case-control関連解析による疾患感受性遺伝子の探索

大橋　順

東京大学医学部人類遺伝

内容

1．候補遺伝子の関連解析

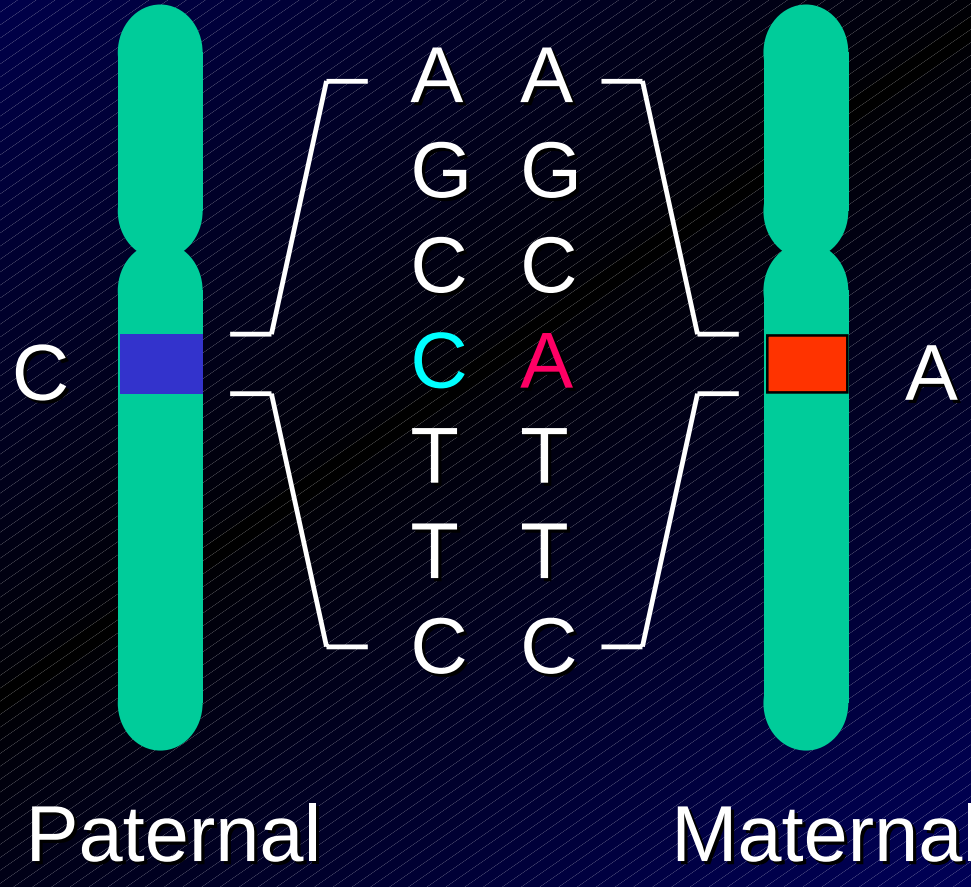
2．連鎖不平衡と連鎖不平衡検定

3．ゲノムワイド関連解析
　　における諸問題

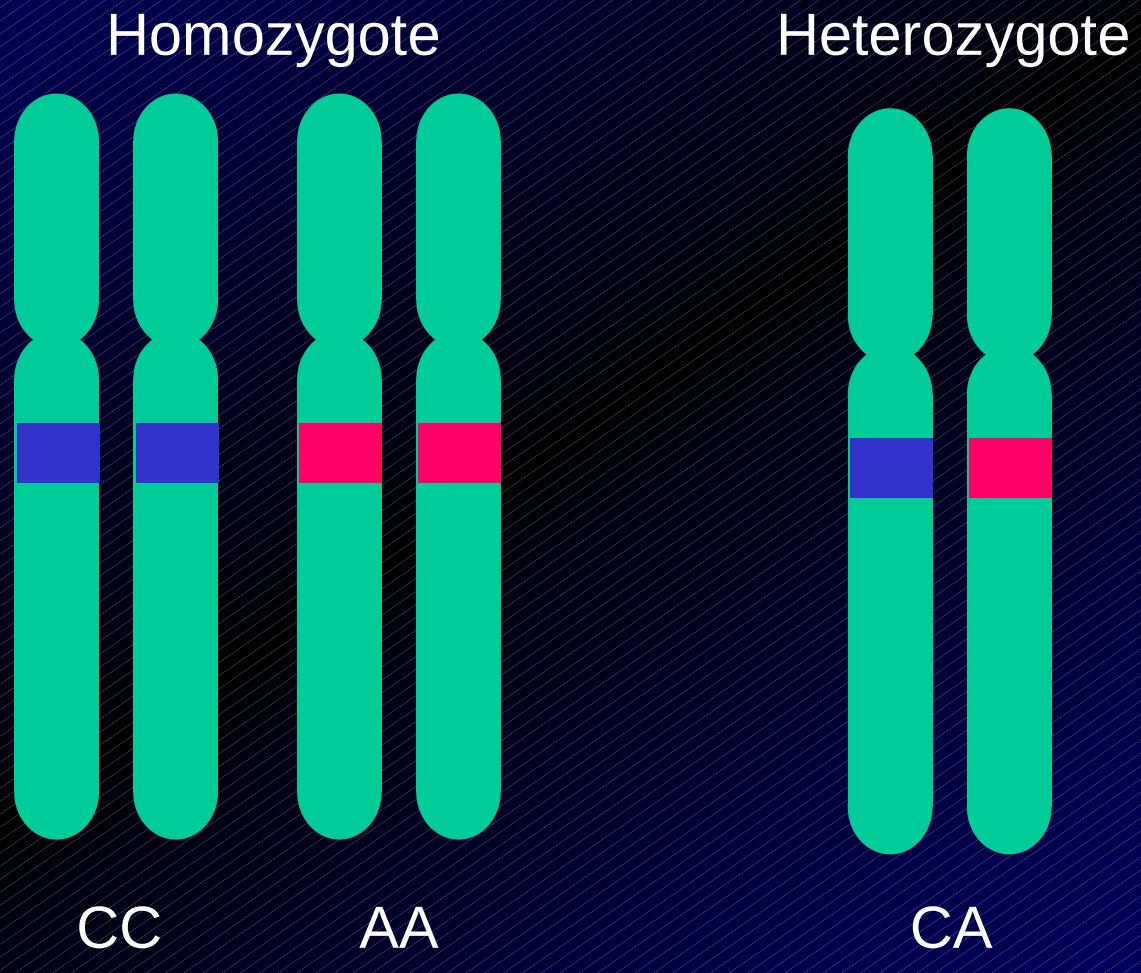
4．関連解析の実際

1） 候補遺伝子の関連解析

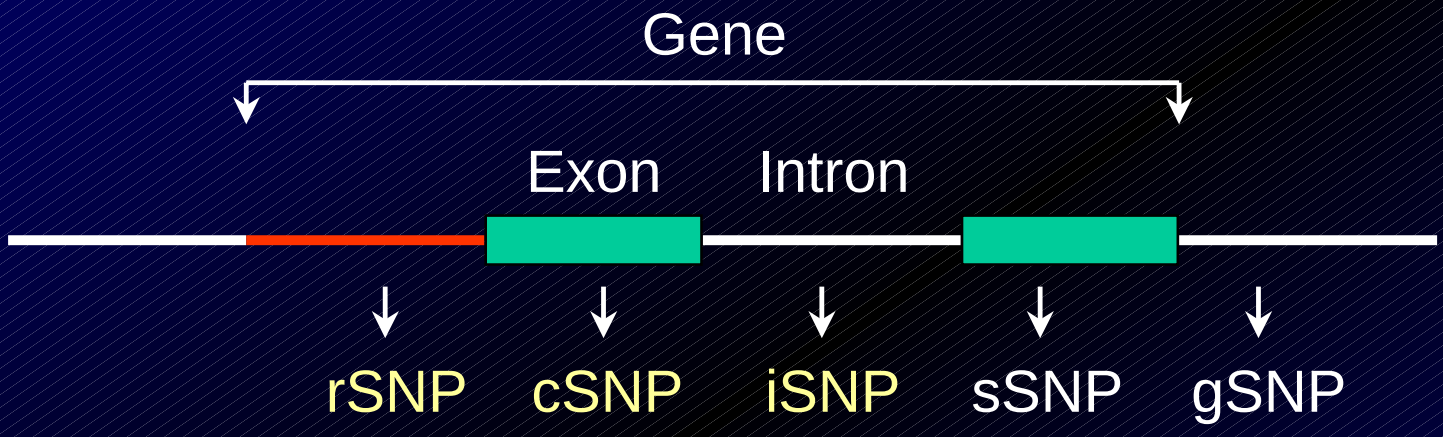
遺伝子座と対立遺伝子
Locus and Allele



遺伝子型



SNPの分類



rSNP; regulatory SNP
cSNP; coding SNP
iSNP; intron SNP
sSNP; silent SNP
gSNP; genome SNP

→ cSNP

→ gSNP

Proposed by
Dr. Yusuke Nakamura

CDCV仮説（Lander, Science 1996）

「ありふれた疾患（common disease）は疾患への寄与の小さい、頻度の高い変異（common variant）によってもたらされる」

しかし、いまだCDCV仮説を裏付けるような感受性遺伝子は見つかっていない。（ApoE-e4であっても、アルツハイマーと関連していない集団はある。）
（Weiss and Clark, Trends in Genetics 2002）

common diseaseでは，各感受性遺伝子の寄与が小さいと考えられ，連鎖解析によって遺伝子を同定することは困難．ただし，候補遺伝子を見つけることは比較的容易．

