関連解析における多変量QTLへの拡張

M2006MM003 橋本登
指導教員　田中豊

1　はじめに

昨今のヒトゲノムワイド研究の主な目的目的の 1 つに, 複雑な疾患の遺伝的基盤を同定することがある. そんな中で通常のハプロタイプと, 突然変異が起きているハプロタイプでどう表現型が変わるかを見るために, 最近遺伝子型と表現型の関連が勢力的に研究されている. 本研究では, '遺伝子型' とは単に自身の遺伝子型だけというわけではなく, 遺伝子型から推定されるハプロタイプやディプロタイプ型も意味する. '表現型' はある疾患に関係あると考えられる数量, または量的変量とする. QTL (quantitative trait locus) と言われる量的変量は BMI や血糖値などを意味する. 遺伝情報と QTL での量的表現型変量との間の関連解析のために, いくつかのアルゴリズムが提案されている (*QTLhaplo*, Shibata *et al*., 2004). *QTLhaplo* アルゴリズムは遺伝子型と正規分布に従うと仮定した一変量表現型との関連を解析するものである. そして, そのアルゴリズムはディプロタイプ型の頻度 (ディプロタイプ型を構成するハプロタイプ頻度の結合確率) と正規分布の密度関数をもとにし, 尤度を計算する. *QTLmarc* (Kamitsuji and Kamatani, 2006) は多変量を扱うことができるが, 遺伝子型から独自に決まるディプロタイプ型の場合しか解析できないため, ドミナントモデルにしか対応しておらず, 新たにレセッシブそしてアディティブモデルといった疾患モデルに対応させた多変量の量的表現型が扱える方法をつくることには意義がある. 本論文では, *QTLhaplo* アルゴリズムを拡張して, 遺伝子型と多変量の量的表現型との間の関連を解析するアルゴリズムを提案した.

2　アルゴリズム

2.1　単一の量的変量のアルゴリズム-*QTLhaplo*

Shibata *et al*., (2004) では *QTLhaplo* のアルゴリズムを以下のように述べている. l 個の連結した SNP 座位があると仮定する. DNA は二重螺旋構造をとり, あるハプロタイプは対をなすものの片方がある. そうすると可能なハプロタイプすべての数は $L = 2^l$ となる. それらのハプロタイプの相対頻度を $\Theta = (\theta_1, \cdots, \theta_j, \cdots, \theta_L)$ とする. ここで θ_j は j 番目のハプロタイプの相対頻度であり, $\theta_j \geq 0, \sum_{j=1}^{L} \theta_j = 1$ である. ハプロタイプ頻度によって, 2 つのハプロタイプの組み合わせは, その集合からの無作為抽出によって各個体に与えられる. この論文において, ディプロタイプ型は 2 つのハプロタイプコピーの組み合わせとする. $a_1, a_2, \cdots, a_{L^2}$ は可能なディプロタイプ型とする. Θ によって決まるディプロタイプ型 a_k を持つ i 番目の個体の確率を $P(d_i = a_k|\Theta) = \theta_l\theta_m$ とし, d_i は i 番目の個体のディプロタイプ型, l, m は a_k を作るハプロタイプのラベル $(1, 2, \cdots, L)$ である. 個体 i は表現型の確率密度関数に従う ψ_i を示す. ここでディプロタイプ型に対する表現型は分散一定, ディプロタイプ型に依存する平均を持つ正規分布に従うと仮定する. この解析の結果は (Θ, D, Ψ) によって決まる. ここで $D = (d_1, \cdots, d_N)$ はディプロタイプ型のベクトルを示す. $\Psi = (\psi_1, \cdots, \psi_N)$ は表現型のベクトルを示す. 表現型の平均 μ は個体がハプロタイプ h_p を持つものと, 持たないものの違いとして推定される. この h_p は他のものとは違った影響を持つとされるハプロタイプである. D_+ はハプロタイプ h_p を含むディプロタイプ型のセットを示す. そして我々は表現型に対する 2 つの正規分布を決める. 1 つは $d_i \in D_+$ の個体に対して $N(\mu_1, \sigma^2)$, もう一つは $d_i \notin D_+$ の個体に対して, $N(\mu_2, \sigma^2)$ とする. $f_{\mu_1}(x)$ は i 番目の個体の $d_i \in D_+$ の時, その個体の表現型 x 確率密度関数である. そして, $f_{\mu_2}(x)$ は $d_i \notin D_+$ の時, i 番目の個体の表現型 x の確率密度関数である. 以上より, もし ψ_i が i 番目の個体の表現型を示すなら確率密度は次のように書ける.

