Likelihood-based method for estimating penetrance and disease allele frequency

Yuki Sugaya and Ritei Shibata

Keio University, Japan

Topic in Japan

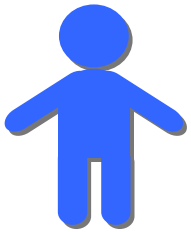
BUNGEISYA CO., LTD.

- 5 millions seller
- Blood type character analysis
- ¥1,050 yen (in tax)

Blood type character analysis in Japan

- Type A
  - polite, cooperative, responsible, oversensitive, and somewhat uptight
- Type B
  - optimistic, social, and easygoing to the point of occasional laziness
- Type O
  - active, strong-willed, friendly, and blunt
- Type AB
  - moody, eccentric, withdrawn, and prone to bouts of brilliance and/or insanity

Blood type

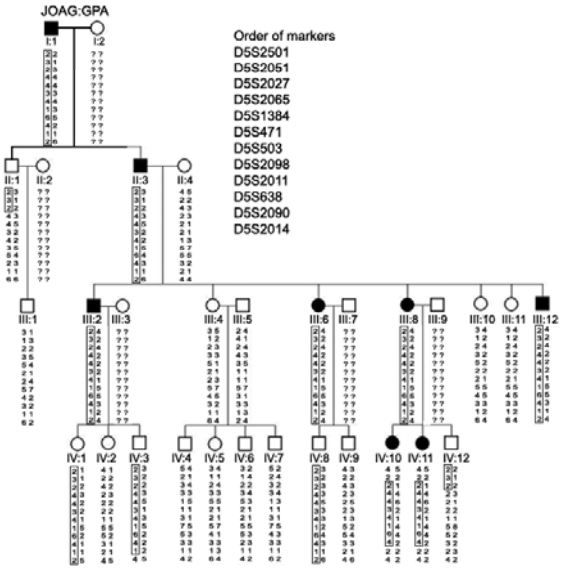
phenotype	Type A	Type B	Type O	Type AB
				
