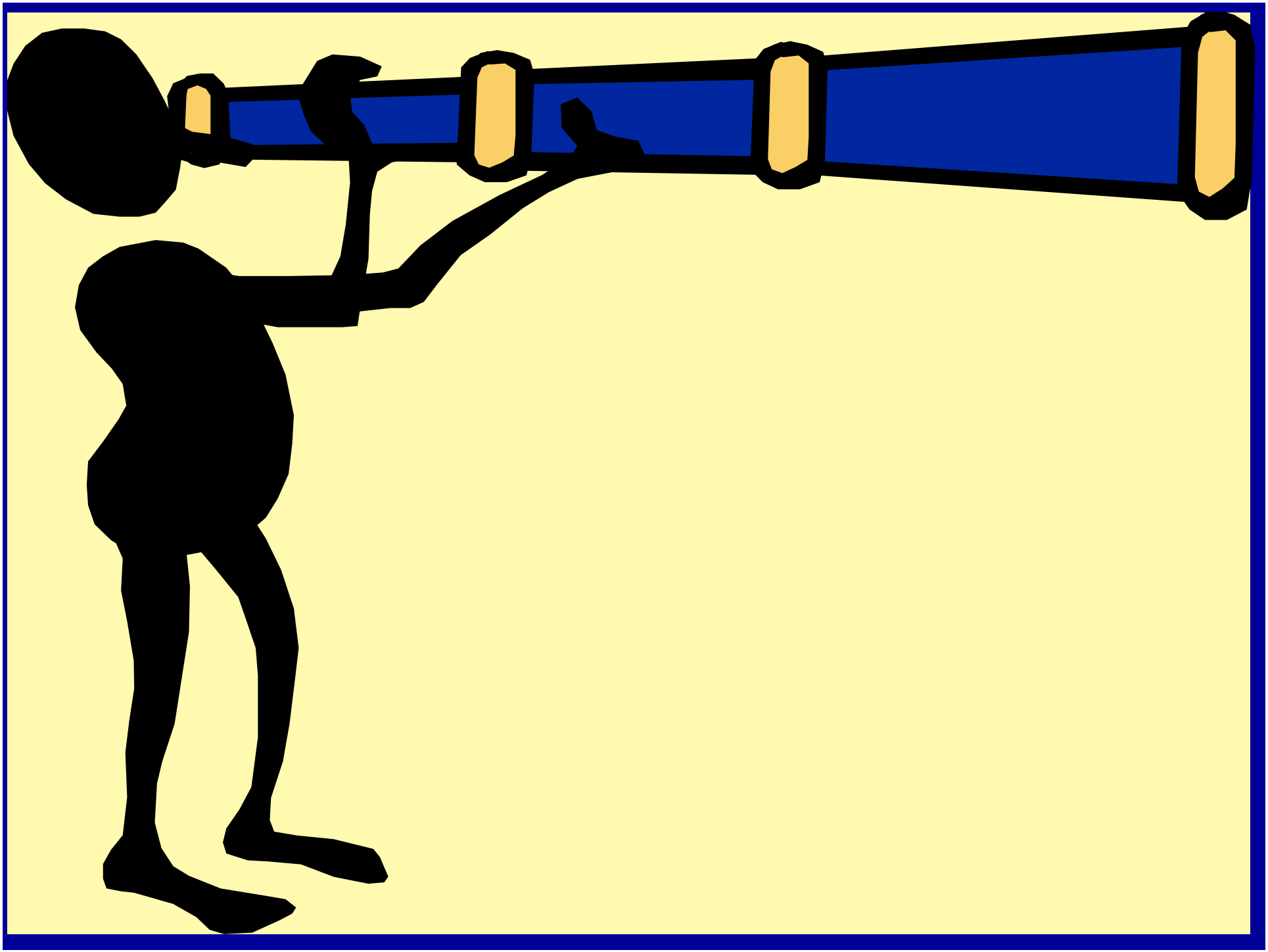
Multi-loci association test in genetic association study using similarity between individuals

Indranil Mukhopadhyay
Human Genetics Unit
Indian Statistical Institute, India


zzzz

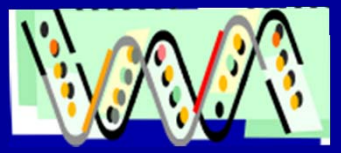


Acknowledgement

- Anbupalam Thalamuthu
- Eleanor Feingold
- Daniel Weeks
- Kushal Dey
- Pronoy Kanti Mondal
- Sarmistha Das

Outline …

- Some prelims
- Disease … Genetics … ??

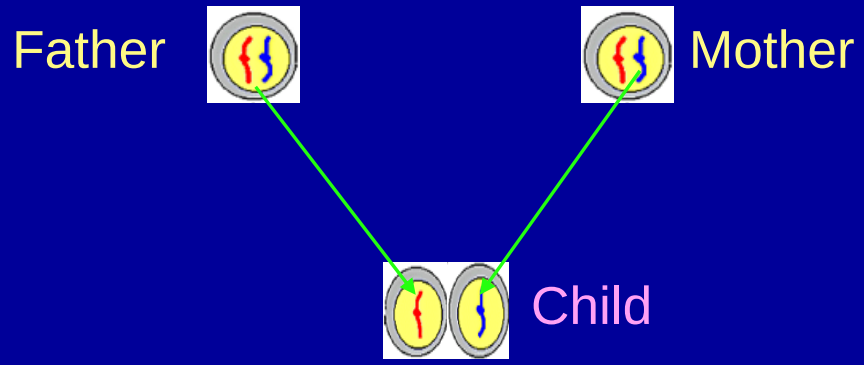


- Finding a disease gene

- A new test for multi-loci association

The Human Genome

- Human genome is *diploid*, meaning we have two copies of each chromosome (one from each parent)



