# 細胞種ごとのDNAメチル化変動を推定する統計手法

竹内史比古
国立国際医療研究センター(NCGM)研究所

2021年3月30日

日本エピジェネティクス研究会年会

http://www.fumihiko.takeuchi.name

# 要旨

【背景】エピゲノムワイド関連解析(EWAS)では、病変組織と正常組織を比較してCpGサイトのメチル化変動を解析する。

【目的】組織丸ごとバルクで計測しつつも、細胞種ごとのメチル化変動を統計的に推定するという、都合の良い話。

【方法】従来法は線形回帰。非線形リッジ回帰によるプログラムを開発した。シミュレーションと実データで検証。

【結果】提案法は、感度と陽性的中率のバランスが取れていた。

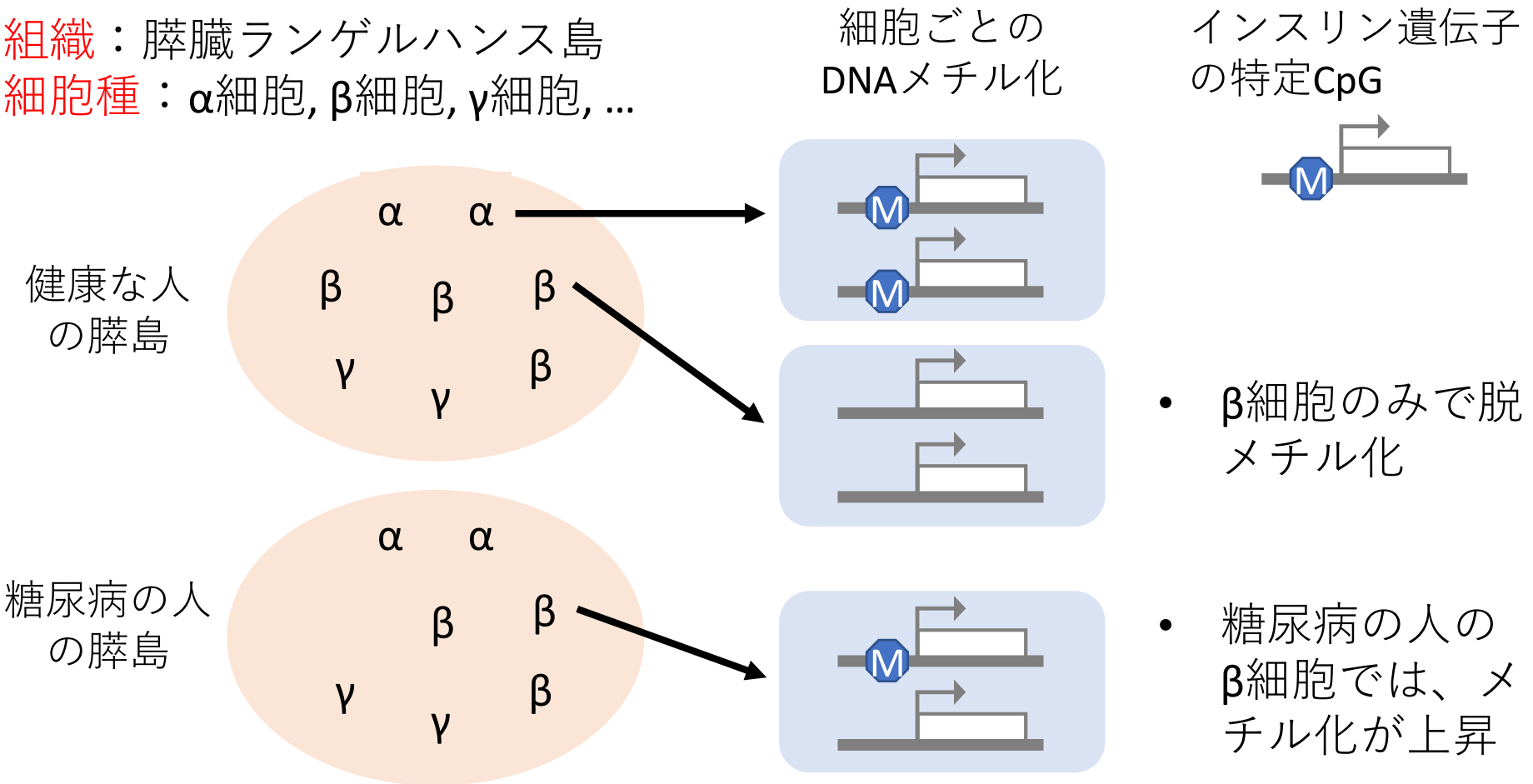
【結論】セルソートした実験には及ばないものの、バルク組織の実験データからもある程度は細胞種ごとのDNAメチル化変動を推定できた。