$$f_{\mu_1}(\mathbf{x}) = f(\psi_i = \mathbf{x}|d_i \in D_+), d_i \in D_+ \text{の場合}$$

$$f_{\mu_2}(\mathbf{x}) = f(\psi_i = \mathbf{x}|d_i \in D_-), \quad d_i \in D_- \text{の場合.}$$

また, ドミナントモデル, レセッシブモデル, アディティブモデルといった発現モデルも考慮できるようにする. A はある表現型に関連するハプロタイプを示す. B は A の補集合, すなわち A 以外のハプロタイプすべてである. 我々は個体のディプロタイプ型に対して, あるハプロタイプの持ちかたで異なる平均ベクトルを個体に与えた. こうして, ドミナントモデルでは AA, AB 両方に対して μ_1 を与え, BB に対して μ_2 を与えた. レセッシブモデルにおいては, AA に対しては μ_1 を与え, AB, BB に対して μ_2 を与えた. アディティブモデルには μ_1 と μ_2 を AA, BB にそれぞれ与え, $\mu_3 = (\mu_1 + \mu_2)/2$ を AB に与えた.

2.2　多変量へ拡張された尤度関数

観測されるデータは個体の遺伝子型と表現型である. $G_{obs} = (g_1, g_2, \cdots, g_N)$ と $\Psi_{obs} = (w_1, w_2, \cdots, w_N)$ は観測された遺伝子型と表現型のそれぞれのベクトルであり, ここで g_i と w_i は i 番目の個人の観測された遺伝子型と, 量的表現型である.

尤度関数は, この観測された遺伝子型から求めたディプロタイプ型の頻度 (構成する 2 つのハプロタイプの頻度の同時確率) と, 量的表現型が従うと仮定される分布の確率密度の積となる. そして, この研究の目的である多変量の量的表現型を扱えるようにするために, その量的表現型が多次元正規分布に従うと仮定する. その時の

尤度関数は次の様になる.

$$L(\Theta,\boldsymbol{\mu},\Sigma)\propto\prod_{i=1}^{N}\sum_{a_k\in A_i}P(d_i{=}a_k|\Theta)f(\Psi_i=\mathbf{x_i}|d_i=a_k,\boldsymbol{\mu},\Sigma)$$

ここで

$$f(\Psi_i=\mathbf{x}|d_i=a_k,\boldsymbol{\mu},\Sigma)=$$
$$\begin{cases}(\frac{1}{(2\pi)^{p/2}|\Sigma|^{-\frac{1}{2}}})e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}\\ \text{ただし } a_k\in D_+,(\text{アディティブの時は } a_k\in D_{AA})\\ (\frac{1}{(2\pi)^{p/2}|\Sigma|^{-\frac{1}{2}}})e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}\\ \text{ただし } a_k\notin D_+,(\text{アディティブの時は } a_k\in D_{BB})\\ (\frac{1}{(2\pi)^{p/2}|\Sigma|^{-\frac{1}{2}}})e^{-\frac{1}{2}(\mathbf{x}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})'\Sigma^{-1}(\mathbf{x}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})}\\ \text{アディティブの時 } a_k\in D_{AB},\end{cases}$$