- 22 pairs of chromosomes + 1 pair of sex chromosome

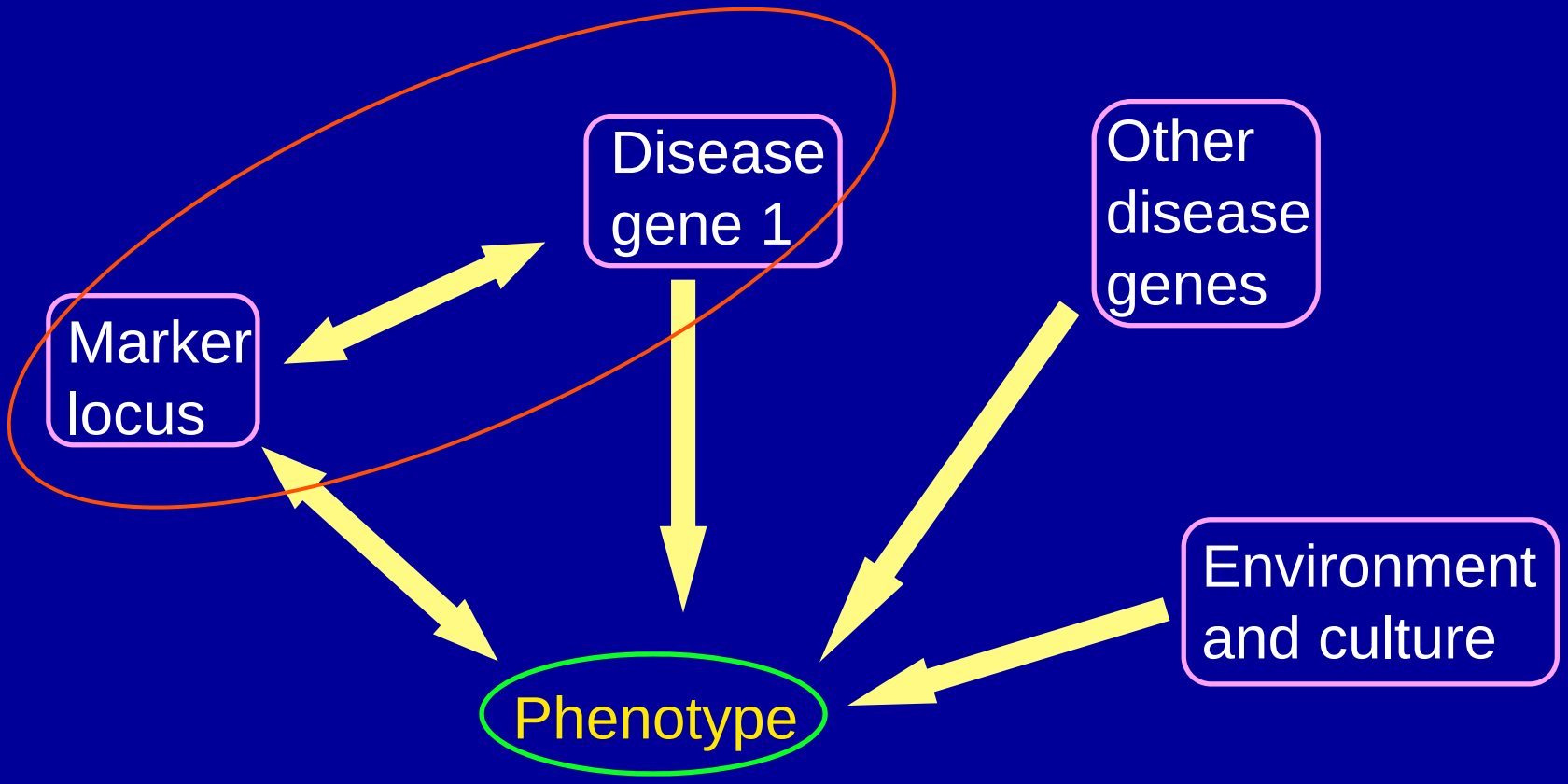
Prelims …

- Gene: Fundamental unit of genetic information that passes from generation to generation
- Allele: One of two or more states in which either copy of a gene can exist
- Marker: A polymorphic entity with known physical location

Genetic Markers

- Known location in genome
  - Human Genome Project tells us precisely where the markers are
- Unchanged from generation to generation
- Follow transmission from parents to offspring
- Be able to distinguish alleles
  - Polymorphic- having more than one state (alleles)

Complex disease



Slide by S Ghosh

SNP *Single Nucleotide Polymorphism*

So an SNP has only two alleles

Marker = SNP Alleles: *a* or *b*

Genotypes: *aa*, *ab*, *bb*

Association: A tendency for a particular genotype to occur more commonly in cases for a disease than expected by chance

Association testing: A testing method to test the possible existence of association between a phenotype and a candidate gene

Basic methods of association

Genotype-based Test

	aa	*ab*	*bb*	Total
Case	n_1	n_2	n_3	S
Control	N_1	N_2	N_3	T

Null hypothesis (H_0): no difference in the genotypic distributions of cases and controls.

$$\chi^2 = \sum_{all\ \ cells} \frac{(O-E)^2}{E}$$

An example

	aa	*ab*	*bb*	Total
Case	50	40	10	100
Control	130	60	10	200

H_0: no difference in genotypic distributions

- Observed frequencies are given
- Calculate expected frequencies under H_0

$$\frac{(50+130)}{(100+200)} \times 100 = 60$$

	aa	*ab*	*bb*	Total
Case	60	33	7	100
Control	120	66	14	200

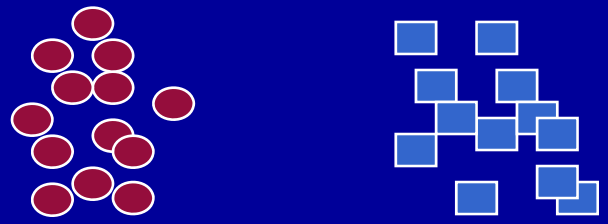
- Calculate chi-square statistic

$$\chi^2 = \sum_{all\ \ cells} \frac{(O-E)^2}{E} = 6.96 > \chi^2_{1,0.05} = 3.84$$

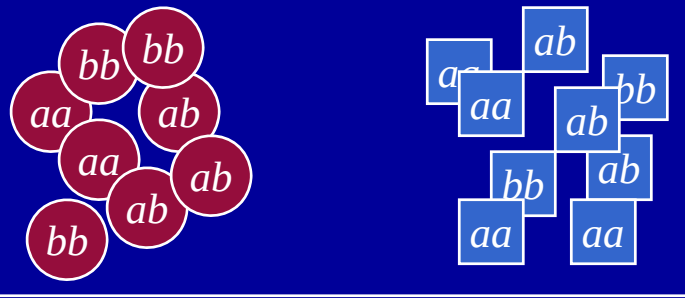
- P-value = 0.008 < 0.05

Genome-wide Association analysis (GWAS)

1) Collect cases and controls.



2) Genotype everyone at a marker.



3) Test genotype/phenotype association.

	aa	*ab*	*bb*
cases	50	40	10
controls	130	60	10

P-value = 0.008 : small enough !!!

4) Genotype everyone at all markers.

- Test at each locus
- Check P-value < 0.05
- Hurray! Found causal locus

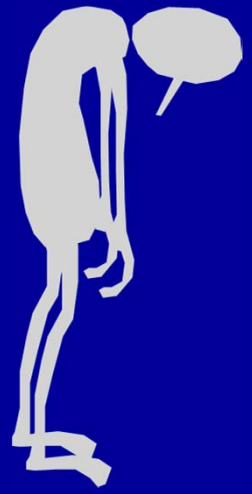
DONE!

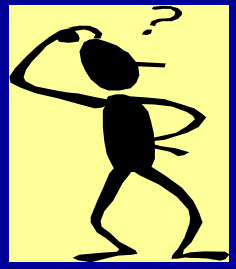
I have found one locus !!!

Write paper, have beer

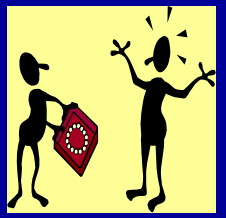
… have fun!