BMC Bioinformaticsに先週出版されました　https://rdcu.be/chh6N
R言語パッケージを公開　https://github.com/fumi-github/omicwas

# 知りたいこと：
# 細胞種ごとのDNAメチル化変動

- 病気で、特定の細胞種のDNAメチル化が変動する

組織：膵臓ランゲルハンス島
細胞種：α細胞, β細胞, γ細胞, …

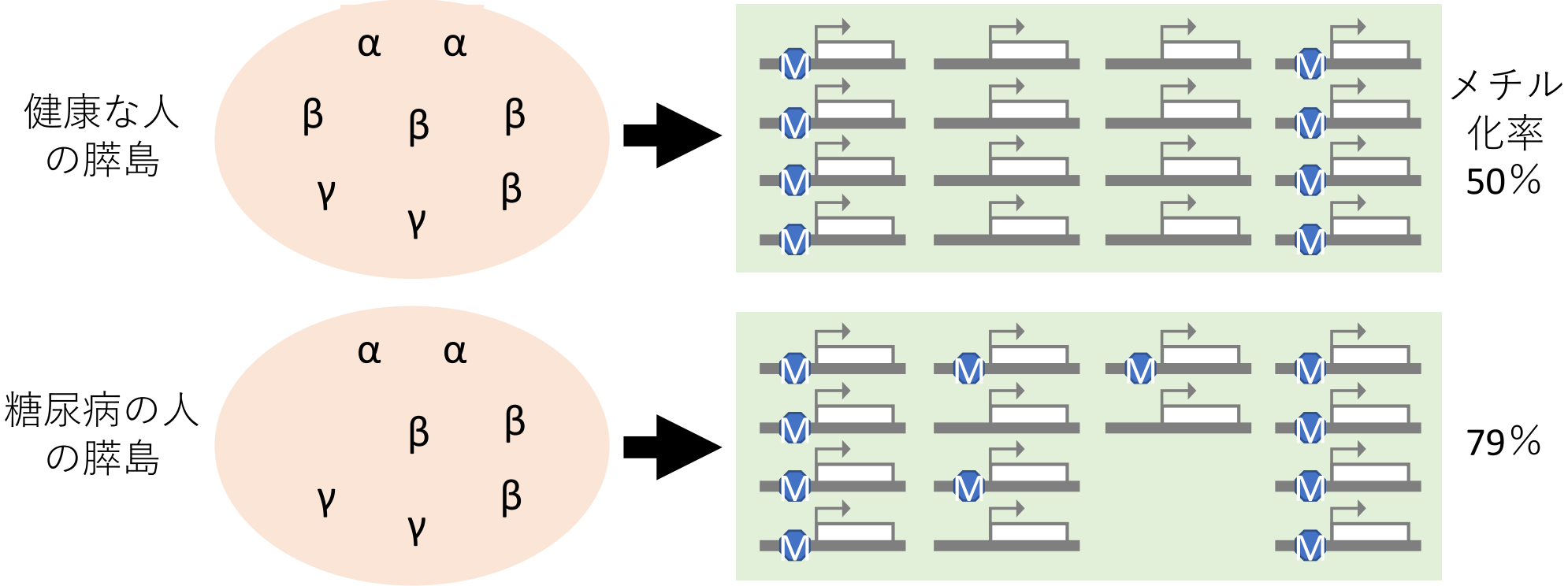



- β細胞のみで脱メチル化
- 糖尿病の人のβ細胞では、メチル化が上昇

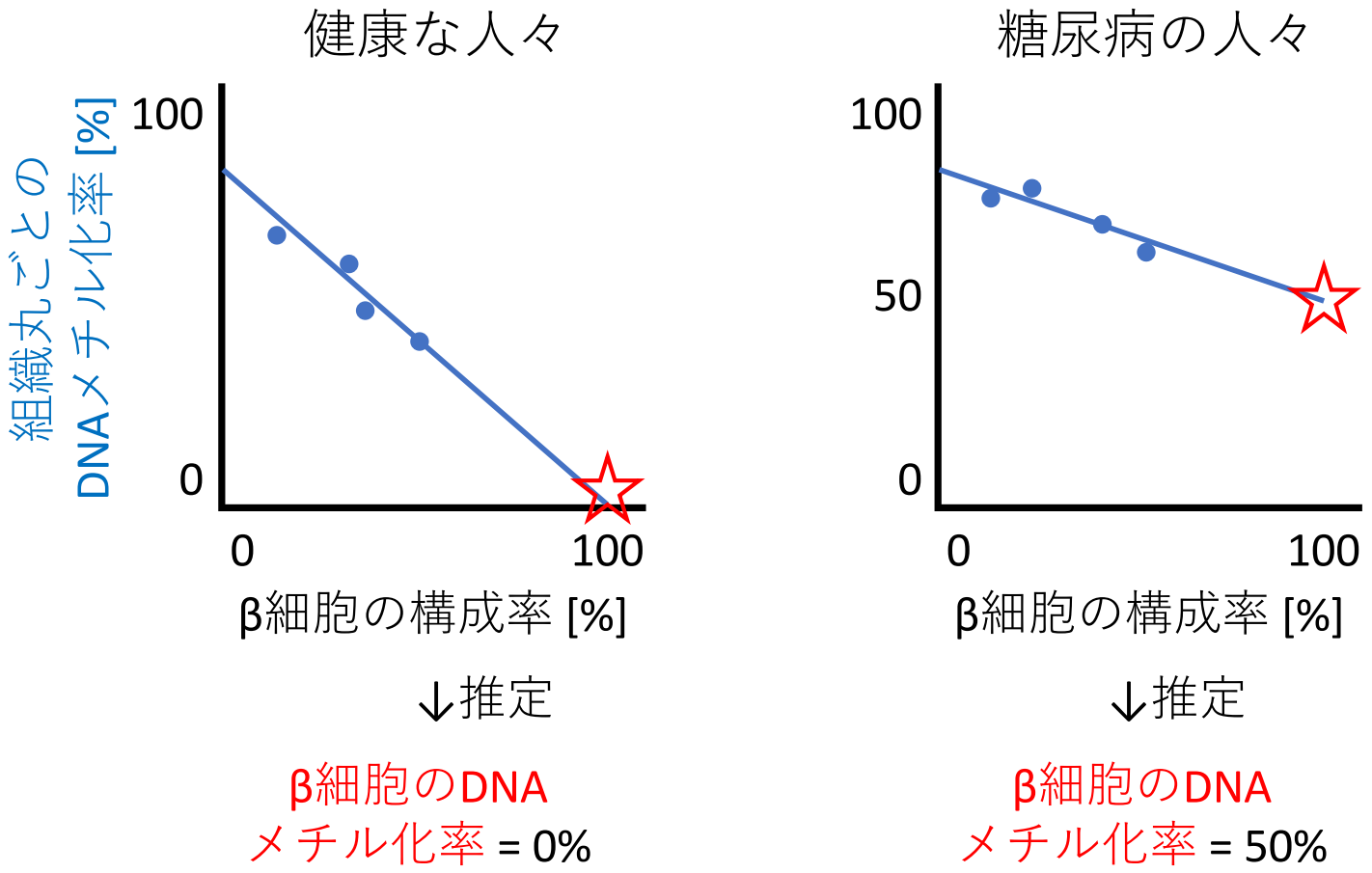
# 簡便に測定できること：
# 組織丸ごとのDNAメチル化変動

- 細胞種ごとのDNAメチル化を推定したい
  - そんなに都合良く行くだろうか…
- 細胞種構成率は、推定済み

「β細胞でのDNAメチル化が糖尿病の人で上昇している」ことを推定したい



# 統計的推定の考え方




β細胞の構成率のバラツキを利用する

# 従来の定式化（線形回帰）

- 以後、1つのCpGに注目
- 指数
  - 細胞種 h, 検体 i
- 入力データ
  - 細胞種構成率 $W_{h\,i}$
  - 罹患状態 $X_i$ （中心化）
    - 細胞種ごとに影響
  - 共変数 $C_i$ （中心化）
    - 細胞種に依らず影響
  - 組織でのメチル化率 $Y_i$

$\varepsilon_i \sim N(0, \sigma^2)$ i.i.d.

- 推定したいパラメータ
  - メチル化率のベースレベル $\alpha_h$
  - $X_i$ の効果 $\beta_h$
  - $C_i$ の効果 $\gamma$

- 細胞種 h でのメチル化率

$$\alpha_h + \beta_h X_i$$

$$\downarrow \sum_h W_{h,i} \times$$

- **全体モデル**

$$Y_i = \sum_h \alpha_h W_{h,i} + \sum_h \beta_h W_{h,i} X_i + \gamma C_i + \varepsilon_i$$

- 細胞種 h のみの**周辺モデル**

$$Y_i = \sum_{h'} \alpha_{h'} W_{h',i} + \beta_h W_{h,i} X_i + \gamma C_i + \varepsilon_i$$