候補遺伝子の選定

↓

多型スクリーニング（少数検体）

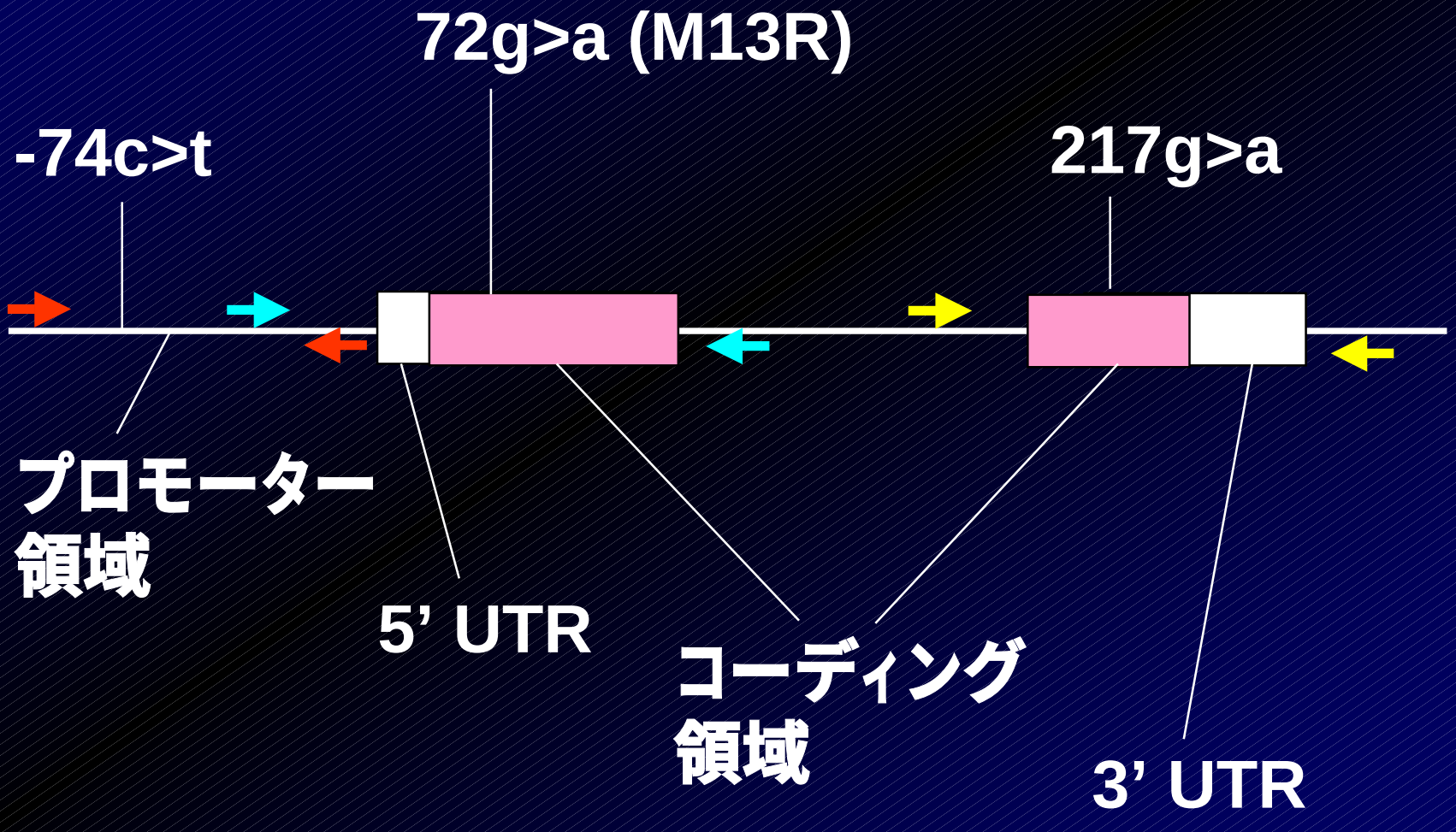
↓

タイピング

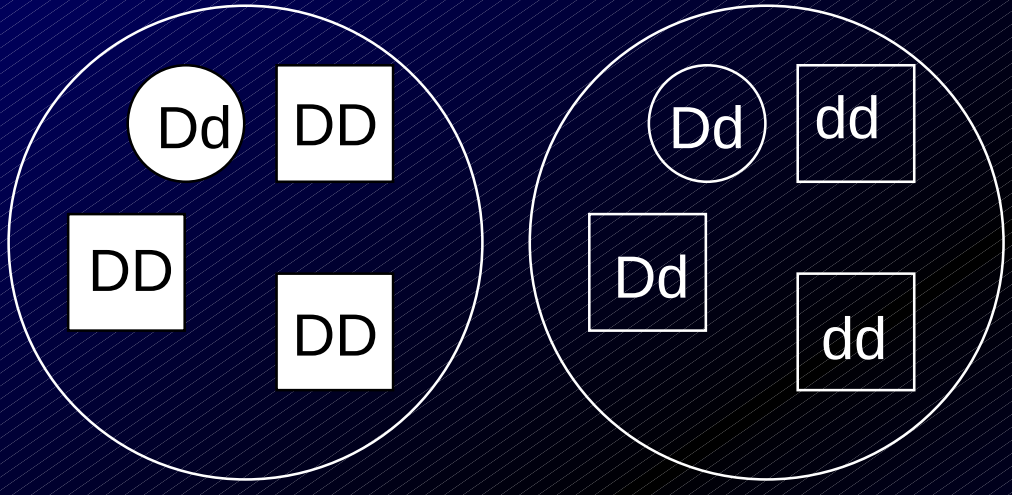
↓

関連検定

多型スクリーニング例



case-control関連検定（χ^2検定）



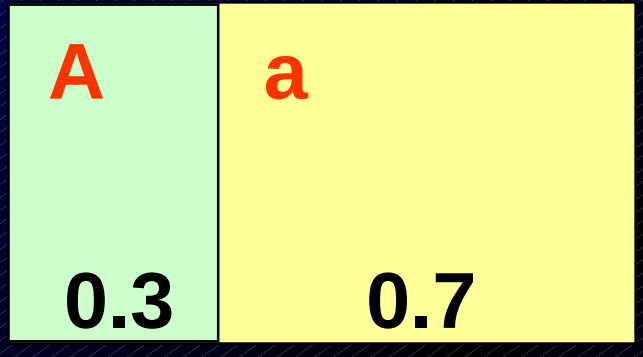
case群　　control群

	case	control
D	a	b
d	c	d

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}$$

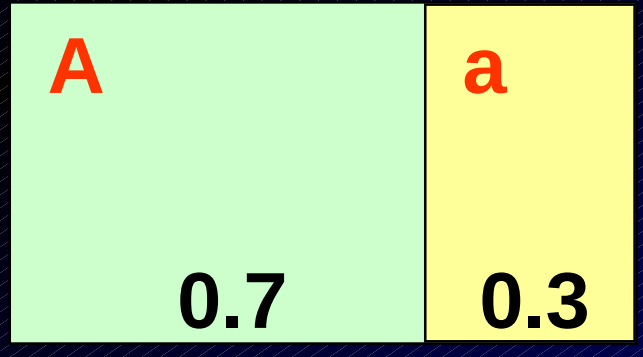
case-control関連検定を構造化のある集団(遺伝的背景の異なる分集団が混在している)で行なうと**偽陽性**が生じる

集団 1



$K_1 = 0.01$

集団 2



$K_2 = 0.03$

罹患率Kが集団2で高いと、集団1よりも集団2で頻度の高い**全ての対立遺伝子**はcase群で頻度が増加してしまう

伝達不平衡検定 (Transmission Disequilibrium Test [TDT])

