ゲノムワイド関連解析における薬物動態関連遺伝子同定のための新しいスクリーニング法

○長島 健悟[1], 佐藤 泰憲[2,3], Nan M. Laird[3]
[1] 城西大学 薬学部 薬科学科
[2] 千葉大学 医学部
[3] ハーバード大学 公衆衛生大学院 生物統計部門

A new statistical screening approach for finding pharmacokinetics-related genes in genome-wide studies

Kengo Nagashima[1], Yasunori Sato[2,3], Nan M. Laird[3]
[1] Department of Parmaceutical Technochemistry, Josai University
[2] School of Medicine, Chiba University
[3] Department of Biostatistics, Harvard School of Public Health

要旨

薬物に対するヒトの反応には，著しい個体差がみられる場合がある．この個体差を生む原因の一つとして，薬物の曝露量に影響する薬物動態関連分子の遺伝子多型が考えられる．薬物動態と関連する遺伝子の同定を目的とした試験では，薬物動態パラメータ (AUC, $C_{\max}$, K_{el} など) と一塩基多型 (SNPs) の関連を調べる方法がよく用いられる．薬物動態パラメータは，左に裾を引いた分布形であり，非負の計量値として測定される．SNPs は，集団の中で高頻度のホモ接合型 (AA)，ヘテロ接合型 (Aa)，低頻度のホモ接合型 (aa) のように，二つの対立遺伝子の対からなる 3 カテゴリの遺伝子型として測定される．そのため，多くの実験家は，薬物動態パラメータと SNPs の関連を調べるために Kruskal-Wallis 検定 (帰無仮説 $H_0 : \mu_{\mathrm{AA}} = \mu_{\mathrm{Aa}} = \mu_{\mathrm{aa}}$) を適用し，各遺伝子型を持つ集団の薬物動態パラメータの母平均の違いを検討する．帰無仮説が棄却された場合，各遺伝子型のいずれかに統計的有意差がある SNPs が検出される．検出された SNPs には，偽陽性が含まれる可能性があるため，実験家は薬物動態パラメータと SNPs の遺伝子型の反応関係が単調であるかどうかを一つずつ目視で確認し，候補遺伝子として選択する．検出される SNPs の数が少ない場合は，目視での確認で問題はなかったが，近年では一度に 10〜100 万の SNPs を調べるゲノム網羅的なアプローチが主流になりつつあり，検出される SNPs の数が膨大なため，目視での確認は困難になった．そのため，薬物動態パラメータと SNPs の遺伝子型の反応関係を考慮した統計学的スクリーニング法を提案した．本論文では，シミュレーションにより提案法の性能評価を行い，ミレニアム・ゲノムプロジェクトにより得られた抗がん剤の薬理ゲノム学研究のデータに対して，提案法を適用した．

キーワード: 薬物動態関連遺伝子; スクリーニング; 対比統計量; 不等標本数; MULTTEST Procedure

1 ゲノムワイド関連解析における薬物動態関連遺伝子のスクリーニング

医薬品に対する患者の反応には，著しい個体差がみられる場合がある．この個体差の原因の一つとして，薬剤の曝露量に影響する薬物動態関連分子や，薬剤の応答性に影響を与える薬効強度関連分子 (例えば，薬剤標的分子) の遺伝子多型が考えられる．多くの医薬品は投与されると循環血液中に吸収され，体内の各組織に分布した後，代謝・排泄という過程をたどる．これら四つの過程は，被験者から血液中薬物濃度を測定することで薬物動態パラメータとして推定することができる．薬物動態のそれぞれの過程は，環境要因や遺伝的要因によって影響を受けるため，近年では個別化医療の実現に向け，薬物動態に関連する薬物動態関連遺伝子の探索が行われるようになった．例えば，UGT1A1 遺伝子の遺伝子多型は，塩酸イリノテカン (CPT-11) の副作用発現に関与しており，UGT1A1 遺伝子に特定の遺伝子型をもつ人はグルクロン酸転移酵素 (UGT) の活性が低下し，好中球減少や重篤な下痢などの副作用の発現が高まることが Innocenti, et al. 2004.[2] 等により報告されている．このような予測に利用できる新規バイ

オマーカーの同定が期待されているものの，臨床応用するには十分な情報が得られていない．現在は新たなバイオマーカーを同定するために，医薬品を投与された集団の薬物動態と遺伝多型の情報を利用した，薬物動態関連遺伝子の探索研究が数多く行われている．薬物動態関連遺伝子に限らず，遺伝多型の探索研究において候補が絞られていない段階では，ゲノム全体からの候補遺伝子多型をスクリーニングする必要がある．そのような研究の事を，ゲノム網羅的関連解析 (Genome-wide associa-tion study) と呼ぶ．薬物動態関連遺伝子を同定するためのゲノム網羅的関連解析では，ある集団のある薬物ついての薬物動態パラメータと，一塩基多型 (Single Nucleotide Polymorphisms，SNPs) の関連を調べる方法がよく用いられる．ここで，薬物動態パラメータと関連する SNPs を，以降では薬物動態関連 SNPs と呼ぶことにする．

SNPs は集団の中で頻度の高い対立遺伝子のホモ接合型 (AA)，ヘテロ接合型 (Aa)，頻度の低いホモ接合型 (aa) の 3 種類の遺伝子型として測定される．遺伝子型間の頻度の違いは，マイナー対立遺伝子頻度 (Minor Allele Frequency, MAF) と呼ばれる指標で表わされる．MAF は集団における a の頻度であり，[0.05, 0.5] の値を取り，集団によって多少の差異が生じるものの，やや左に裾を引いた分布である事が知られている [3]．薬物動態パラメータには，吸収過程を要約する薬物血中濃度－時間曲線下面積 (AUC)，最高血中濃度 ($C_{\max}$)，代謝・排泄を表わす消失速度定数 (K_{el}) などがある．これらのパラメータは，通常図 1 に示したように右に裾を引いた分布であり，非負の計量値として測定される．生物学的には遺伝形質は，対立遺伝子頻度に比例する (対立遺伝子モデル) か，劣性遺伝する (劣性遺伝モデル) か，優性遺伝する (優性遺伝モデル) の 3 通りが自然であり，薬物動態パラメータと遺伝子型の間には単調な反応関係を持つものが多いと考えられている．分子生物学的にも，3 通り以外の反応関係を示す遺伝子は，検出されたとしても真の薬物動態関連 SNP である可能性が低いといわれている [4].