# 従来法の問題点、本研究の提案

- DNAメチル化比較はlogitスケールで行うべき

➔非線形回帰

- 相互作用項 $W_{h,i}X_i$ の多重共線性

➔リッジ回帰

$\sum_i {\varepsilon_i}^2 + \lambda \sum_h {\beta_h}^2$ を最小化

正則化パラメータ $\lambda$

- 細胞種 h でのメチル化率

$\alpha_h + \beta_h X_i$

- 線形モデル

$$\mu_i = \sum_h W_{h,i}(\alpha_h + \beta_h X_i) + \gamma C_i; \quad Y_i = \mu_i + \varepsilon_i$$

- 非線形モデル

$$\mu_i = f\left(\sum_h W_{h,i}\, g(\alpha_h + \beta_h X_i)\right) + \gamma C_i; \quad f(Y_i) = \mu_i + \varepsilon_i$$

線形スケール ⇄ logitスケール

正規化 $f$ = logit

逆 $g$ = logistic

$\varepsilon_i \sim N(0, \sigma^2)$ i.i.d.

計画行列

$$\frac{\partial \mu_i}{\partial \beta_h} = W_{h,i}X_i$$

細胞種 $h$ の割合

$$\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \beta_h} = \widetilde{W_{h,i}}(\boldsymbol{\beta})X_i$$

メチル化CpG中の細胞種 $h$ の寄与分

➔ 多重共線性弱まる

# 相互作用項の多重共線性

- 関節リウマチ
  - 罹患者336名
  - 健常者322名
- 末梢血の白血球7種
- 細胞種構成率 $W_{h,i}$
  - 細胞種間で弱い相関
  - 変動係数小さい

| | 好中球 | CD4+T | CD8+T | NK | 単球 | B細胞 | 好酸球 | |
|---|---|---|---|---|---|---|---|---|
| 平均 | 0.59 | 0.10 | 0.08 | 0.08 | 0.07 | 0.07 | 0.01 | |
| 標準偏差 | 0.11 | 0.06 | 0.05 | 0.04 | 0.02 | 0.03 | 0.02 | |
| 変動係数 | 0.2 | 0.6 | 0.6 | 0.5 | 0.3 | 0.4 | 2.7 | |
| | | | | | | | | |
| *r* | 好中球 | CD4+T | CD8+T | NK | 単球 | B細胞 | 好酸球 | 疾患 X |
| 好中球 | 1 | -0.68 | -0.60 | -0.46 | -0.06 | -0.49 | -0.48 | 0.44 |
| CD4+T | -0.68 | 1 | 0.14 | 0.05 | -0.17 | 0.38 | 0.26 | -0.33 |
| CD8+T | -0.60 | 0.14 | 1 | 0.08 | -0.05 | 0.19 | 0.13 | -0.27 |
| NK | -0.46 | 0.05 | 0.08 | 1 | -0.04 | 0.01 | 0.11 | -0.27 |
| 単球 | -0.06 | -0.17 | -0.05 | -0.04 | 1 | -0.17 | 0.05 | 0.10 |
| B細胞 | -0.49 | 0.38 | 0.19 | 0.01 | -0.17 | 1 | 0.11 | -0.22 |

ほぼ定数・$X_i$

- 疾患との相互作用項 $W_{h,i}X_i$ 同士は強い相関

➔ 回帰の精度を下げる

| *r* | 好中球*X | CD4+*X | CD8+*X | NK*X | 単球*X | B細胞*X | 好酸球*X |
|---|---|---|---|---|---|---|---|
| 好中球*X | 1 | 0.83 | 0.80 | 0.85 | 0.93 | 0.90 | 0.27 |
| CD4+*X | 0.83 | 1 | 0.78 | 0.78 | 0.83 | 0.88 | 0.42 |
| CD8+*X | 0.80 | 0.78 | 1 | 0.77 | 0.82 | 0.83 | 0.35 |
| NK*X | 0.85 | 0.78 | 0.77 | 1 | 0.85 | 0.83 | 0.35 |
| 単球*X | 0.93 | 0.83 | 0.82 | 0.85 | 1 | 0.88 | 0.35 |
| B細胞*X | 0.90 | 0.88 | 0.83 | 0.83 | 0.88 | 1 | 0.36 |
| 好酸球*X | 0.27 | 0.42 | 0.35 | 0.35 | 0.35 | 0.36 | 1 |