genotype	$A/A \;\; A/O$	$B/B \;\; B/O$	O/O	A/B

$$P(TypeA \mid A/A) = 1$$
$$P(TypeB \mid A/A) = 0$$
$$\vdots$$

penetrance

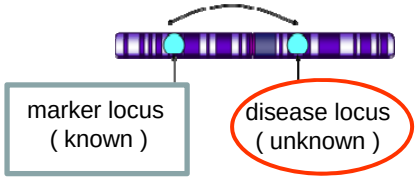
Linkage analysis

pedigree data

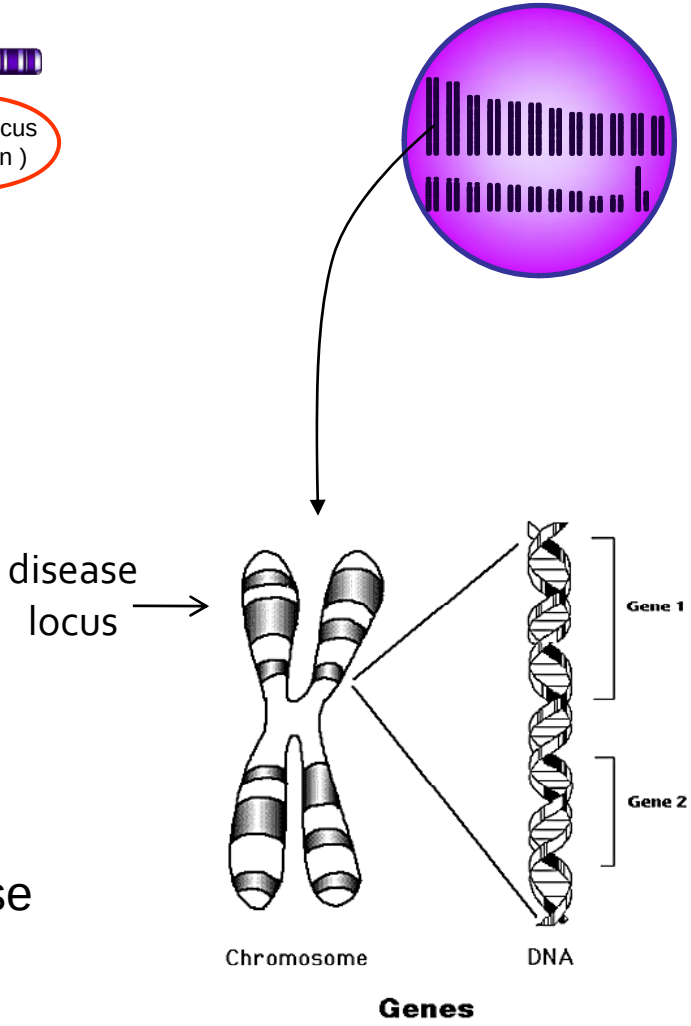


(PANG et al., 2006)

linkage



where?

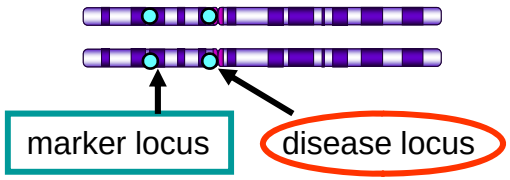


Objective

To detect a gene associated with a disease from pedigree data

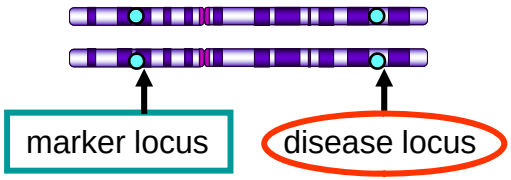
Recombination fraction

- Two loci are close to

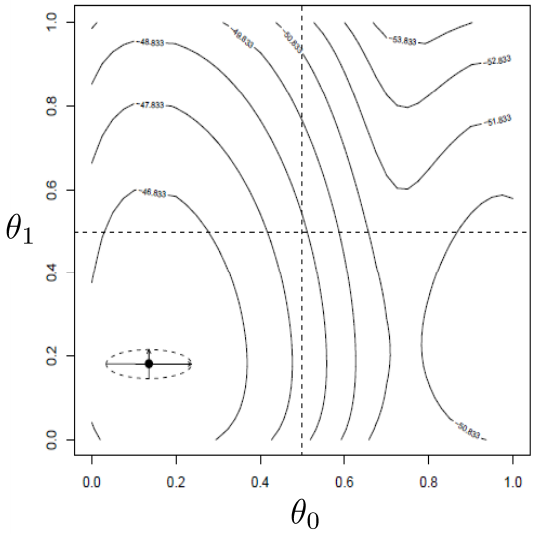


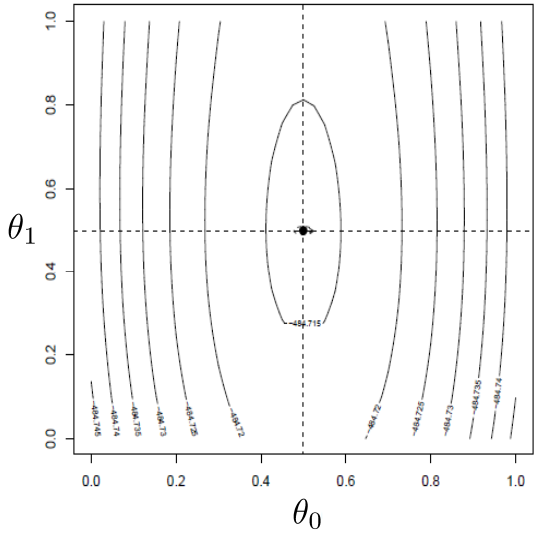
$\theta < 0.5$

- Two loci are far from



$\theta \approx 0.5$





We detect the disease gene through recombination fraction.

