

SNPの数理解析とSNP利用研究のための新規統計解析手法の開発

●大橋 順

東京大学大学院医学系研究科

<研究の目的と進め方>

common disease（ありふれた疾患）の感受性遺伝子変異の探索法として、ゲノムワイドに膨大な数の多型マーカーを設定して行なうゲノムワイド関連解析が注目されている。ゲノムワイド関連解析でゲノム全体をもれなく解析するためには、ヒト集団の連鎖不平衡構造から考えると、少なくとも数十万個のSNPマーカーを調べる必要がある。SNP数の増加は検討する帰無仮説数（検定数）の増加を招き、不十分な研究デザインでは多くの偽陽性を許容することになる。研究効率（検出力）を高め、さらに偽陽性を減らすためには、適切なサンプルサイズで有意水準を低く抑えて研究を行うことが必要である。しかしながら、どの程度のサンプルサイズで研究を行うべきか、また、浸透率が低いと考えられているありふれた疾患の感受性変異をいかに同定するかについての議論は十分になされているとはいえない。そこで、疾患感受性遺伝子変異を効率的に探索するためのガイドラインを作成するとともに、実在集団の連鎖不平衡の程度を解析し、関連解析のための基礎情報の収集を行うことを本研究の目的とした。具体的には次の二点である。

（１）SNPマーカーを用いたゲノムワイド関連解析および候補遺伝子アプローチにおける統計学上の問題点を分析し、疾患感受性遺伝子を効率的に探索するためのガイドラインの作成（サンプルサイズの計算や使用する各種マーカーの選択や配置も含む）および新規統計解析手法の開発を行なう。

（２）実在集団におけるSNP間連鎖不平衡を調べ、実データと理論予測との比較を行う。さらに、連鎖不平衡構造情報を利用して疾患関連変異を探索するモデル研究として、マラリア感受性関連変異の同定を試みる。

<研究開始時の研究計画>

（１）SNP マーカーを用いたゲノムワイド関連解析および候補遺伝子アプローチにおける統計学上の問題点を分析し、疾患感受性遺伝子を効率的に探索するためのガイドラインの作成および新規統計解析手法の開発を行なうことを計画した。具体的には以下の通りである。

（i）候補遺伝子アプローチを行うための適切な研究デザイン（サンプルサイズ）の設定

（ii）間接的ゲノムワイド関連解析を行うための適切な研究デザイン（サンプルサイズ）の設定

（iii）最も効率よく疾患感受性遺伝子を同定しうる（検出力の高い）SNP マーカー数とサンプルサイズを求めるアルゴリズムの開発

（iv）骨髄移植後移植片対宿主病の原因となるマイナー組織適合性抗原（mHa）を検出するための遺伝統計学的手法の開発

（２）ゲノムワイド関連研究を成功させるためには、基礎情報として解析対象集団における SNP 連鎖不平衡構造が必要である。そこで、実在集団における SNP 間連鎖不平衡構造の解析と疾患遺伝子モデル研究も計画した。

（１）（i）候補遺伝子アプローチを行うための適切な研究デザイン（サンプルサイズ）の設定

二種類の対立遺伝子 D と d を有する遺伝子座に対して、割合の差の検定を利用して D と d の対立遺伝子頻度を、患者群と対照群とで調べるとする。患者を n 個体サンプルした場合に、遺伝子型 DD、遺伝子型 Dd、遺伝子型 dd の個体数が n_2, n_1, n_0 であり ($n = n_2 + n_1 + n_0$)、対照を m 個体サンプルした場合に、遺伝子型 DD、遺伝子型 Dd、遺伝子型 dd の個体数が m_2, m_1, m_0 であつたとすると ($m = m_2 + m_1 + m_0$)、患者群中の D 対立遺伝子の推定遺伝子頻度 $\hat{P}_1$ は $\frac{2n_2 + n_1}{2n}$ 、対照群中の D 対立遺伝子の推定遺伝子頻度

$\hat{P}_2$ は $\frac{2m_2 + m_1}{2m}$ である（表１と表２）。

表1. 遺伝子型頻度表

	患者群	対照群
DD	n_2	m_2
Dd	n_1	m_1
dd	n_0	m_0
計	n	m

表2. 遺伝子頻度表

	患者群	対照群
D	$2n_2 + n_1$	$2m_2 + m_1$
d	$n_1 + 2n_0$	$m_1 + 2m_0$
計	$2n$	$2m$

帰無仮説 H_0 は「2 群の遺伝子頻度 P_1, P_2 には差が無い ($P_1 = P_2$)」

とし、対立仮説 H_1 は「患者群の遺伝子頻度 P_1 は対照群の P_2 より大きい ($P_1 > P_2$)」とする。 P_1 が P_2 より増加しているかどうか

