Whole-Genome scan data の解析方法

東京大学医学部生物統計学教室
田中　紀子

今日のお話

- Whole-genome scan とは？
- Whole-genome scan data の解析方法とその問題点
- まとめ

遺伝子の研究　-遺伝疫学-

病気に関連する遺伝子を見つけたい！

↓

マーカーを使った研究　　アレイを使った研究

↓

ケース･コントロール研究

家系分析

など･･･

遺伝疫学研究で用いられる研究デザイン

家系データ

- **大家系を使った連鎖解析**
- Affected sib pair analysis (ASP)**による連鎖解析**
- TDT (Transmission/Disequilibrium Test) **デザイン**
  → S-TDT
  RC-TDT
- **双子研究**
- **ケース･コントロール研究**　　**など。。。**

研究デザイン

発症年齢の比較的高い生活習慣病などを対象とした場合は。。。

家系データ

- ~~**大家系を使った連鎖解析**~~
- △ Affected sib pair analysis (ASP)**による連鎖解析**
- ~~TDT (Transmission/Disequilibrium Test) **デザイン**~~

→ S-TDT

RC-TDT

- **双子研究**
- **ケース･コントロール研究**

今までの研究方法

- **対象疾患：**
  稀な疾患 (遺伝性疾患など)
  Complex disease **だと、乳がん、アルツハイマー病など**
- **研究デザイン**
  家系分析 (ASP TDT) 主流
- **対象とするマーカーの種類とallele数**
  VNTRs **数個～数十個**
- **対象者数**
  数人 (家系) ～数十人 (家系)

Whole-genome scan

- **対象疾患：**
  稀な疾患よりComplex disease, Common disease
- **研究デザイン**
  家系分析（ASP**など**）**だけではなく、**Population based design (TDT, **ケース・コントロール研究など**) **も多く行われるようになる。**

- **対象とするマーカーの種類と**allele**数**
  SNPs **数万個～　ある領域に絞っても数百個**
- **対象者数**
  数十人～数千人？

Whole-genome scan data を解析するときの問題点

1. いずれのデザインでも仮説検定でマーカーと疾患との関連があるかどうかが判断されている
2. マーカーの数が膨大

↓

検定の多重性の問題

検定の多重性について

- 多重性の問題とは…

  検定を複数回行った場合に全体としての結論の第一種の過誤の確率が1つずつの検定の有意水準より大きくなってしまう。

多重比較法

複数の検定の結果を合わせて1つの結論を導きたいような場合に、正しい帰無仮説のうち少なくとも1つが誤って棄却される確率の最大値を有意水準α（通常5%）以下に保つように一つ一つの検定の第一種の過誤を調整する方法

多重比較法

- 検定したい仮説や状況によって、適用可能な方法は異なる。例えば、k個の検定を行った場合

方法	比較あたりの有意水準α'	欠点	適用条件
Bonferroni	$\alpha' = \alpha/k$	K が多いとかなり保守的	たいてい適用可能
Sidak	$\alpha' = 1-(1-\alpha)^{1/k}$	精度よい (適用条件のもとで)	相関がないとき

多重性を考慮した有意水準－連鎖解析・TDTの場合－

- 有意水準の調整
  - ボンフェローニの方法
    - 検定回数（調べたマーカーの数）による
  - Lander and Kruglyak(1995)
    - 無限に緻密なマーカーマップができたとして
    - ASPで2.5×10^{-5}
  - Morton(1955, 1998)
    - 連鎖解析で1×10^{-4}で大丈夫
  - Risch and Merikangas(1996)
    - ヒト遺伝子が10万個あるとして　TDTで5×10^{-8}

多重性を考慮した有意水準

- **有意水準の調整**
  - Lander and Kruglyak(1995)
  - Risch and Merikangas(1996)
  - Morton(1955, 1998)
  - **ボンフェローニの方法**

↓

どれをとっても非常に厳しい値

Complex diseaseに関与する遺伝子一つ一つのリスクは小さいと考えられることから、検出するには相当のサンプルサイズが必要になる

多重性問題を解決するためには・・・

1.仮説検証型の研究にしてマーカーの数を減らす

→仮説の構築方法を考える必要有り

・分子生物学的アプローチ

・バイオインフォマティックス的アプローチ

2.マーカーの数はそのままで統計的に調整を行う

→調整の方法を考える必要有り

・permutation　test

・ closed min-p method

・combining p-value　method

Combined P-Value method

- 複数の帰無仮説について検定を行ったときに計算される複数のp値から、全体としての結論の第一種の過誤の確率 (タイプ1FWE)を計算する方法

例えば…

制御しにくい測定不能な環境下において複数回繰り返された生物学的実験で複数回検定を行ったとき

raw　dataは得られないがp値だけが得られるような状況でメタアナリシス を行いたいとき

などの状況で適用されてきている

Combined P-Value method

今日紹介する方法

- Fisher's combined probability test (Fisher, 1932)
- P_{sum} test (Edgington, 1972)
- 標準正規スコアに変換する方法 (Stouffer et al., 1949)
- Simesの方法 (Simes, 1986)
- Wilkinsonの方法 (Wilkinson,1951)
- Truncated Probability Method (TPM) (Zaykin et al.,2002)