Assumption

- The disease is caused by a pair of disease susceptibility allele A

a : normal allele

genotype A/A A/a a/a

Penetrance

- Conditional probability of observing the corresponding phenotype (affected status) given the specified genotype

A/A

$$P(Affected \mid A/A) = \alpha$$

A/a

$$P(Affected \mid A/a) = \beta$$

a/a

$$P(Affected \mid a/a) = \gamma$$

$$\alpha > \beta > \gamma$$

Normal linkage analysis

- Penetrance
  - $\alpha = 1,\ \beta = 1,\ \gamma = 0$ $\Rightarrow$ dominant model
  - $\alpha = 1,\ \beta = 0,\ \gamma = 0$ $\Rightarrow$ recessive model
  - $\vdots$
- Disease susceptibility allele frequency f_A

$$f_A = 0.0001,\ 0.001,\ 0.1 \cdots$$

There are a lot of choices of parameters.
Different assumption may lead us to different result.

Previous works

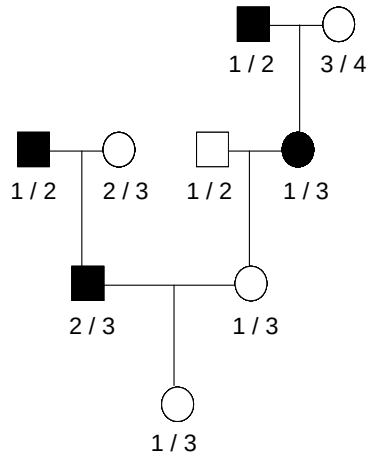
- Swart et al. (2004)
  - penetrance and allele frequency
  - $\alpha = \beta$ or $\beta = \gamma$

- Wang et al. (2006)
  - penetrance
  - nuclear family

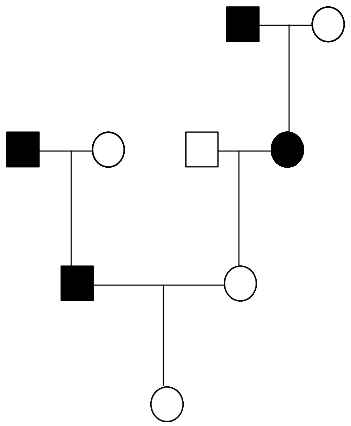
Likelihood of pedigree data

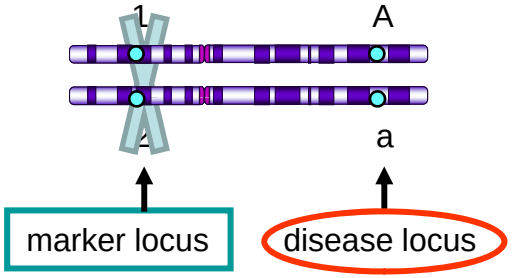
linkage analysis

$P(\mathscr{A}(V) \cap \mathscr{M}(V))$



$P(\mathscr{A}(V))$





A_v : affected status

M_v : marker genotypes

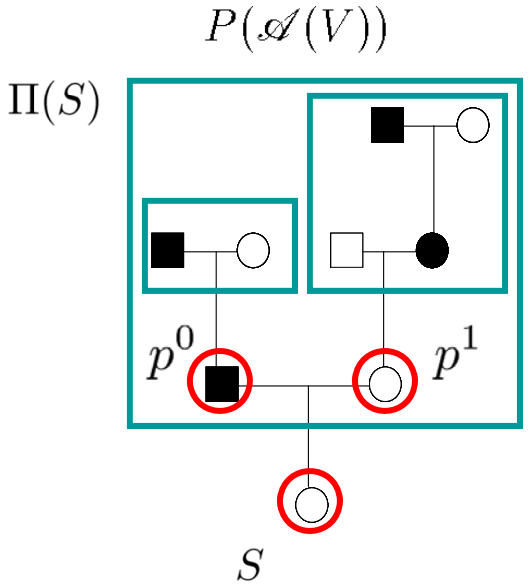
marker information is not necessary

Calculation of the likelihood (1)

$$\xi_S(d) = \sum_{(d_0,d_1)} \underbrace{\rho_p((d_0,d_1))}_{\text{penetrance}} \underbrace{P(D_S = d \mid D_p = (d_0,d_1))}_{\text{inheritance probability}} \xi_{p^0}(d_0)\xi_{p^1}(d_1)$$