$\boldsymbol{\mu}$, Σ はそれぞれ, 量的表現型の平均ベクトルと分散共分散行列. $\mathbf{x}$ は個体の量的表現型のベクトルである. 式 (1) は対立仮説のもとでは Θ,$\boldsymbol{\mu}$,Σ 上で最大化され, こうして得られた最大尤度は L_{max} と書かれる. そして, 帰無仮説のもとでは $\Theta,\boldsymbol{\mu}_0,\sigma$ 上で最大化され, こうして得られた最大尤度は L_{0max} と書かれる. 尤度比 L_{0max}/L_{max} はハプロタイプの存在と表現型分布との間の関連の検定に使われる. もし, $d_1,d_2,\cdots,d_N$ と $\Psi_1,\Psi_2,\cdots,\Psi_N$ の完全データが利用可能ならば, $\Theta=(\theta_1,\theta_2,\cdots,\theta_L)$, $\boldsymbol{\mu}$, Σ の最尤推定量は次のように簡単に求められる. $\hat{\theta_j}=n_j/(2N)$ $(j=1,2,\cdots,L)$, $\hat{\boldsymbol{\mu}_1}=\sum_{d_i\in D_+}\psi_i/N_+$, $\hat{\boldsymbol{\mu}_2}=\sum_{d_i\notin D_+}\psi_i/N_-$, $\hat{\Sigma}=[\sum_{d_i\in D_+}(\Psi_i-\boldsymbol{\mu}_1)(\Psi_i-\boldsymbol{\mu}_1)'+$
$\sum_{d_i\notin D_+}(\Psi_i-\boldsymbol{\mu}_2)(\Psi_i-\boldsymbol{\mu}_2)']/N$. ここで n_i は N 個体における j 番目のハプロタイプのコピーの数であり, N_+ はハプロタイプ hp を持つ個体の数を示し, N_- はハプロタイプ hp を持たない個体の数を示す. しかしながら, 以上のことには問題がある. 現実的には, 個体からは遺伝子型と表現型しか観測できないから, 完全データの使用はできない. 個体の観測された遺伝子型の組み合わせからハプロタイプはできるので, 個体のディプロタイプ型の候補は, 1 個体に付つき 1 つとは限らない. 一般的には, ディプロタイプ型の候補の集合ができ, それらは確率的に個体に与えられる. これらの確率を求める手法は EM アルゴリズムなどを使って求めるハプロタイプ頻度推定アルゴリズムとして種々実装されている. このハプロタイプ頻度の推定は, この研究目的とは違うので詳しい説明は省く. 3 節で詳しく説明するが, ハプロタイプ頻度の推定をする時は, R パッケージ *haplo.stats* のハプロタイプ推定関数 *haplo.em* を使って頻度を推定する.

2.3 パラメータの推定

平均ベクトル, 分散共分散行列は式 (1) の尤度関数より次のように求まる. まずドミナントモデルとレセッシブモデルの時,D_+ と D_- の平均ベクトルは

$$\hat{\boldsymbol{\mu}_1}=\frac{\sum_{i=1}^{N}\psi_i(u_b/u_0)}{\sum_{i=1}^{N}(u_b/u_0)},\quad \hat{\boldsymbol{\mu}_2}=\frac{\sum_{i=1}^{N}\psi_i(v_b/v_0)}{\sum_{i=1}^{N}(v_b/v_0)},$$

共通の分散共分散行列は

$$\hat{\Sigma}=\frac{1}{n}\Big[\sum_{i=1}^{N}(\psi_i-\boldsymbol{\mu}_1)(\psi_i-\boldsymbol{\mu}_1)^T(u_b/u_0)+\sum_{i=1}^{N}(\psi_i-\boldsymbol{\mu}_2)(\psi_i-\boldsymbol{\mu}_2)^T(v_b/v_0)\Big].$$

ただし, n は $\sum_{i=1}^{N}(u_b/u_0)+\sum_{i=1}^{N}(v_b/v_0)$ である. また u, v は次の様に定義される.

$$u_b=\sum_{a_k\in D_+\cap A_i}P(d_i=a_k|\Theta)f(\psi_i|d_i=a_k,\boldsymbol{\mu}_1,\Sigma)$$
$$u_0=\sum_{a_k\in A_i}P(d_i=a_k|\Theta)f(\psi_i|d_i=a_k,\boldsymbol{\mu},\Sigma)$$
$$v_b=\sum_{a_k\in A_i\cap D-}P(d_i=a_k|\Theta)f(\psi_i|d_i=a_k,\boldsymbol{\mu}_2,\Sigma)$$
$$v_0=\sum_{a_k\in A_i}P(d_i=a_k|\Theta)f(\psi_i|d_i=a_k,\boldsymbol{\mu},\Sigma).$$

ただし $\boldsymbol{\mu}=(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2)$ で, D_+ は発現モデルをみたすハプロタイプ hp をもつ個体のディプロタイプ型の集合である. アディティブモデルはあるハプロタイプの効果を加算的に評価するために, $\hat{\boldsymbol{\mu}_3}=(\hat{\boldsymbol{\mu}_1}+\hat{\boldsymbol{\mu}_2})/2$ を加え, 同様に推定すると次の式を解くことによって得られる.