Fisher's combined probability test (Fisher, 1932)

L個の帰無仮説 $H_i, i = 1, \cdots, L$ に関して計算されたp値

p_i ～U[0,1] （L個の帰無仮説が真の場合）

$$t = -2\sum_{i=1}^{L} \ln p_i = -2\ln\left(\prod_i^L p_i\right)$$

は自由度2Lのχ2分布に従うので、タイプ1FWEは

$$P_{Fisher} = \Pr(\chi^2_{2L} \geq t)$$

P_{sum} test (Edgington, 1972)

L個のp値 p_i(i=1,…,L)が区間[0,1]の一様分布に従うとき、その和$S_L=p_1+p_2+\ldots+p_L$がS以下である確率は、一様分布にしたがう確率変数の和の分布であるから

$$P_{sum} = 1 - \Pr(S_L \le s)$$

$$= 1 - \sum_{r=0}^{L} (-1)^r \binom{L}{r} \frac{(s-L)_+^L}{L!}$$

但し $x_+ = \dfrac{x + |x|}{2}$

標準正規スコアに変換する方法（Stouffer et al., 1949）

帰無仮説H_iが真のとき、p_iはそれぞれ標準正規スコア

$$z_i = \Phi^{-1}(1 - p_i)$$

と変換できるので、

$$P_{normal} = 1 - \Phi\left(\frac{1}{L}\sum_{i=1}^{L}\Phi^{-1}(1 - p_i)\right)$$

Simesの方法（Simes, 1986）

L個のp値 $p_i (i=1,\ldots,L)$ を小さい順に並べ替えたものを $p^{(i)} (i=1,\ldots,L)$ とすると、Simesの不等式

$$\Pr\left\{\bigcup_{i}^{L}(p^{(i)} \le i\alpha / L)\right\} \le \alpha$$

より、

$$P_{Simes} = \min\{Lp^{(i)} / i\}$$

Wilkinsonの方法（Wilkinson,1951）

全ての帰無仮説が真のとき、ある値τ以下をとるp値の数は二項分布$B(L,\tau)$に従うので、

比較あたりの有意水準は少なくともk個の値がτ以下になる確率

$$\sum_{i=k}^{L}\binom{L}{i}\tau^{i}(1-\tau)^{L-i}$$

となり、k=1のとき

$$\alpha_{\text{wilkinson}}=1-(1-\alpha)^{L}$$

方法間の比較

- Little とFolks[1971]
  Fisherの方法がもっとも相対効率がよい
- Naik[1969]
  Wilkinsonの方法の方がFisherの方法より検出力が高い
- HedgeとOlkin[1985]
  P_{sum} testは大きいp値に影響を受けやすい
- Rice[1990]
  P_{sum} testや正規スコアに変換する方法の方が検出力は劣るが個々の試験 (実験) の重みが等しいという点で優れている

TPM (Zaykin et al., 2002)

L個のp値 $p_i (i=1,\ldots,L)$が観測されたとき、ある値τ未満のp値について掛け合わせた値

$$W = \prod_{i=1}^{L} p_i^{I(p_i \le \tau)}$$

の確率分布

$$\Pr(W \le w) = \sum_{k=1}^{L} \binom{L}{k} (1-\tau)^{L-k} \left(w \sum_{s=0}^{k-1} \frac{(k \ln \tau - \ln w)^s}{s!} I(w \le \tau^k) + \tau^k I(w > \tau^k) \right)$$