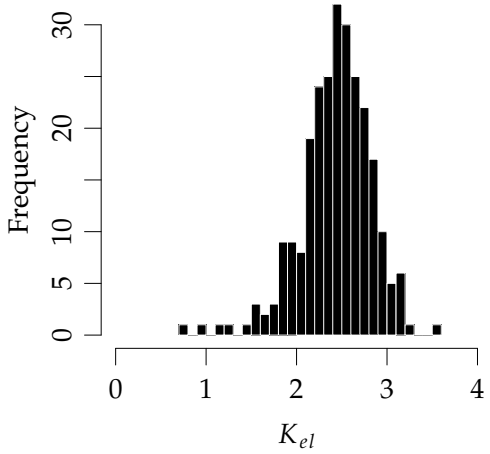

図 1: 薬物動態パラメータ K_{el} のヒストグラム

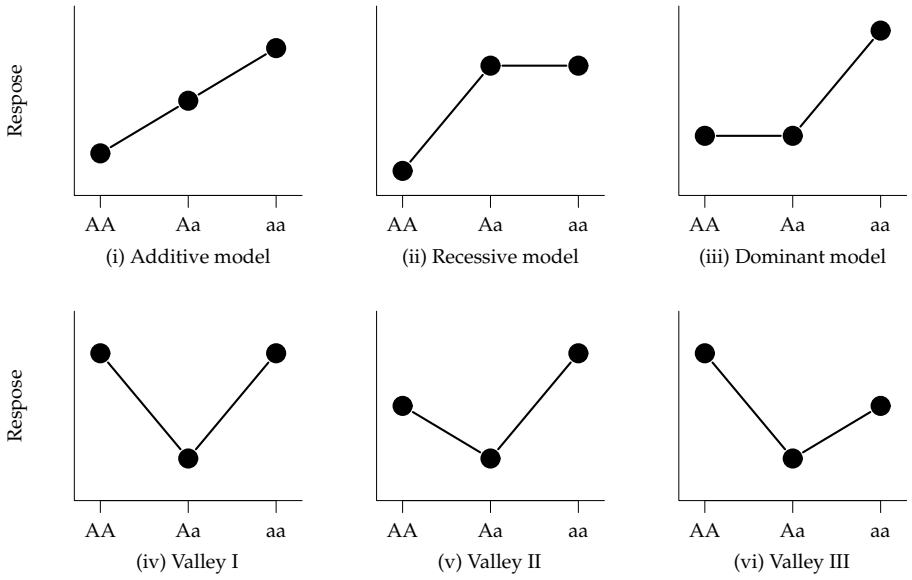

図 2: 薬物動態パラメータと遺伝子型の反応関係

図 2 に薬物動態パラメータと遺伝子型の反応関係の例を示した．(i)，(ii)，(iii) の様な単調な反応関係を示すものは薬物動態関連 SNPs である可能性が高く，(iv)，(v)，(vi) の様な谷型の反応関係を示すもの

は薬物動態関連 SNPs である可能性が低いといえる．

以上を踏まえると，薬物動態パラメータと SNPs を用いたゲノム網羅的関連解析を行う上では，以下の 2 点に注意すべきである．

1. 多くの SNP で各遺伝子型の頻度が異なるため，不等標本数を前提とした手法の適用を検討すべき．
2. 生物学的には単調な反応関係を優先してスクリーニングした方がよい．

2　統計的スクリーニング法

2.1　Kruskall-Wallis 検定 [6]

本稿で想定するデータ構造は，薬物動態関連 SNPs を探索するゲノム網羅的関連解析である．ここで，記号の定義を行う．ある SNP について，i 群の j 番目の個体の薬物動態パラメータの測定値を Y_{ij} とする．i は遺伝子型を表わす添え字で $i = 1, 2, 3$ はそれぞれ，遺伝子型 AA, Aa, aa に対応する．j は遺伝子型間の頻度が異なるため，それぞれ n_i までとなっている．多数の SNPs から薬物動態関連 SNPs をスクリーニングする場合，SNP それぞれについて「各遺伝子型における薬物動態パラメータの母平均に差がない」という帰無仮説 H_0 の仮説検定が行われる．ここで，各遺伝子型における薬物動態パラメータの母平均を μ_i とおくと，

$$H_0 : \mu_1 = \mu_2 = \mu_3 \tag{1}$$

と表わされる．

薬物動態パラメータの分布形状から，正規性の仮定を必要としないノンパラメトリック検定を用いたスクリーニングがしばしば行われる．解析手法には，(1) 式の帰無仮説 H_0 と，対立仮説 $H_1 : \text{not } H_0$ の Kruskal-Wallis 検定が用いられる．

