原　　著

SNPデータに基づく疾患感受性遺伝子同定のための多段階症例－対照研究デザインの特性

北　村　信　隆
新潟大学大学院医歯学総合研究科
情報科学・統計学分野
（主任：赤澤宏平教授）

Properties of Multistage Case - Control Association Study Designs Based on SNP Data for Identifying Disease Susceptibility Genes

Nobutaka KITAMURA
Division of Information Science and Biostatistics,
Niigata University Graduate School of Medical and Dental Sciences
(*Director: Prof. Kouhei AKAZAWA*)

Abstract

A multistage case - control association study is one of the study designs to efficiently identify disease susceptibility genes. The purpose of this study was to clarify the properties of multistage case - control association studies. For this purpose, new programs for n - stage designs (n being an arbitrary integer larger than 1) were developed to calculate statistical powers, positive predictive values (PPVs) and numbers of genotypings of replication - based analysis (RBA) and joint analysis (JA). Assuming that the numbers of alleles remaining at the final stage or the numbers of genotypings were fixed, the powers and PPVs of RBA and JA in three - stage designs were higher than those indicators in two - stage designs when a proportion of the sample size at the 1st stage ($\pi_{s,1}$) was smaller than 0.5. On the other hand, when $\pi_{s,1}$ was larger than 0.5, the indicators in the two - stage designs were higher than those in the three - stage designs. This tendency was more remarkable the closer the proportions of sample sizes of the 2nd and 3rd stages were and was more remarkable in JA than in RBA. It may be useful to conduct a prior simulation under

Reprint requests to: Nobutaka KITAMURA
Division of Information Science and Biostatistics
Niigata University Graduate School of
Medical and Dental Sciences
1 - 754 Asahimachi - dori Chuo - ku,
Niigata 951 - 8520 Japan

別刷請求先：〒951 - 8520　新潟市中央区旭町通 1 - 754
新潟大学医歯学総合病院医療情報部　　北 村 信 隆

various conditions by our programs for multistage designs at the planning stage of a genome-wide association study.

Key words: case-control association study, multistage design, single nucleotide polymorphisms (SNPs), replication-based analysis, joint analysis

はじめに

1塩基多型（single nucleotide polymorphism, SNP）やマイクロサテライトマーカー等のDNA多型データを用いた症例ー対照関連解析は，疾患感受性遺伝子を見出す解析法として頻繁に用いられる．SNPベースの症例ー対照関連解析は，疾患群と対照群を用いて，どのSNPが疾患群と有意に関連しているのかを遺伝統計学的に調べ，最終的には有意なSNPの近傍に位置する疾患感受性遺伝子を同定する方法である[1)2)]．SNP解析法の向上に伴い，現在では1症例につき数十万個のSNPデータを迅速に採取できるようになった．仮に，100,000個のSNP解析を1,000例（疾患群500例，対照群500例）で行なうとすると，研究全体で1億SNPのデータ採取が必要となる．SNP解析機器の性能が向上したとはいえ，SNP解析には多額の費用がかかる．そこで，遺伝統計学の分野では，より少ない検体数で，高い検出力を保ちつつ疾患関連遺伝子を検出する方法が議論されてきた．多段階デザイン（Multistage design）による関連解析法もそのひとつである．この解析法では，症例数を複数のステージに分け，第1ステージである程度有意なSNPをスクリーニングした後，第2ステージ以降で疾患感受性遺伝子かどうかの確証的な検定を行なう[3)-7)]．実際のSNP解析で用いられる多段階関連解析法には，Replication-based analysis（RBA）とJoint analysis（JA）とがある[8)9)]．

多段階関連解析で用いられる研究デザインの多くは2段階法であるが，最近，極めて多数の候補因子を対象として，3段階法を用いた疾患感受性遺伝子同定の研究が行なわれた[10)11)]．Prenticeら[10)]は，冠動脈心疾患，脳血管障害ならびに乳癌に感受性があり，同時にそれらの疾患に及ぼす閉経後のホルモン併用療法の影響の強さにも関連する多次元的な感受性遺伝子を同定する研究を3段階関連解析法で解析した．同研究計画は，まず第1ステージで，ベースライン時にホルモン単独療法を行っていない症例と対照を用いて，DNAプーリング法により250,000タグSNPsから2,500SNPsまで絞り込む．次に，第2ステージでベースライン時にホルモン単独療法を行っていない症例と対照を用いた関連解析をさらに行い，2,500SNPsから50SNPsに絞り込む．さらに，第3ステージでホルモン併用療法の無作為化比較臨床試験用コホートの症例と対照を用いた関連解析によって，2.5個の多次元的な感受性SNPsの検出を行うというデザインである．またKathiresanら[11)]は，互いに独立な3つのコホートを用いて3段階法により疾患感受性遺伝子に関与するSNPsの再現性を検証した．Kathiresanら[11)]の研究では，血中脂質レベルを表すHDLコレステロール，LDLコレステロールならびに中性脂肪に関連する感受性遺伝子を同定するために，まず第1ステージでFramingham Heart Study（FHS）コホートの家族ベースの症例に関する100,000個のSNPsデータベースを用いて287SNPsまで候補SNPsの絞り込みが行われ，ついで第2ステージでFHSと関連のない症例セットを用いて40SNPsが選択され，さらに第3ステージで別のコホート集団を用いた感受性SNPsの同定が行われた．