# リッジ回帰の正則化パラメータ選択

- 何を指標にするか？
  - MSE, AIC, BIC, GCV, CV, …
  - 病気の遺伝子発現への効果を予測したいので Mean Squared Error, $\mathrm{MSE}[\hat{\boldsymbol{\beta}}(\lambda)]$ を最小化する
- どのように選択するか？
  - 計画行列を特異値分解 $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = UDV^T$
  - $\mathrm{MSE}[\hat{\boldsymbol{\beta}}(\lambda)] = \left\|\mathrm{Bias}[\hat{\boldsymbol{\beta}}(\lambda)]\right\|^2 + \mathrm{tr}(\mathrm{Var}[\hat{\boldsymbol{\beta}}(\lambda)])$
    $= \sum_{m=1}^{M} \left(\frac{\lambda}{{d_m}^2+\lambda}\right)^2 (\boldsymbol{v}_m{}^T\boldsymbol{\beta})^2 + \left(\frac{{d_m}^2}{{d_m}^2+\lambda}\right)^2 \left(\frac{\sigma^2}{{d_m}^2}\right)$
  - 各$m$については $\lambda_m = \sigma^2/(\boldsymbol{v}_m{}^T\boldsymbol{\beta})^2$ が最小化
  - [Hoerl 他 1975] $\lambda_m$ の調和平均（逆数の平均の逆数）
  - [Lawless 他 1976] 逆数を精度 ${d_m}^2$ で重みづけ
  - [本研究] さらにバイアスを引く

- $\lambda$ ↑に伴い、
  Bias ↑
  Var ↓
- バランスを取る

# 実装

- Rパッケージomicwasとして実装した
- PORTライブラリのNL2SOLで最小化

- 非線形モデル

$$\mu_i = f\left(\sum_h W_{h,i}\, g(\alpha_h + \beta_h X_i)\right) + \gamma C_i$$

$$f(Y_i) = \mu_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2) \text{ i.i.d.}$$

- リッジ回帰

$\sum_i {\varepsilon_i}^2 + \lambda \sum_h {\beta_h}^2$ を最小化

メチル化率のベースレベル $\boldsymbol{\alpha}$
>> $X_i$ の効果 $\boldsymbol{\beta}$
$C_i$ の効果 $\boldsymbol{\gamma}$

1. 最小2乗回帰（1回目）
   - $\boldsymbol{\beta} = \boldsymbol{\gamma} = \mathbf{0}$ として $\hat{\boldsymbol{\alpha}}(0)$ を推定
   - $\widehat{\sigma^2}$ を推定
2. 最小2乗回帰（2回目）
   - $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}(0)$ として $\hat{\boldsymbol{\beta}}(0)$ , $\hat{\boldsymbol{\gamma}}(0)$ を推定
   - 正則化パラメータ $\lambda$ を決定
3. リッジ回帰
   - $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}(0)$ として $\hat{\boldsymbol{\beta}}(\lambda)$ , $\hat{\boldsymbol{\gamma}}(\lambda)$ を推定
   - “non-exact” t-type test (Wald検定と同じ式)で検定

# 比較検討した検定アルゴリズム

- 周辺モデル
  - 周辺モデル
  - TCA
- 非線形回帰・リッジ回帰（本研究）
  - logit.ridge
  - logit
  - ridge
- 全体モデル
  - 全体モデル
  - $f = g$ = identity
  - TOAST
  - CellDMC
- 周辺+全体のハイブリッドモデル（本研究）
  - 周辺モデルで $P < 2.4 \times 10^{-7}$ かつ
    全体モデルで $P < 0.05$ かつ同じ向きの効果