を検討するので、このような検定は「片側検定」とよばれる。 H_0 より、平均の推定遺伝子頻度

<研究期間の成果>

$$\hat{p} = \frac{2n_2 + n_1 + 2m_2 + m_1}{2n + 2m}$$

が集団全体での遺伝子頻度 P ($= P_1 = P_2$) の推定値であると仮定する。いま、データ数が大きく、 $\hat{P}$ があまり小さくないとき、 H_0 が正しいとすると、 $\hat{P}_1 - \hat{P}_2$ は平均 0、標準偏差 $\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{2n} + \frac{1}{2m}\right)}$ の正規分布に従う。したがって、2 群の遺伝子頻度の差が $\hat{P}_1 - \hat{P}_2$ となる確率は、それを次式により標準化して (これを平均 0、標準偏差 1 となるように調整して)、標準正規分布表から求めることができる。

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{2n} + \frac{1}{2m}\right)}}$$

P_1 が P_2 より大きいほど、 Z は 0 より大きくなる。 z_α は標準正規分布の右片側 α の確率を与える Z の値であり、 H_0 の下で Z が z_α を越える確率が α である Z の値のことである。有意水準 α を 0.05 に設定して片側検定を 1 回行なうとき、 $z_\alpha = z_{0.05} = 1.64$ よりも Z が大きければ、 H_0 の下で確率 0.05 以下でしか生じない稀なことが起きたことを示している。そこで、稀なことが起きたのではなく、 H_0 が間違っていると判断し、有意水準 0.05 で帰無仮説 H_0 を棄却し、対立仮説 H_1 を採択する。すなわち、D は疾患と関連していると結論する。しかし、D と d のどちらが疾患感受性対立遺伝子であるかを、研究以前に通常は予想できないので、両方の対立遺伝子に対して片側検定が 2 回行なわれると考えるべきである。すなわち、患者群中の d 対立遺伝子の推定遺伝子頻度 $\hat{P}_1$ を $\frac{n_1 + 2n_0}{2n}$ 、対照群中の d 対立遺伝子の推定遺伝子頻度 $\hat{P}_2$ を $\frac{m_1 + 2m_0}{2m}$ 、平均の推定遺伝子頻度は $\hat{p} = \frac{n_1 + 2n_0 + m_1 + 2m_0}{2n + 2m}$ として (表 1 と表 2)、再度検定することになる。もし、有意水準を α として 2 度検定を行えば、第一種の過誤が起る確率は 2α になってしまう。そのため、1 つの遺伝子座の 2 つの対立遺伝子を調べる研究では、どちらかの対立遺伝子について計算した Z の値が $z_{0.025}$ を越えた場合に、有意水準 0.05 で当該対立遺伝子が疾患と関連すると結論する。

・サンプルサイズの決定

H_0 の下で期待される $\hat{P}_1 - \hat{P}_2$ を考える。D の集団内頻度を p 、d の集団内遺伝子頻度を $1-p$ とし、これらの接合体頻度は集団全体でハーディ・ワインバーク平衡状態にあると仮定する (すなわち、DD の頻度は p^2 、Dd の頻度は $2p(1-p)$ 、dd の頻度は $(1-p)^2$ である)。

患者を n 個体、対照を m 個体サンプルしたとすると、 H_0 の下で

$\hat{P}_1 - \hat{P}_2$ は、平均 0、標準偏差

$$\sigma_{(n)} = \sqrt{p(1-p)\left(\frac{1}{2n} + \frac{1}{2m}\right)}$$
 の正規分布に従う。次に、 H_1

の下で期待される $\hat{P}_1 - \hat{P}_2$ を考える。D を感受性対立遺伝子とする。遺伝子型 dd の個体の浸透率 (患者になる確率) を f_0 とし

(complex disease では一般に $f_0 > 0$ である)、遺伝子型 DD の個体および遺伝子型 Dd の個体の浸透率をそれぞれ f_2 および f_1 とおく (一般に $1 > f_2 \geq f_1 \geq f_0 > 0$ である)。集団の罹患率 K は $K = p^2 f_2 + 2p(1-p) f_1 + (1-p)^2 f_0$ であらわされる。患者群の中から一

個体を選び、その遺伝子型を調べたときに、その個体が DD、Dd、

dd である確率は、それぞれ、 $\frac{p^2 f_2}{K}$ 、 $\frac{2p(1-p) f_1}{K}$ 、 $\frac{(1-p)^2 f_0}{K}$

