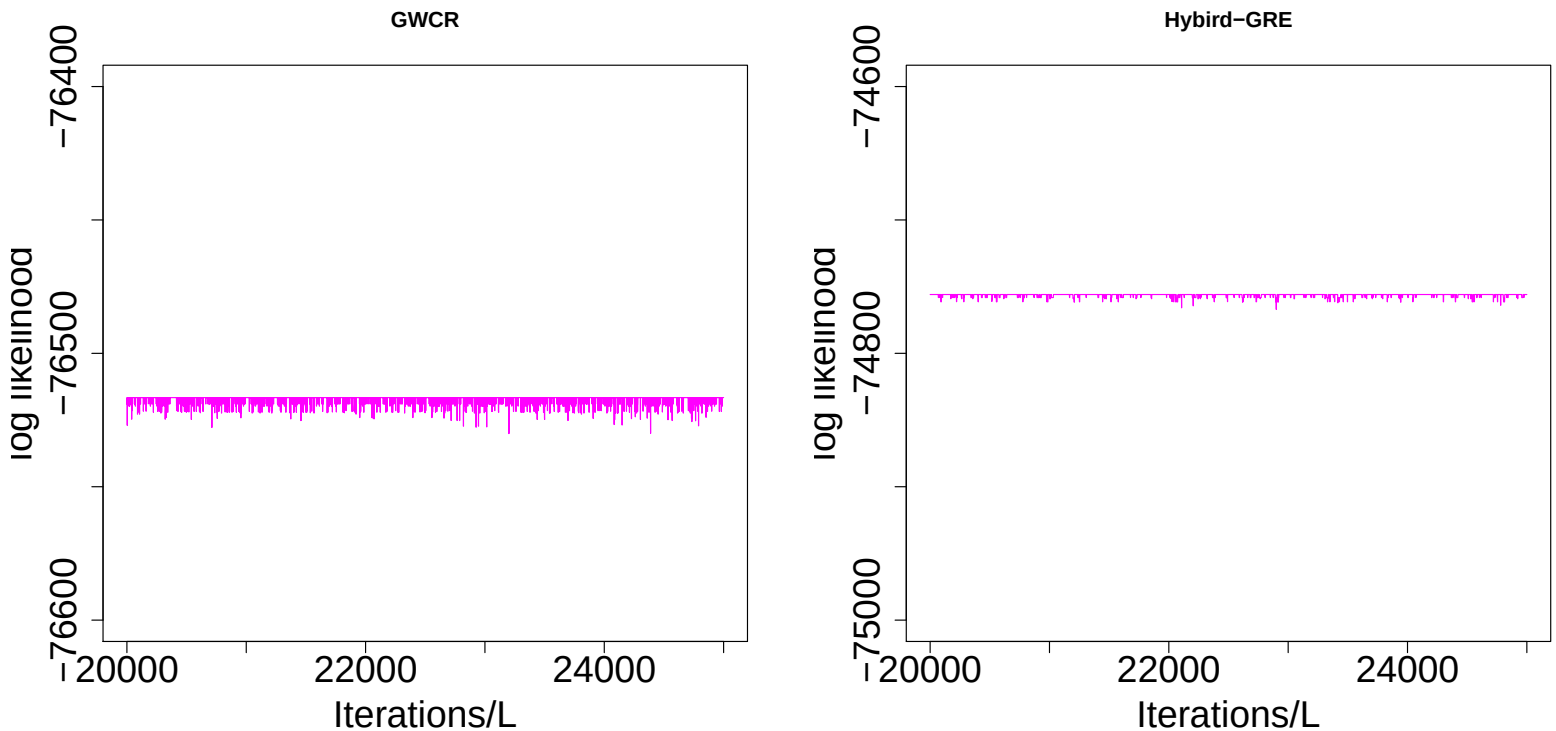
Hybird-GRE sampler consists of:

1. Global move: EE sampler.
2. Local move: GWCR(1)+RGSM(1).

- Chain H^K: only local move.
- Other chain: prob for the global move is increasing.



Result for Hybird-GRE sampler



Outline

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.