一方，上述のように，3段階以上のスクリーニングが必要な関連解析法があるにもかかわらず，3段階法による疾患感受性遺伝子の統計学的検出力や陽性反応適中度（PPV）に関する研究は国際的にも殆ど行われていない．

そこで本研究では，多段階デザインにおけるReplication-based analysis（RBA）とJoint analysis（JA）について，種々の条件下での検出力なら

びにPPVの特性を汎用的なプログラムを開発し詳細に検討した．

方　法

1．RBAとJAにおける棄却域の上限値（もしくは下限値）と検出力の求め方

1）n段階法による replication - based analysis (RBA)

症例－対照研究の症例群の合計数と対照群の合計数をそれぞれN，第kステージにおける sample size の割合を$\pi_{s,k}$（$k=1,\cdots,n$）とする．母集団における疾患感受性アリルの症例群，対照群の遺伝子型頻度の真値を，それぞれ，p'，pとして，標本から得られたそれらの推定値をそれぞれ，$\hat{p}'$，$\hat{p}$とする．本論文では，遺伝子型モデルとして，allelic model を仮定する．

ここで第kステージにおいては，第1から第$k-1$（$1 \leqq k \leqq n$）番目の各ステージまでに用いなかった症例だけを使用して検定統計量z_kを求め，棄却域の上限値（または下限値）C_kと比較する．この時，

$$z_k = \frac{\hat{p}' - \hat{p}}{\sqrt{\left[\hat{p}'(1-\hat{p}') + \hat{p}(1-\hat{p})\right]/2N\pi_{s,k}}}$$

は，帰無仮説，すなわち$p'=p$の下で標準正規分布に従う．

一方，対立仮説の下では，検定統計量z_kはNが十分に大きい場合，平均

$$\mu_k = \frac{p' - p}{\sqrt{\left(p'(1-p') + p(1-p)\right)/2N\pi_{s,k}}}、$$

分散1の正規分布に従う．

ここで，$\Phi[\mathbf{x}]$を標準正規分布の分布関数とし，第kステージで選択されるアリルの割合を$\pi_{m,k}$（mは marker の略，$k=1,\cdots,n-1$）とすると，

$$\pi_{m,k} = P\left\{z_k \middle| |z_k| > C_k\right\} \quad (k=1,\cdots,n-1)$$

となり，両側検定の場合，$C_1 = \Phi^{-1}[1-\pi_{m,1}/2]$，$C_k = \Phi^{-1}[1-\pi_{m,k}/2]$として求められる．ただし，$\Phi^{-1}[\mathbf{x}]$は，累積確率$\mathbf{x}$に対応する$z$値を求める関数である．

さらに実験全体の有意水準をα_{genome}，候補アリルの数をM，個々のアリルに関する比率の差の検定において，多重比較検定での有意水準補正後の有意水準をα_{marker}，第nステージにおける棄却域をC_nとすると，C_nは帰無仮説の下で

$$\pi_{m,n} = P\left\{z_n \middle| |z_n| > C_n, sign(z_1) = \cdots = sign(z_n)\right\}$$
$$= \frac{\alpha_{genome}}{P\left\{(z_1,\cdots,z_{n-1}) \middle| |z_1| > C_1,\cdots,|z_{n-1}| > C_{n-1}, sign(z_1) = \cdots = sign(z_{n-1})\right\} \times M}$$

を満たすように定める．

この時，RBAにおける第1ステージならびに第kステージ（$n \geqq 2$）の検出力$P_{r,n}$は以下のようになる．

$$P_1 = 1 - \Phi[C_1 - \mu_1] + \Phi[-C_1 - \mu_1]$$

$$P_{r,n} = \frac{\prod_{k=1}^{n}\left(1 - \Phi[C_k - \mu_k]\right) + \prod_{k=1}^{n}\Phi[-C_k - \mu_k]}{1 - \Phi[C_1 - \mu_1] + \Phi[-C_1 - \mu_1]}.$$

従って，n段階法における RBA 全体としての検出力は，$P_1 P_{r,n}$となる．

2）n段階法による joint analysis (JA)

第1ステージから第nステージの検定統計量の重み付け合計を

$$z_{j,k} = \sum_{k=1}^{n}\sqrt{\pi_{s,k}}\, z_k$$

とする．

ここで$z_k = a_k$（a_kは定数，$k=1,\cdots,n-1$）とすると，$z_{j,n}$は，平均

$$\mu_{j,n} = \frac{p' - p}{\sqrt{\left[p'(1-p') + p(1-p)\right]/(2N)}} + \sum_{k=1}^{n-1}\sqrt{\pi_{s,k}}(a_k - \mu_k)、$$

分散が$1 - \sum_{k=1}^{n-1}\pi_{s,k}$の正規分布を示す．

この時，帰無仮説において次の重積分

$$\begin{aligned}
P_{j,n} &= \int_{-\infty}^{-C(n-1)}\cdots\int_{-\infty}^{-C2}\int_{-\infty}^{-C1}\left\{P\left[z_{j,n} > C_{j,n} \middle| \bigcap_{k=1}^{n-1}(z_k = x_k)\right] + P\left[z_{j,n} < -C_{j,n} \middle| \bigcap_{k=1}^{n-1}(z_k = x_k)\right]\right\} \\
&\quad f(x_1 \| z_1| > C_1)\cdots f(x_{n-1} \| z_{n-1}| > C_{n-1})\,dx_1\cdots dx_{n-1} \\
&\quad + \int_{C(n-1)}^{\infty}\cdots\int_{C2}^{\infty}\int_{C1}^{\infty}\left\{P\left[z_{j,n} > C_{j,n} \middle| \bigcap_{k=1}^{n-1}(z_k = x_k)\right] + P\left[z_{j,n} < -C_{j,n} \middle| \bigcap_{k=1}^{n-1}(z_k = x_k)\right]\right\} \\
&\quad f(z_1 \| z_1| > C_1)\cdots f(z_{n-1} \| z_{n-1}| > C_{n-1})\,dx_1\cdots dx_{n-1} \\
&= \alpha_{marker} = \alpha_{genome}/M
\end{aligned}$$