$$\begin{pmatrix} \alpha, & 1-\alpha, \\ \beta, & 1-\beta, \\ \gamma, & 1-\gamma \end{pmatrix}$$

$$(0, \tfrac{1}{2}, 1)$$



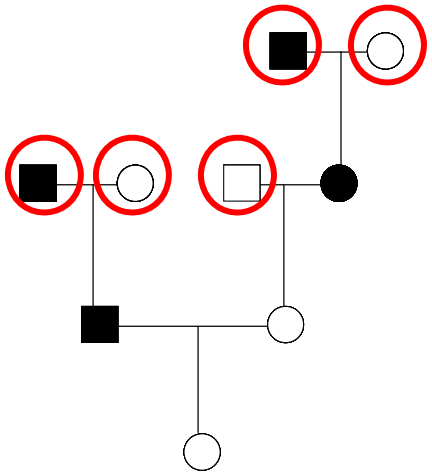
$$\xi_U(d) = P(\{D_U = d\} \cap \mathscr{A}(\Pi(U)))$$

$$\rho_U(d) = P(\mathscr{A}(U) \mid D_U = d)$$

$\Pi(U)$: ancesters of U

D_U : genotypes $(u \in U)$

Calculation of the likelihood (2)



v : founder

$$\xi_v(d) = P(\{D_v = d\} \cap \underline{\mathscr{A}(\Pi(v))})$$

empty set

$$\xi_v(d) = P(\{D_v = d\})$$
$$= \underline{P(\{D_v^0 = d_0\})} \underline{P(\{D_v^1 = d_1\})}$$

disease susceptibility allele frequency

$$(f_A,\ 1 - f_A)$$

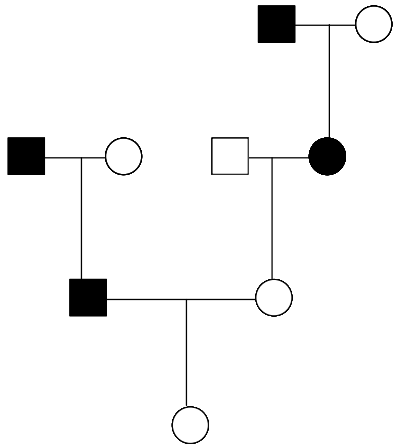
Parameter estimation

$$P(\mathscr{A}(V)) = L(\alpha, \beta, \gamma, f_A) \xrightarrow[\alpha, \beta, \gamma, f_A]{} \max$$

$$L(\alpha, \beta, \gamma, f_A) = \sum c_{ijkl} \alpha^i \beta^j \gamma^k f_A^l$$

$max(i + j + k)$: informative individuals

$max(l)$: 2 × founders



Optimization problem

maximize $L(\boldsymbol{x})$ $\qquad \boldsymbol{x} = (\alpha, \beta, \gamma, f_A)$

subject to $\gamma \leq \beta \leq \alpha,\ 0 \leq \alpha \leq 1,\ 0 \leq \beta \leq 1,$
$0 \leq \gamma \leq 1$ and $0 \leq f_A \leq 1$