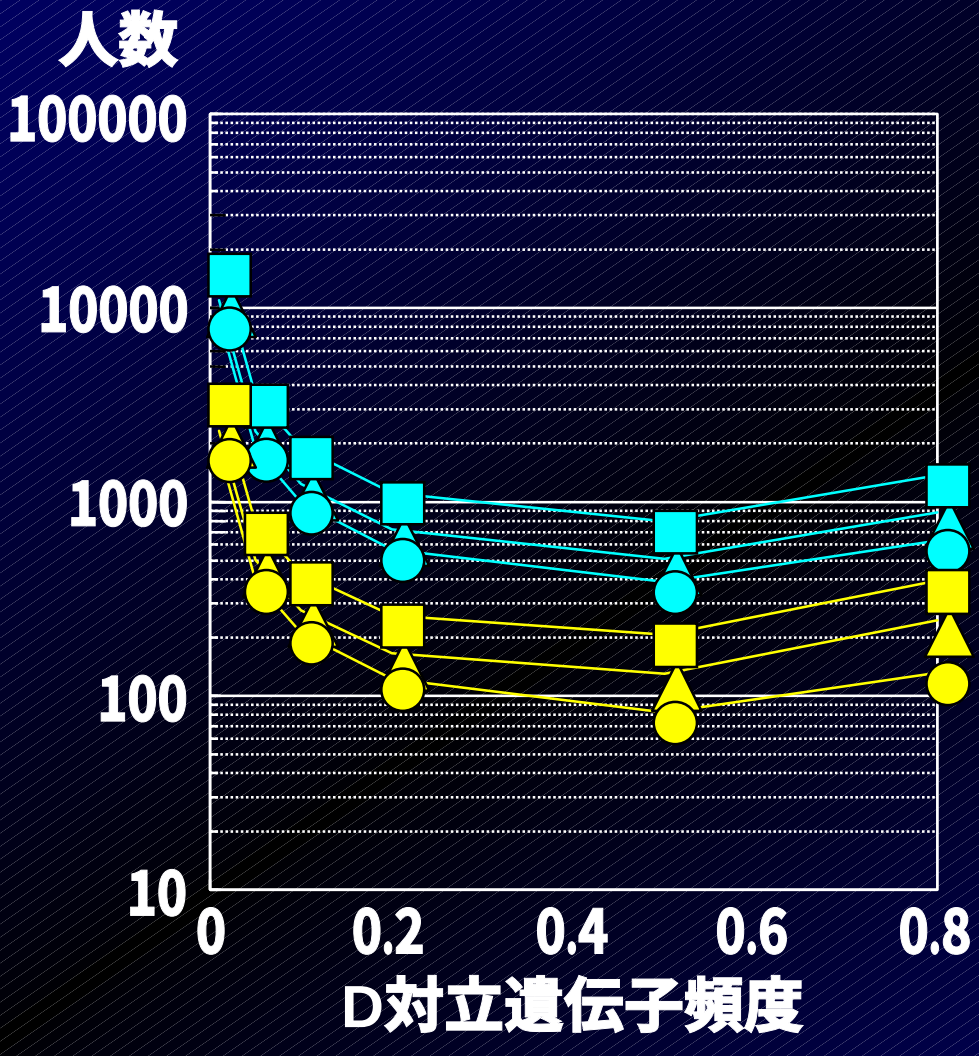
$\chi^2_{TDT} = (b-c)^2/(b+c)$

$df = 1$

		非伝達 M	非伝達 m
伝達	M	a	b
伝達	m	c	d

(a + b + c + d = 2n,
n: number of families)

0.8の検出力を達成するのに必要なサンプル数



- case-control関連検定 (γ = 1.5, f = 0.1)
- case-control関連検定 (γ = 1.5, f = 0.01)
- TDT (γ = 1.5)
- case-control関連検定 (γ = 2, f = 0.1)
- case-control関連検定 (γ = 2, f = 0.01)
- TDT (γ = 2)

理想的な関連研究とは

1. 大きなサンプルサイズ
2. 小さな*P*値
3. 遺伝子の生物学的意義
4. 関連のある変異が発現や
   産物に影響を与える
5. 最初の研究
6. 関連の再現性を確認
7. 家系研究と集団研究の
   両方で関連が確認

3) 連鎖不平衡と連鎖不平衡検定

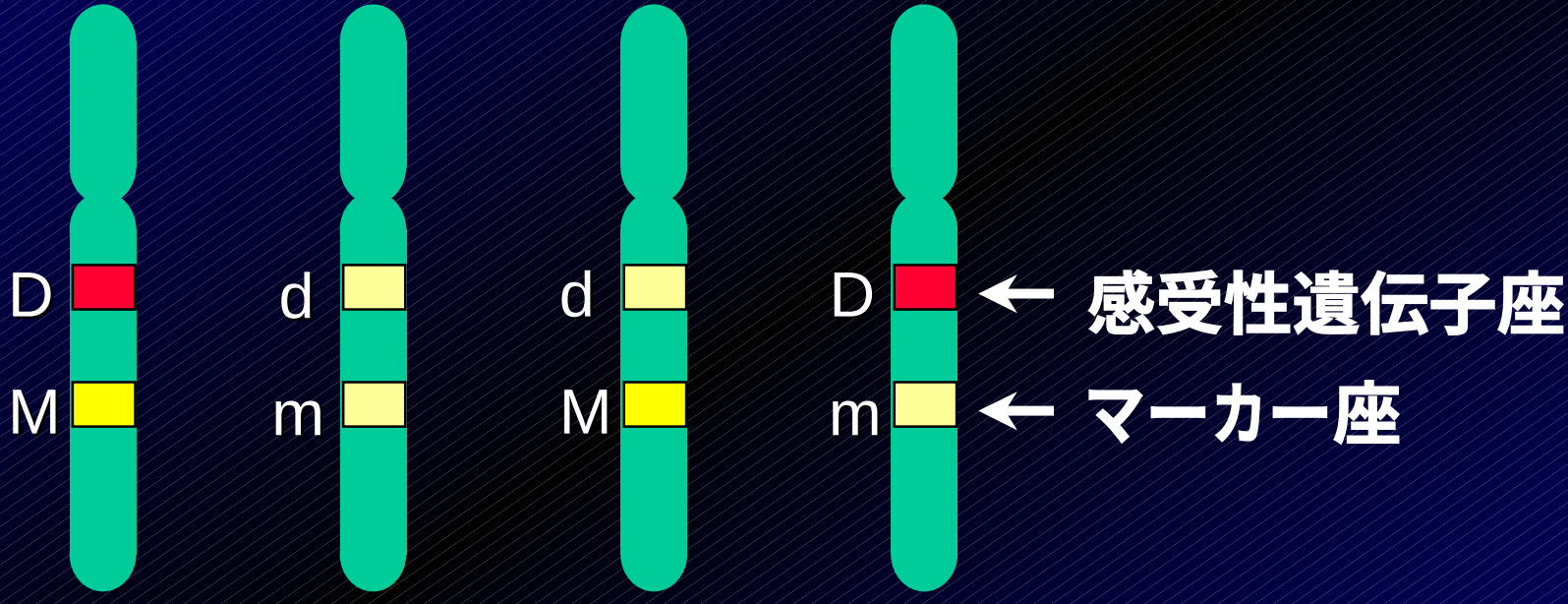
遺伝地図の作成や疾患の原因遺伝子を同定するために用いられるDNAマーカー

Type of marker	No. of loci	Features
DNA VNTRs (microsatellites) (di-, tri- and tetranucleotide repeats) 1989-	> 10^5 (potentially)	Many alleles Highly informative
DNA SNPs (single nucleotide polymorphisms) 1998-	> 10^6 (potentially)	Usually 2 alleles Less informative than microsatellites

連鎖不平衡 (linkage disequilibrium) とは，
(連鎖する) 2つ以上の遺伝子座における
対立遺伝子間に関連 (非独立性) が存在する
ことをいう．