ここで, 全ての測定値 Y_{ij} を値の小さい方から順に並べ，その順位を R_{ij} とし，同順位のときは中間順位で置き換えるとする．Kruskal-Wallis 検定の検定統計量 H は，

$$H = \left[\frac{N-1}{N} \sum_{i=1}^{k} \frac{n_i \left\{ \bar{R_i}^2 - \frac{1}{2}(N+1) \right\}}{(N^2-1)/12} \right] \Bigg/ \left[1 - \frac{\sum_{l=1}^{h} (t_l - 1) t_l (t_l + 1)}{(N-1)N(N+1)} \right] \tag{2}$$

N : 全サンプルサイズ, $N = \sum_{i=1}^{k} n_k$
$\bar{R_i}$: 第 i 群の平均順位, $\bar{R_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}$
t_l : タイの長さ, $R^{(m)}$ を第 m 番目とすると, $R^{(1)} = R^{(2)} = \ldots = R^{(t_1)} < R^{(t_1+1)} = \ldots = R^{(t_1+t_2)} < \ldots < \ldots < R^{(\sum_{l=1}^{h-1} t_l + 1)} = \ldots = R^{(\sum_{l=1}^{h} t_l)}$

である．統計量 H は帰無仮説 H_0 のもとで漸近的に自由度 2 (群の数 -1) の χ^2 分布に従うことが知られている．Kruskal-Wallis 検定による P 値は，χ^2 分布による近似を用いて求めることができる．

しかし，Kruskal-Wallis 検定に基づいて有意差がついた SNPs を薬物動態関連 SNPs として判定を行うと，対立仮説に (iv)，(v)，(vi) のような谷型の反応関係を表わす仮説が含まれるため，真の薬物動態関連 SNPs である可能性が低い SNPs も検出されてしまうことがある．そのため，実験家は検出された SNPs の遺伝子型と薬物動態パラメータの反応関係を目視によりチェックし，遺伝子型と薬物動態パラメータが単調に変化する図 2 の (i)∼(iii) のような反応関係を持つ SNPs を選択する．検出される SNPs の数が少ない場合には，実験家が目視で確認することはそれほど困難ではない．しかし，近年では一度に 10 万 ∼100 万の SNPs を調べることが可能となったため，検出される SNPs も膨大になり，実験家が目視により反応関係を一つずつ確認することは困難になっている．本稿が対象とするようなデータに対して，Kruskal-Wallis 検定は理論的にも効率が悪いため，ゲノム網羅的関連解析において，遺伝子型の反応関係を考慮した薬物動態関連 SNPs を検出するための統計学的スクリーニング法が必要であるといえる．

2.2 最大対比法

Yoshimura, et al. 1997.[7] は毒性試験において，用量反応関係を検出するために最大対比法を適用することを提案している．ゲノム網羅的関連解析においても，最大対比法を適用することで，特定の反応関係をもつ薬物動態関連 SNPs の検出に利用できると考えられる．

最大対比法について定式化を行う．最大対比法は，各群の測定値 Y_{ij} が独立に母平均 $\mathrm{E}(Y_{ij}) = \mu_i$，母分散 $\mathrm{Var}(Y_{ij}) = \sigma^2$ の正規分布に従うことを仮定する．ここで，第 i 群の標本平均を $\bar{Y}_i$，平均ベクトルを $\bar{\mathbf{Y}}$，母平均ベクトルを $\boldsymbol{\mu}$，各群の標本数の逆数を要素に持つ対角行列 $\mathbf{D}$ を

$$\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij},\ \ \bar{\mathbf{Y}} = (\bar{Y}_1,\, \bar{Y}_2,\, \bar{Y}_3)^{\mathrm{t}},\ \ \bar{\boldsymbol{\mu}} = (\mu_1,\, \mu_2,\, \mu_3)^{\mathrm{t}},\ \ \mathbf{D} = \mathrm{diag}\left(\frac{1}{n_1},\, \frac{1}{n_2},\, \frac{1}{n_3}\right) \tag{3}$$

と定義する．diag は対角行列, $^{\mathrm{t}}$ は行列またはベクトルの転置をあらわす. (1) 式の帰無仮説に対し，対立仮説 H_1 として，$\mu_1 \le \mu_2 \le \mu_3$ や $\mu_1 = \mu_2 \le \mu_3$ といった特定の反応関係を指定した検定を行うことを考える．指定を行いたい反応関係が m 個あるならば，対比係数ベクトル $\mathbf{c}_k$ を要素として持つ，対比係数行列

$$\mathbf{C} = (\mathbf{c}_1\ \mathbf{c}_2\ \ldots\ \mathbf{c}_k\ \ldots\ \mathbf{c}_m)^{\mathrm{t}} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ \vdots & \vdots & \vdots \\ c_{m1} & c_{m2} & c_{m3} \end{pmatrix},\ \ \sum_{i=1}^{3} c_{ki} = 0 \tag{4}$$

を用いて，対立仮説は $H_1 : \mathbf{C}\boldsymbol{\mu} > 0$ と表現できる．本稿では，(i) に対応する対比係数ベクトル $\mathbf{c}_1 = (-1,\, 0,\, 1)^{\mathrm{t}}$，(ii) に対応する $\mathbf{c}_2 = (-2,\, 1,\, 1)^{\mathrm{t}}$，(iii) に対応する $\mathbf{c}_3 = (-1,\, -1,\, 2)^{\mathrm{t}}$ の三つの対比係数ベクトルを用いて最大対比法を適用した．対比係数ベクトル $\mathbf{c}_k$ に対応する統計量 (対比統計量と呼ぶ) を

$$T_k = \frac{\mathbf{c}_k^{\mathrm{t}}\bar{\mathbf{Y}}}{\sqrt{S^2\mathbf{c}_k^{\mathrm{t}}\mathbf{D}\mathbf{c}_k}}$$

$$S^2 = \frac{1}{\gamma}\sum_{i=1}^{3}\sum_{j=1}^{n_j}(Y_{ij} - \bar{Y}_i)^2,\ \ \gamma = \sum_{i=1}^{3} n_j - 3$$