But this ‘world is not enough’



Why???  : let’s look carefully …

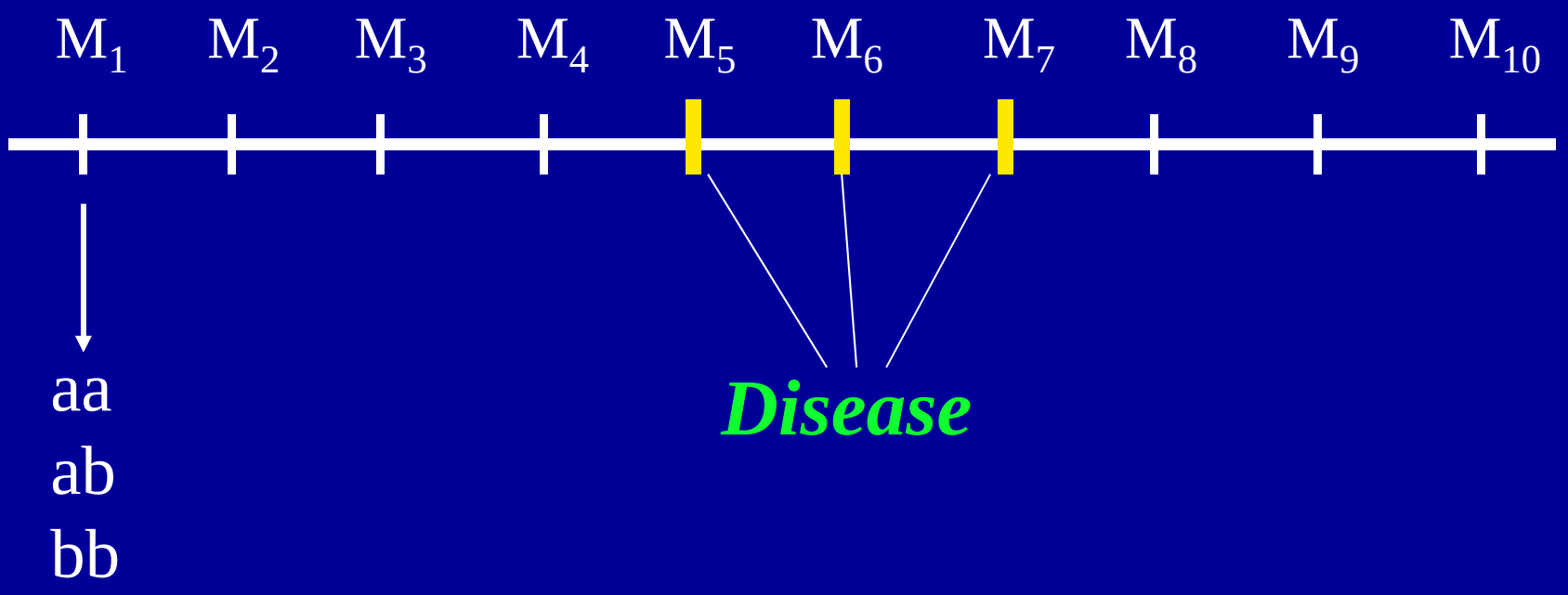
Simple, good, … but…



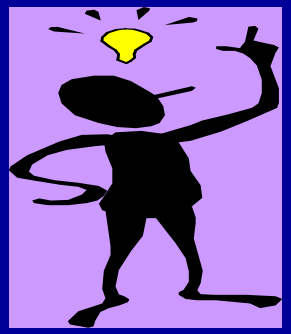
- Millions of SNPs
- Need for multiple comparison
- … … …