である。また、対照群の中から一個体を選び、その遺伝子型を調べたときに、その個体が DD、Dd、dd である確率は、それぞれ、

$\frac{p^2(1-f_2)}{1-K}$ 、 $\frac{2p(1-p)(1-f_1)}{1-K}$ 、 $\frac{(1-p)^2(1-f_0)}{1-K}$ である。

H_1 の下で、患者群中での D の頻度が $P_{1(a)} = \frac{p^2 f_2 + p(1-p) f_1}{K}$ 、

対 照 群 中 で の D の 頻 度 が

$P_{2(a)} = \frac{p^2(1-f_2) + p(1-p)(1-f_1)}{1-K}$ となることに注意して、

H_1 の下で $\hat{P}_1 - \hat{P}_2$ は、平均 $P_{1(a)} - P_{2(a)}$ 、標準偏差

$\sigma_{(a)} = \sqrt{\frac{P_{1(a)}(1-P_{1(a)})}{2n} + \frac{P_{2(a)}(1-P_{2(a)})}{2m}}$ の正規分布に従う。第

一種の過誤 (有意水準) ・と第 2 種の過誤 ・に対して、 $z_\alpha \sigma_{(n)} = P_{1(a)} - P_{2(a)} + z_{1-\beta} \sigma_{(a)}$ なる関係式が成り立つ。有意水準 ・で片側検定を行なった場合に、 $1-\cdot$ の検出力を達成するのに必要なサンプルサイズ n は、 $m = kn$ とおけば、以下の式で求められる。

$$n = \frac{1}{2} \left[\frac{z_{\alpha} \sqrt{p(1-p) \left(1 + \frac{1}{k}\right)} - z_{1-\beta} \sqrt{P_{1(a)}(1 - P_{1(a)}) + \frac{P_{2(a)}(1 - P_{2(a)})}{k}}}{P_{1(a)} - P_{2(a)}} \right]^2$$

D と d の両方を検定するので、有意水準を 0.05 に設定するには、 $z_{\alpha} = z_{\alpha/2} = 1.96$ 、検出力を 0.8 に設定するのであれば $z_{1-\beta} = z_{1-\beta} = -0.84$ ($z_{1-\beta} = 0.84$ であることに注意されたい) を採用する。なお、サンプルサイズが決まっている場合の研究の検出力は、 $z_{1-\beta} = [z_{\alpha} \sigma_{(n)} - (P_{1(a)} - P_{2(a)})] / \sigma_{(a)}$ であらわされる。以上の成果は論文として発表した[2]。

(ii) 間接的ゲノムワイド関連解析を行うための適切な研究デザイン (サンプルサイズ) の設定

浸透率の低い多因子疾患の感受性遺伝子を同定する方法の一つに、ゲノムワイドに多数の多型マーカーを調べる連鎖不平衡検定法がある。患者-対照関連研究における連鎖不平衡検定法とは、ある多型マーカーと疾患感受性遺伝子とが連鎖不平衡状態にある場合には、患者群と対照群の当該マーカー対立遺伝子頻度に差が生じるという原理に基づいている。連鎖不平衡の強さは、世代の経過とともに減衰し、その減衰率はマーカーと疾患感受性遺伝子座との間の遺伝距離 (組換え率) に比例するため、できるだけ多くのマーカーを設定して (疾患遺伝子との期待遺伝距離を小さくして) 研究を行なう必要がある。しかし、多数のマーカーについて統計検定を繰り返せば多くの偽陽性が起こるため、検定した回数で有意水準を調整する Bonferroni の補正が一般的に行われる。Bonferroni の補正を行なうと、マーカー数を増やせば必ずしも連鎖不平衡 (マーカーと疾患との関連) が検出されるとは限らず、調べるべき最適なマーカー数 (最適なマーカー密度) が存在する。また、ゲノムワイドに行なう連鎖不平衡検定法では、SNP マーカーやマイクロサテライトマーカーが用いられる。これらの要因を考慮に入れ、決定論的方程式を用いて世代の経過に伴う連鎖不平衡の減衰をモデル化し、疾患感受性遺伝子とマーカーとの間の連鎖不平衡が、カイ 2 乗検定によって検出される確率、すなわちゲノムワイド連鎖不平衡検定の検出力の計算を行なった。