を満たす$C_{j,n}$がJAの棄却域となり，$P_{j,n}$はn段階法によるJA全体の検出力となる．

さらに，評価値として陽性反応適中度（positive predictive value, PPV）を算出した．ここで候補アリルM個中の真のアリルの個数をm個とすると，第kステージで残る真陽性ならびに偽陽性の個数は以下のようになる．

第kステージの真陽性の個数をtp_kとすると，

$$tp_k = \begin{cases} m \times P_1 \times P_{r,k} & \text{（RBAの場合）} \\ m \times P_{J,k} & \text{（JAの場合）} \end{cases}.$$

第kステージの偽陽性の個数をfp_kとすると，

$$fp_k = (M - m) \times \prod_{k=1}^{n} \pi_{m,k}.$$

従って，第kステージにおける陽性反応適中度PPV_kは，$PPV_k = tp_k / (tp_k + fp_k)$となる．

また多段階デザインのコストに関連する指標として，各多段階デザインによるタイピング数を以下のように算出した．

1段階法のタイピング数$= M \times 2 \times N$

2段階法のタイピング数$= M \times 2 \times N \times \pi_{s,1} + M \times \pi_{m,1} \times 2 \times N \times (1 - \pi_{s,1})$

3段階法のタイピング数$= M \times 2 \times N \times \pi_{s,1} + M \times \pi_{m,1} \times 2 \times N \times \pi_{s,2} + M \times \pi_{m,1} \times \pi_{m,2} \times 2 \times N \times (1 - \pi_{s,1} - \pi_{s,2})$．

2．プログラム

前節で導出したRBAとJAにおける検出力を算出するプログラムを，数値解析用ソフトウエアMathematicaならびにRを用いて作成した．JAにおける棄却域の上限値または下限値を表すパラメータ$C_{j,n}$は解析的には陽に求めることは出来ない．従って本プログラムでは，$C_{j,n}$を逐次計算法によって求める．本プログラムの精度を確認するため，2段階法によるJAに関して，MathematicaならびにRによる数値計算結果とSkolら[8]の論文と比較し正しいことを確認した．

3．多段階法の特性を把握するためのシミュレーション

1段階法，2段階法，3段階法におけるRBAならびにJAの統計学的検出力とPPVの特性を把握するためのシミュレーションを以下の条件で行なった．

シミュレーション1.

3段階法において，$\pi_{s,1} = 0.4$，$0.1 \leqq \pi_{s,2} \leqq 0.5$でのRBAとJAの検出力ならびにタイピング数を算出した．同様に，$\pi_{s,2} = 0.4$，$0.1 \leqq \pi_{s,1} \leqq 0.5$ならびに$\pi_{s,3} = 0.4$，$0.1 \leqq \pi_{s,1} \leqq 0.5$におけるRBAとJAの検出力ならびにタイピング数を算出した．各ステージで選択されるアリルの割合は，$\pi_{m,1} = 0.01$，$\pi_{m,2} = 0.01$とした．その他のパラメータは，$N = 1{,}000$，$M = 500{,}000$，$\alpha_{genome} = 0.05$，allelic model，odds ratio（OR）$= 1.5$とした．

シミュレーション2.

症例数の配分割合を変化させたときのRBAとJAによる検出力，PPVならびにタイピング数を比較した．2段階法では，$0.1 \leqq \pi_{s,1} \leqq 0.9$，残りの症例数の割合を$\pi_{s,2}$とした．3段階法においては，$0.1 \leqq \pi_{s,1} \leqq 0.9$に設定して，残りの症例数を種々の割合に分割して$\pi_{s,2}$ならびに$\pi_{s,3}$とした．また2段階法ならびに3段階法の各ステージで選択されるアリルの割合の設定は，最終ステージに残るアリルの個数が等しくなる，もしくは，2つのデザインのタイピング数が等しくなるように設定した．前者の場合，2段階法では，$\pi_{m,1} = 0.0001$，3段階法では，$\pi_{m,1} = 0.01$，$\pi_{m,2} = 0.01$（すなわち$\pi_{m,1} \times \pi_{m,2} = 0.0001$となる）と設定した．後者の場合は，3段階法では$\pi_{m,1} = 0.01$，$\pi_{m,2} = 0.01$とし，そのときのタイピング数と等しくなるように2段階法での$\pi_{m,1}$を設定した．その他のパラメータの設定条件はシミュレーション1と同様とした．

シミュレーション3.