ハプロタイプ頻度 = 対立遺伝子頻度の積？

連鎖した2遺伝子座2対立遺伝子



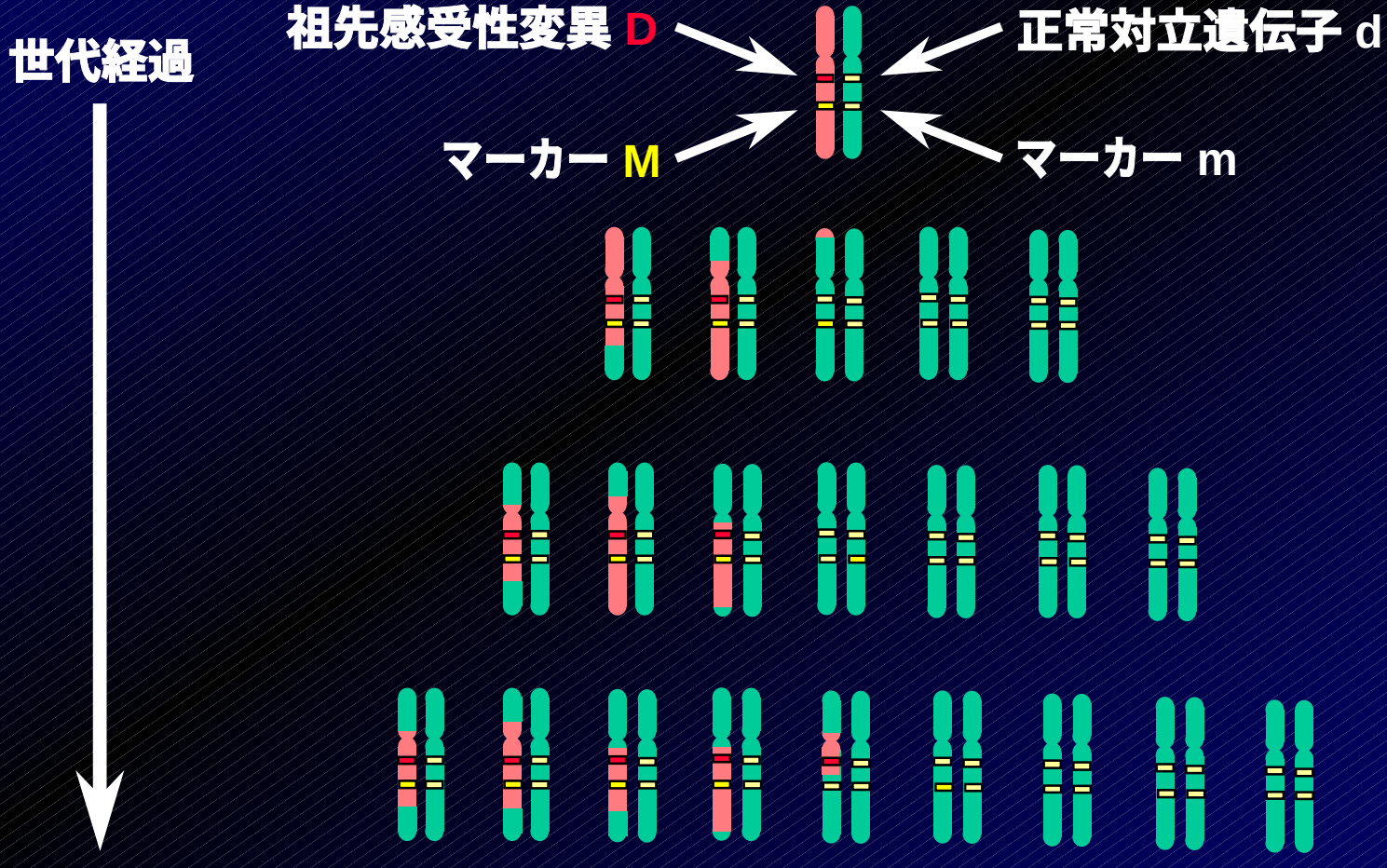
4つのハプロタイプが理論上存在する

連鎖不平衡係数：$D = H_{DM} - p_D \times p_M$

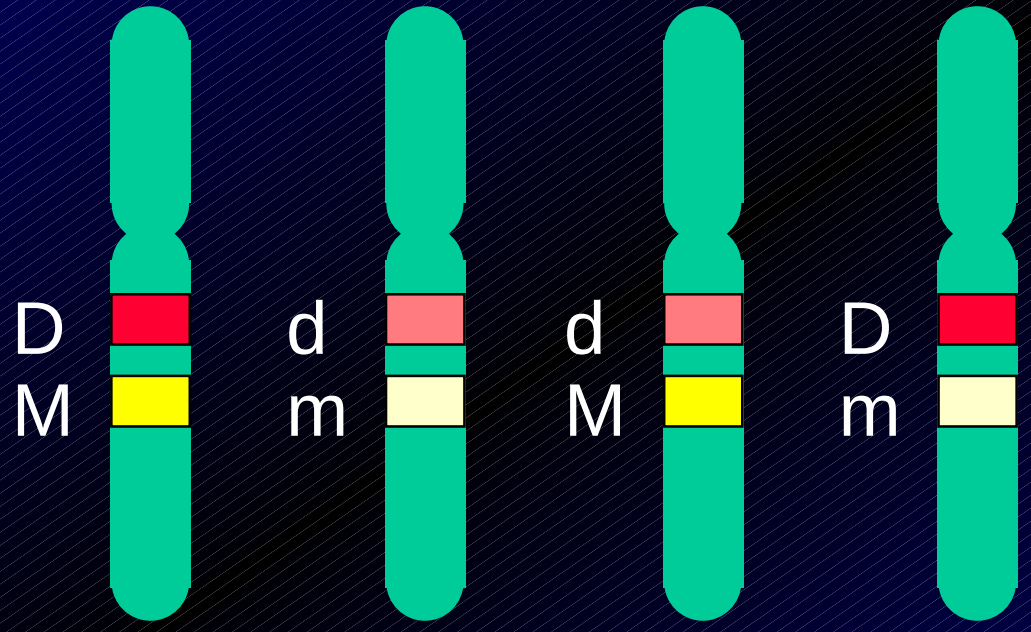
相対連鎖不平衡係数：$D' = D / D_{max}$

$= D / \{\min(p_D, p_M) - p_D \times p_M\}$

感受性変異の近傍に位置していたマーカーは，
感受性変異と共に次世代へと伝達される．



具体例)



ハプロタイプ頻度
(集団A)

0.25　　0.25　　0.25　　0.25

D = 0.25 - (0.5 x 0.5) = 0

D' = 0 / {0.5 - (0.5 x 0.5)} = 0

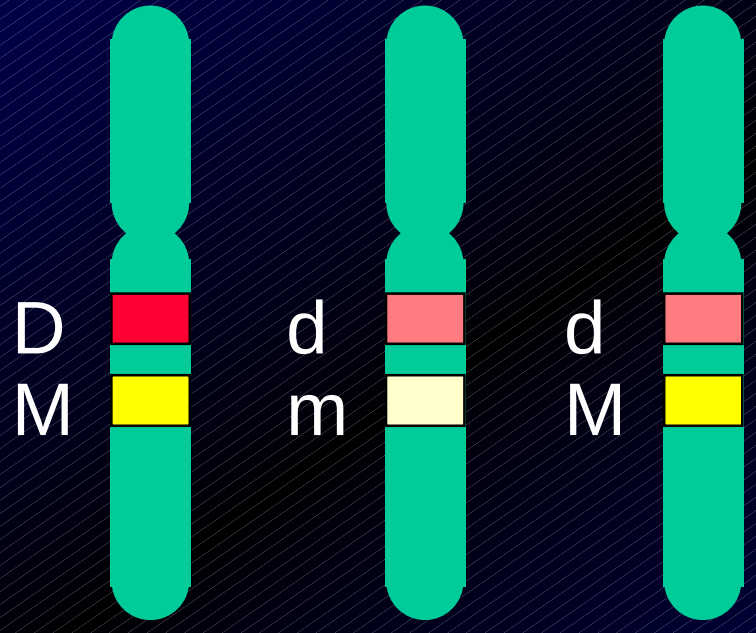
ハプロタイプ頻度
(集団B)

0.4　　0.4　　0.1　　0.1

D = 0.4 - (0.5 x 0.5) = 0.15 > 0

D' = 0.15 / {0.5 - (0.5 x 0.5)} = 0.6

突然変異直後)