 May miss some true signals
 Need extremely large sample
 many other issues …

Let's give a fresh look …

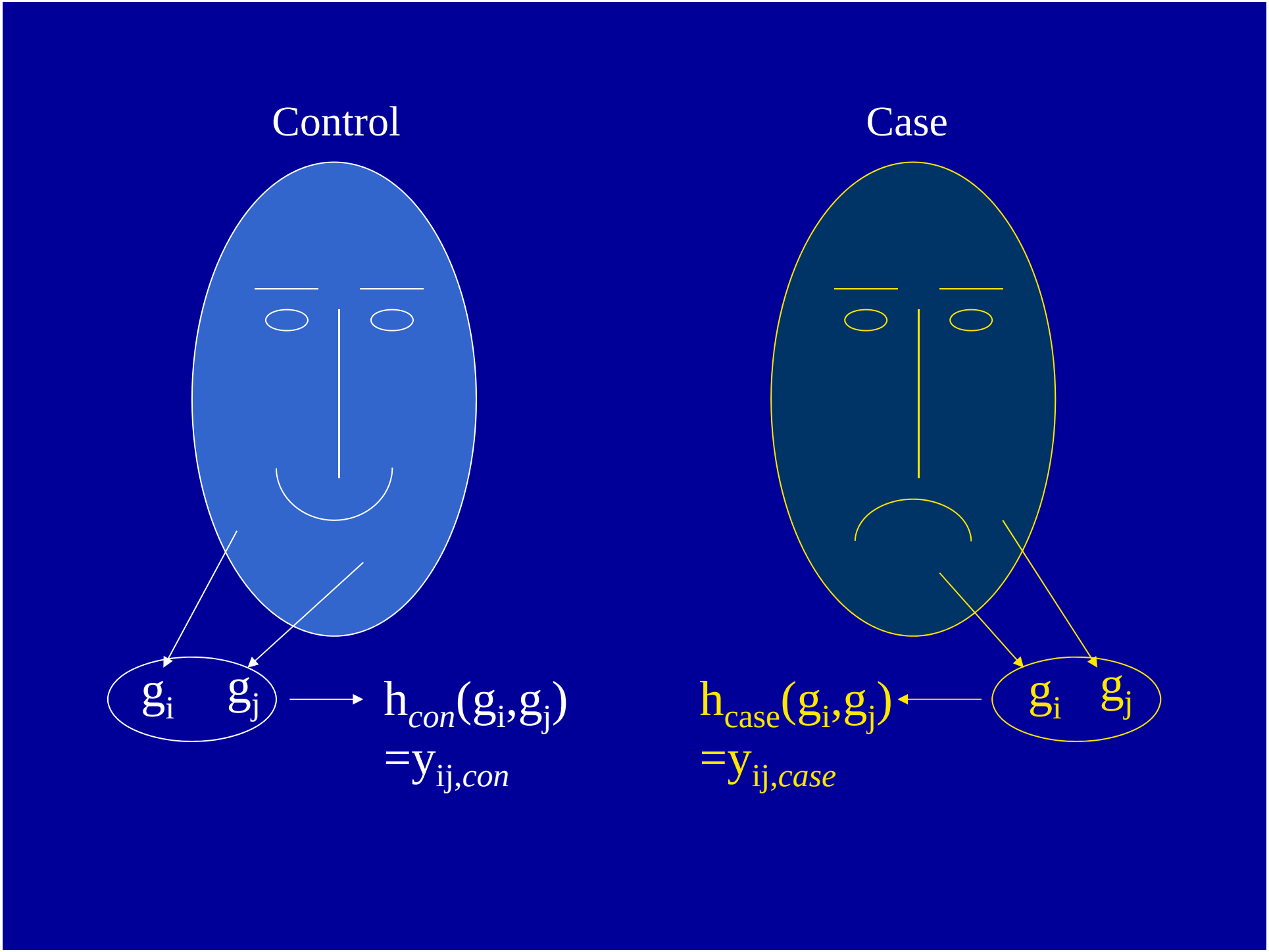


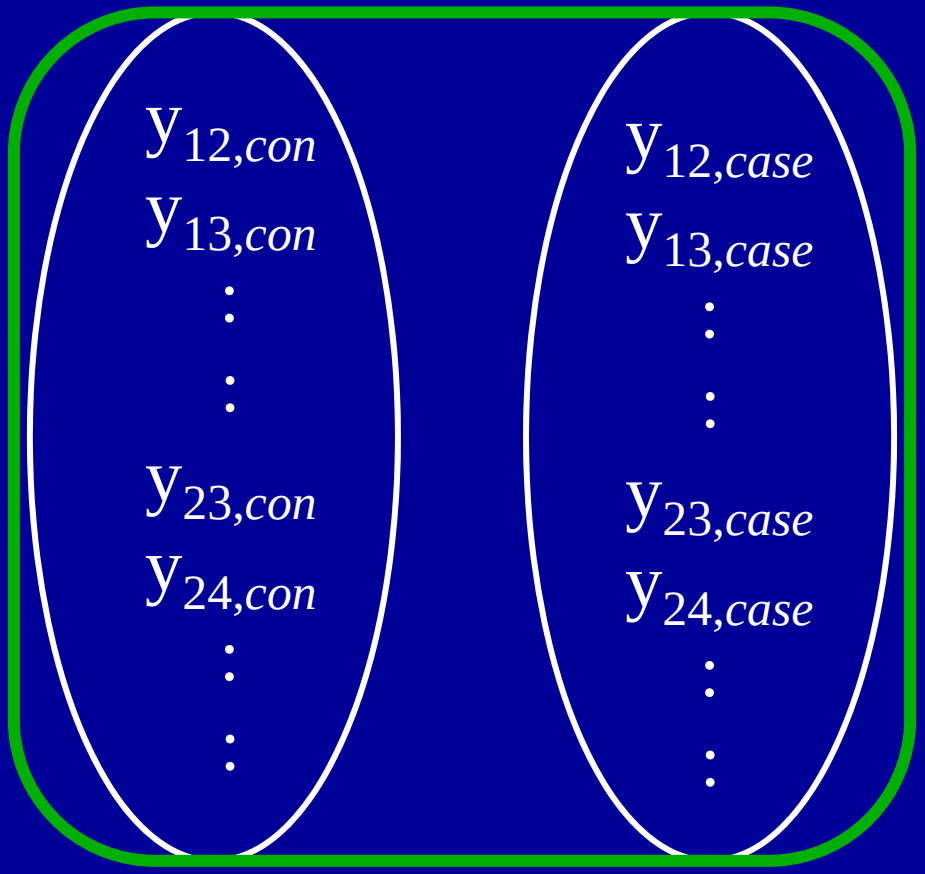
H_0: no association



Idea

- Individuals belonging to control group form a class, those having the disease (cases) form another class
- Use variation between cases and controls and variation within each class
- Similarity scores or values based on the genotype of each marker
- We study each marker separately and combine them to get a global statistic that is finally used to detect disease-marker association


Control
Case
g_i g_j
$h_{con}(g_i,g_j)$
$=y_{ij,con}$
$h_{case}(g_i,g_j)$
$=y_{ij,case}$
g_i g_j

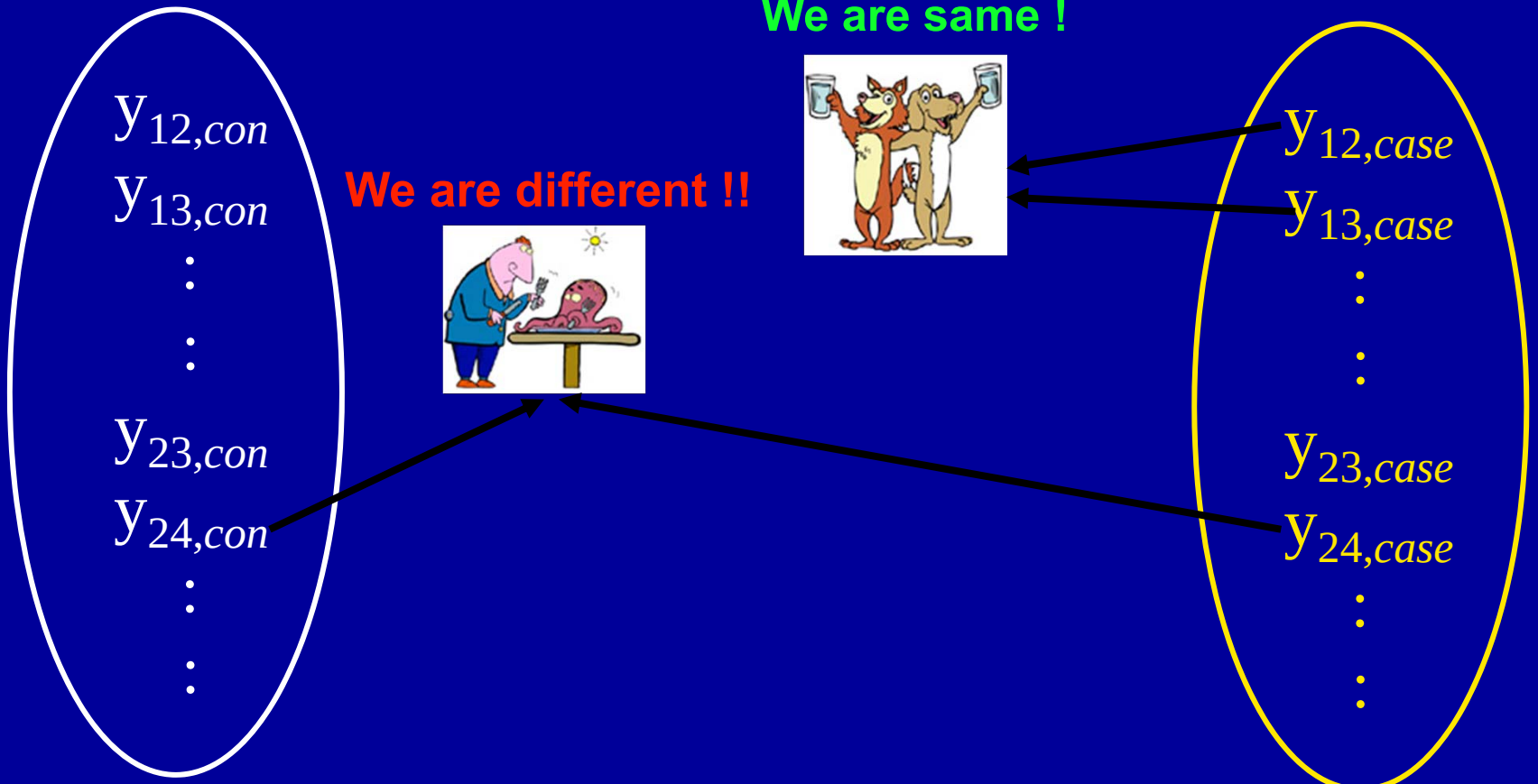


H_0: no association

&

H_0 is true

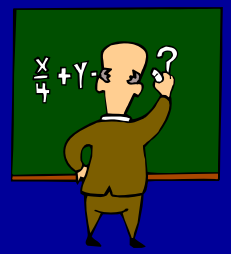
$$y_{lij} = \mu + e_{lij} \quad i < j = 1,2,\ldots,n_l; \;\; l = case, \;\; control$$



$$y_{lij} = \mu + \alpha_l + e_{lij} \quad i < j = 1,2,\ldots,n_l; \;\; l = case, \;\; control$$