疾患群と対照群における症例数をそれぞれ500，1000ならびに2000とした場合のRBAとJAにおける検出力ならびにタイピング数を各デザイン間で比較した．症例数の配分方法は，$0.1 \leqq \pi_{s,1} \leqq 0.9$，残りの症例を2等分してそれぞれ第2ステージと第3ステージでの症例となるように$\pi_{s,2}$を定めた．その他のパラメータの設定条件はシミュレーション1と同様とした．

結　　果

1．3段階法における症例数の配分割合とRBAならびにJAの検出力との関連

図1は，第1，第2，第3のステージの症例割合をそれぞれ固定して，他のステージの症例割合を変化させたときの3段階法のRBAとJAの検出力とタイピング数の変化を示す．結果は，いずれの場合もJAがRBAよりも高い検出力を示し，また3段階法によるタイピング数は1段階法によるタイピング数よりもかなり小さい値を示した．$\pi_{s,1}=0.4$の場合は，$\pi_{s,2}$の上昇とともにRBAの検出力が上昇し，$\pi_{s,2}$が0.3のときに最大値をとりその後減少した．即ち，検出力曲線は上に凸のほぼ左右対称な曲線となった（**図1A**）．一方，JAの検出力曲線は，$\pi_{s,2}$の上昇とともに単調に上昇し，$\pi_{s,2}=0.5$で1段階法の検出力とほぼ同じ値となった．また，3段階法によるタイピング数は，$\pi_{s,2}$の上昇とは無関係で殆ど変化しなかった．$\pi_{s,2}=0.4$の場合，$\pi_{s,1}$の上昇によるRBAならびにJAの検出力の変化は，$\pi_{s,1}$を固定した場合と同じ変化を示した．一方，タイピング数は，$\pi_{s,1}$の上昇とともに直線的に上昇していた（**図1B**）．$\pi_{s,3}=0.4$については，RBAとJAはほぼ同様の変化を示し，$\pi_{s,1}$の上昇とともにRBAならびにJAの検出力は$\pi_{s,1}$が0.1から0.3にかけて上昇し，0.3のときに最大値をとった後減少した．検出力曲線は上に凸のほぼ左右対称な曲線となっていた．またタイピング数は$\pi_{s,2}$を固定した場合と同じ変化を示した（**図1C**）．

2．最終のステージに残す候補アリル数もしくは全体のタイピング数を一定とした場合の検出力，PPVならびにタイピング数について

表1は，最終のステージに残す候補アリルの数を一定に設定した条件下における各多段階デザインによる検出力，PPVならびにタイピング数を比較したものである．$\pi_{s,1}$が0.1から0.5の範囲では，RBAならびにJAによる3段階法の検出力ならびにPPVは2段階法よりも高い値を示す傾向が認められた．特に$\pi_{s,1}$が小さいほど，また，$\pi_{s,2}$と$\pi_{s,3}$が同じであるとき，その傾向が強くみられた．また，RBAよりもJAの方が，3段階法の検出力やPPVが2段階法のそれらよりも高くなる範囲が広くなっていた．一方，$\pi_{s,1}$が0.5以上の範囲では，RBAとJAの検出力ならびにPPVは，3段階法よりも2段階法の方が常に高い値を示していた．**表2**は，2段階法と3段階法のタイピング数を一定に設定した条件下における各多段階法による検出力ならびにPPVを比較したものである．**表1**と同様，$\pi_{s,1}$が小さい範囲では，RBAならびにJAによる3段階法の検出力ならびにPPVは2段階法よりも高い値を示す傾向が認められた．しかし，**表1**に比べてその傾向を示す$\pi_{s,1}$の範囲は狭くなり，3段階法と2段階法の検出力の差も小さくなっていた．

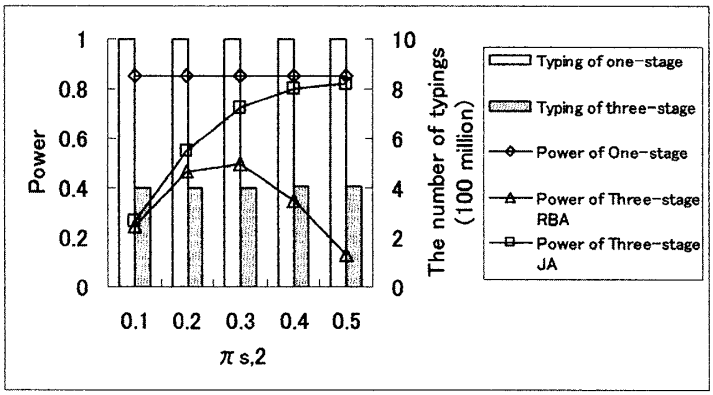
3．症例数の検出力，タイピング数に与える影響

図2は，$N=500$，1000，2000と変化させたときのRBA（**図2A**）とJA（**図2B**）の検出力ならびにタイピング数（**図2C**）の変化を示す．症例数の増加とともに各多段階デザインの検出力ならびにタイピング数は上昇していた．しかし，検出力の変化のパターンはいずれの症例数においても同様の傾向を示し，$\pi_{s,1}$が0.1から0.5の範囲においては，3段階法によるRBAやJAの検出力は2段階法の検出力よりも高く，$\pi_{s,1}$が0.5よりも大きい範囲では3段階法よりも2段階法の方が高い値を示していた．なお，JAにおいて，各症例数毎に多段階デザインの検出力とタイピング数を比較すると，いずれの症例数においても$\pi_{s,1}=0.3$における3段階法の検出力は，$\pi_{s,1}=0.5$における2段階法の検出力とほぼ同程度の検出力を示しているが，前者のタイピング数は後者のタイピング数の約60％に抑えられていた．