と定義する．T_k は仮定のもとで，帰無仮説 $H_0 : \mathbf{c}_k^{\mathrm{t}}\boldsymbol{\mu} = 0$ を対立仮説 $H_1 : \mathbf{c}_k^{\mathrm{t}}\boldsymbol{\mu} > 0$ に対して検定するための検定統計量である．最大対比法の検定統計量は，これら m 個の対比統計量の最大値

$$T_{\max} = \max\{T_1,\, T_2,\, \ldots,\, T_k,\, \ldots,\, T_m\} \tag{5}$$

で定義されている．$t^*_{\max}$ を最大対比統計量の観測値とすると，P 値は

$$\begin{aligned} P\text{-value} &= \Pr(T_{\max} > t^*_{\max} \mid H_0) = 1 - \Pr(T_{\max} \le t^*_{\max} \mid H_0) \\ &= 1 - \Pr(T_1 \le t^*_{\max},\, T_2 \le t^*_{\max},\, \ldots,\, T_k \le t^*_{\max},\, \ldots,\, T_m \le t^*_{\max} \mid H_0) \end{aligned} \tag{6}$$

により求められる．したがって，$\mathbf{T} = (T_1,\, T_2,\, \ldots,\, T_k,\, \ldots,\, T_m)^{\mathrm{t}}$ の同時分布を，適切な領域を積分する事が必要である．

以下で $\mathbf{T}$ の同時分布を導出する．仮定から, $\bar{Y}_i \sim \mathrm{N}(\mu_i,\, \sigma^2/n_i)$ であるから, $\mathrm{E}(\mathbf{c}_k^{\mathrm{t}}\bar{\mathbf{Y}}) = \mathbf{c}_k^{\mathrm{t}}\boldsymbol{\mu}$, $\mathrm{Var}(\mathbf{c}_k^{\mathrm{t}}\bar{\mathbf{Y}}) = \sigma^2\mathbf{c}_k^{\mathrm{t}}\mathbf{D}\mathbf{c}_k$ となる．以上から，統計量 $Z_k = \mathbf{c}_k^{\mathrm{t}}\bar{\mathbf{Y}}\Big/\sqrt{\mathrm{Var}(\mathbf{c}_k^{\mathrm{t}}\bar{\mathbf{Y}})} = \mathbf{c}_k^{\mathrm{t}}\bar{\mathbf{Y}}\Big/\sqrt{\sigma^2\mathbf{c}_k^{\mathrm{t}}\mathbf{D}\mathbf{c}_k}$ が従う分布を求めると，平均 $\mathbf{c}_k^{\mathrm{t}}\boldsymbol{\mu}\Big/\sqrt{\sigma^2\mathbf{c}_k^{\mathrm{t}}\mathbf{D}\mathbf{c}_k}$，分散 1^2 の正規分布に従うことがわかる．不偏分散の推定量を S^2 とすると，χ^2 分布の再帰性より，

$$V = \gamma\frac{S^2}{\sigma^2} = \sum_{i=1}^{3}\sum_{j=1}^{n_i}\left(\frac{Y_{ij} - \bar{Y}_i}{\sigma}\right)^2 \sim \chi^2(\gamma) \tag{7}$$

である．したがって，対比係数ベクトル $\mathbf{c}_k$ に対応する対比統計量 T_k は，

$$T_k = \frac{Z_k}{\sqrt{\gamma \frac{S^2}{\sigma^2}/\gamma}} = \frac{\mathbf{c}_k^{\mathrm{t}} \bar{\mathbf{Y}}}{\sqrt{S^2 \mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_k}} \sim t\left(\gamma,\ \frac{\mathbf{c}_k^{\mathrm{t}} \boldsymbol{\mu}}{\sqrt{\sigma^2 \mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_k}}\right) \tag{8}$$

である (自由度 γ, 非心度 $\lambda = \mathbf{c}_k^{\mathrm{t}} \boldsymbol{\mu} / \sqrt{\sigma^2 \mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_k}$ の非心 t 分布に従う)．また，T_k は帰無仮説 H_0 のもとで自由度 γ の t 分布に従う．T_k の分母の分布は共通であるから，分子にあたる $\mathbf{Z} = (Z_1,\ Z_2,\ \ldots,\ Z_k,\ \ldots,\ Z_m)^{\mathrm{t}}$ の同時分布を求めればよい．$\bar{Y}_i$ は互いに独立であるから，$\mathbf{Z}$ は平均ベクトルが $\mathbf{C}\boldsymbol{\mu}$，共分散 (相関係数) が

$$\mathrm{Cov}(Z_k,\ Z_l) = \rho_{k,\,l} = \frac{\mathrm{Cov}(\mathbf{c}_k^{\mathrm{t}} \bar{\mathbf{Y}},\ \mathbf{c}_l^{\mathrm{t}} \bar{\mathbf{Y}})}{\sqrt{\sigma^2 \mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_k} \sqrt{\sigma^2 \mathbf{c}_l^{\mathrm{t}} \mathbf{D} \mathbf{c}_l}} = \frac{\mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_l}{\sqrt{\mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_k} \sqrt{\mathbf{c}_l^{\mathrm{t}} \mathbf{D} \mathbf{c}_l}} \tag{9}$$