ハプロタイプ頻度

0.01　　0.5　　0.49

D = 0.01 - (0.01 x 0.5) = 0.005

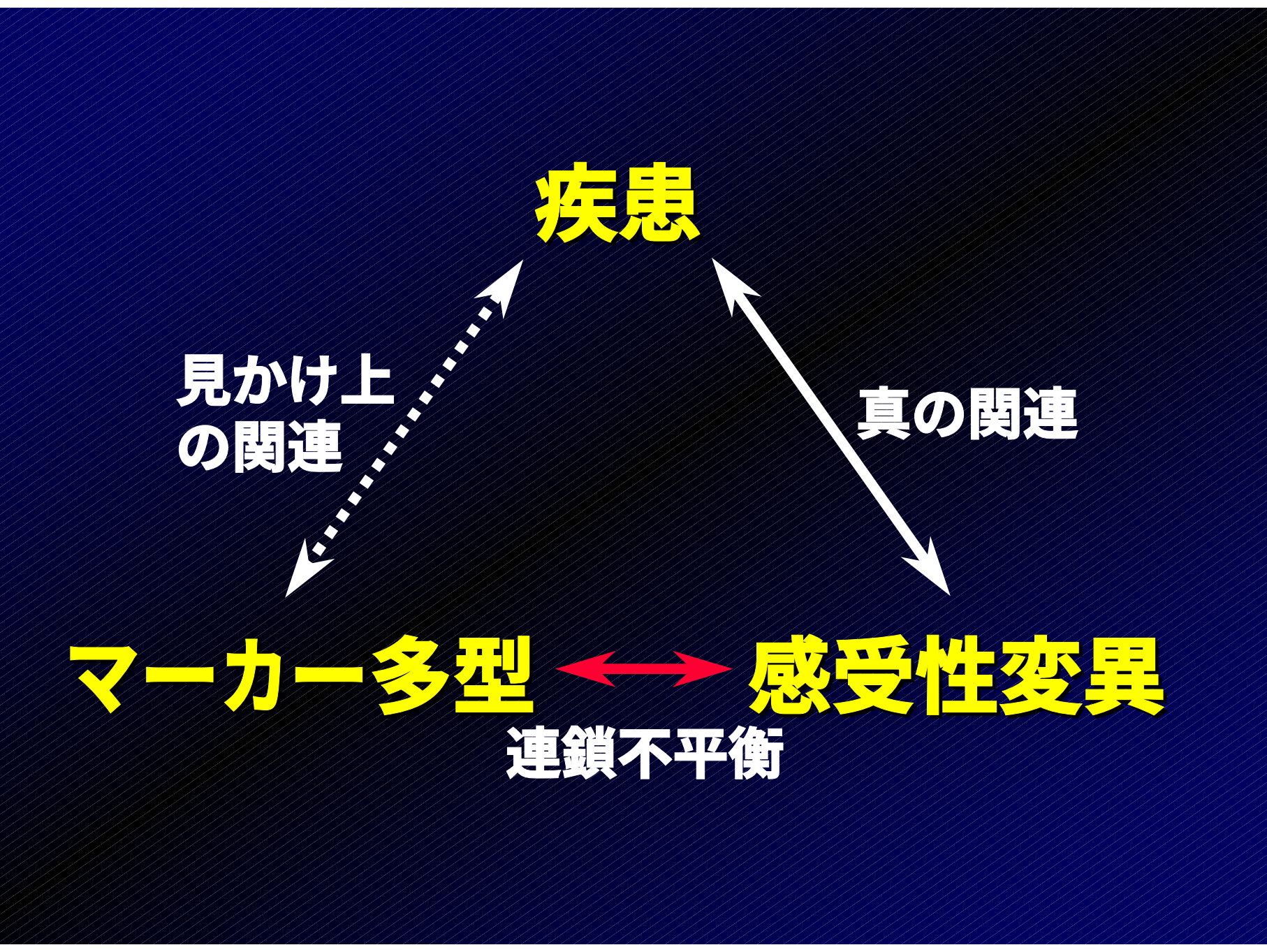
D' = 0.005 / {0.01 - (0.01 x 0.5)}

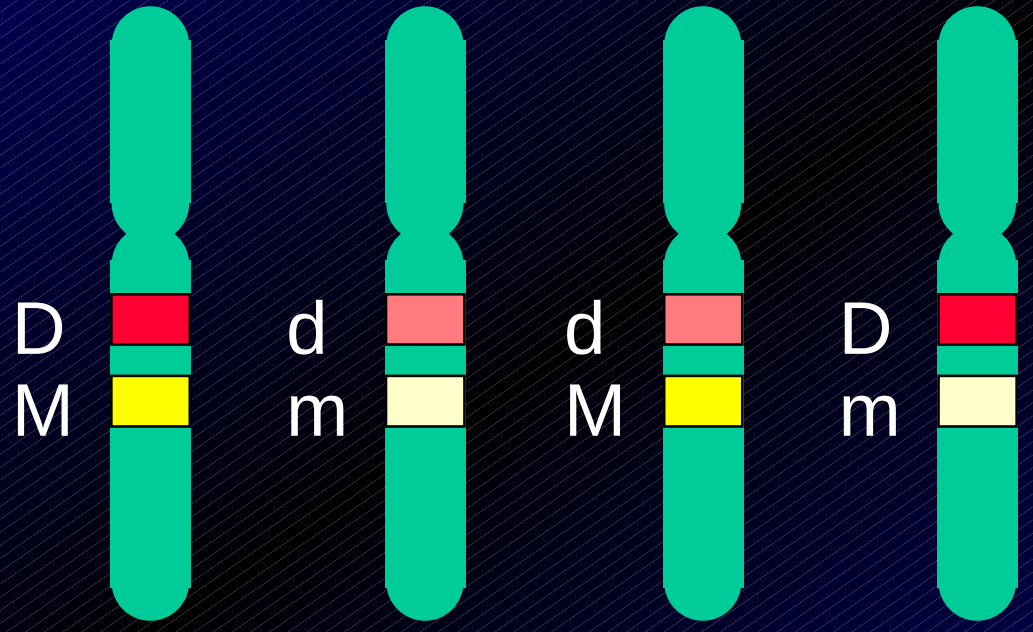
= 1

もしDとMとの間に正の連鎖不平衡が存在すれば，case群ではDの頻度が増加するため，それにつられてcase群ではMの頻度も増加する．

↓

マーカー座でもχ^2検定によって有意差が検出される（連鎖不平衡検定）．


疾患
見かけ上
の関連
真の関連
マーカー多型
感受性変異
連鎖不平衡



ハプロタイプ頻度
(集団A)

0.25　　0.25　　0.25　　0.25

D = 0.25 - (0.5 x 0.5) = 0

D' = 0 / {0.5 - (0.5 x 0.5)} = 0

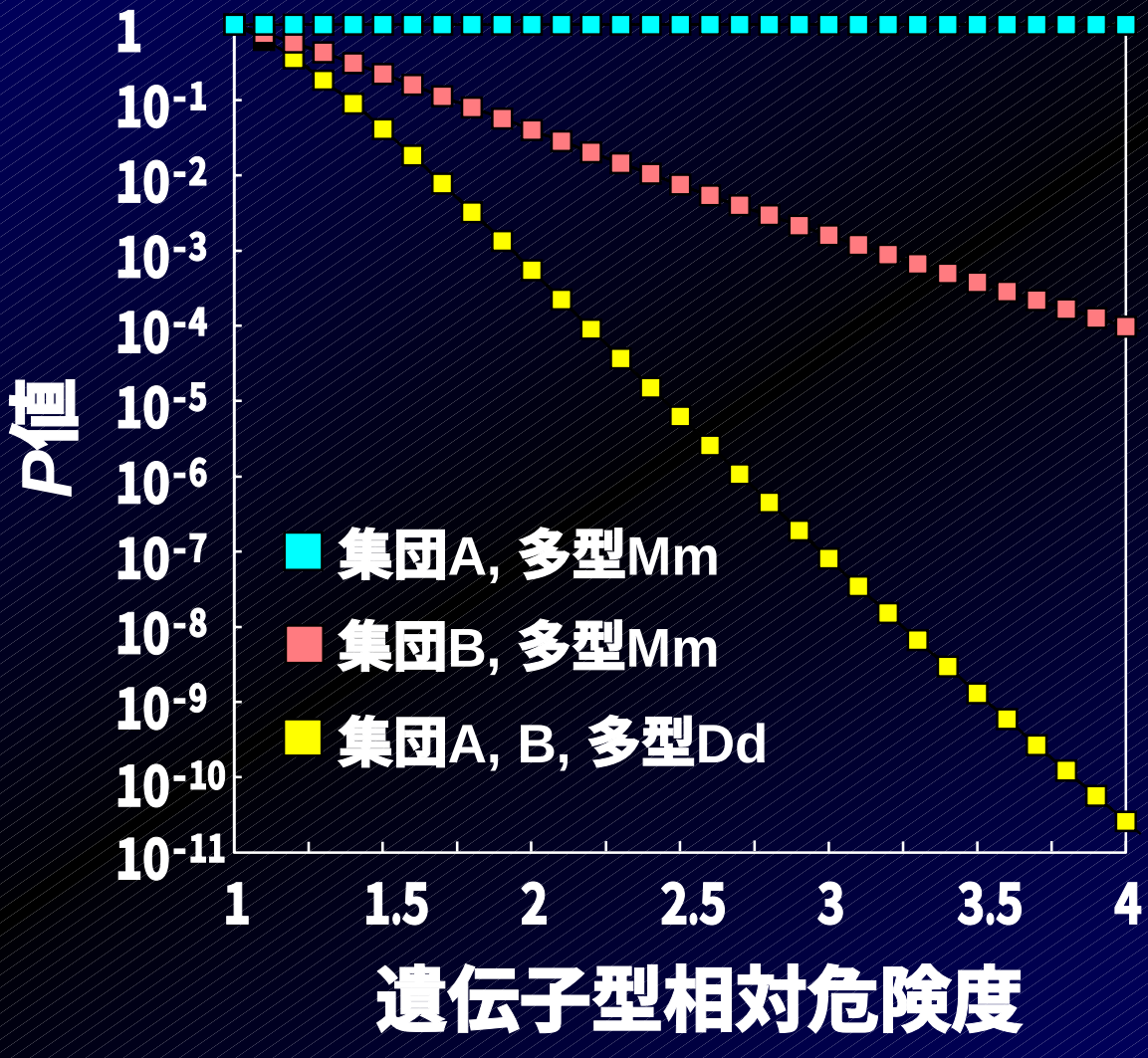
ハプロタイプ頻度
(集団B)