・遺伝学的モデルの構築

疾患感受性遺伝子座には、高リスク対立遺伝子 D と低リスク対立遺伝子 d の 2 個が存在し、それぞれの対立遺伝子頻度を p と $1-p$ とおく。マーカー遺伝子座には m 個の対立遺伝子 (i 番目の対立遺伝子を M_i と呼ぶ) が存在し、 M_i の頻度を q_i とおく。集団はハーディー・ワインバーク平衡状態にあり、これらの対立遺伝子頻度は世代の経過により変化しないものとする。疾患感受性遺伝子座とマーカー遺伝子座との間の組換え率は θ とする。世代 t において D と M_i から構成されるハプロタイプ頻度を $H_{Di}(t)$ とおく。同様に、世代 t において d と M_i から構成されるハプロタイプ頻度を $H_{di}(t)$ とおく。したがって、 $q_i = H_{Di}(t) + H_{di}(t)$ が常に成立する。世代当たりの突然変異率は、マイクロサテライトマーカーの場合が $10^{-2} \sim 10^{-5}$ であり、SNP の場合が $10^{-7} \sim 10^{-8}$ と考えられるため、本研究ではマイクロサテライトマーカーについてのみ突然変異を考慮する。なお、疾患感受性遺伝子座での突然変異は考慮しない。マイクロサテライトマーカーの突然変異に関しては 1 段階ステップ突然変異モデル (SSM) を採用し、 M_i は、次に大きい M_{i+1} 、または次に小さい M_{i-1} にそれぞれ確率 u_i で突然変異すると仮定する (図 1)。最も小さな M_1 は、確率 u_1 で一つ大きな M_2 にのみ突然変異し、最も大きな M_m は、確率 u_m で一つ小さな M_{m-1} にのみ突然変異する。

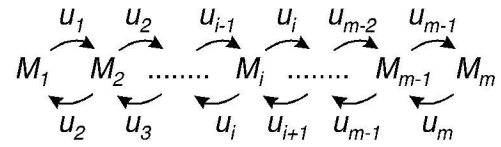


図1. 1段階ステップ突然変異モデル(SSM)

以上の仮定より、ハプロタイプ頻度変化に関する次の決定論的方程式 (頻度変化の漸化式) がえられる。詳細は記さないが、初期頻度 ($t = 0$) さえ与えれば、任意の世代におけるハプロタイプ頻度を計算することが可能である。SNP マーカー ($m = 2$ および $u_i = 0$ が成立) に対しては、漸化式は以下になる。

$$H_{D1}(t) = (H_{D1}(0) - pq_1)(1 - \theta)^t + pq_1,$$

$$H_{D2}(t) = (H_{D2}(0) - pq_2)(1 - \theta)^t + pq_2 = p - H_{D1}(t),$$

$$H_{d1}(t) = (H_{d1}(0) - (1-p)q_1)(1 - \theta)^t + (1-p)q_1,$$

$$H_{d2}(t) = (H_{d2}(0) - (1-p)q_2)(1 - \theta)^t + (1-p)q_2 = 1 - p - H_{d1}(t)$$

ここで $q_1 + q_2 = 1$ である。

マーカー遺伝子座における対立遺伝子の平衡頻度分布

マーカー遺伝子座における対立遺伝子頻度はマーカー毎に大きく異なるが、3734 個の (CA)_n リピートマーカーの対立遺伝子頻度分布から、平均的な対立遺伝子頻度分布 (これを平衡頻度分布とみなす) が明らかにされている (<http://www.well.ox.ac.uk/~mfarrall/microsatellite.html>) においてデータが公開されている)。1 段階ステップ突然変異モデルにおいて、マイクロサテライトマーカーの対立遺伝子数および対立遺伝子頻度は変化しないように仮定するため、対立遺伝子の平衡頻度分布と突然変異率との間に以下のような関係式が成立する必要がある。

$$\bar{q}_1 u_1 + \bar{q}_m u_m + \sum_{i=2}^{m-1} 2\bar{q}_i u_i = u.$$

ここで、 M_i の平衡頻度分布は $\bar{q}_i$ 、マーカー遺伝子座における総突然変異率は u である。これより、 M_i の突然変異率は

$$u_i = \frac{u}{2(m-1)\bar{q}_i} \text{ と設定される。} (\text{CA})_n \text{ リピートマーカーの}$$

中で最も多く観察されるものは対立遺伝子を 6 個含むものなので、本研究では、マイクロサテライトマーカーは 6 個の対立遺伝子を有し、それぞれの対立遺伝子平衡頻度は、 $\bar{q}_1 = 0.160$ 、 $\bar{q}_2 = 0.199$ 、 $\bar{q}_3 = 0.214$ 、 $\bar{q}_4 = 0.205$ 、 $\bar{q}_5 = 0.158$ 、 $\bar{q}_6 = 0.064$ とする。SNP マーカーは淘汰上中立であり、SNP マーカーでは再突然変異が起こらないと仮定すると、ある SNP 対立遺伝子の頻度が q から $q+dq$ の間にある確率を、

$$C_A \left(\frac{1}{q} + \frac{1}{1-q} \right) dq \text{ で表すことができる。ここで、} C_A \text{ は規格化}$$