であるような多変量正規分布に従う．(9) 式の要素を持つ分散共分散行列を $\mathbf{R}$，非心ベクトルを $\boldsymbol{\lambda}$ とおくと，

$$\mathbf{T} = \frac{\mathbf{Z}}{\sqrt{\gamma \frac{S^2}{\sigma^2}/\gamma}} \sim t(\gamma,\ \mathbf{R},\ \boldsymbol{\lambda}) \tag{10}$$

$$\boldsymbol{\lambda} = (\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_k,\ \ldots,\ \lambda_m)^{\mathrm{t}},\ \ \lambda_k = \frac{\mathbf{c}_k^{\mathrm{t}} \boldsymbol{\mu}}{\sqrt{\sigma^2 \mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_k}}$$

である (自由度 γ，分散共分散行列 $\mathbf{R}$，非心ベクトル $\boldsymbol{\lambda}$ の多変量非心 t 分布に従う). また，$\mathbf{T}$ は帰無仮説 H_0 のもとで自由度 γ，分散共分散行列 $\mathbf{R}$ の多変量 t 分布に従う．

ゲノム網羅的関連解析において，最大対比法を用いる場合，以下の問題点が存在する．多くの SNP では各遺伝子型の頻度が異なるため，対比統計量 T_k は各群の標本数の相違が大きいほど，分子の標準誤差の推定値にあたる分母

$$\sqrt{S^2 \mathbf{c}_k^{\mathrm{t}} \mathbf{D} \mathbf{c}_k} = \sqrt{S^2 \left(\frac{c_{k1}^2}{n_1} + \frac{c_{k2}^2}{n_2} + \frac{c_{k3}^2}{n_3}\right)} \tag{11}$$

の値を過大推定してしまうことがある．例えば，$c_{k3}^2 \neq 0$ の対比係数ベクトルを与えた仮説は n_3 が n_1, n_2 と比較して極端に小さい場合には，$c_{k3}^2 = 0$ の対比係数ベクトルを与えた仮説よりも，分母が大きくなりやすい事が分かる．

2.3 修正最大対比法 (Modified Maximum Contrast Method)[1]

Kruskall-Wallis 検定は不要な反応関係を持つ SNPs を検出しやすく，最大対比法は不等標本数の場合に標準誤差を過大推定することがあるという問題があった．そこで，我々は最大対比法をもとに，不等標本数の場合に (11) 式について過大推定が起こらないように変更を加えた修正最大対比統計量

$$T'_{\max} = \max\{T'_1,\ T'_2,\ \ldots,\ T'_k,\ \ldots,\ T'_m\},\ \ T'_k = \frac{\mathbf{c}_k^{\mathrm{t}} \bar{\mathbf{Y}}}{\sqrt{\mathbf{c}_k^{\mathrm{t}} \mathbf{c}_k}} \tag{12}$$

を提案した．最大対比法と同様に，対比係数ベクトルとして，$\mathbf{c}_1 = (-1,\ 0,\ 1)^{\mathrm{t}}$，$\mathbf{c}_2 = (-2,\ 1,\ 1)^{\mathrm{t}}$，$\mathbf{c}_3 = (-1,\ -1,\ 2)^{\mathrm{t}}$ を用いた．また，標本再抽出に基づいた並べ替え検定の近似 P 値を求めた [9]．計算方法を以下に示す．

1. リサンプリング回数 $NRESAMP$ を設定する．カウント用変数 $COUNT = 0$ をセットする．標本の修正最大対比統計量 $t'^{*}_{\max}$ を求める．
2. データ y_{ij} から復元抽出を行う．抽出した標本を $y^{(r)}_{ij}$ とする．
3. $y^{(r)}_{ij}$ から $t'^{(r)}_{\max}$ を求め，$t'^{(r)}_{\max} \geq t'^{*}_{\max}$ であれば $COUNT = COUNT + 1$ とする．
4. 2–3 を $NRESAMP$ 回反復する．P 値は

$$P\text{-value} = \frac{COUNT}{NRESAMP}$$

である．

3 シミュレーションによる性能評価

3.1 目的と方法

本シミュレーションの目的は，Kruskal-Wallis 検定，最大対比法，修正最大対比法の検出性能の評価である．実データを参考に，各群のサンプルサイズ，図 2 に対応する反応関係を持つような各群の薬物動態パラメータを設定した上で測定値 y_{ij} を生成し，モンテカルロシミュレーションを行った．サンプルサイズについては，全サンプルサイズを 300 と固定し，MAF を 0.5，0.25，0.12 の 3 通り設定することで，各群の標本数が近い状況から異なる状況までを検討した．薬物動態パラメータについては，一般に非負で右に裾を引いた分布であり，実データを考慮して次のパラメータを持つ対数正規分布，$\bar{Y}_i \sim \mathrm{LN}(\mu_i, 1^2)$，$\mu_i = \Delta \cdot c_{ki}/4$ に従う乱数をシミュレーションデータとして発生させた．条件 $\Delta = 0$ は帰無仮説を表わし，第一種の過誤の制御の評価を，条件 $\Delta = 0.5, 1.0$ では検出性能の評価を行った．c_{ki} については，検出すべき反応関係である (i)，(ii)，(iii) の場合と，検出する必要のない反応関係である (iv)，(v)，(vi) の場合を設定した．シミュレーション回数は 10,000 回とし，有意水準 として各手法で検定を行った．最大対比法，および修正最大対比法では，統計量が最大となる対比係数ベクトルが観測値にもっとも適合する反応関係であると判定する事ができる．そのため，シミュレーション条件で設定した反応関係を正しく判定できているかどうかを評価する事ができる．これは R_{TP} という指標を用いて評価を行った．