0.4　　0.4　　0.1　　0.1

D = 0.4 - (0.5 x 0.5) = 0.15 > 0

D' = 0.15 / {0.5 - (0.5 x 0.5)} = 0.6

関連検定において期待されるP値



100 cases
100 controls

$D(t+1) = (1 - \theta) \times D(t)$

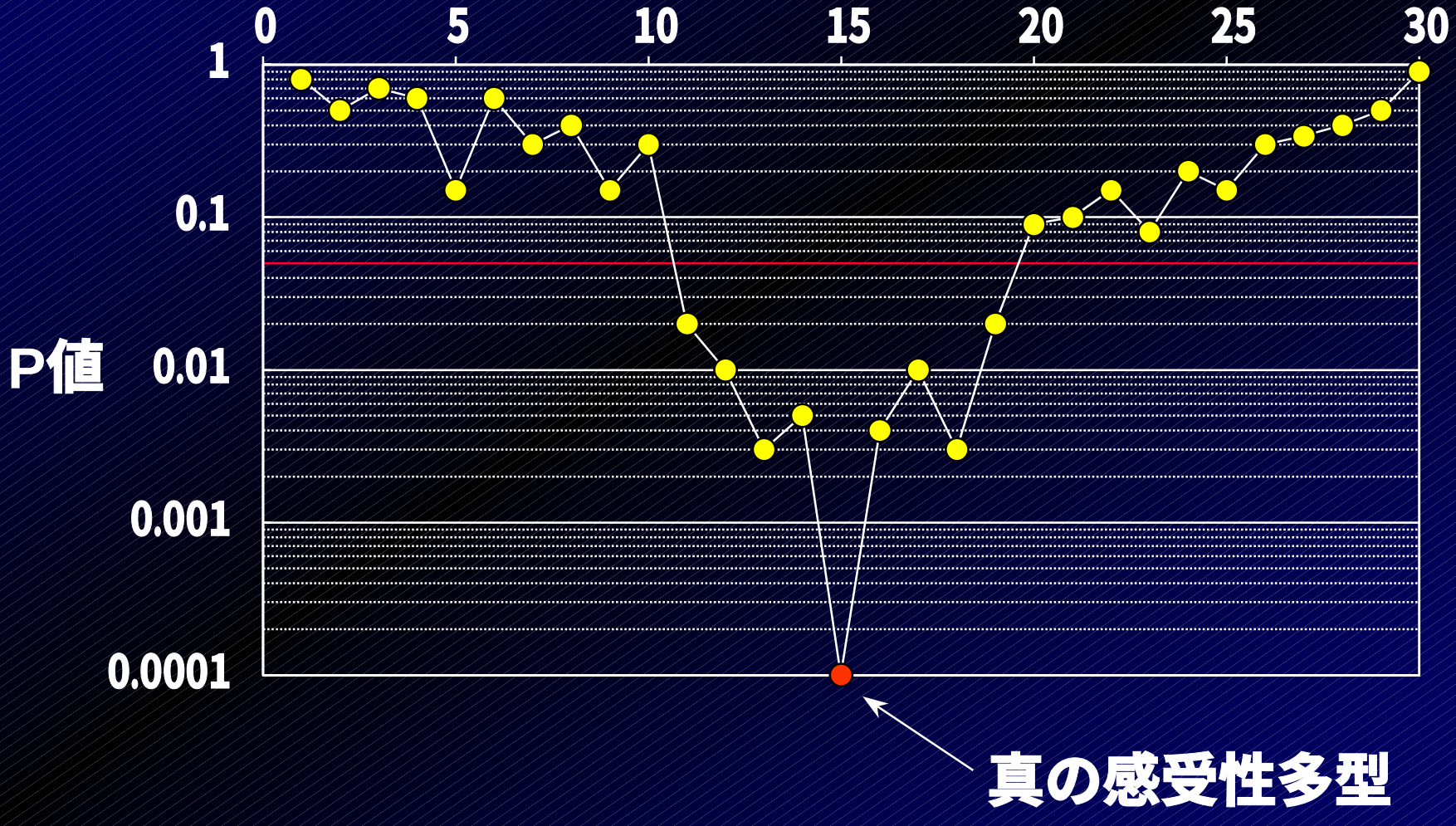
$D(t) = (1 - \theta)^t \times D(0)$

DとMとの連鎖不平衡は，組換えによって毎世代減少するため，Dとの遺伝距離が遠いマーカー座では有意差が検出されず，近いマーカー座でのみ有意差は検出される．

↓

有意差を示したマーカー座の近傍に真の疾患感受性遺伝子座が存在する．

イメージ



3) ゲノムワイド関連解析における諸問題

多重検定の補正

ゲノムワイドに設定した多型マーカーを、有意水準5%でそれぞれ統計検定を行なえば、5%のマーカーにおいて偽陽性が起こる.

そのため、有意水準αを5%より厳しく設定し、偽陽性が起こるマーカー数を調整する必要がある.

(例　Bonferroniの補正)

$\alpha = 0.05 / n$　　n: マーカー数

case-control関連研究において0.8の検出力を期待するために必要なcase数（同数のcontrol必要）

浸透率（遺伝モデル）	p	case数 α=0.05	case数 α=2.5X10^{-7}
f_2=0.04, f_1=0.04, f_0=0.01（優性モデル）	0.01	230	1059
	0.05	63	291
	0.1	45	210
	0.5	136	624
f_2=0.04, f_1=0.01, f_0=0.01（劣性モデル）	0.01	875463	4019983
	0.05	7778	35718
	0.1	1119	5141
	0.5	39	180
f_2=0.04, f_1=0.02, f_0=0.01（相乗モデル）	0.01	1169	5369
	0.05	251	1154
	0.1	137	632
	0.5	65	300

SNPマーカーを利用するゲノムワイド連鎖不平衡検定の検出力

問題

1)対立遺伝子頻度の低い(<5%)感受性変異をゲノムワイド連鎖不平衡検定で検出できるか？

2)どのようなSNPマーカーを選ぶべきか？
マーカー密度は？
マーカーの対立遺伝子頻度は？

遺伝学的モデル

1) ハーディ･ワインバーグ平衡状態にある集団を仮定。

2) 感受性遺伝子座には、淘汰上中立な
2つの対立遺伝子*D*と*d*が存在する。
（世代が経過しても、対立遺伝子頻度は変化しない。）

3) 突然変異は考慮しない。

使用されるパラメタ

D: 疾患感受性対立遺伝子
d: 正常対立遺伝子
M: *D*と正の連鎖不平衡にあるマーカー対立遺伝子
m: *d*と正の連鎖不平衡にあるマーカー対立遺伝子
p: *D*の対立遺伝子頻度
q: *M*の対立遺伝子頻度
t: 経過世代数
$H_{DM}(t)$: 世代*t*における*DM*ハプロタイプの頻度
θ: 組換え率。Haldaneのマップ関数により与える。

最近接マーカーまでの遺伝的距離と組換え率

$g = L / 2n$

g: 疾患感受性遺伝子座と最も近接するマーカーとの間の遺伝的距離

L: マーカーを設定する領域の長さ
(ゲノムワイド連鎖不平衡検定であれば 3000 cM)

n: 使用するマーカー数

θ: $\theta = \{1-\exp(-2g)\}/2$

これまでの仮定より、次の漸化式が成立する.

$$H_{DM}(t+1) = (1-\theta)H_{DM}(t) + \theta pq\,,$$

$$H_{Dm}(t+1) = (1-\theta)H_{Dm}(t) + \theta p(1-q)\,,$$

$$H_{dM}(t+1) = (1-\theta)H_{dM}(t) + \theta(1-p)q\,,$$

and

$$H_{dm}(t+1) = (1-\theta)H_{dm}(t) + \theta(1-p)(1-q)\,.$$

ハプロタイプの初期頻度を与えると、以下の解をえる。

$$H_{DM}(t) = \left(H_{DM}(0) - pq\right)(1-\theta)^t + pq\,,$$

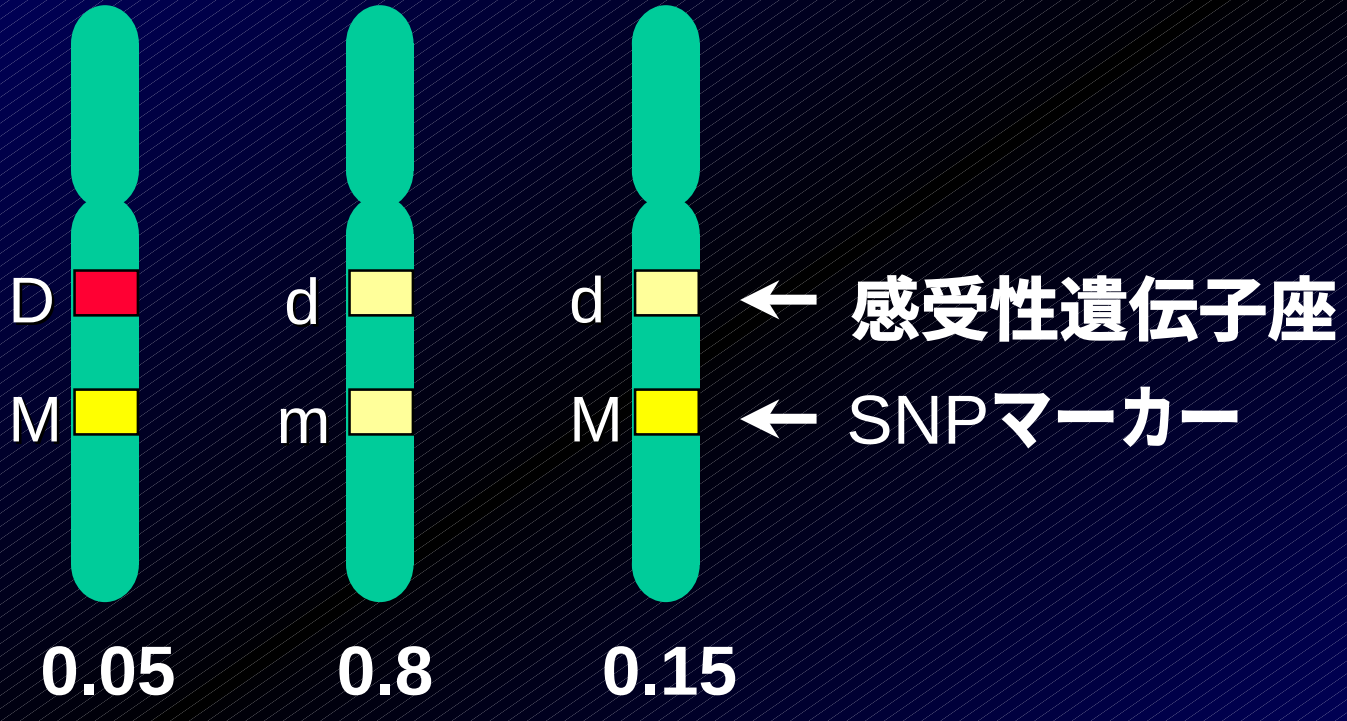
$$H_{Dm}(t) = \left(H_{Dm}(0) - p(1-q)\right)(1-\theta)^t + p(1-q)\,,$$