$$\Big(\sum_{i=1}^{N}\frac{u_b}{u_0}+\frac{1}{4}\sum_{i=1}^{N}\frac{w_b}{w_0}\Big)\mu_1+\frac{1}{4}\sum_{i=1}^{N}\frac{w_b}{w_0}\mu_2=\sum_{i=1}^{N}\Big(\frac{u_b}{u_0}+\frac{1}{2}(\frac{w_b}{w_0})\Big)\mathbf{x}_i$$
$$\frac{1}{4}\sum_{i=1}^{N}\frac{w_b}{w_0}\mu_1+\Big(\sum_{i=1}^{N}\frac{v_b}{v_0}+\frac{1}{4}\sum_{i=1}^{N}\frac{w_b}{w_0}\Big)\mu_2=\sum_{i=1}^{N}\Big(\frac{v_b}{v_0}+\frac{1}{2}(\frac{w_b}{w_0})\Big)\mathbf{x}_i$$

$$\hat{\Sigma}=\frac{1}{n}\Big[\sum_{i=1}^{N}(\psi_i-\mu_1)(\psi_i-\mu_1)^T(u_b/u_0)+\sum_{i=1}^{N}(\psi_i-\mu_2)(\psi_i-\mu_2)^T(v_b/v_0)+\sum_{i=1}^{N}(\psi_i-(\mu_1-\mu_2)/2)(\psi_i-(\mu_1-\mu_2)/2)^T(w_b/w_0)\Big]$$

上式において, n は $\sum_{i=1}^{N}(u_b/u_0)+\sum_{i=1}^{N}(v_b/v_0)+\sum_{i=1}^{N}(w_b/w_0)$. ただし u, v, w は次の様に定義される.

$$u_b=\sum_{a_k\in A_i\cap AA}P(d_i=a_k|\Theta)f(\psi_i|d_i=a_k,\boldsymbol{\mu}_1,\Sigma)$$
$$u_0=\sum_{a_k\in A_i}P(d_i=a_k|\Theta)f(\psi_i|d_i=a_k,\boldsymbol{\mu},\Sigma)$$
$$v_b=\sum_{a_k\in A_i\cap BB}P(d_i=a_k|\Theta)f(\psi_i|d_i=a_k,\boldsymbol{\mu}_2,\Sigma)$$
$$v_0=\sum_{a_k\in A_i}P(d_i=a_k|\Theta)f(\psi_i|d_i=a_k,\boldsymbol{\mu},\Sigma)$$
$$w_b=\sum_{a_k\in A_i\cap AB}P(d_i=a_k|\Theta)f(\psi_i|d_i=a_k,(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2,\Sigma)$$
$$w_0=\sum_{a_k\in A_i}P(d_i=a_k|\Theta)f(\psi_i|d_i=a_k,\boldsymbol{\mu},\Sigma)$$

ただし $\boldsymbol{\mu}=\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\mu}_3\}$ である. 以上の推定量はそのままでは解くことができないので, 初期値を与え反復法を用いてそれぞれのパラメータを推定する.

2.4 尤度比検定

量的表現型が従う多変量正規分布が p 次元の場合, 帰無仮説の下では $-2log(L_{0max}/L_{max})$ は自由度 p の χ^2 分布に従うと期待される. 前節で求めた最尤推定量を使いハプロタイプと量的表現型との関連解析に χ^2 検定を使う. 帰無仮説 ($H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_0$) の下で最大化された最大尤度は L_{0max}, 対立仮説 ($H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ または $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \boldsymbol{\mu}_3 = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$) の下で最大化された最大尤度は L_{max} である. そして検定統計量 $-2\log(L_{0max}/L_{max})$ を計算する. その統計量はデータ数が十分あるとき, χ^2 分布に従うことが知られている. その時の分布の自由度はデータ内の変量の数となる.

2.5 仮説検定の多重性

関連解析では検定される多くのハプロタイプ候補があるので, 多重性の問題が出てくる. 我々の解析では, FDR を制御するために, *B-H* 法 (Benjamini and Hochberg, 1995) を使った. 具体的には, ハプロタイプ数が m の時, 仮説数も m となる. その検定結果の m 個の P 値を昇順に並び替え, i 番目に P 値が小さいものを有意水準 $\alpha^* i/m$ で評価する. ただし α^* は任意の有意水準であり, 本論文では $\alpha^* = 0.05$ とする. これらの値を *q-value* とし, P 値がその値を下回った時, 仮説は棄却される.