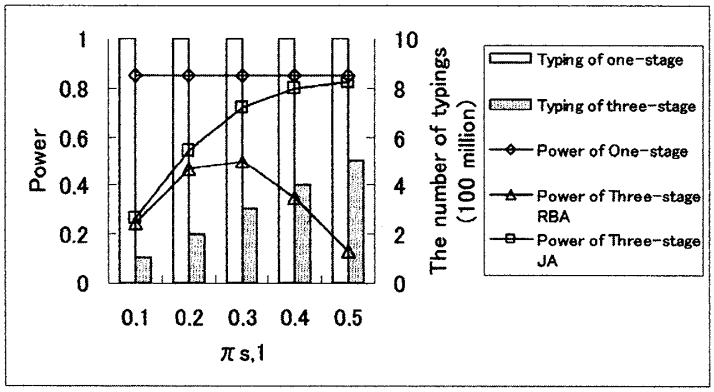
考　　察

効率的な遺伝子解析を行うためには，研究デザインや対象症例数等，実験を開始する前の研究計画の立案が重要となる．とりわけ，極めて多くの候補アリルの中から遺伝性疾患の感受性遺伝子を

A. $\pi_{s,1}$ を 0.4 に固定して、$\pi_{s,2}$ を 0.1〜0.5 に変化させた場合



B. $\pi_{s,2}$ を 0.4 に固定して、$\pi_{s,1}$ を 0.1〜0.5 に変化させた場合



C. $\pi_{s,3}$ を 0.4 に固定して、$\pi_{s,1}$ を 0.1〜0.5 に変化させた場合

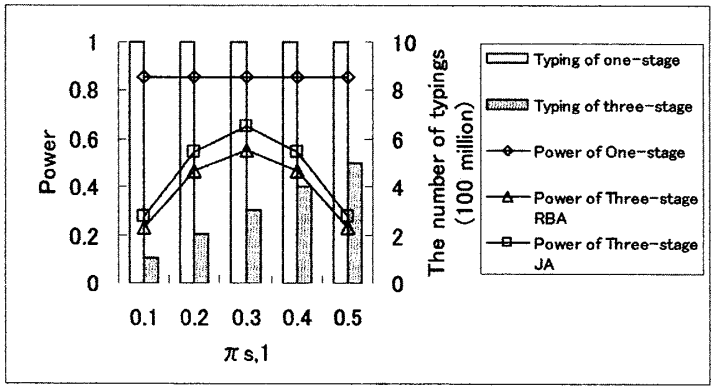


図1 種々の症例割合における3段階法による検出力ならびにタイピング数の変化

A: 3段階法を用いて，$\pi_{s,1} = 0.4$，$0.1 \leqq \pi_{s,2} \leqq 0.5$ での RBA と JA の検出力ならびにタイピング数を示す．B: $\pi_{s,2} = 0.4$，$0.1 \leqq \pi_{s,1} \leqq 0.5$ での RBA と JA の検出力ならびにタイピング数を示す．C: $\pi_{s,3} = 0.4$，$0.1 \leqq \pi_{s,1} \leqq 0.5$ での RBA と JA の検出力ならびにタイピング数を示す．各ステージで選択されるアリルの割合は，$\pi_{m,1} = 0.01$，$\pi_{m,2} = 0.01$ とした．その他のパラメータの設定条件は，症例群＝1000例，対照群＝1000例，$M =$ 500,000，$\alpha_{\text{genome}} = 0.05$，allelic model，odds ratio (OR) ＝1.5とした．

表1　種々の症例割合における1段階法，2段階法ならびに3段階法による検出力ならびにPPV（症例数，候補アリル数ならびに最終ステージに残す候補アリル数を一定にした場合）

Design	Proportion or samples			Power		PPV		Number of
	π s,1	π s,2	π s,3	RBA	JA	RBA	JA	Typings
One-stage				0.852		0.988		1E+09
Two-stage	0.1	0.9		0.030	0.030	0.752	0.752	1E+08
Three-stage	0.1	0.1	0.8	**0.082**	**0.083**	**0.943**	**0.892**	1E+08
	0.1	0.3	0.6	**0.228**	**0.234**	**0.979**	**0.959**	1E+08
	0.1	0.5	0.4	**0.231**	**0.274**	**0.979**	**0.965**	1E+08
	0.1	0.7	0.2	**0.116**	**0.278**	**0.959**	**0.965**	1E+08
Two-stage	0.2	0.8		0.148	0.149	0.937	0.937	2E+08
Three-stage	0.2	0.1	0.7	**0.173**	**0.174**	**0.972**	**0.946**	2E+08
	0.2	0.3	0.5	**0.460**	**0.490**	**0.989**	**0.980**	2E+08
	0.2	0.4	0.4	**0.466**	**0.548**	**0.989**	**0.982**	2E+08
	0.2	0.5	0.3	**0.388**	**0.569**	**0.987**	**0.983**	2E+08
	0.2	0.7	0.1	0.086	**0.574**	**0.945**	**0.983**	2E+08
Two-stage	0.3	0.7		0.340	0.342	0.971	0.972	3E+08
Three-stage	0.3	0.1	0.6	0.228	0.234	**0.979**	0.959	3E+08
	0.3	0.3	0.4	**0.556**	**0.652**	**0.991**	**0.985**	3E+08
	0.3	0.5	0.2	0.323	**0.747**	**0.985**	**0.987**	3E+08
Two-stage	0.4	0.6		0.538	0.545	0.982	0.982	4E+08
Three-stage	0.4	0.1	0.5	0.246	0.263	0.980	0.963	4E+08
	0.4	0.3	0.3	0.499	**0.724**	**0.990**	**0.986**	4E+08
	0.4	0.5	0.1	0.127	**0.822**	0.962	**0.988**	4E+08
Two-stage	0.5	0.5		0.674	0.705	0.985	0.986	5E+08
Three-stage	0.5	0.1	0.4	0.231	0.274	**0.979**	0.965	5E+08
	0.5	0.3	0.2	0.323	**0.747**	0.985	**0.987**	5E+08
Two-stage	0.6	0.4		0.705	0.800	0.986	0.988	6E+08
Three-stage	0.6	0.1	0.3	0.187	0.278	0.974	0.965	6E+08
	0.6	0.3	0.1	0.115	0.752	0.958	0.987	6E+08
Two-stage	0.7	0.3		0.607	0.841	0.984	0.988	7E+08
Three-stage	0.7	0.1	0.2	0.116	0.278	0.959	0.965	7E+08
Two-stage	0.8	0.2		0.391	0.852	0.975	0.988	8E+08
Three-stage	0.8	0.1	0.1	0.041	0.278	0.890	0.965	8E+08