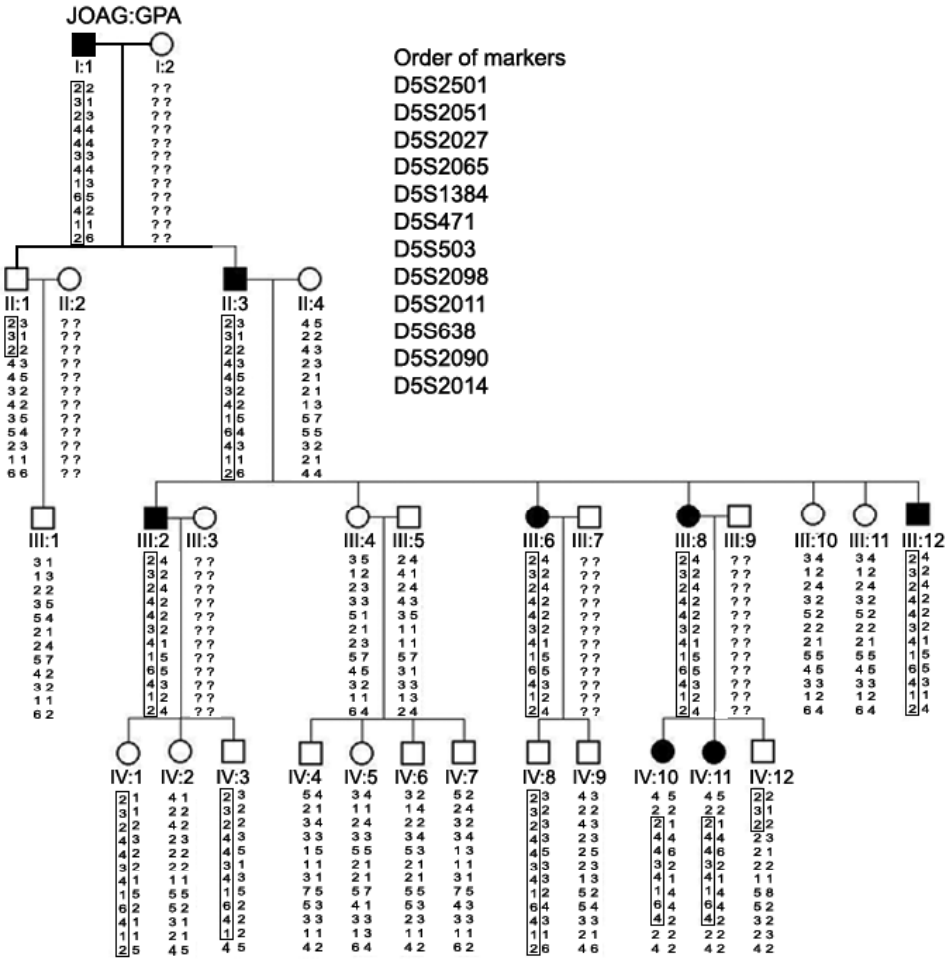
$$\Leftrightarrow \begin{bmatrix} \alpha - \beta \geq 0 : g_1(\boldsymbol{x}) \\ \beta - \gamma \geq 0 : g_2(\boldsymbol{x}) \\ \alpha \geq 0 : g_3(\boldsymbol{x}) \\ \beta \geq 0 : g_4(\boldsymbol{x}) \\ \gamma \geq 0 : g_5(\boldsymbol{x}) \\ f_A \geq 0 : g_6(\boldsymbol{x}) \\ \alpha \geq 0 : g_7(\boldsymbol{x}) \\ \beta \geq 0 : g_8(\boldsymbol{x}) \\ \gamma \geq 0 : g_9(\boldsymbol{x}) \\ 1 - f_A \geq 0 : g_{10}(\boldsymbol{x}) \end{bmatrix}$$

POAG pedigree data (PANG et al., 2006)

Primary Open Angle Glaucoma

30 pedigree members

- 8 affected individuals
- 22 non-affected individuals



PANG et al. assume

$$(\alpha,\ \beta,\ \gamma) = (1,\ 1,\ 0)$$

(dominant model)