定数であり、 $C_A^{-1} = \int_0^1 \left(\frac{1}{q} + \frac{1}{1-q} \right) dq$ である。これを SNP マー
 カーでの平衡頻度分布とみなす。ゲノムワイドな連鎖不平衡検定
 研究では、マイナー対立遺伝子頻度が高い SNP が使用されるので、
 その頻度のカットオフ値を Q とおく。 Q 以上の対立遺伝子頻度を
 有する SNP がマーカーとしてランダムに選ばれりと仮定すると、
 ある SNP マーカの対立遺伝子頻度が q から $q+dq$ の間にある確
 率は、 $C_B \left(\frac{1}{q} + \frac{1}{1-q} \right) dq$ で表すことができる。ここで、

$$C_B^{-1} = \int_Q^{1-Q} \left(\frac{1}{q} + \frac{1}{1-q} \right) dq \text{ である。}$$

・疾患モデル

患者一対照研究により、疾患遺伝子探索を目的とする連鎖不平衡
 検定を行なうとする。遺伝子型 DD, Dd, dd に対する浸透率を f_2, f_1, f_0 とおく（一般的に、 $1 \geq f_2 \geq f_1 \geq f_0 \geq 0$ が成立する）。
 患者または対照という表現型が与えられた条件下での各遺伝子
 型頻度は、

$$P(DD|case) = p^2 f_2 / e,$$

$$P(Dd|case) = 2p(1-p)f_1 / e,$$

$$P(dd|case) = (1-p)^2 f_0 / e,$$

$$P(DD|control) = p^2 (1-f_2) / (1-e),$$

$$P(Dd|control) = 2p(1-p)(1-f_1) / (1-e),$$

$$P(dd|control) = (1-p)^2 (1-f_0) / (1-e)$$

で与えられる。ここで、 e は疾患の罹患率
 $p^2 f_2 + 2p(1-p)f_1 + (1-p)^2 f_0$ を示している。ベイズの定理
 より、世代 t において患者が $M_i M_j$ という遺伝子型をとる確率は

$$P(M_i M_j | case, t) = \frac{1}{e} \{ f_2 H_{Di}(t)^2 + 2f_1 H_{Di}(t) H_{Dj}(t) + f_0 H_{di}(t)^2 \}$$

、世代 t において患者が $M_i M_j$ (ここで、 $i \neq j$) という遺伝子型
 をとる確率は

$$P(M_i M_j | case, t) =$$

$$\frac{1}{e} \{ 2f_2 H_{Di}(t) H_{Dj}(t) + 2f_1 (H_{Di}(t) H_{Dj}(t) + H_{di}(t) H_{dj}(t)) + 2f_0 H_{di}(t) H_{dj}(t) \}$$

である。同様に、世代 t において対照が $M_i M_j$ という遺伝子型を
 とる確率は

$$P(M_i M_j | control, t) =$$

$$\frac{1}{1-e} \{ (1-f_2) H_{Di}(t)^2 + 2(1-f_1) H_{Di}(t) H_{Dj}(t) + (1-f_0) H_{di}(t)^2 \}$$

、世代 t において対照が $M_i M_j$ (ここで、 $i \neq j$) という遺伝子型
 をとる確率は

$$P(M_i M_j | control, t) =$$

$$\frac{1}{1-e} \{ 2(1-f_2) H_{Di}(t) H_{Dj}(t) + 2(1-f_1) (H_{Di}(t) H_{Dj}(t) + H_{di}(t) H_{dj}(t)) + 2(1-f_0) H_{di}(t) H_{dj}(t) \}$$

である。

・検定統計量

患者一対照研究で連鎖不平衡検定を行なう場合は、あるマーカー
 遺伝子座における対立遺伝子頻度分布を患者群と対照群とで比
 較することになる。 m 個の対立遺伝子を有するマーカーを調べれ
 ば、表 3 のような $m \times 2$ の分割表がえられる。