$$R_{TP} = \frac{\text{正しい反応関係を検出した上で有意と判定された回数}}{\text{シミュレーション回数}}$$

が高いほど，真の関係を正しく判定し，検出する事が出来ている事を意味する．一般的な第一種の過誤確率と検出力については R_P という指標を用いて評価を行った．

$$R_P = \frac{\text{有意と判定された回数}}{\text{シミュレーション回数}}$$

3.2 結果

真の状態として (i)〜(iii) の単調な反応関係を設定した場合の結果を表 1 に示した．表の左から 2 番目の列にはシミュレーション条件として設定した真の反応関係を示しており，$\delta = 0.5, 1.0$ の場合には真の反応関係を正しく判定できているかどうか評価するため，表中の灰色のセルに R_{TP} を示した．Kruskal-Wallis 検定では R_{TP} 求める事ができないため表示していない．$\delta = 0$ の場合，すべての条件・手法において，第一種の過誤確率は名目の有意水準付近に保たれている事が分かる．次に $\delta = 0.5, 1.0$ の場合は，MAF = 0.5 の場合，最大対比法・修正最大対比法の R_P および R_{TP} はそれほど大きな差が見られなかった．Kruskal-Wallis 検定の R_P は他二つの手法よりも小さい事が分かった．MAF = 0.25, 0.12 の場合，真の反応関係が (i) 対立遺伝子モデルの場合は，修正最大対比法の R_{TP} が最も高く，最大対比法と比較すると 0.01〜0.3 程度高いことが分かった．(ii) 劣性の場合は，最大対比法の R_{TP} が最も高く，修正最大対比法

と比較すると 0.05〜0.5 程度高いことが分かった．(iii) 優性の場合は，修正最大対比法の R_{TP} が最も高く，最大対比法と比較すると 0.1〜0.3 程度高いことが分かった．また，MAF が小さくなり，群間の標本数が異なるほど手法間の差が大きくなることが分かった．

次に，真の状態として (iv)〜(vi) の谷型の反応関係を設定した場合の結果を表 2 に示した．こちらは検出する必要が無い反応関係であるため，R_P が低くなることが望ましい．MAF = 0.5 の場合，対立仮説に (iv)〜(vi) の反応関係を含むため，Kruskal-Wallis 検定の は他の二つの手法よりも大きくなることが分かった．MAF = 0.25, 0.12 の場合も同様に，Kruskal-Wallis 検定の は他の二つの手法よりも大きくなることが分かった．真の反応関係が (v) の場合は，最大対比法の がもっとも低く，修正最大対比法と比較すると 0.02〜0.05 程度小さいことが分かった．(iv) や (vi) の場合は，修正最大対比法の がもっとも低く，最大対比法と比較すると 0.1〜0.5 程度小さいことが分かった．

表 1: 単調な反応関係を指定した条件における R_P および R_{TP}

			$\Delta = 0$				$\Delta = 0.5$				$\Delta = 1.0$
MAF[1]	真の関係	手法	判定								
			R_P	(i)	(ii)	(iii)	R_P	(i)	(ii)	(iii)	R_P
0.5	(i)	MMCM[2]	0.050	0.405	0.140	0.144	0.689	0.966	0.019	0.012	0.997
		MCM[3]	0.050	0.408	0.138	0.144	0.690	0.968	0.018	0.011	0.997
		K-W[4]	0.049	–	–	–	0.628	–	–	–	0.998
	(ii)	MMCM	–	0.130	0.672	0.005	0.807	0.004	0.996	0.000	1.000
		MCM	–	0.128	0.674	0.005	0.807	0.004	0.996	0.000	1.000
		K-W	–	–	–	–	0.772	–	–	–	1.000
	(iii)	MMCM	–	0.134	0.012	0.679	0.825	0.002	0.000	0.998	1.000
		MCM	–	0.133	0.011	0.682	0.826	0.002	0.000	0.998	1.000
		K-W	–	–	–	–	0.769	–	–	–	1.000
0.25	(i)	MMCM	0.055	0.211	0.041	0.139	0.391	0.764	0.133	0.053	0.950
		MCM	0.055	0.135	0.373	0.030	0.538	0.395	0.592	0.001	0.988
		K-W	0.046	–	–	–	0.452	–	–	–	0.980
	(ii)	MMCM	–	0.173	0.308	0.009	0.490	0.052	0.944	0.000	0.996
		MCM	–	0.009	0.747	0.001	0.757	0.000	1.000	0.000	1.000
		K-W	–	–	–	–	0.789	–	–	–	1.000
	(iii)	MMCM	–	0.063	0.000	0.413	0.476	0.042	0.000	0.934	0.976
		MCM	–	0.134	0.066	0.238	0.438	0.195	0.041	0.724	0.960
		K-W	–	–	–	–	0.381	–	–	–	0.920
0.12	(i)	MMCM	0.043	0.036	0.002	0.106	0.144	0.390	0.014	0.121	0.525
		MCM	0.048	0.043	0.192	0.010	0.245	0.071	0.708	0.003	0.782
		K-W	0.049	–	–	–	0.267	–	–	–	0.818
	(ii)	MMCM	–	0.089	0.017	0.027	0.133	0.185	0.485	0.001	0.671
		MCM	–	0.004	0.443	0.001	0.448	0.000	0.964	0.000	0.964
		K-W	–	–	–	–	0.593	–	–	–	0.994
	(iii)	MMCM	–	0.010	0.000	0.192	0.202	0.033	0.000	0.624	0.657
		MCM	–	0.058	0.058	0.072	0.188	0.175	0.151	0.319	0.645
		K-W	–	–	–	–	0.165	–	–	–	0.483

[1]MAF: Minor allele frequency. [2]MMCM: 修正最大対比法. [3]MCM: 最大対比法. [4]K-W: Kruskal-Wallis 検定.
灰色のセルは R_{TP} を表わす.