$$H_{dM}(t) = \left(H_{dM}(0) - (1-p)q\right)(1-\theta)^t + (1-p)q\,,$$

and

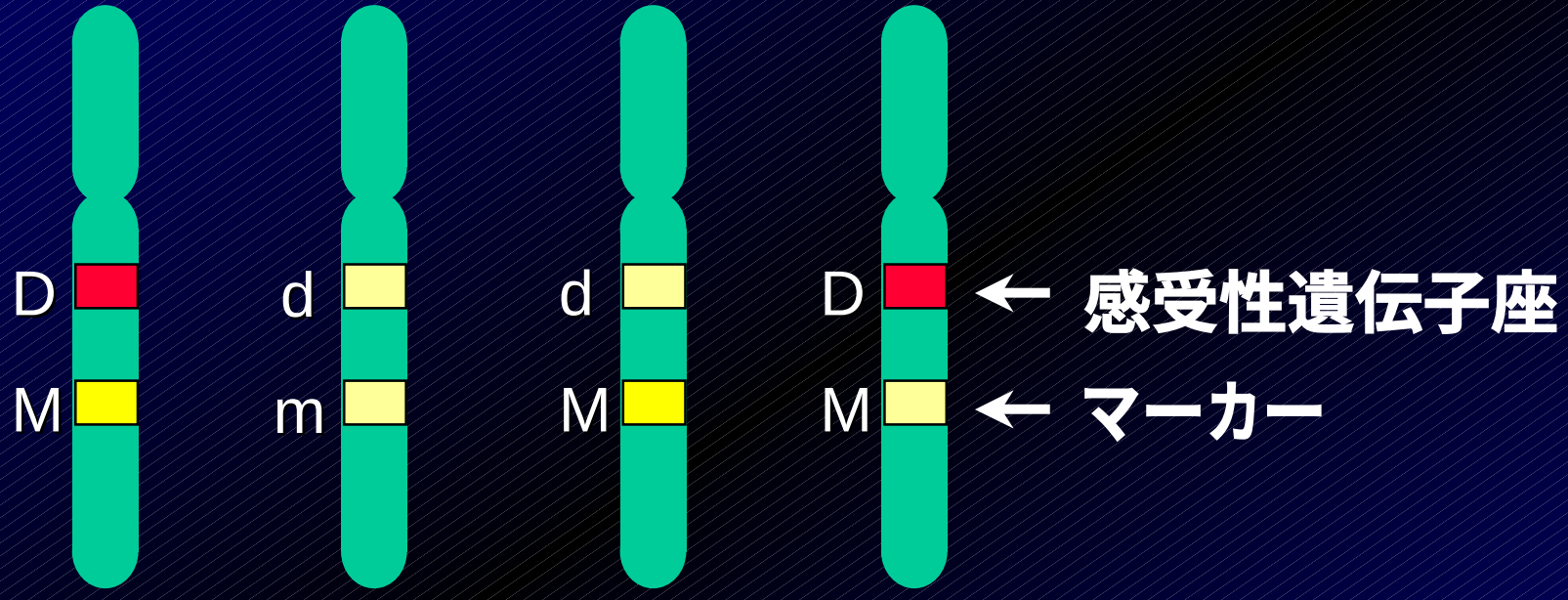
$$H_{dm}(t) = \left(H_{dm}(0) - (1-p)(1-q)\right)(1-\theta)^t + (1-p)(1-q)\,.$$

初期状態



初期状態では、3種類のハプロタイプのみが存在すると仮定。
感受性変異のアリル頻度よりも大きいminorアリル頻度をもつSNPマーカーを利用する。

数世代経過後



数世代経過すると、組換えによって4種類のハプロタイプが存在するようになる。

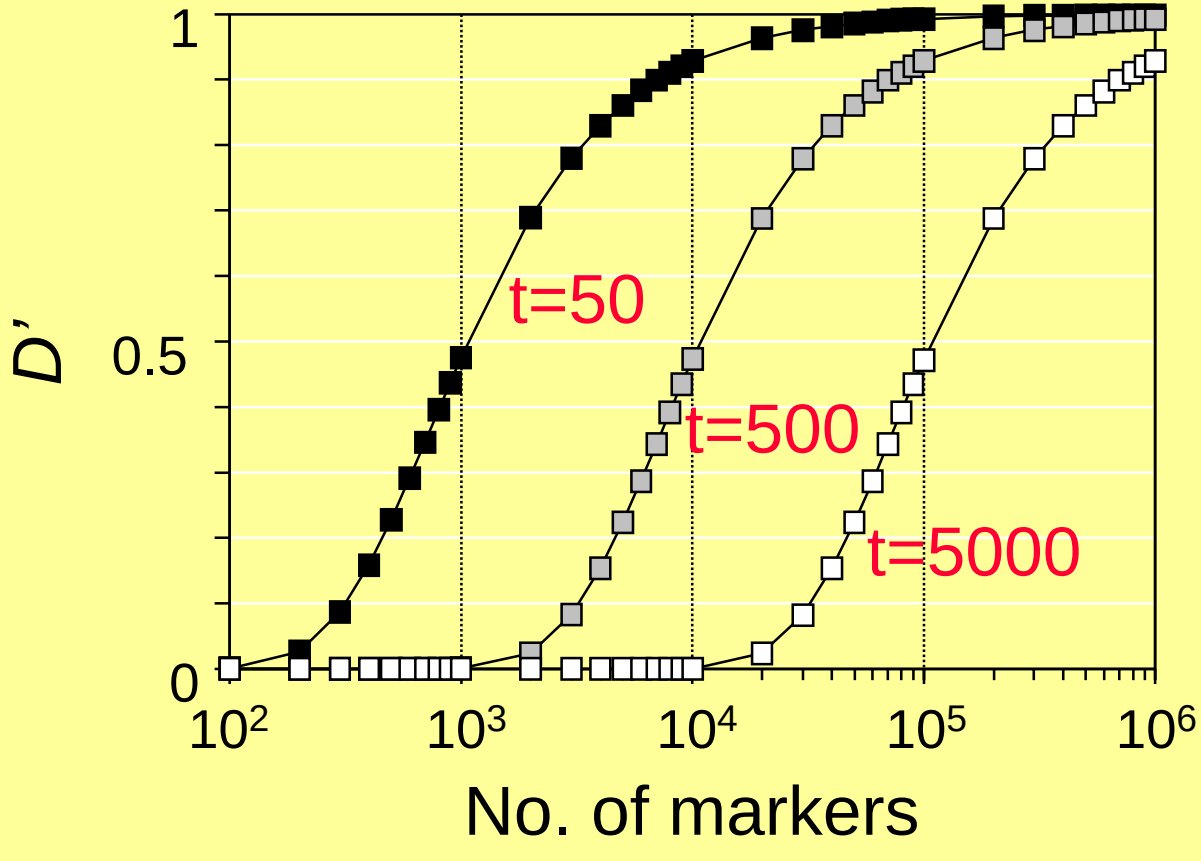


Figure 1. Relative value of linkage disequilibrium between the disease variant and the most adjacent marker allele. *D'* values for t = 50, 500, and 5000 are indicated by ■, ■, and □, respectively.

疾患モデル

f_2: *DD*の浸透率 (*DD*の個体が病気になる確率)
f_1: *Dd*の浸透率 (*Dd*の個体が病気になる確率)
f_0: *dd*の浸透率 (*dd*の個体が病気になる確率)

集団中の*DD*個体の割合は p^2,
集団中の*Dd*個体の割合は $2p(1-p)^2$,
集団中の*dd*個体の割合は $(1-p)^2$ なので,
罹患率は $p^2f_2 + 2p(1-p)f_1 + (1-p)^2f_0$ である.

割合の差の検定により、caseとcontrolとの間のSNPマーカー対立遺伝子頻度の差を評価する

検定統計量

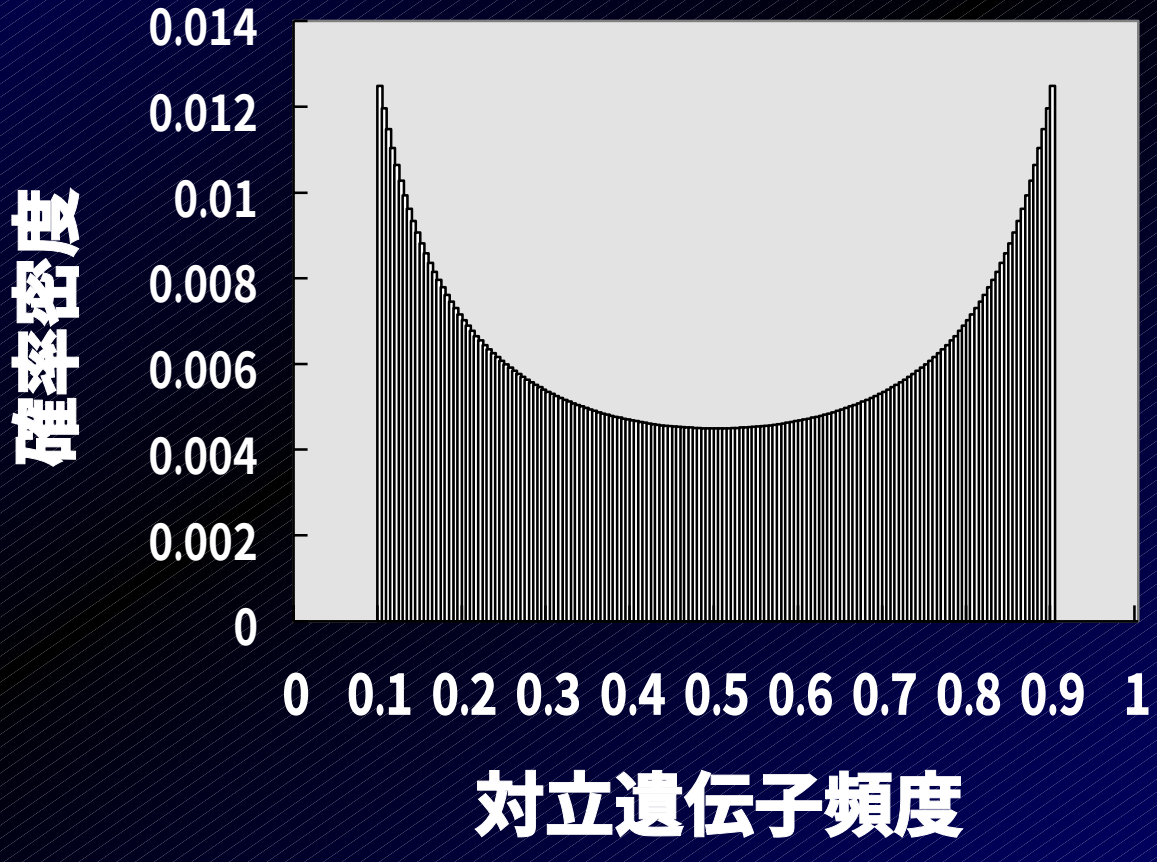
$$Z = \frac{\bar{P}_1 - \bar{P}_2}{\sqrt{\bar{P}(1-\bar{P})\left(\frac{1}{2N_1} + \frac{1}{2N_2}\right)}}$$