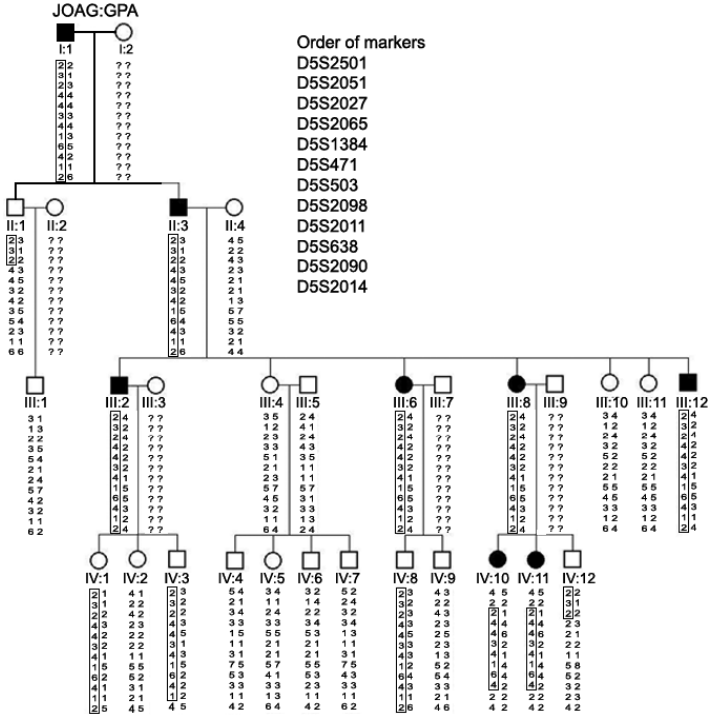
$$f_A = 0.0001$$

Maximum likelihood estimates

Kuhn-Tucker Conditions

$\hat{\boldsymbol{x}}$: local maximum

$$\Rightarrow \quad \exists \mu_i \leq 0 \qquad (i = 1, \cdots, 10)$$
$$\nabla L(\hat{\boldsymbol{x}}) + \sum_{i=1}^{10} \mu_i \nabla g_i(\hat{\boldsymbol{x}}) = \mathbf{0}$$
$$\mu_i g_i(\hat{\boldsymbol{x}}) = \mathbf{0} \qquad (i = 1, \cdots, 10)$$



$$\hat{\boldsymbol{x}} = (0.921,\ 0.921,\ 0,\ 0.065)$$

$$\mu_1 = -1.510 \times 10^{-2}$$
$$\mu_5 = -7.351$$
$$\mu_2 = \mu_3 = \mu_4 = \mu_6 = \mu_7 = \mu_8 = \mu_9 = \mu_{10} = 0$$

$\hat{\boldsymbol{x}}$ satisfy Kuhn-Tucker condition

$$\nabla L(\hat{\boldsymbol{x}}) = (-1.510 \times 10^{-2},\ 1.510 \times 10^{-2},\ -7.351,\ 7.064 \times 10^{-5})$$

Probability inheritance algorithm (Sugaya and Shibata)

- Probability inheritance algorithm
  - polynomial form

$$L(\boldsymbol{\theta}) = \sum_{i,j=0}^{k} \alpha_{ij} \theta_0^i \theta_1^j$$

$\boldsymbol{\theta} = (\theta_0, \theta_1)$
θ_0 : paternal recombination fraction
θ_1 : maternal recombination fraction

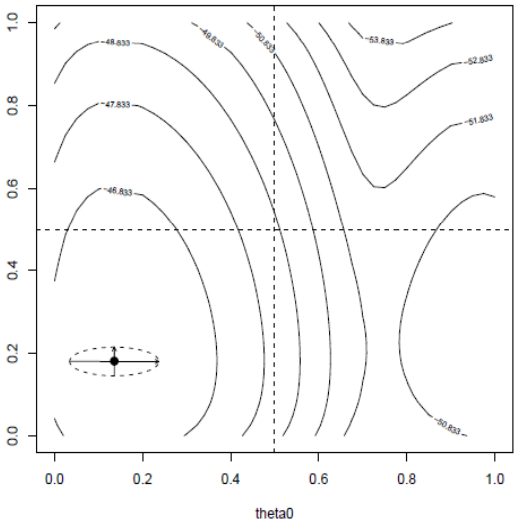
k : non-founder

- Elston-Stewart algorithm（1971）
  - backward
  - large pedigree，small number of markers
  - Linkage package
- Lander-Green algorithm（1987）
  - inheritance vector
  - hidden markov model
  - small pedigree，large number of markers
  - Genehunter