表 3. m 個の対立遺伝子を有するマーカー
 から選られる分割表

	case	control
M_1	x_1	y_1
M_i	x_i	y_i
M_m	x_m	y_m
Total	1	1

もし患者と対照とを N 個体ずつサンプルするとすると、 χ^2 統計

$$\text{量 } X^2 = 2N \sum_{i=1}^m \frac{(\tilde{x}_i - \tilde{y}_i)^2}{\tilde{x}_i + \tilde{y}_i} \text{ を用いて、帰無仮説「每一マーカー対}$$

立遺伝子頻度に関して患者群と対照群との間に差は無い（調べた
 マーカーと疾患感受性遺伝子との間に連鎖不平衡は存在しない）
 を検定することができる。ここで、 $\tilde{x}_i$ と $\tilde{y}_i$ はそれぞれ

case 群と control 群で観察された M_i の頻度を表す。帰無仮説下
 で、 X^2 は近似的に自由度 $m-1$ の χ^2 分布に従うと期待される。対立
 仮説「マーカー対立遺伝子頻度に関して患者群と対照群との間に
 差がある（調べたマーカーと疾患感受性遺伝子との間に連鎖不平
 衡が存在する）」の下では、 X^2 は近似的に非心パラメタが

$$\gamma = 2NG^2 = 2N \sum_{i=1}^m \frac{(x_i - y_i)^2}{x_i + y_i}, \text{ 自由度 } m-1 \text{ の } \chi^2 \text{ 分布に従う}$$

と期待される。世代 t 後の非心パラメタは

$$\gamma(t) = 2NG^2(t) = 2N \sum_{i=1}^m \frac{(x_i(t) - y_i(t))^2}{x_i(t) + y_i(t)}$$

と $y_i(t)$ はそれぞれ

$$\left(P(M_i M_i | case, t) + \sum_{j=1}^m P(M_i M_j | case, t) \right) / 2 \quad \text{と}$$

$$\left(P(M_i M_i | control, t) + \sum_{j=1}^m P(M_i M_j | control, t) \right) / 2 \quad \text{で}$$

ある。

簡単のため、 $t=0$ においては、マーカー対立遺伝子 M_k のみがハイリスク対立遺伝子 D と正の連鎖不平衡状態にあると仮定する。数学的には、 $H_{Dk}(0) = p$ 、 $H_i(0) = 0$ ($i \neq k$)、 $H_{dk}(0) = q_k - p$ 、 $H_{di}(0) = q_i$ ($i \neq k$) で表される。当然、 q_k は p より大きい必要がある。したがって、本研究はハイリスク対立遺伝子が比較的低頻度で存在する場合のみしか扱えないことになる。 $t=0$ において、あるマーカー対立遺伝子がハイリスク対立遺伝子と正の連鎖不平衡状態にある確率は、そのマーカー対立遺伝子頻度に比例すると仮定すれば (M_k が D と正の連鎖不平衡にある確率は q_k に比例すれば)、マイクロサテライトマーカーの場合の $G(t)$ の期待

値は $G^2(t) = \sum_{k=1}^m q_k G_k^2(t)$ で表される。ここで、 $G_k(t)$ は $t=0$

において M_k が D と正の連鎖不平衡にある場合の t 世代後の G の値である。SNP マーカーでも同様に、ハイリスク対立遺伝子と正の連鎖不平衡にある確率がマーカー対立遺伝子頻度に依存するとすると、頻度が Q 以上の SNP マーカー中からランダムに選ばれたマーカーが、 $t=0$ において対立遺伝子頻度が q と $q+dq$ の間の値をとり、ハイリスク対立遺伝子と正の連鎖不平衡状態にある

確率は、 $C_c q \left(\frac{1}{q} + \frac{1}{1-q} \right) dq$ で表される。ここで、

$C_c^{-1} = \int_0^Q q \left(\frac{1}{q} + \frac{1}{1-q} \right) dq$ である。SNP では、 M_1 と M_2 のうち

M_1 のみが D と正の連鎖不平衡にあると仮定しても一般性を失

わないので、 $G^2(t) = C_c^{-1} \int_0^Q G_1^2(t) q \left(\frac{1}{q} + \frac{1}{1-q} \right) dq$ が成

立する。

・検出力の計算

ゲノムワイド連鎖不平衡検定では、多数の遺伝マーカーで検定が行なわれる。ここで問題となるのが検定の多重性である。検定を多数回行なうことで第一種の過誤が起こる確率が上昇するため、一般的には、Bonferroni の補正を行なう必要がある。いま n 個の遺伝マーカーを調べるとすると、言い換えれば、 n 回の独立な検定を行うと、研究の有意水準・は $0.05/n$ に設定されるべきである。 n 個の遺伝マーカーをゲノムの L cM 上に均等に配置させたとすると (例えば、ゲノム全体であれば L は 3000 である)、疾患感受性遺伝子とそれに最も近接した遺伝マーカーとの間の遺伝距離はせいぜい $L/2n$ cM である (図 2)。