- 有意水準
  - 45万個のCpGの多重検定補整が必要
  - $P < 2.4 \times 10^{-7}$

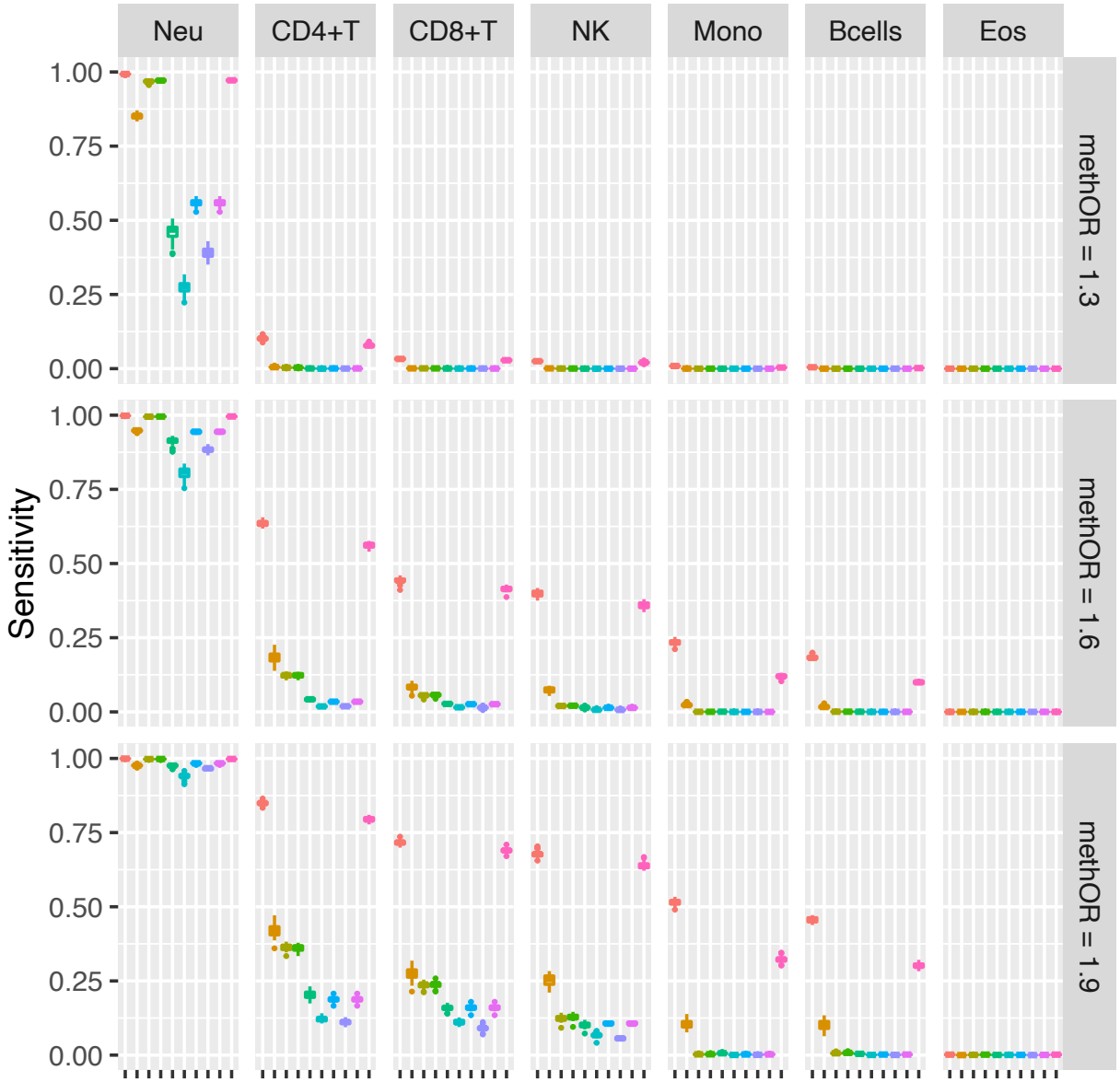
# [方法] 実データに基づくシミュレーション

- 関節リウマチの末梢血白血球EWAS
  - 罹患者336名, 健常者322名
  - 451,725 CpG
- 白血球7種の構成率 $W_{hi}$ はGLINTプログラムで推定
- 3シナリオ×50回の試行
- 検体の半々を無作為に罹患者・健常者に割当

- CpGを無作為に分類
  - [95%] 元データのまま
    - 疾患と無関係
  - [2.5%] 単一細胞種が罹患者でメチル化上昇
    - 各細胞種につき、2.5%÷7 = 1613 CpGが上昇
  - [2.5%] 単一細胞種が罹患者でメチル化低下
- メチル化変動はオッズ比1.3, 1.6, 1.9のいずれかに固定
- 各人各細胞種のメチル化率をランダム生成
  - 平均と標準偏差は元データから

# 感度 Sensitivity

$$\frac{TP}{TP + FN}$$




強い効果(OR)➔高
高割合の細胞種➔高

Algorithm

- Marginal
- TCA
- omicwas.logit.ridge
- omicwas.identity.ridge
- omicwas.logit
- omicwas.identity
- Full
- TOAST
- CellDMC
- Marginal.Full005