$\theta_0 \neq \theta_1$

×polynomial form

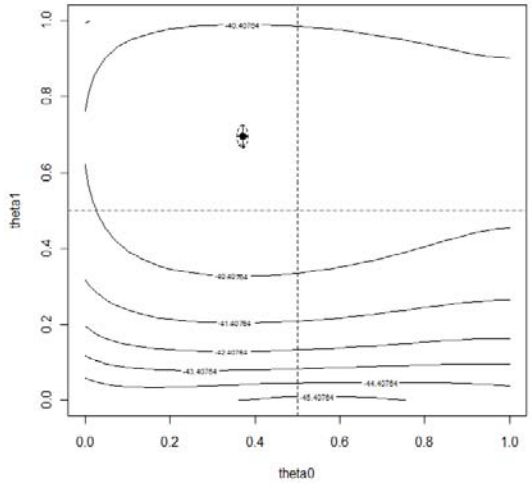
Visual interpretation



- Contour plot
- Fisher information matrix

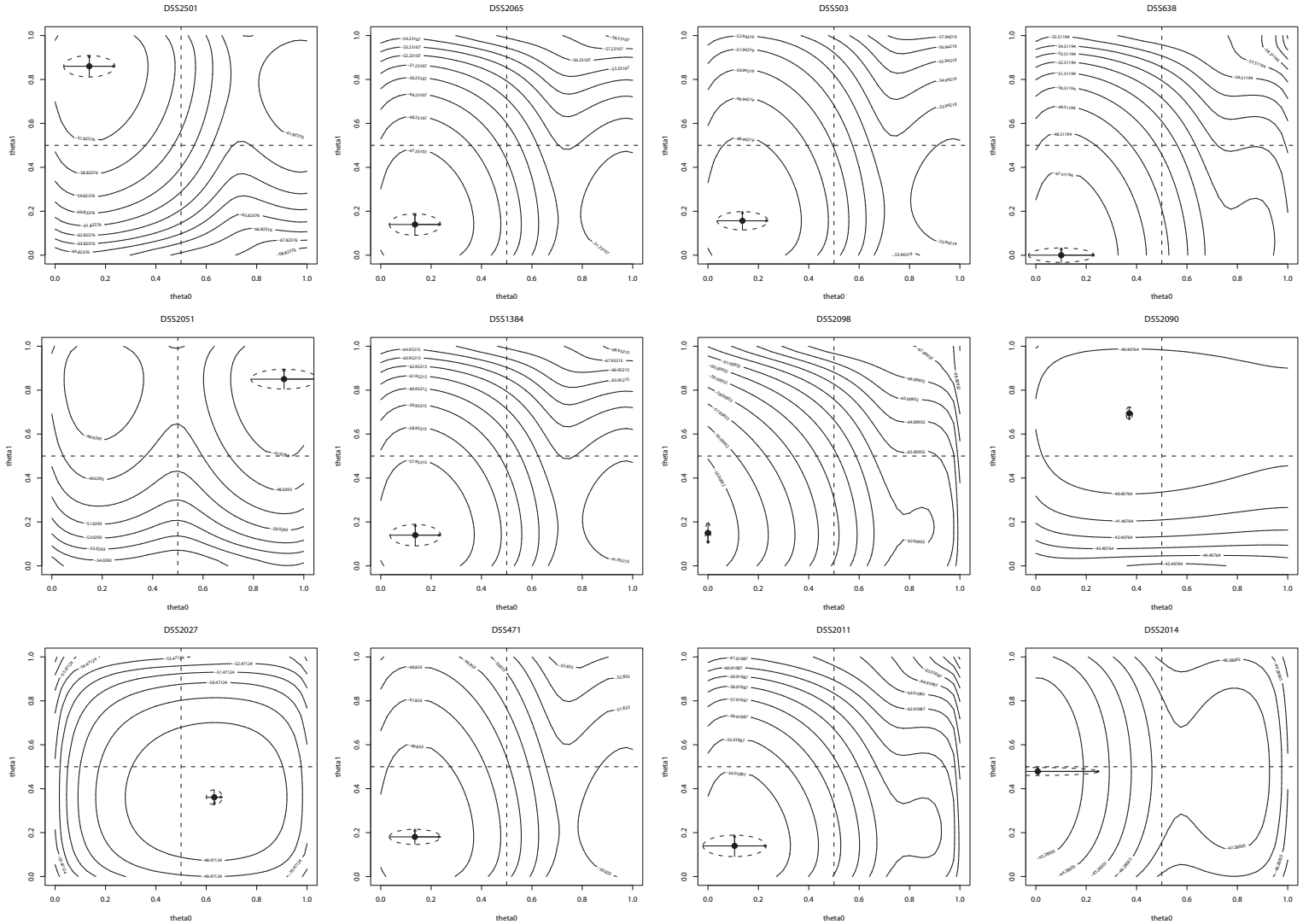
$$I(\boldsymbol{\theta}) = E\Big(-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L(\boldsymbol{\theta})\Big)$$