3.3 考察

シミュレーションの結果から，薬物動態関連 SNPs を探索するゲノム網羅的関連解析における適切なスクリーニング方法について考察する．各群の標本数が均一になる SNPs については，ほぼ性能が変わらないため，最大対比法・修正最大対比のどちらかを用いればよいと考えられる．各群の標本数が不均一な場合については，反応関係が (i) 対立遺伝子モデルと (iii) 優性遺伝モデルの場合は，修正最大対比法を用いた方が良いと考えられる．反応関係が (ii) 劣性遺伝モデルの場合は最大対比法を用いた方が良いと考えられる．Kruskal-Wallis 検定は検出力が他の二手法と比較して低く，谷型の反応関係を検出して

表 2: 谷型の反応関係を指定した条件における R_P

MAF[1]	真の関係	手法	$\Delta = 0.5$	$\Delta = 1.0$
0.5	(iv)	MMCM[2]	0.455	0.987
		MCM[3]	0.467	0.989
		K-W[4]	0.780	1.000
	(v)	MMCM	0.513	0.995
		MCM	0.519	0.995
		K-W	0.617	0.999
	(vi)	MMCM	0.540	0.991
		MCM	0.543	0.992
		K-W	0.623	0.998
0.25	(iv)	MMCM	0.177	0.774
		MCM	0.304	0.937
		K-W	0.765	1.000
	(v)	MMCM	0.311	0.797
		MCM	0.259	0.746
		K-W	0.355	0.924
	(vi)	MMCM	0.227	0.890
		MCM	0.493	0.988
		K-W	0.713	1.000
0.12	(iv)	MMCM	0.087	0.243
		MCM	0.184	0.579
		K-W	0.515	0.991
	(v)	MMCM	0.149	0.404
		MCM	0.124	0.346
		K-W	0.213	0.687
	(vi)	MMCM	0.071	0.298
		MCM	0.272	0.800
		K-W	0.546	0.996

[1]MAF: Minor allele frequency. [2]MMCM: 修正最大対比法.
[3]MCM: 最大対比法. [4]K-W: Kruskal-Wallis 検定.

やすいため，適切な手法ではないと考えられる．シミュレーションでは 1 つの SNP の検出における性能を評価したが，実際には 10 万〜100 万 SNPs から検出を行う必要があり，仮に有意水準 $\alpha = 0.05$ で検出を行うと，単純計算で 5000〜50000 SNPs が検出される．実際には，より有望な SNPs を優先的に検出したいため，有意水準を調整するが，多くの SNPs の反応関係を目視でチェックする必要がある．したがって，上で述べたような，谷型の反応関係を検出しにくく， が高くなる最大対比法および修正最大対比法を組み合わせて適用することが望ましいといえる．

参考文献

[1] Sato Y, Laird NM, Nagashima K, Kato R, Hamano H, Yafune A, Kaniwa N, Saito Y, Sugiyama E, Kim S-R, Furuse J, Ishii H, Ueno H, Okusaka T, Saijo N, Sawada J, Yoshida T. A new statistical screening approach for finding pharmacokinetics-related genes in genome-wide studies. *The Pharmacogenomics Journal* 2009; **9**: 137–146.

[2] Innocenti F, Undevia SD, Iyer L, Chen PX, Das S, Kocherginsky M, Karrison T, Janisch L, Ramirez J, Rudin CM, Vokes EE, Ratain MJ. Genetic variants in the UDP-glucuronosyltransferase 1A1 gene predict the risk of severe neutropenia of Irinotecan. *Journal of Clinical Oncology* 2004; **22**(8): 1382–1388.

[3] The International HapMap Consortium. The international HapMap project. *Nature* 2003; **426**: 789–796.

[4] Evans WE, McLeod HL. Pharmacogenomics — drug disposition, drug targets, and side effects. *The New England Journal of Medicine* 2003; **348**(6): 538–549.

[5] Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y. JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Research* 2002; **30**: 158–162.

[6] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 1952; **260**(47): 583–621.

[7] Yoshimura I, Wakana A, Hamada C. A performance comparison of maximum contrast methods to detect dose dependency. *Drug Information Journal* 1997; **31**: 423–432.

[8] 西山智, 栁原宏和, 吉村功. 最大対比法を活用するための SAS/IML プログラム. 計量生物学 2004; **24**(2): 57–70.

[9] Westfall PH, Young SS. *Resampling-based multiple testing: Examples and methods for p-Value adjustment (Wiley series in probability and statistics)*. New York: John Wiley & Sons, Inc. 1993.

[10] Sugiyama E, Kaniwa N, Kim S R, Kikura-Hanajiri R, Hasegawa R, Maekawa K, Saito Y, Ozawa S, Sawada J, Kamatani N, Furuse J, Ishii H, Yoshida T, Ueno H, Okusaka T, Saijo N. Pharmacokinetics of gemcitabine in Japanese cancer patients: the impact of a cytidine deaminase polymorphism. *Journal of Clinical Oncology* 2007; **25**(1): 32–42.