2.6 R 関数

本研究で実装した関連解析の関数は, *association* と *association.full* である. 1 つ目の関数は, あるハプロタイプについて, 検定するものである. 2 つ目の関数は, *association* をすべてのハプロタイプ候補について適用するものであり, 結果はハプロタイプが P 値の昇順に従って並ぶ. ここの hapfre. は, 推定されたハプロタイプ頻度に従って決まったディプロタイプ型におけるそれぞれのハプロタイプの相対頻度である.

3 実データの解析

R パッケージ *SNPassoc* (Gonzalez *et al*., 2007) に入っている case-control study 用のサンプルデータ SNPs を実データとして解析する. 全個体数 157 だが, 罹患状態である case 群 109 個体のみを使う. 量的変量は動脈血圧 (blood.pre) とタンパク値 (protein) の 2 つである.

3.1 解析準備

SNPs には 35 座位あるが, 遺伝子型データを SNP データ解析環境統合システム (Tomita *et al*., 2004) で解析した結果より, 座位 {snp1001, snp1002, snp1005, snp1008} で解析する. アレルは A,T,G,C をそれぞれ 1,2,3,4 と置き換えた. このデータに *SNPassoc* に含まれている R パッケージ *haplo.stats* のハプロタイプ頻度推定関数 *haplo.em* を適用し, その頻度からディプロタイプ型の頻度を求め, 複数ディプロタイプ型の候補がある場合はディプロタイプ頻度を元に個体内の頻度を求めた. これらのデータをまとめたものが pre ファイルとなり, それが表 1 である. 本論文での量的データは正規分布に従うことを仮定するものなので, まずは SNPs データの量的変量 (動脈血圧, タンパク値) の正規性を Shapiro-Wilk 検定で確かめる. 動脈血圧は P 値が 0.1937 となり正規性の仮説は棄却されず, 正規分布に従っているといえる. 次に, タンパク値は P 値が 0.0002655 となり正規性があるとは言えない. そのタンパク値の対数をとったものは, P 値が 3.924e-07 となり, これも正規性があるとは言えない. 平方根をとったものは, P 値が 0.6397 となり正規性の仮説は棄却されず, 正規分布に従っているといえる. この結果より, 量的表現型として, 動脈血圧とタンパク値の平方根をとったものを使う. これを dat ファイルとし, 一部を表 2 にのせる.

3.2 解析

本論文で提案した関連解析で, 表 1,2 のデータを解析した. 2 変量, 血圧, タンパク値について, 3 つの発現モデルそれぞれで解析し, 有意水準 5% を下回ったハプロタイプを含む解析結果を以下の表に載せた. P 値が 5% を下回ったハプロタイプは, 表 3 で 2434, 2114, 4434, 表 4 で 4434, 2434, 表 6 で 4434, 表 7 で 4434, 2114, 表 8 で 4434, 2434 である.

表 1　pre ファイル

No	probability	pop.freq.	diplo.1	diplo.2
1	1.000000000	0.214906416	2434	2434
2	0.998551802	0.108278381	2114	2434
2	0.001448198	0.000157036	2414	2134
⋮	⋮	⋮	⋮	⋮
108	0.998551802	0.108278381	2114	2434
108	0.001448198	0.000157036	2134	2414
109	1.000000000	0.008425164	4433	2134

表 2　dat ファイル

No	動脈血圧	タンパク値の平方根
1	13.7	275.02822
2	12.7	169.37596
3	12.9	131.45186
⋮	⋮	⋮
107	13.1	233.62107
108	12.7	209.63738
109	15.0	60.03503

表 3　2 変量へのドミナントモデル適用結果

haplo.	hapfre.	Xsquared	Pvalue	qvalue	df
2434	0.4631	6.4001	0.04076	0.00714	2
2114	0.2406	6.2955	0.04295	0.01429	2
4434	0.0503	6.2168	0.04467	0.02143	2
4433	0.1860	3.5884	0.16626	0.02857	2
2134	0.0458	2.3288	0.31212	0.03571	2
2414	0.0049	2.1037	0.34929	0.04286	2
2433	0.0094	0.2623	0.87710	0.05000	2