$$\Rightarrow \quad -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$



- $\theta_i > 0.5$
- Fisher information is low

⇒ NA



recombination fraction	D5S2501	D5S2051	D5S2027	D5S2065	D5S1384	D5S471	D5S563	D5S2098	D5S2011	D5S638	D5S2090	D5S2014
θ_0	0.136	NA	NA	0.136	0.137	0.136	0.137	NA	0.106	0.101	NA	0.008
θ_1	NA	NA	NA	0.140	0.140	0.181	0.157	NA	0.140	0	NA	NA

disease locus?

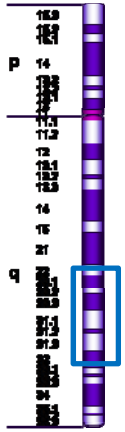
Comparison to dominant model

marker	dominant model		estimated values	
	$(\hat{\theta}_0, \hat{\theta}_1)$	log likelihood	$(\hat{\theta}_0, \hat{\theta}_1)$	log likelihood
D5S2501	(0.200, NA)	-62.787	(0.136, NA)	-56.824
D5S2051	(NA, NA)	-51.858	(NA, NA)	-46.029
D5S2027	(NA, NA)	-53.023	(NA, NA)	-47.471
D5S2065	(0.200, 0.200)	-52.195	(0.136, 0.140)	-46.231
D5S1384	(0.200, 0.200)	-62.887	(0.137, 0.140)	-56.952
D5S471	(0.200, 0.239)	-51.766	(0.136, 0.181)	-45.833
D5S503	(0.200, 0.216)	-53.861	(0.137, 0.157)	-47.942
D5S2098	(NA, NA)	-59.411	(NA, NA)	-54.009
D5S2011	(0.174, 0.200)	-59.047	(0.106, 0.140)	-53.011
D5S638	(0.174, 0.067)	-52.556	(0.101, 0)	-46.512
D5S2090	(NA, NA)	-44.993	(NA, NA)	-39.408
D5S2014	(0.085, NA)	-48.257	(0.008, NA)	-42.280

Linkage analysis by PANG et al.

- Linkage Package (MLINK)
- dominant model
- disease allele susceptibility frequency：0.0001
- 5q22.1 – q32
- $\theta_0 = \theta_1$

Chromosome 5



$$\log \frac{L(\theta)}{L(1/2)}$$

Marker	position(Mb)	0	0.01	0.05	0.1	0.2	0.3	0.4	θ max
D5S2501	110.1	−Inf	−7.52	−3.49	−1.9	−0.55	−0.02	0.13	0.4
D5S2051	111	−Inf	−7.08	−3.55	−2.08	−0.79	−0.26	−0.1	0.4
D5S2027	–	−Inf	−7.05	−3.64	−2.25	−1	−0.4	−0.1	0.4
D5S2065	113.7	−Inf	−1.54	0.35	0.97	1.26	1.09	0.66	0.2
D5S1384	118.9	−Inf	−1.54	0.34	0.97	1.26	1.09	0.66	0.2
D5S471	119.1	−Inf	−1.81	0.09	0.72	1.05	0.93	0.57	0.2
D5S503	120.3	−Inf	−1.67	0.21	0.84	1.15	1	0.61	0.2
D5S2098	–	−Inf	1.6	2.06	2.07	1.78	1.33	0.74	0.1
D5S2011	141.3	−Inf	−1.24	0.6	1.18	1.39	1.15	0.68	0.2
D5S638	146.7	−Inf	−0.12	1.12	1.48	1.53	1.22	0.7	0.2
D5S2090	147.2	−Inf	−5.78	−3.05	−1.93	−0.91	−0.41	−0.13	0.4
D5S2014	149.9	−Inf	−0.56	0.7	1.1	1.2	0.96	0.52	0.2