を計算して、タイプ1FWEとする

ただし、$I(\cdot)$は指示関数

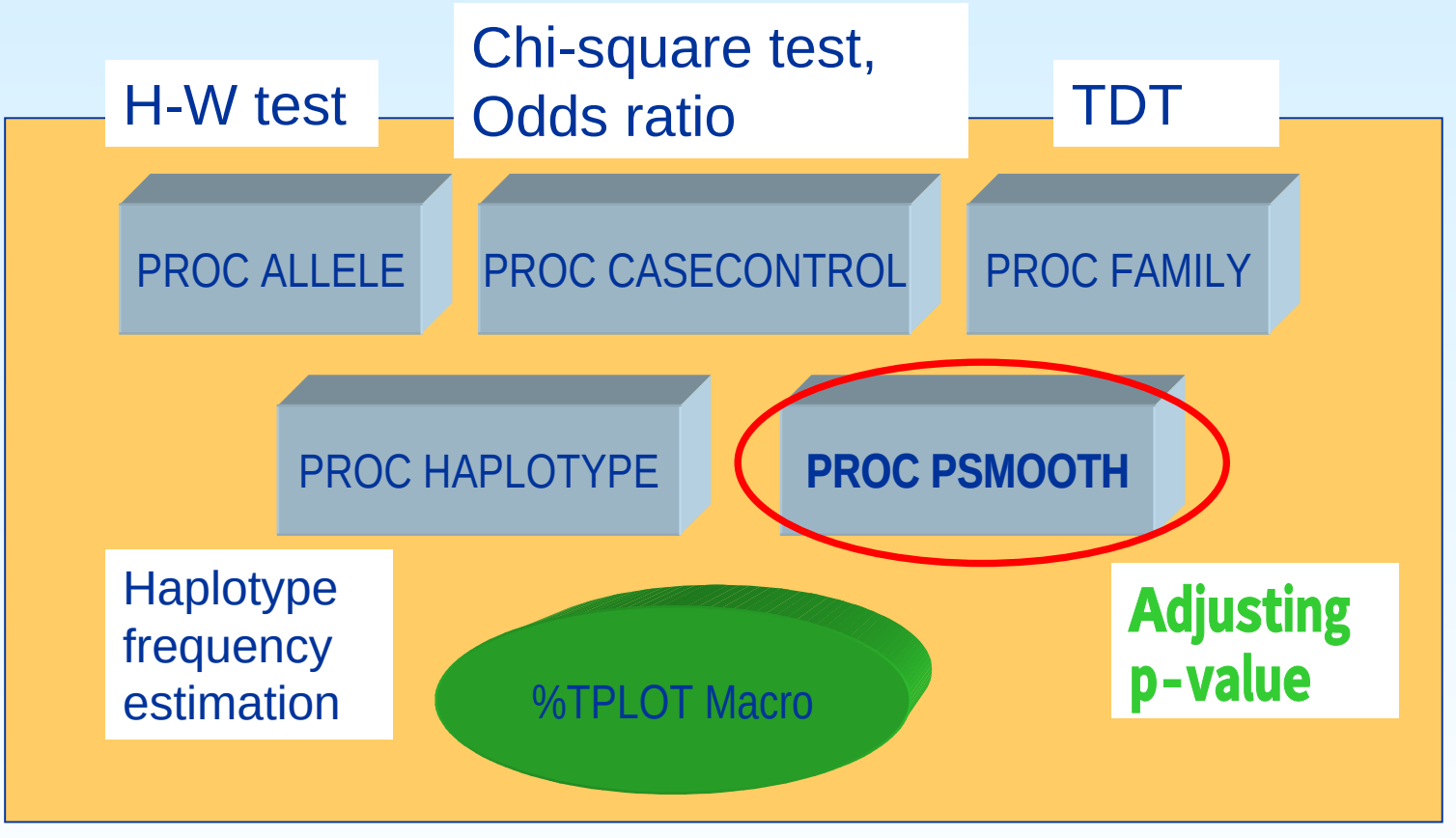
P値に相関がある場合の*w*の計算方法

- データから相関行列が推定された場合、相関行列からraw p値を変換したp’値で計算する
- 式から計算するのでなくモンテカルロシミュレーションにより計算する
- ResamplingによりWの分布を推定して計算する

TPMの利点

- Wを計算するのに小さいp値だけで計算を行うことで高い検出力が得られる
- 任意にτを設定できる。
- P値間の相関を考慮することができる
- 重み付けなどの拡張が容易である

SAS/Genetics Software



SAS/Genetics psmooth　プロシジャ

- FWE: Bonferroni もしくは　Sidakの方法で調整
- マーカー間の相関を考慮した調整:
  Simes もしくはFisher の方法で調整
- 対数変換した調整済みp値をプロットしグラフ表示

まとめ

- Whole-Genome scan data **を解析するときに問題となる検定の多重性問題について説明した**

- **多重性問題に対処するために、**SAS/Genetics **では、**PSMOOTH**プロシジャで**Simes, Fisher**の方法を用いた調整を実行することができる。**