表 4　タンパク値の平方根へのドミナントモデル適用結果

haplo.	hapfre.	Xsquared	Pvalue	qvalue	df
4434	0.0503	5.6357	0.01760	0.00714	1
2434	0.4631	5.4910	0.01911	0.01429	1
2114	0.2406	3.7168	0.05387	0.02143	1
4433	0.1860	3.2370	0.07199	0.02857	1
2414	0.0049	1.9104	0.16692	0.03572	1
2134	0.0458	1.9001	0.16807	0.04286	1
2433	0.0094	0.2367	0.62662	0.05000	1

表 5　2 変量へのレセッシブモデル適用結果

haplo.	hapfre.	Xsquared	Pvalue	qvalue	df
4434	0.0503	5.6858	0.05826	0.01	2
4433	0.1860	2.4836	0.28886	0.02	2
2114	0.2406	1.7883	0.40896	0.03	2
2434	0.4631	1.4215	0.49127	0.04	2
2134	0.0458	0.8350	0.65870	0.05	2

表 6　タンパク値の平方根へのレセッシブモデル適用結果

haplo.	hapfre.	Xsquared	Pvalue	qvalue	df
4434	0.0503	4.4229	0.03546	0.01	1
4433	0.1860	1.6579	0.19789	0.02	1
2434	0.4631	1.3309	0.24865	0.03	1
2114	0.2406	0.5124	0.47409	0.04	1
2134	0.0458	0.0281	0.86686	0.05	1

表 7　2 変量へのアディティブモデル適用結果

haplo.	hapfre.	Xsquared	Pvalue	qvalue	df
4434	0.0503	7.4349	0.02430	0.01	2
2114	0.2406	7.0612	0.02929	0.02	2
2434	0.4631	5.3402	0.06925	0.03	2
4433	0.1860	4.4291	0.10920	0.04	2
2134	0.0458	1.6185	0.44520	0.05	2

表 8　タンパク値の平方根へのアディティブモデル適用結果

haplo.	hapfre.	Xsquared	Pvalue	qvalue	df
4434	0.0503	7.3180	0.00683	0.01	1
2434	0.4631	4.7248	0.02973	0.02	1
2114	0.2406	3.8178	0.05071	0.03	1
4433	0.1860	3.7369	0.05322	0.04	1
2134	0.0458	1.5339	0.21552	0.05	1

4　結果の考察

P 値が q-value を下回ったのは, アディティブモデルの場合でタンパク値の平方根に関する解析でのハプロタイプ 4434 である. その時の q-value は 0.01 で, そのハプロタイプの P 値は 0.00683 となった. よって, タンパク値について, ハプロタイプ 4434 を 2 つ持つ群と持たない群の平均値は等しいという帰無仮説は棄却された. このことより, ハプロタイプ 4434 は他のハプロタイプよりもタンパク値に影響するものだと言う事ができる. ただ, ハプロタイプ頻度は約 0.05 と低く, 原因ハプロタイプと言い切るのはデータ数が少ないかもしれない.

5　結論

我々は多変量での関連解析の手法を提案した. この手法は, 3 つの発現モデルに対応する *QTLhaplo* の拡張である. また FDR を 制御するため q-value で結果を評価するようにし, R に実装した. 実データ解析より, 1 変量であるハプロタイプに差がある場合, 2 変量で見た場合も, ある程度差が出ることが解った. その上, 1 変量では差がないとなる場合でも, 多変量では差がないという仮説が棄却される場合がある (論文数値例参照). 以上の点で多変量へ拡張することにより有利になるだろう. また, ディプロタイプ型の候補を確率的に扱うことができ, 頻度の少ないハプロタイプの影響を小さくするといった点では, *QTLmarc* よりも有利である.

参考文献

[1] Shibata K., *et al*. (2004): Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. *Genetics* 168:525-539

[2] Kamitsuji S. and Kamatani N. (2006): Estimation of haplotype associated with several quantitative phenotypes based on maximization of are under a receiver operating characteristic (ROC) curve. *J Hum Genet.* ,51(4),314-325.

[3] Benjamini Y. and Hochberg Y. (1995): Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B, 57, 289-300

[4] Gonzalez J.R., *et al.*(2007) SNPassoc: an R package to perform whole genome association studies *Bioinformatics.*;23(5):644-645

[5] Gonzalez J.R., *et al*. (2007): SNPs, `http://davinci.crg.es/estivill_lab/snpassoc`.

[6] Tomita M., *et al*. (2004): Genome Analysis and Medicine, Program & Abstracts, p.92, The 13th Takeda Science Foundation Symposium on Bioscience, Hotel Okura, Tokyo.