P_1: case**群中の**D**の頻度**
P_2: control**群中の**D**の頻度**
P: **全体での**D**の頻度**
N_1: case**数**
N_2: control**数**
(1000**人ずつと仮定**)

α : **有意水準** (0.05/n)
(Bonferroni**の補正**)

対立遺伝子頻度の平衡分布

・淘汰上中立
・マイナーアリルが0.1以上の頻度もつSNPをランダムに選択



マーカー対立遺伝子頻度により重み付けして検出力を計算する。

$$E(z_{1-\beta}) = \int_{Q}^{1-Q} 2Cq\left(\frac{1}{q}+\frac{1}{1-q}\right)\left[\frac{\sqrt{q_0(1-q_0)\left(\frac{1}{2R}+\frac{1}{2S}\right)}z_\alpha - (q_{\text{case}}-q_{\text{control}})}{\sqrt{\frac{q_{\text{case}}(1-q_{\text{case}})}{2R}+\frac{q_{\text{control}}(1-q_{\text{control}})}{2S}}}\right]dq$$

ここでCは定数である。

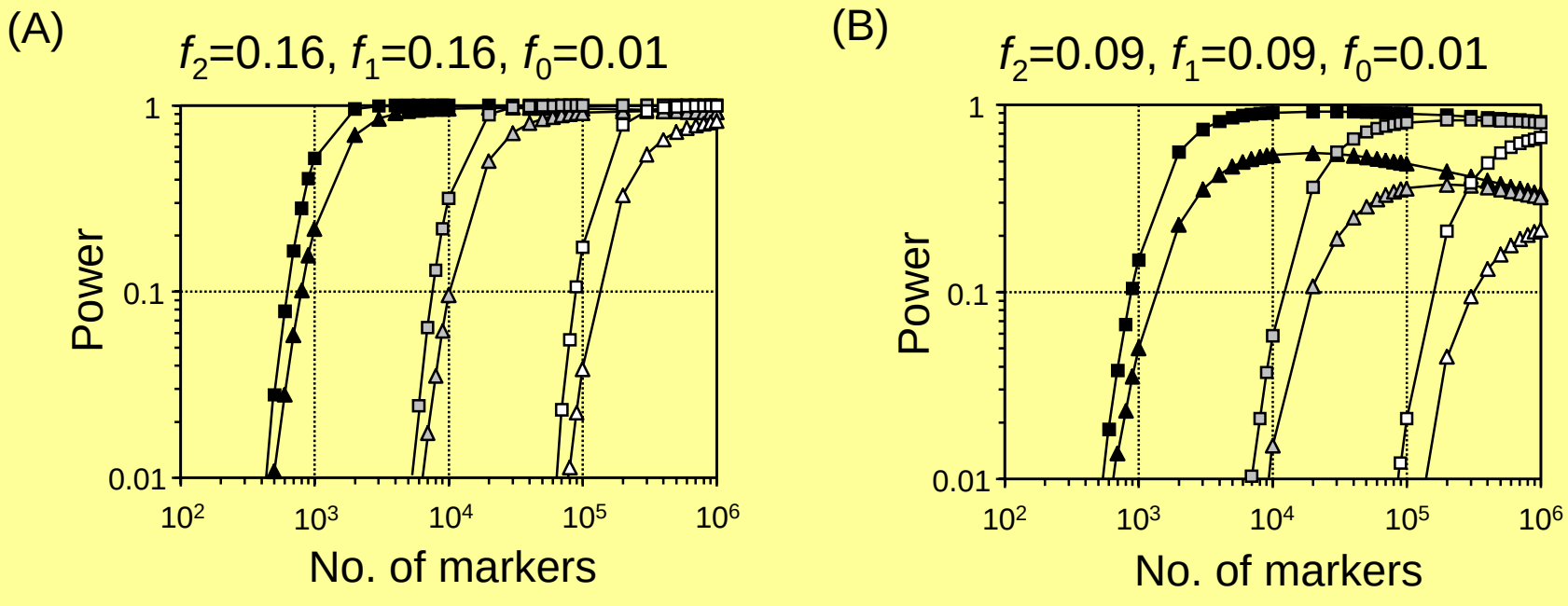


Figure 2. Conditional expected power in genome-wide LD testing for the detection of the disease variant showing a dominant mode of inheritance. (A) f_2 = 0.16, f_1 = 0.16, and f_0 = 0.01. (B) f_2 = 0.09, f_1 = 0.09, and f_0 = 0.01. Powers in the case of Q = 0.2 for t = 50, 500, and 5000 are indicated by ■, ▩, and □, respectively. Powers in the case of Q = 0.05 for t = 50, 500, and 5000 are indicated by ▲, △, and △, respectively. It is assumed that p = 0.05 and R = S = 1000.

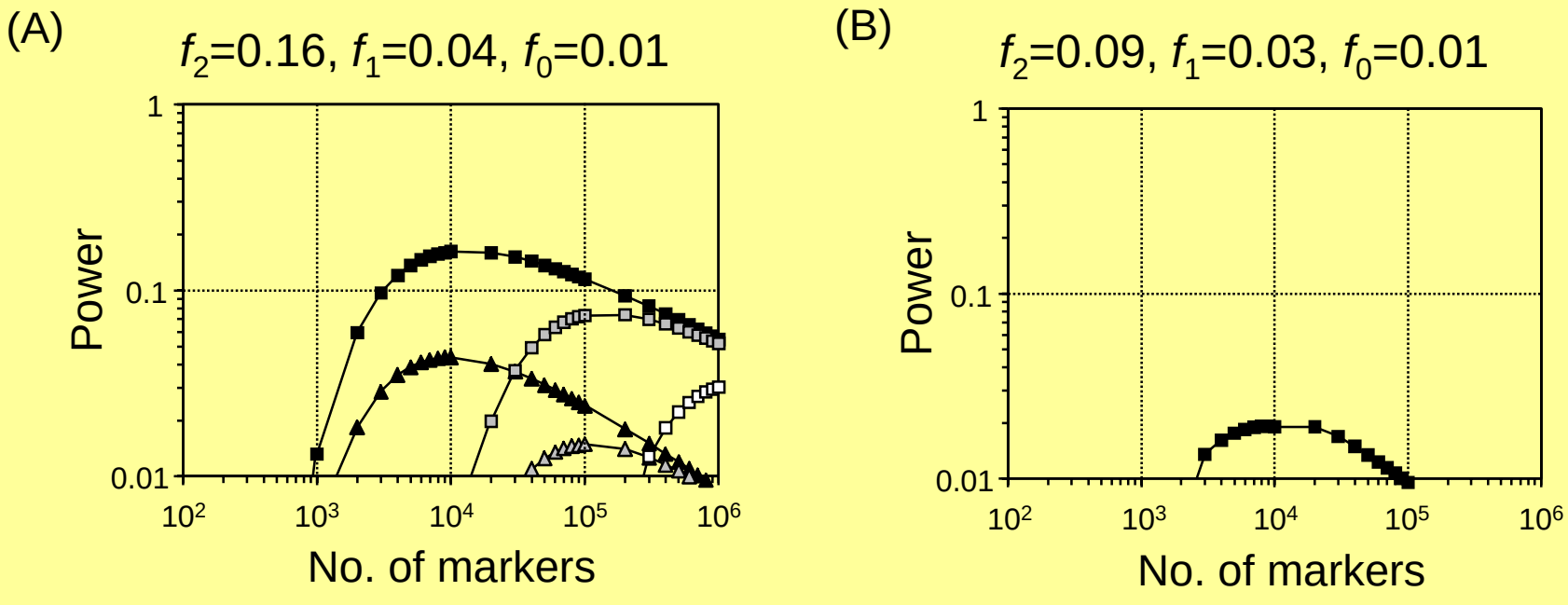


Figure 3. Conditional expected power in genome-wide LD testing for the detection of the disease variant showing a multiplicative mode of inheritance. (A) f_2 = 0.16, f_1 = 0.04, and f_0 = 0.01. (B) f_2 = 0.09, f_1 = 0.03, and f_0 = 0.01. Powers in the case of Q = 0.2 for t = 50, 500, and 5000 are indicated by ■, ■, and □, respectively. Powers in the case of Q = 0.05 for t = 50, 500, and 5000 are indicated by ▲, △, and △, respectively. Only cases showing a power higher than 0.01 are shown. It is assumed that p = 0.05 and R = S = 1000.

解答

1)浸透率が高くなければ、頻度の低い感受性変異を検出することは極めて困難である

2) マイナーアリルの対立遺伝子頻度が高いSNPを用いるべき
マーカー密度は（見つけたい）感受性変異の特性を考慮して決定すべき