→ additional effect over general effect

Model



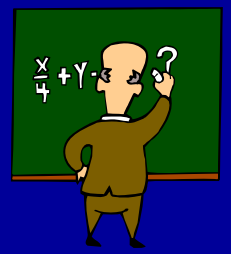
Let $y_{lij} = h_l(g_i, g_j)$ denote the kernel score between (i,j)-th pair in the l-th group

TABLE 1. Kernel scores corresponding to different choices of additive kernels associated with pair of genotypes g_i and g_j.

	Allele match			Allele share			Linear dosage			Recessive			Quadratic		
g_i / g_j	a/a	a/b	b/b	a/a	a/b	b/b	a/a	a/b	b/b	a/a	a/b	b/b	a/a	a/b	b/b
a/a	4	2	0	0	0	0	0	1	2	0	0	1	2	3	5
a/b	2	4	2	0	1	1	1	2	3	0	0	1	3	4	6
b/b	0	2	4	0	1	2	2	3	4	1	1	2	5	6	8

$y_{lij} = h_l(g_i, g_j)$: not uncorrelated

Model



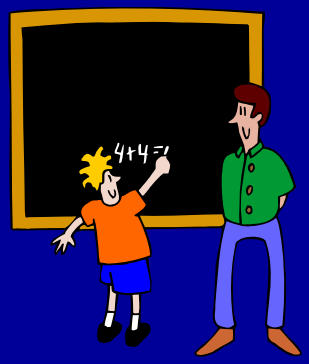
$$y_{lij} = \mu + \alpha_l + e_{lij} \quad i < j = 1,2,...,n_l; \quad l = 1,2$$

$$(i) \quad \alpha_1 + \alpha_2 = 0$$

$$(ii) \quad V(y_{lij}) = \sigma^2$$

$$(iii) \quad Cov(y_{lij}, y_{l'i'j'}) = \begin{cases} \rho\sigma^2 & for \;\; i \neq i' \;\; or \;\; j \neq j' \;\; if \;\; l = l' \\ 0 & if \;\; l \neq l' \end{cases}$$

$\{l=1\} \Rightarrow$ case, $\{l=2\} \Rightarrow$ control



- Consider *each* marker separately
- Combine them to get a statistic
- SSW_k = Within class variation
- SSB_k = Between class variation

$$y_{lij} = \mu + \alpha_l + e_{lij} \quad i < j = 1,2,...,n_l; \; l = 1,2$$

H_0: no association

⇕

H_0: $\alpha_{case} = \alpha_{con} = 0$

SSB_k / SSW_k : *for a single marker*

Test statistic :

$$\Im = \frac{\sum_{k=1}^{K} SSB_k}{\sum_{k=1}^{K} SSW_k}$$

- If observed $\Im$ is small we can think that H_0 is true
- If observed $\Im$ is large we can think that H_0 is not true

$$P(\Im > \Im_\gamma | H_0) = \gamma = P(Type \;\; I \;\; error)$$

$$P - value = P(\Im > Obsd.\Im | H_0)$$

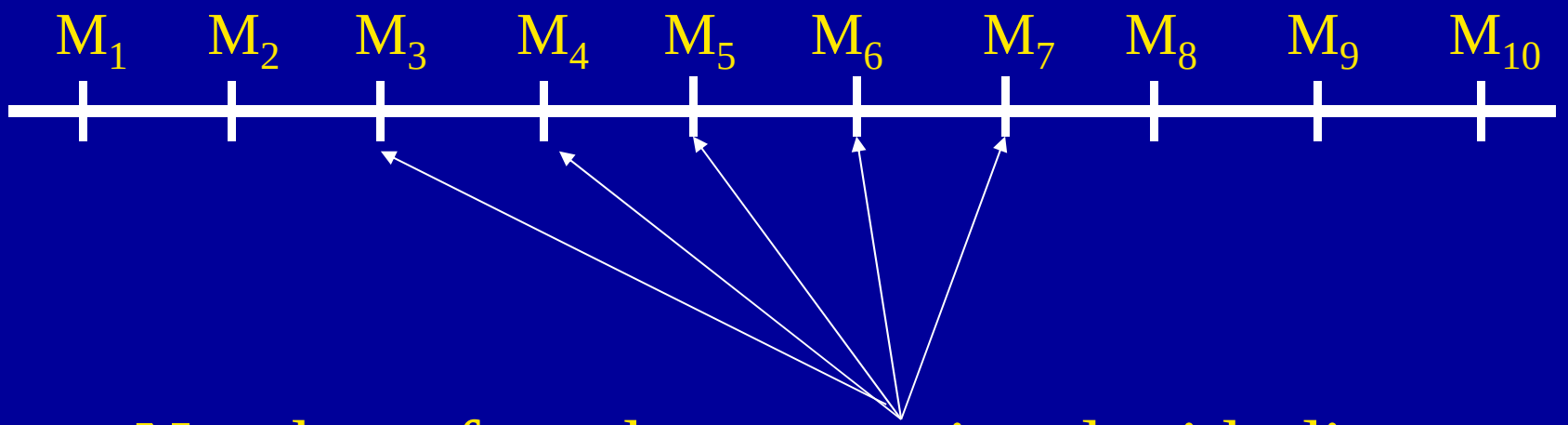
$$P(\Im > \Im_\gamma | H_0) = \gamma = P(Type\ \ I\ \ error)$$

$$Power = P(\Im > \Im_\gamma | H_1)$$

- The test is one-sided to the right
- The distribution of the test statistic is not known
- We calculate Power by simulation/permutation