周辺モデル 高

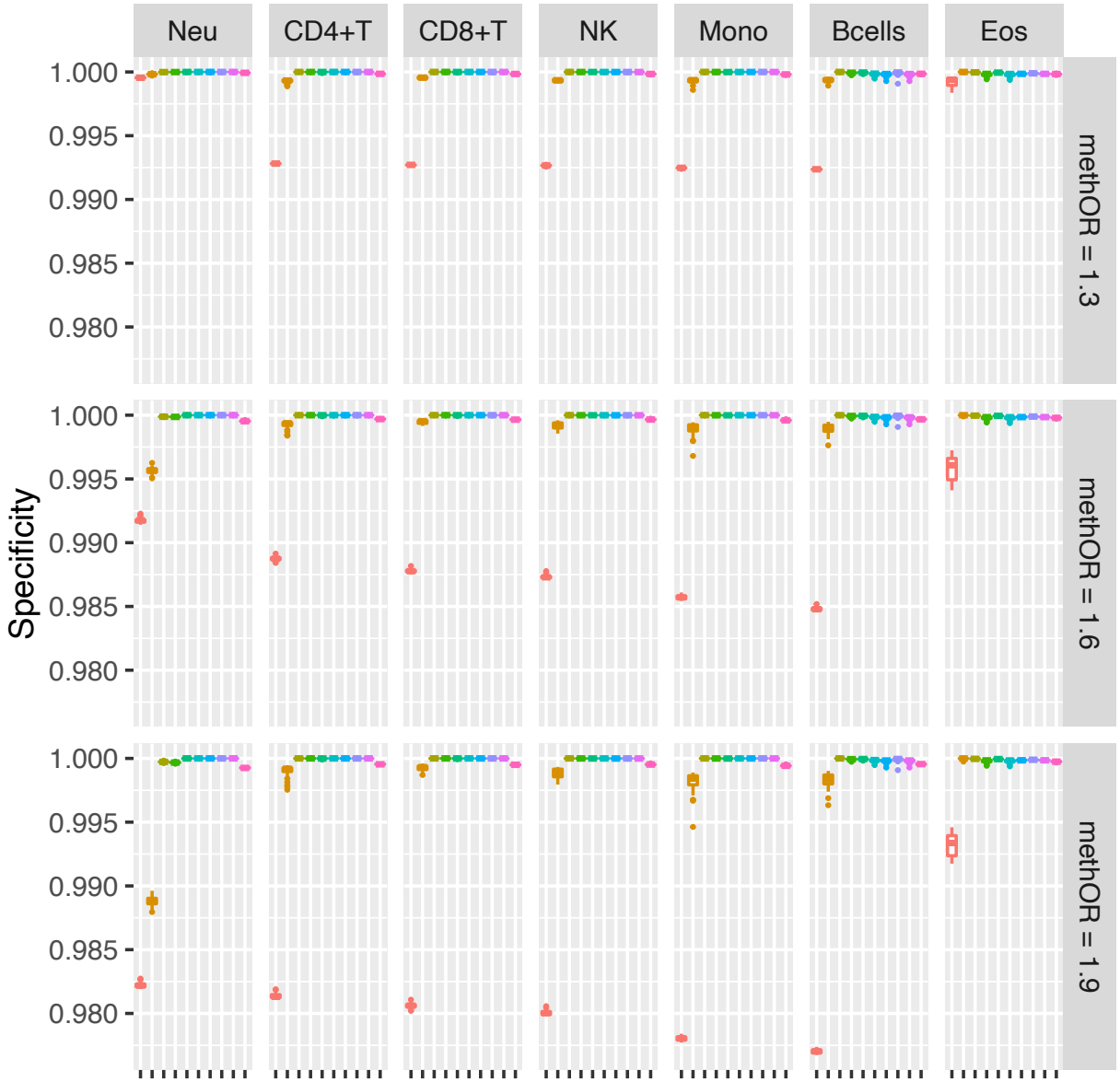
logit, ridge 中

全体モデル 低

周辺+全体モデル 高

# 特異度 Specificity

$$\frac{TN}{TN+FP}$$




どれも平均>0.97

Algorithm

- Marginal
- TCA
- omicwas.logit.ridge
- omicwas.identity.ridge
- omicwas.logit
- omicwas.identity
- Full
- TOAST
- CellDMC
- Marginal.Full005