表中の太字の数字は、2段階法より3段階法の方が大きい値を示す.

同定するための症例－対照研究においては，検出力の低下と偽陽性率の上昇を防ぎつつ，候補アリルの数と症例－対照数との積によって決まる総タイピング数を制限することが重要である．これらを実現するために多段階デザインによる関連解析が行われるようになった．

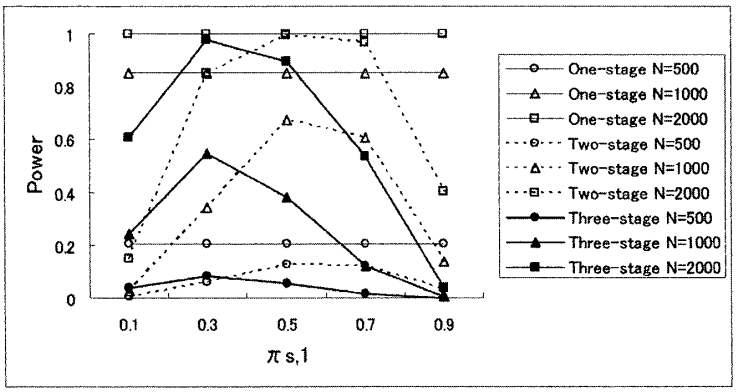
段階的関連解析の検出力やコストに影響し得る要因としては，利用可能な症例数と各ステージへの配分割合，各ステージで検証される候補アリルの数，ターゲットとなる遺伝子座位の領域の長さ，一アリル当たりのタイピングコスト，等があげられる．タイピング数と候補アリル数を一定にして1段階法と2段階法とを比較した場合，2段階法の方が少ない症例数で目標とする値に近い検出力を得ることが出来る[3]．一方，症例数と候補アリル数を一定にして1段階法と2段階法とを比較すると，2段階法の方が少ないタイピング数で目標に近い検出力が得られる[4)－6)]．さらに，タイピン

表2 種々の症例割合における1段階法，2段階法ならびに3段階法による検出力ならびにPPV（症例数、候補アリル数ならびに2段階法と3段階法のタイピング数を一定にした場合）