Simulation

- Genotypes of 10 independent markers

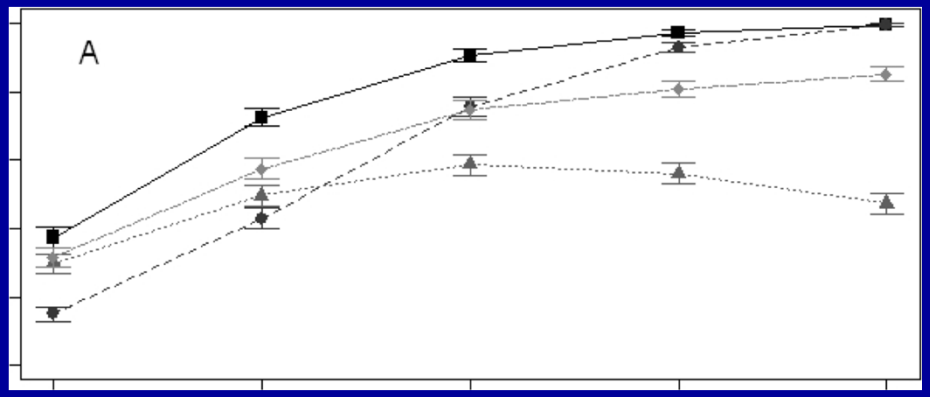


- Number of markers associated with disease ranges from 1 to 5

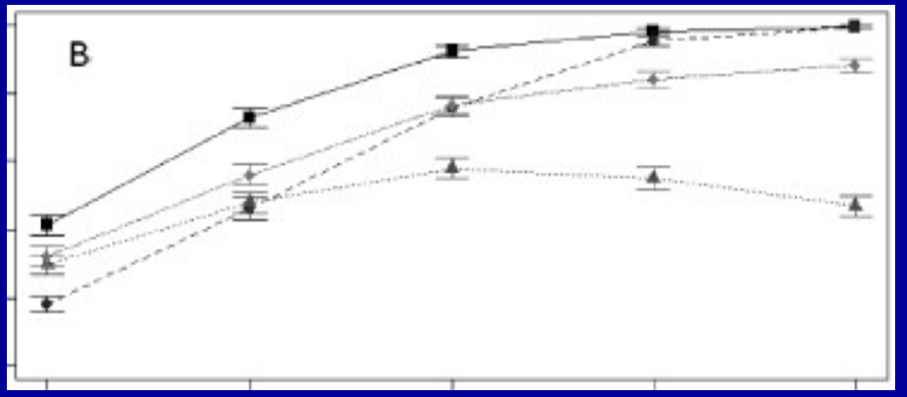
- High-risk allele frequency is 0.05
- Relative risk is 1.5 and assume multiplicative model
- Sample size for each group is 500
- $\mathfrak{I}_{\gamma}$ is calculated based on 10000 simulations
- Power is calculated based on 1000 simulations

POWER STUDY

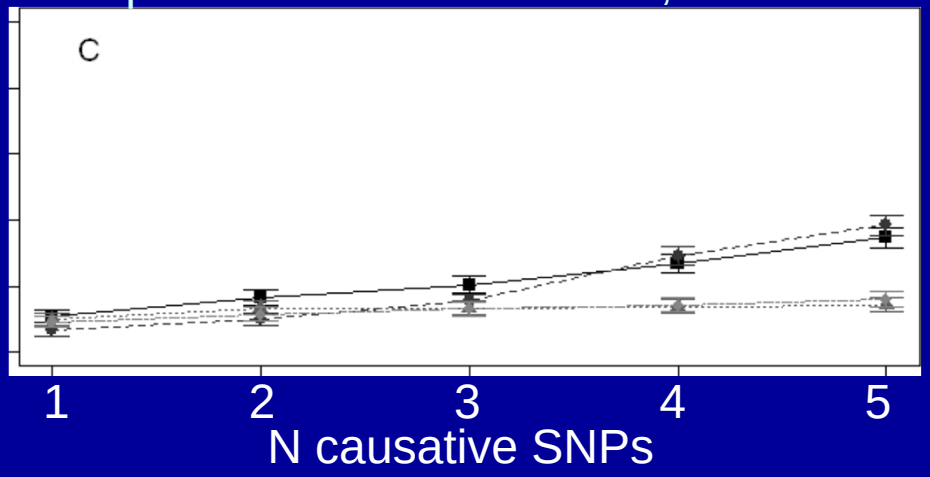
Additive model: RR=1.25, MAF=0.05



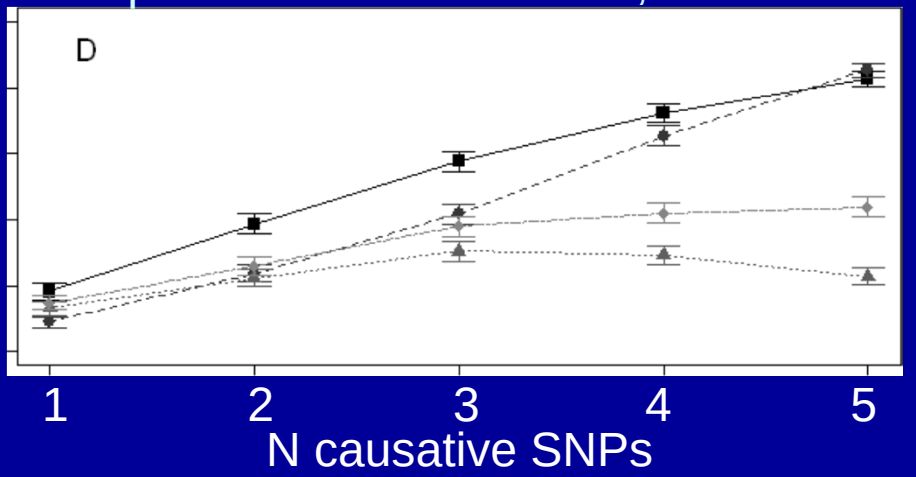
Additive model: RR=1.5, MAF=0.05



Multiplicative model: RR=1.25, MAF=0.05



Multiplicative model: RR=1.5, MAF=0.05



KBAT (Mukhopadhyay et al (2010))

Zglobal (Schaid et al (2005))

MDMR (Wessel & Schork (2006))

MDMR+ (Modified MDMR)

→ Other competitive tests

Asymptotic distribution of KBAT statistic

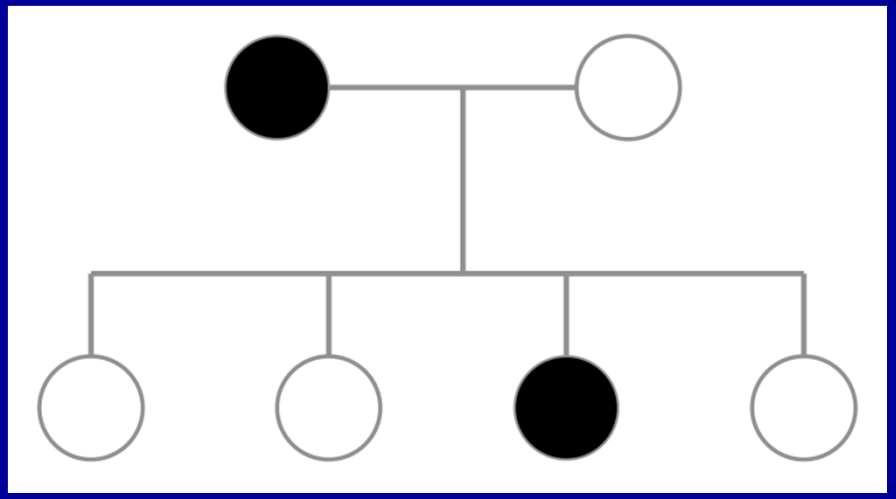
$$T = \beta(n_1, n_2)\frac{K(1+\nu^2)}{2\nu(1+\nu)}\frac{\sum_{k=1}^{K} SSB_k / \hat{\sigma}_{1k}^2}{\sum_{k=1}^{K} SSW_k / \hat{\sigma}_k^2} \xrightarrow{L} \chi_K^2 \ as \ (n_1, n_2) \to \infty$$

$$where \quad \beta(n_1, n_2) = \frac{n_1(n_1-1)+n_2(n_2-1)}{2n_1}$$

Family based KBAT

Notations

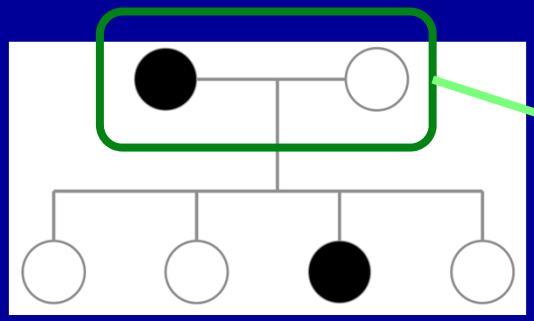
- SNP marker: *aa*, *ab*, *bb*
- No. of markers in a gene: *L*
- Phenotype: qualitative – affected or unaffected
- Nuclear families with at least one affected sib
- No. of families: *n*