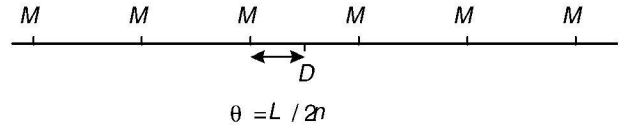


図2. 疾患感受性遺伝子座と遺伝マーカーとの間の遺伝距離

遺伝マーカー間にも連鎖不平衡は存在するため、検定の独立性は成り立たないが、その点についてはここでは考慮しない。また本研究では、疾患感受性遺伝子に最も近接するマーカーで期待される検出力のみを解析対象とし、他のマーカーでの検出力については考察しない。有意水準・で検定を行なった場合の検出力 $1-\beta$ は、近似的に $\text{Prob}[\chi^2_{m-1}(\gamma) \geq \chi^2_{m-1, 1-\alpha}(\gamma)]$ により計算される。以上の成果は、論文として発表した[1, 4, 6]。

(iii) 最も効率よく疾患感受性遺伝子を同定しうる (検出力の高い) SNP マーカー数とサンプルサイズを求めるアルゴリズムの開発

Saito and Kamatani (2002)が開発した2段階による段階的絞込み法を応用し、研究予算が限定されている中で、もっとも高い検出力を達成するための、表現型タイピング数と遺伝子型タイピング数を算出した。第1段階で必要なサンプルサイズと有意水準を N_1 および α_1 とし、第2段階で必要なサンプルサイズと有意水準を N_2 および α_2 とする。また、ゲノム中の候補 SNP 総数を n とし、実際に調べる SNP 数を m とする。第1段階の検出力を P_1 、第2段階の検出力を P_2 とする。さらに、遺伝子タイピングコスト、サンプル収集コスト、研究の総資金を、それぞれ C_g 、 C_p 、 B とおく。このような状況下では、 $Q = mP_1P_2/n$ を最大化することが研究の最大効率となる。そこで、研究全体の有意水準を 0.05 に保ち、 Q が最大となるように、

$$0.05/m = \alpha_1\alpha_2 \quad (1)$$

$$mC_g(N_1 + \alpha_1N_2) + C_p(N_1 + N_2) \leq B \quad (2)$$

を満たすパラメタ範囲で、マーカー数およびサンプルサイズを求めるアルゴリズム開発を行った。以上の成果は、論文として発表した[8]。

(iv) 骨髄移植後移植片対宿主病の原因となるマイナー組織適合性抗原 (mHa) を検出するための遺伝統計学的手法の開発

骨髄移植ペアの HLA 遺伝子のタイプを完全にマッチさせた場合であっても、骨髄移植後移植片対宿主病が起こることがある。この要因の一つとして、HLA 遺伝子によって提示される自己抗原に多型が存在し、その抗原提示能が移植ペア間で異なることが知られている。このような自己抗原は、マイナー組織適合性抗原 (mHa) とよばれ、これまでに多数の mHa が同定されている。しかしながら、それらの多くは分子生物学的手法によって同定されたものであり、統計遺伝学的手法によってものは少ない。そこで、骨髄移植後移植片対宿主病を起こしたペアと起こさなかったペアの遺伝子型を比較し、mHa 抗原を同定するための新規統計解析手法の開発を行った。以上の成果は、論文として発表した[5, 7]。

(2) 実在集団における SNP 間連鎖不平衡構造の解析と疾患遺伝子モデル研究

タイ人集団での CD36 遺伝子領域の連鎖不平衡構造を明らかにするとともに、CD36 遺伝子多型とマラリア重症化との関連を見出し報告した。具体的には、以下の通りである。CD36 遺伝子の多型スクリーニングによって 14 種類の多型を見出した。イントロン 3 内にある TG マイクロサテライトの 12 回繰返しアリルが、もっとも強く脳性マラリア抵抗性

と関連することを見出し、さらにこの 12 回繰返しアレルを有する個体では、CD36 遺伝子のエクソン 4 と 5 を欠くアインフォームが産生されないことを見出した。これらのエクソンは、マラリア感染赤血球との接着に必須のアミノ酸をコードすることが知られている。以上より、我々は「完全な CD36 分子を発現することがマラリア原虫の排除に重要な役割を果たし、その効率的な発現と関連する 12 回繰返し TG マイクロサテライトアレル自体、もしくはそのアレルと同一のハプロタイプ上にある未知の変異アレルが、脳性マラリアに対する抵抗性と関連する」という仮説を提唱した。以上の成果は、論文として発表した[3]。