A SAS プログラム

```
%macro mmcm_resamp(
  data, var, class,
  outd,
  nrep = 20000
);

/*** 並べ替え分布からのリサンプリング ***/
proc multtest noprint
  data = &data permutation nsample = &nrep
  outsamp = Resamp(keep=_sample_ _class_ &var.);
  class &class;
  test mean(&var);
run;

/*****************************************/
/*** 独自の統計量を扱う場合は以下を修正 ***/
/*****************************************/

/* 標本の統計量 */
proc summary data = &data;
  var &var; output out = Samp(keep= m &class) mean=m;
  by &class;
proc transpose data = Samp out = Samp(drop = _name_)
     prefix=m;
  var m; id &class;
data Samp;
  set Samp;
  /* 修正最大対比統計量 */
  T1 = abs((m1*(-1) + m2*(0)  + m3*(1)) / sqrt(2));
  T2 = abs((m1*(-2) + m2*(1)  + m3*(1)) / sqrt(6));
  T3 = abs((m1*(-1) + m2*(-1) + m3*(2)) / sqrt(6));
  Tmax = max(T1, T2, T3);
  keep Tmax;
run;

/* リサンプリングによる統計量の帰無分布の生成 */
proc means data = Resamp noprint;
  var &var;
  by _sample_ _class_;
  output out = Resamp(drop = _TYPE_ _FREQ_) mean = mean;
proc transpose data = Resamp prefix = mean
  out = Resamp(drop = _NAME_ _LABEL_);
  by _sample_; id _class_; var mean;
data Resamp;
  set Resamp;
  if _n_=1 then set Samp;
  Ta = abs((Mean1*(-1) + Mean2*(0)  + Mean3*(1)) / sqrt
       (2));
  Tb = abs((Mean1*(-2) + Mean2*(1)  + Mean3*(1)) / sqrt
       (6));
  Tc = abs((Mean1*(-1) + Mean2*(-1) + Mean3*(2)) / sqrt
       (6));
  if Ta >= Tmax then a = 1; else a = 0;
  if Tb >= Tmax then b = 1; else b = 0;
  if Tc >= Tmax then c = 1; else c = 0;
/* P値の計算 */
proc means data = Resamp noprint;
  var a b c; output out = Resamp(where=(_STAT_='MEAN'));
data &outd;
  set Resamp;
  p_value = min(a, b, c);
  contrast = '-';
  if a = b & b = c then contrast = 'a, b, c';
  else if a = b then contrast = 'a, b';
  else if a = c then contrast = 'a, c';
  else if b = c then contrast = 'b, c';
  else if p_value = a then contrast = 'a';
  else if p_value = b then contrast = 'b';
  else if p_value = c then contrast = 'c';
  keep p_value contrast;
run;

/****************/
/*** ここまで ***/
/****************/
%mend mmcm_resamp;


/* サンプルデータの生成 */
data datax;
  call streaminit(5436984);
  g = 1; do i = 1 to 180; y = rand('Normal', 2.30, 0.1);
       output; end;
  g = 2; do i = 1 to  55; y = rand('Normal', 2.30, 0.1);
       output; end;
  g = 3; do i = 1 to   5; y = rand('Normal', 2.20, 0.1);
       output; end;
  drop i;
run;


/*******************************/
/*  マクロ名   mmcm_resamp      */
/*-----------------------------*/
/*  定義                       */
/*  %mmcm_resamp(              */
/*    data, var, class,        */
/*    outd,                    */
/*    nrep = 20000             */
/*  )                          */
/*                             */
/*  変数                       */
/*  data  : 入力データセット名 */
/*  var   : 解析対象変数名     */
/*  class : 群変数名           */
/*  outd  : 出力データセット名 */
/*  nrep  : リサンプリング回数 */
/*******************************/

%mmcm_resamp(
  datax, y, g,
  outd,
  nrep = 20000
);
proc print data = outd; run;
```

B 修正対比統計量の分布

B.1 修正対比統計量の分布 (A)

測定値 Y_{ij} が母平均 $\mathrm{E}(Y_{ij}) = \mu_i$, 母分散 $\mathrm{Var}(Y_{ij}) = \sigma^2$ の正規分布に従うことを仮定する. このとき, $\bar{\mathbf{Y}} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2\mathbf{D})$ である. i 次元の確率変数が $\bar{\mathbf{X}} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ の場合, $\mathbf{Z} = \mathbf{A}\bar{\mathbf{X}}$ という変換を考える. ただし, $\mathbf{Z}$, $\mathbf{A}$, $\bar{\mathbf{X}}$ はそれぞれ $m \times 1$, $m \times i$, $i \times 1$ の行列とする. 一般に, 変換後の行列の従う分布は,

$$\mathbf{Z} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu},\ \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{t}})$$

である. この性質を用いれば, 修正対比統計量 $\mathbf{T}' = \mathbf{C}^{\mathrm{t}}\bar{\mathbf{Y}}/\sqrt{\mathbf{C}^{\mathrm{t}}\mathbf{C}}$ の分布は,

$$\mathbf{T}' \sim \mathrm{N}(\mathbf{B}\boldsymbol{\mu},\ \sigma^2\mathbf{B}\mathbf{D}\mathbf{B}^{\mathrm{t}}) \tag{13}$$

$$\mathbf{B} = \left(\frac{\mathbf{c}_1^{\mathrm{t}}}{\sqrt{\mathbf{c}_1\mathbf{c}_1^{\mathrm{t}}}},\ \frac{\mathbf{c}_2^{\mathrm{t}}}{\sqrt{\mathbf{c}_2\mathbf{c}_2^{\mathrm{t}}}},\ \ldots,\ \frac{\mathbf{c}_k^{\mathrm{t}}}{\sqrt{\mathbf{c}_k\mathbf{c}_k^{\mathrm{t}}}},\ \ldots,\ \frac{\mathbf{c}_m^{\mathrm{t}}}{\sqrt{\mathbf{c}_m\mathbf{c}_m^{\mathrm{t}}}} \right)^{\mathrm{t}}$$

となる事が分かる.