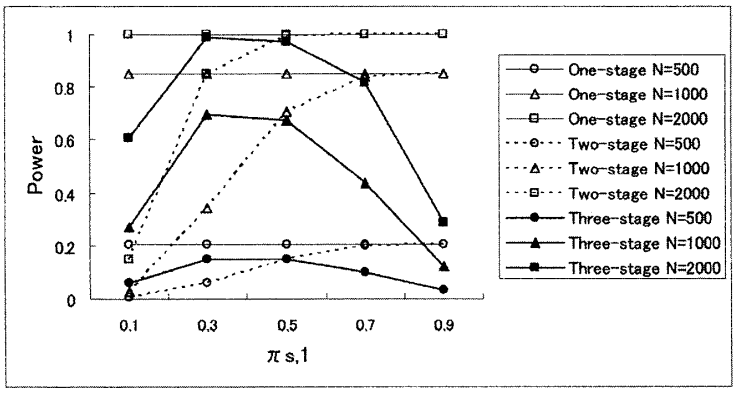
Design	Proportion or samples			Power		PPV		Number of
	$\pi_{s,1}$	$\pi_{s,2}$	$\pi_{s,3}$	RBA	JA	RBA	JA	Typings
One-stage				0.852		0.988		1E+09
Two-stage	0.1	0.9		0.109	0.109	0.916	0.916	1E+08
Three-stage	0.1	0.1	0.8	0.082	0.083	**0.943**	0.892	1E+08
Two-stage	0.1	0.9		0.148	0.149	0.937	0.937	1E+08
Three-stage	0.1	0.3	0.6	**0.228**	**0.234**	**0.979**	**0.959**	1E+08
Two-stage	0.1	0.9		0.219	0.220	0.956	0.956	1E+08
Three-stage	0.1	0.5	0.4	**0.231**	**0.274**	**0.979**	**0.965**	1E+08
Two-stage	0.1	0.9		0.236	0.237	0.959	0.959	1E+08
Three-stage	0.1	0.7	0.2	0.116	**0.278**	0.959	**0.965**	1E+08
Two-stage	0.2	0.8		0.350	0.352	0.972	0.972	2E+08
Three-stage	0.2	0.1	0.7	0.173	0.174	0.972	0.946	2E+08
Two-stage	0.2	0.8		0.459	0.465	0.979	0.979	2E+08
Three-stage	0.2	0.3	0.5	**0.460**	**0.490**	**0.989**	**0.980**	2E+08
Two-stage	0.2	0.8		0.490	0.497	0.980	0.980	2E+08
Three-stage	0.2	0.4	0.4	0.466	**0.548**	**0.989**	**0.982**	2E+08
Two-stage	0.2	0.8		0.490	0.497	0.980	0.980	2E+08
Three-stage	0.2	0.5	0.3	0.388	**0.569**	**0.987**	**0.983**	2E+08
Two-stage	0.2	0.8		0.549	0.559	0.982	0.982	2E+08
Three-stage	0.2	0.7	0.1	0.086	**0.574**	0.945	**0.983**	2E+08
Two-stage	0.3	0.7		0.584	0.599	0.983	0.984	3E+08
Three-stage	0.3	0.1	0.6	0.228	0.234	0.979	0.959	3E+08
Two-stage	0.3	0.7		0.635	0.659	0.985	0.985	3E+08
Three-stage	0.3	0.3	0.4	0.556	0.652	**0.991**	0.985	3E+08
Two-stage	0.3	0.7		0.689	0.730	0.986	0.986	3E+08
Three-stage	0.3	0.5	0.2	0.323	**0.747**	0.985	**0.987**	3E+08
Two-stage	0.4	0.6		0.702	0.759	0.986	0.987	4E+08
Three-stage	0.4	0.1	0.5	0.246	0.263	0.980	0.963	4E+08
Two-stage	0.4	0.6		0.709	0.806	0.986	0.988	4E+08
Three-stage	0.4	0.3	0.3	0.499	0.724	**0.990**	0.986	4E+08
Two-stage	0.4	0.6		0.700	0.822	0.986	0.988	4E+08
Three-stage	0.4	0.5	0.1	0.127	0.822	0.962	0.988	4E+08
Two-stage	0.5	0.5		0.672	0.830	0.985	0.988	5E+08
Three-stage	0.5	0.1	0.4	0.231	0.274	0.979	0.965	5E+08
Two-stage	0.5	0.5		0.614	0.844	0.984	0.988	5E+08
Three-stage	0.5	0.3	0.2	0.323	0.747	**0.985**	0.987	5E+08
Two-stage	0.6	0.4		0.517	0.850	0.981	0.988	6E+08
Three-stage	0.6	0.1	0.3	0.187	0.278	0.974	0.965	6E+08
Two-stage	0.6	0.4		0.428	0.852	0.977	0.988	6E+08
Three-stage	0.6	0.3	0.1	0.115	0.752	0.958	0.987	6E+08
Two-stage	0.7	0.3		0.297	0.852	0.967	0.988	7E+08
Three-stage	0.7	0.1	0.2	0.116	0.278	0.959	0.965	7E+08
Two-stage	0.8	0.2		0.104	0.852	0.912	0.988	8E+08
Three-stage	0.8	0.1	0.1	0.041	0.278	0.890	0.965	8E+08

表中の太字の数字は、2段階法より3段階法の方が大きい値を示す.

A.　N = 500, 1000, 2000 における各多段階デザインによる RBA の検出力



B.　N = 500, 1000, 2000 における各多段階デザインによる JA の検出力



C.　N = 500, 1000, 2000 における各多段階デザインによるタイピング数

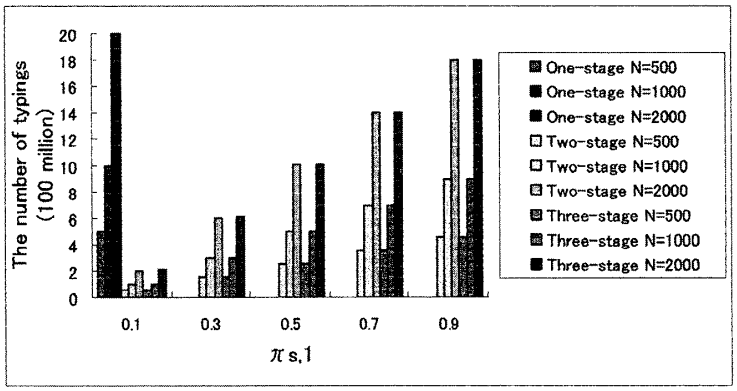


図 2　種々の症例数における 1 段階法，2 段階法ならびに 3 段階法による検出力とタイピング数の変化

症例群と対照群における症例数をそれぞれ，500，1000 ならびに 2000 とした場合における RBA（図 2A）と JA（図 2B）による検出力ならびにタイピング数（図 2C）の変化を示す．症例の分割の方法は，$\pi_{s,1}$ を 0.1 ～ 0.9 に設定して第 1 段階での症例とし，残りの症例を 2 等分してそれぞれ第 2 ステージと第 3 ステージでの症例となるように $\pi_{s,2}$ を定めた．その他のパラメータの設定条件は図 1 と同様とした．

グ数を一定にして1段階法と2段階法とを比較すると，2段階法の検出力が1段階法に較べて高くなる[7]．さらにSkolら[8]は，2段階法においてRBAとは異なる検定手法を提唱した．即ち，各ステージの症例数割合の平方根で重み付けした検定統計量の一次結合による新たな検定統計量を用いたJoint analysisである．Joint analysisは従来のReplication-based analysisよりも高い検出力が得られると報告した[8)9)]．