<国内外での成果の位置づけ>

本研究の成果として発表した論文 8 編は、これまでに合計して 70 回以上他の国際誌上の論文に引用されており、特に、SNP 関連研究の研究デザインに関する論文の被引用回数が高いことから、国内外において評価されており、十分な成果をあげられたと考えている。

<達成できなかったこと、予想外の困難、その理由>

集団遺伝学理論に基づき、復帰突然変異の非存在下における、2SNP 間連鎖不平衡の程度を予測する理論式を解析的に導くことができなかった[文献4]。そのため、研究の検出力を計算するために、連鎖不平衡の期待値を用いることができず、計算時に与える必要が生じてしまった。これは、ともに 2 対立遺伝子を有する 2SNP 間の連鎖不平衡を理論予測するためには、4 種類の染色体の集団内頻度変化を記述する必要があるが(変数としては 3 つ)、遺伝的組換えと機会的遺伝子浮動の影響によって、染色体の種類数が 3 種類しか存在しない場合と 4 種類存在する場合とがあり、その両者を同時に扱う理論を構築することが困難であったためである。

<今後の課題>

統計解析手法の開発に関しては、case-control データに対する新規尤度比検定法の開発を行う。特に、感受性候補領域内の多数の SNP を調べる場合に、当該領域の連鎖不平衡の程度を考慮した尤度比検定法の開発を試みる。

連鎖不平衡の予測式に関しては、SNP サイト間の連鎖不平衡を把握するために、2 遺伝子座 2 対立遺伝子モデルを、拡散方程式を用いて数理的に解析する。数値解を求め、シミュレーションによって結果を確認する。重要なことは、2 つのサイトがともに多型的であるという条件下で解を与えることと、 $T=0$ の各ハプロタイプの初期頻度も確率的に与えることである。事前情報が全く無い条件下で、マッピングできる必要があるからである。さらに、ヒト集団で実際に観察される連鎖不平衡を再現しうるコンピュータ・シミュレーションを行い、連鎖不平衡の程度や遺伝距離との関係についてさらに詳細に解析する。

日本人を含めた複数の集団において、様々なゲノム領域の SNP を多数解析し、集団間の連鎖不平衡の差を理解し、さらに、日本人集団との比較から、日本人集団を対象に設定される連鎖不平衡検定用 SNP マーカーの他集団における利用可能性について検討する。

<研究期間の全成果公表リスト>

論文／プロシーディング(査読付きのものに限る)

[1] 0301182115 Ohashi J, Tokunaga K (2001) Power of genomewide association studies of complex disease genes: Statistical limitation of indirect approaches using SNP markers. *Journal of Human Genetics* 46: 478-482.

[2] 0301192119 Ohashi J, Yamamoto S, Tsuchiya N, Hatta Y, Komata T, Matsushita M, Tokunaga K (2001) Comparison of statistical power between 2×2 allele frequency and allele positivity tables in case-control studies of complex disease genes. *Annals of Human Genetics* 65: 197-206.

[3] 0301241749 Omi K, Ohashi J, Patarapotikul J, Hananantachai H, Naka I, Looareesuwan S, Tokunaga K (2003) CD36 polymorphism is associated with protection from cerebral malaria. *American Journal of Human Genetics* 72: 364-374.

[4] 0303261252 Ohashi J, Tokunaga K (2002) The expected power of genome-wide linkage disequilibrium testing using single nucleotide polymorphism markers for detecting a low-frequency disease variant. *Annals of Human Genetics* 66: 297-306.

[5] 0407081906 Ohashi J, Maruya E, Tokunaga K, Saji H (2003) Power of association test for detection minor histocompatibility gene causing graft-versus-host disease following bone marrow transplantation. *Journal of Human Genetics* 48: 502-507.

[6] 0407082039 Ohashi J, Tokunaga K (2003) The power of genome-wide linkage disequilibrium (LD) testing using microsatellite markers. *Journal of Human Genetics* 48: 487-491.

[7] 0408131759 Araki J, Ohashi J, Muramatsu M. (2004) Application of discordant sib-pair linkage analysis for mapping minor histocompatibility antigen loci in a novel graft-versus-host-disease model. *Tissue Antigens* 64: 243-250.

[8] 0601281856 Ohashi J, Clark AG (2005) Application of the stepwise focusing method to optimization of the cost-effectiveness of genome-wide association study with limited research budget for genotyping and phenotyping. *Annals of Human Genetics* 69: 323-328.