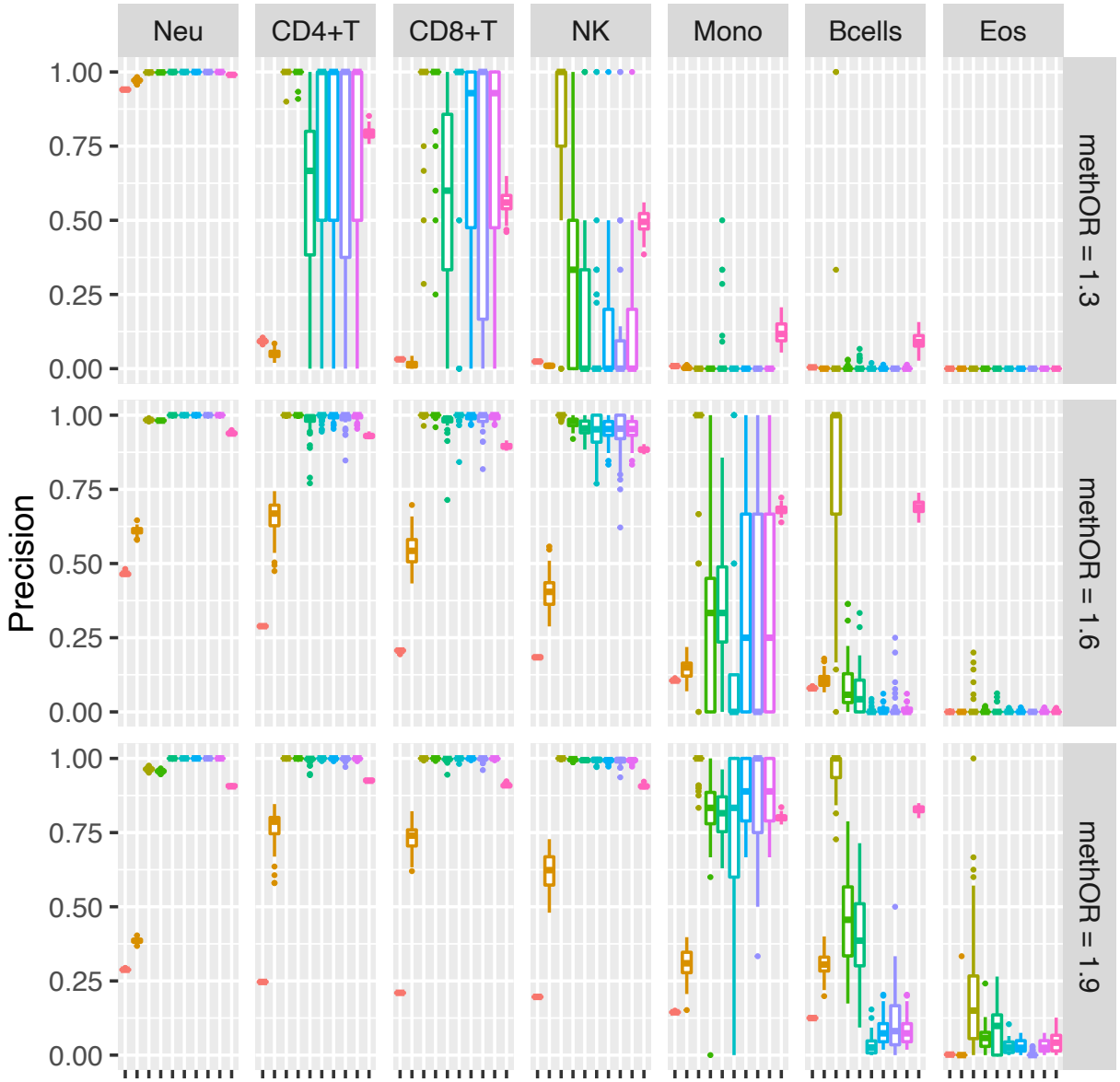
周辺モデル >0.97

logit, ridge >0.999

全体モデル >0.999

周辺+全体モデル >0.999

# 陽性的中率 PPV, Precision $\frac{TP}{TP + FP}$




Algorithm

- Marginal
- TCA
- omicwas.logit.ridge
- omicwas.identity.ridge
- omicwas.logit
- omicwas.identity
- Full
- TOAST
- CellDMC
- Marginal.Full005

周辺モデル 低

logit, ridge 高

全体モデル

周辺+全体モデル

どのアルゴリズムも感度0のケースを除外すると、omicwas.logit.ridgeは>0.79で、13/16のケースでベスト

# シミュレーション結果のまとめ

| モデル | 感度 | 特異度 | 陽性的中率 | 総合評価 |
|---|---|---|---|---|
| 周辺<br>TCA | 高 | >0.97 | | △ |
| logistic.ridge<br>ridge<br>logistic | 中 | >0.999 | 13/16でベスト | △ |
| 全体<br>identity<br>TOAST<br>CellDMC | 低 | >0.999 | 3/16でベスト | × |
| 周辺+全体 | 高 | >0.999 | | △ |

- 周辺モデル; 特定細胞種 h のみ

$$Y_i = \sum_{h'} \alpha_{h'} W_{h',i} + \beta_h W_{h,i} X_i + \gamma C_i + \varepsilon_i$$

  - h が当たりなら高感度で検出
  - h が外れでも多重共線性から誤検出
- 全体モデル

$$Y_i = \sum_h \alpha_h W_{h,i} + \sum_h \beta_h W_{h,i} X_i + \gamma C_i + \varepsilon_i$$

  - 多重共線性により低感度
- logistic, ridge
  - 周辺モデルと全体モデルの中間
- 周辺+全体のハイブリッドモデル
  - 周辺モデルで$P < 2.4 \times 10^{-7}$かつ ➔感度
  - 全体モデルで$P < 0.05$ ➔特異度