しかし段階的関連解析で用いられるデザインは主として2段階法であり，3段階以上の多段階デザインの統計学的性質に関する評価報告は極めて少ない[10)-12)]．Satoら[12)]は，第1ステージ終了後さらに疾患関連候補SNPの絞り込みを行う場合に，2段階法と3段階法のどちらがより適切であるかについて，モンテカルロシミュレーションによりPPVと検出力を指標として比較検討した．その結果，3段階法による検出力は2段階法の検出力よりもわずかに高かったと報告している．しかし同報告では第1ステージの症例数が940例（症例群188例，対照群752例）に固定されており，シミュレーションのパラメータ変数として各ステージの症例数の割合（$\pi_{s,1}$や$\pi_{s,2}$等）が用いられていないため，一般的にどのような条件下において3段階法の検出力が2段階法の検出力よりも高くなるかについては明らかにされていない．また著者らが調べたその他の3段階法に関する報告では，RBAおよびJAについて2段階法との比較は行われていない．

本研究では，我々の作成した多段階デザインプログラムを用いて，RBAならびにJAの検出力ならびにPPVに関して2段階法と3段階法との比較を行った．その結果，第1ステージにおける症例割合である$\pi_{s,1}$が0.5よりも小さい範囲において，3段階法による検出力ならびにPPVが2段階法よりも高くなる傾向があることが示された．さらにJAにおいては，$\pi_{s,1}=0.3$における3段階法の検出力が，$\pi_{s,1}=0.5$における2段階法の検出力とほぼ同程度の検出力を示しているにもかかわらず，前者のタイピング数は後者のタイピング数の約60％に抑えられていた．このようなことから，タイピングコストにも依存するが，第1ステージにおける症例数の割合が0.5よりも小さい場合には，3段階法の方がやや効率的になり得るものと思われた．

本研究により，2段階法だけでなく，3段階法によるJAの有用性も示唆された．今後，SNPに関するgenome-wide関連解析を行うに際し，本プログラムにより事前のシミュレーションを行うことが有用である．

結　　語

疾患感受性遺伝子に関する関連解析の計画を行うに際し，今回我々の作成した多段階デザインによるプログラムにより，種々の条件下におけるシミュレーションを行うことが可能となった．今後，本プログラムを用いて実際の疾患感受性遺伝子に関するデータベースを用いた検証を行い，個々の感受性遺伝子の特性に合わせたより精度の高い遺伝子同定法の開発を進める必要がある．

謝　辞

本研究を行うにあたり種々ご指導ご協力を賜りました新潟大学脳研究所附属生命科学リソース研究センターバイオリソース研究部門の桑野良三先生および宮下哲典先生，新潟大学医歯学総合病院危機管理室の鳥谷部真一先生，ならびに新潟大学医歯学総合病院医療情報部の赤澤宏平先生と他の皆様に深甚なる感謝の意を捧げます．

参考文献

1）The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-683, 2007.

2）Samani NJ, Erdmann J, Hall AS, Hengsternberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, König I R, Stevens SE, Szymczak S, Tregouet DA, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Path FR, Blankenberg S,

Baimforth AJ, Baessier A, Ball SG, Strom TM, Brænne I, Gieger C, Deloukas P, Tobin M D, Ziegler A, Thompson JR and Schunkert H: Genomewide association analysis of coronary artery disease. N Engl J Med 357;5 www.NejM. ORG: 443 - 453, 2007.

3) Satagopan JM, Verbel DA, Venkatraman ES, Offit KE and Begg CB: Two - stage designs for gene - disease association studies. Biometrics 58: 163 - 170, 2002.

4) Satagopan JM and Elston RC: Optimal two - stage genotyping in population - based association studies. Genet Epidemiol 25: 149 - 157, 2003.

5) Satagopan JM, Venkatraman ES and Begg CB: Two - stage designs for gene - disease association studies with sample size constraints. Biometrics 60: 589 - 597, 2004.

6) Thomas D, Xie R and Gebregziabher M: Two - stage sampling designs for gene association studies. Genet Epidemiol 27: 401 - 414, 2004.

7) Zehetmayer S, Bauer P and Posch M: Two - stage designs for experiments with a large number of hypotheses. Bioinformatics 21: 3771 - 3777, 2005.

8) Skol AD, Scott LJ, Abecasis GR and Boehnke M: Joint analysis is more efficient than replication - based analysis for two - stage genome - wide association studies. Nat Genet 38: 209 - 213, 2006.

9) Skol AD, Scott LJ, Abecasis GR and Boehnke M: Optimal designs for two - stage genome - wide association studies. Genet Epidemiol DOI 10.1002/gepi: 1 - 13, 2007.

10) Prentice RL and Lihong Qi: Aspects of the design and analysis of high - dimensional SNP studies for disease risk estimation. Biostatistics 7: 339 - 354, 2006.

11) Kathiresan S, Manning AK, Demissie S, D'Agostino RB, Surti A, Guiducci C, Gianniny L, Burtt NP, Melander O, Orho - Melander M, Arnett DK, Peloso GM, Oardovas JM and Cupples LA: A genome - wide association study for blood lipid phenotypes in the Framingham Heart Study. BMC Med Genet 8 (Supple 1): S17 doi: 10.1186/1471 - 2350 - 8 - S1 - S17. Available from: http://www.biomedcetral.com/1471 - 2350 /8/S1/S17 2007.

12) Sato Y, Suganami H, Hamada C, Yoshimura I, Yoshida T and Yoshimura K: Designing a multi - stage, SNP - based, genome screen for common diseases. J Hum Genet 49: 669 - 676, 2004.

(平成19年12月25日受付)