Towards test statistic…

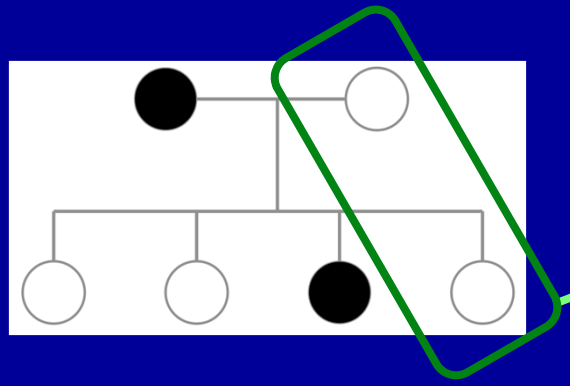
Consider *l*-th locus, *r*-th family

①



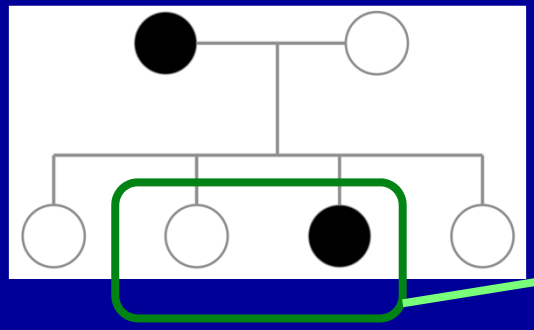
$$h_r(g^l_{P1}, g^l_{P2})$$

②



$$\frac{1}{2n_r}\sum_{j=1}^{n_r} h_r(g^l_{P1}, g^l_{S_j}) + \frac{1}{2n_r}\sum_{j=1}^{n_r} h_r(g^l_{P2}, g^l_{S_j})$$

③



$$\frac{2}{n_r(n_r-1)}\sum_{i<j} h_r(g^l_{S_i}, g^l_{S_j})$$

Towards test statistic…

- Propose a 3-dimensional statistic using three statistics:

$$U_{rl} = \Sigma_{rl}^{-\frac{1}{2}}(T_{rl} - \mu_l)$$

where $T_{rl} = (T_{1,rl}, T_{2,rl}, T_{3,rl})'$ and Σ_{rl} is the var-cov matrix of T_{rl}; $r = 1,\ldots,n$; $l = 1,\ldots,L$.

- Combine genetic information from *L* loci at a time for all *n* families to get the final statistic:

Kernel based association test for family data

F-KBAT: $U_n = \overline{\hat{U}}_n' \overline{\hat{U}}_n$

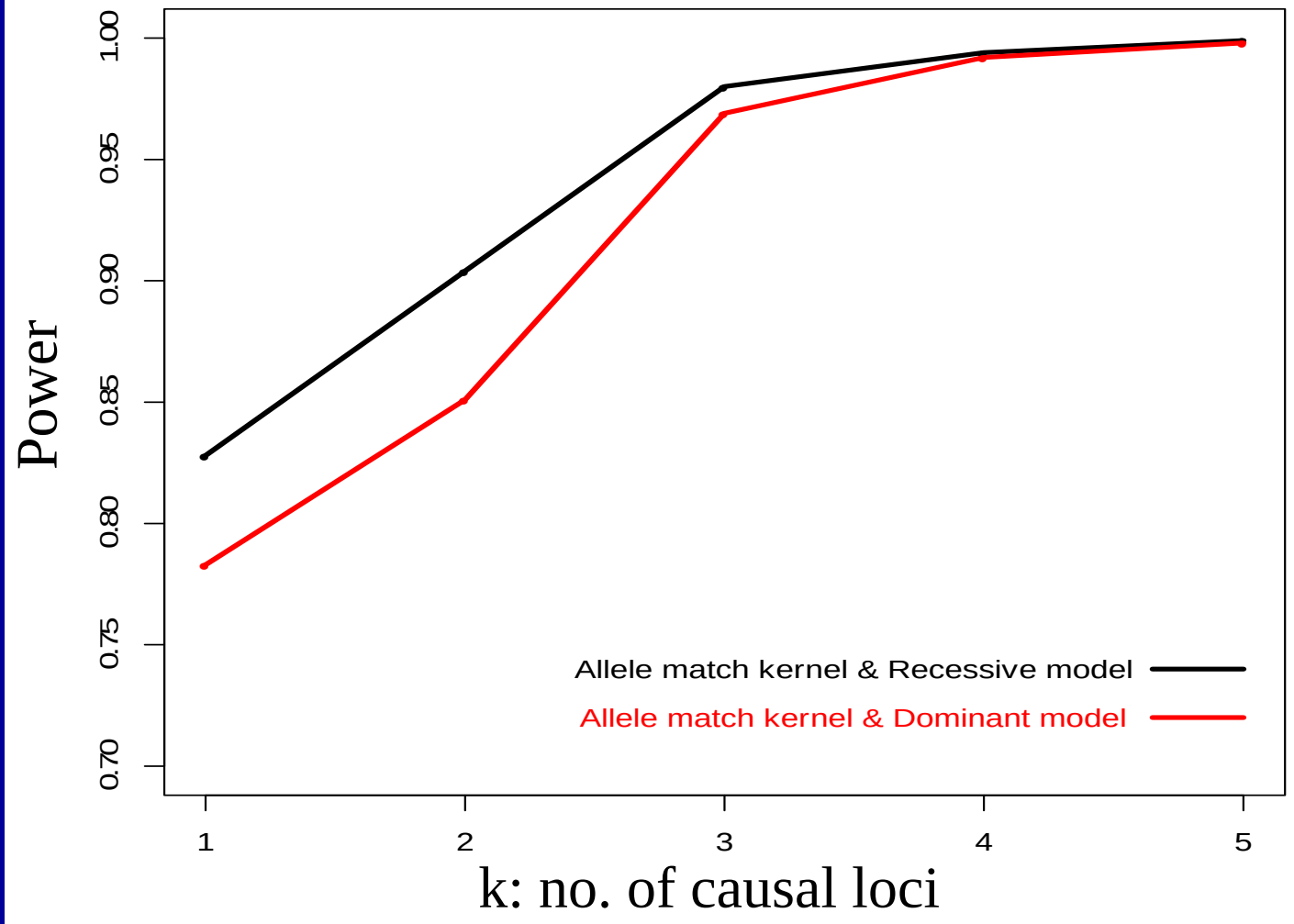
Theorem : *Let* $\overline{\hat{U}}_n$ *be the mean of all estimated scaled score vectors* $\hat{U}_{rl}$ *over all families and for all* l*, replace* μ_l *and* Σ_{rl} *by their consistent estimators. Assume* $\forall r \ \forall l$*,* $j = (1,1,1)'$, $\left\| \Sigma_{rl}^{-\frac{1}{2}} j \right\| \leq M < \infty$. *Then under* H_0 *(no assoc.),*

$$Ln\overline{\hat{U}}' \overline{\hat{U}} \xrightarrow{d} \chi_3^2 \ \text{ as } n \to \infty.$$

Simulation

- 10 SNPs; causal markers k=1,2,3,4,5
- MAF = 0.1+i/100, i=1,2,…,10
- Genetic model: recessive, dominant
- No. of sibs per family (X) ~ Poisson(3|X>1)
- n = 200 families
- Average p-value over 1000 simulations
- Disease model:
  - Model 1: affected if at least one of *k* causal loci has risk genotype
  - Model 2: affected if all *k* causal loci have risk genotypes

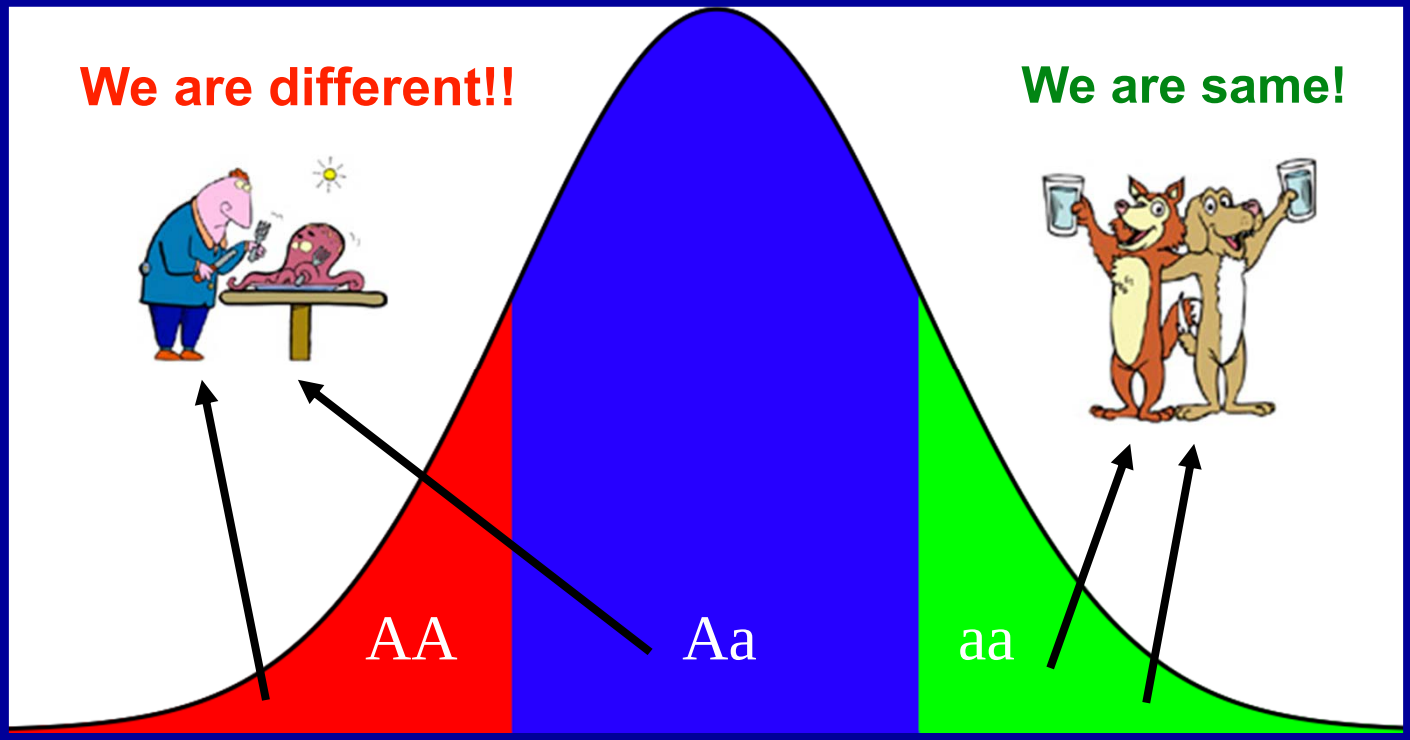
Power against no. of causal loci



Qt-KBAT

QT-KBAT: using quantitative trait

- Phenotype similarity ⇔ Genotype similarity
- People who have similar phenotype trait values should have higher sharing of genetic material near the genes that influence those traits



But are we genetically same (with respect to trait)??

MODEL

Phenotype similarity: $P_{ij} = | z_i - z_j |$

Genotype similarity: 3 possible groups based on 3 possible similarity values

$$G_1 = \{(g_i, g_i) : g_i = \text{a/a, a/b \& b/b}\}$$

$$G_2 = \left\{ \begin{array}{r} (g_i, g_j) : [g_i = \text{a/a} \,\&\, \text{g}_\text{j} = \text{a/b}] \\ \text{or } [g_i = \text{a/b} \,\&\, \text{g}_\text{j} = \text{b/b}] \end{array} \right\}$$

$$G_3 = \{(g_i, g_j) : g_i = \text{a/a and } \text{g}_\text{j} = \text{b/b}\}$$

Total Number of markers: K

Model

$$P_{l(ij)} = \mu + \beta_l G_{l(ij)} + e_{l(ij)}; i < j = 1,\ldots,n; l = 1,\ldots,K$$

$(i)\quad V(e_{l(ij)}) = \sigma^2$

(ii) *Errors* $(e_{l(ij)})$ *are correlated*

(iii) *Errors are not Normally distributed*

Test Statistic

$$\Im = \sum_{l=1}^{K} \Im_l \quad where \quad \Im_l = \frac{SSE_{\beta_l=0} - SSE}{SSE}$$

Asymptotic distribution of Qt-KBAT statistic

$$\Im = \sum_{k=1}^{K} \Im_k \xrightarrow{L} \sum_{k=1}^{K} w_k \chi_1^2 \; as \; n \to \infty$$

Conclusion, Future & ongoing works

- Our method is generally more powerful
- Significance may be determined by permutation
- Asymptotic distn helps in computing p-value fast
- Choice / effects of kernels and models
- Asymptotic distn when markers are not independent

Conclusion, Future & ongoing works

- KBAT for case-control data & Qt-KBAT for quantitative phenotype
- KBAT for family data
- Develop gene-gene interaction test
- Develop gene-environment test
- Asymptotic distns in all above cases …


HSBC
THE FULLERTON HOTEL
THANK YOU