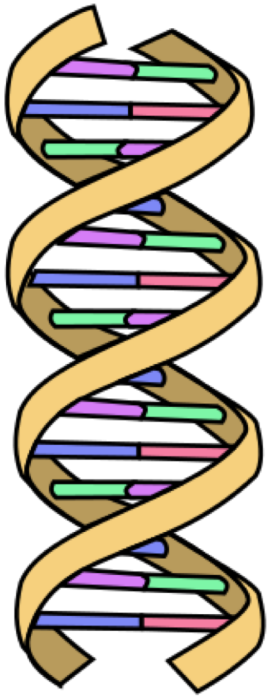
[3AAM-01] 疫学統計

非線形リッジ回帰による、細胞種ごとの遺伝子発現変動の推定

竹内史比古
国立国際医療研究センター(NCGM)研究所

2020年9月11日
統計関連学会連合大会　@オンライン

http://www.fumihiko.takeuchi.name



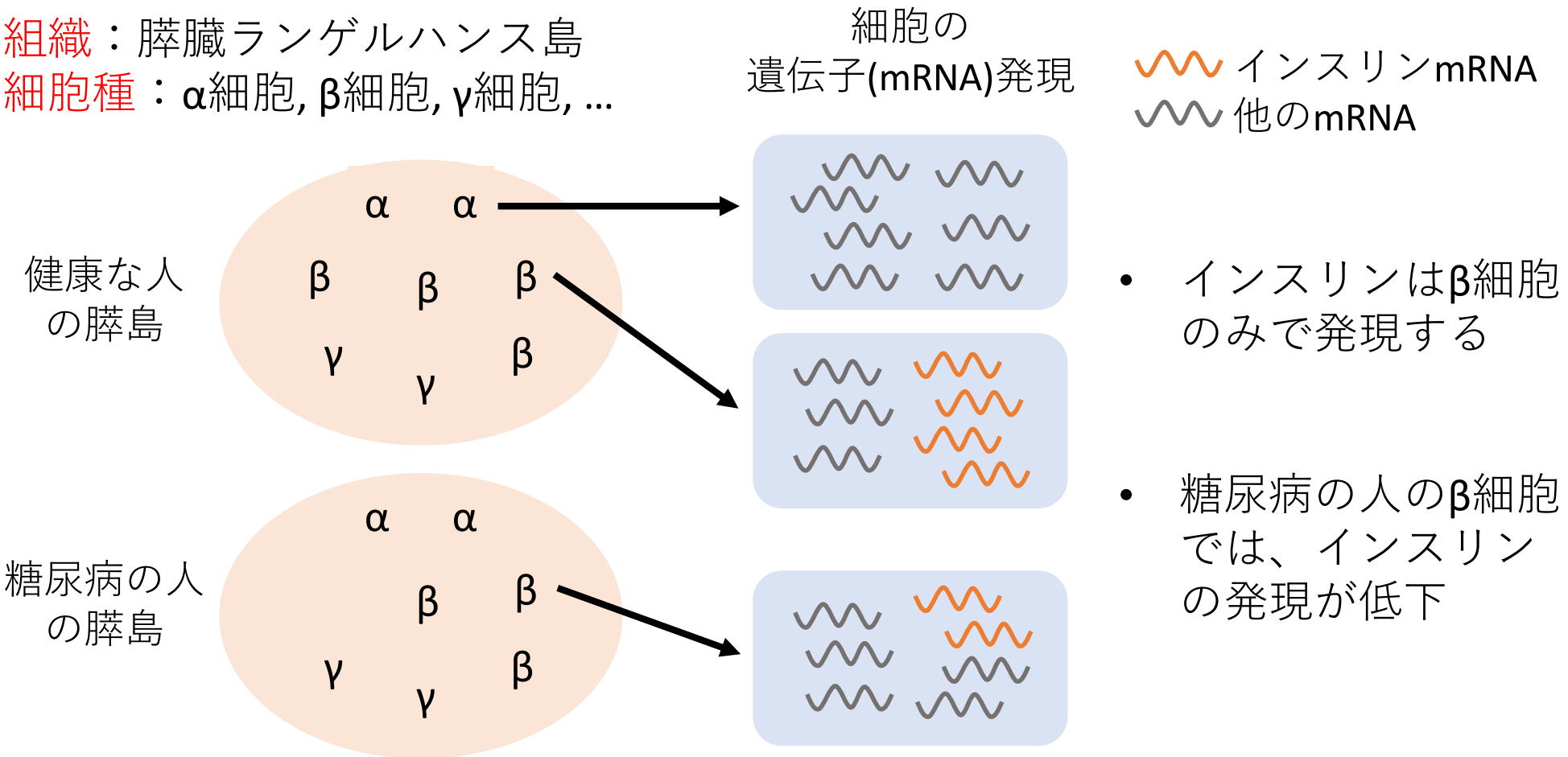
- 対象とする問題
- 従来法の問題点
- 提案法

- シミュレーションによる検証
- 実データでの検証

知りたいこと：
細胞種ごとの遺伝子発現変動

- 病気で、特定の細胞種の遺伝子発現が変動する

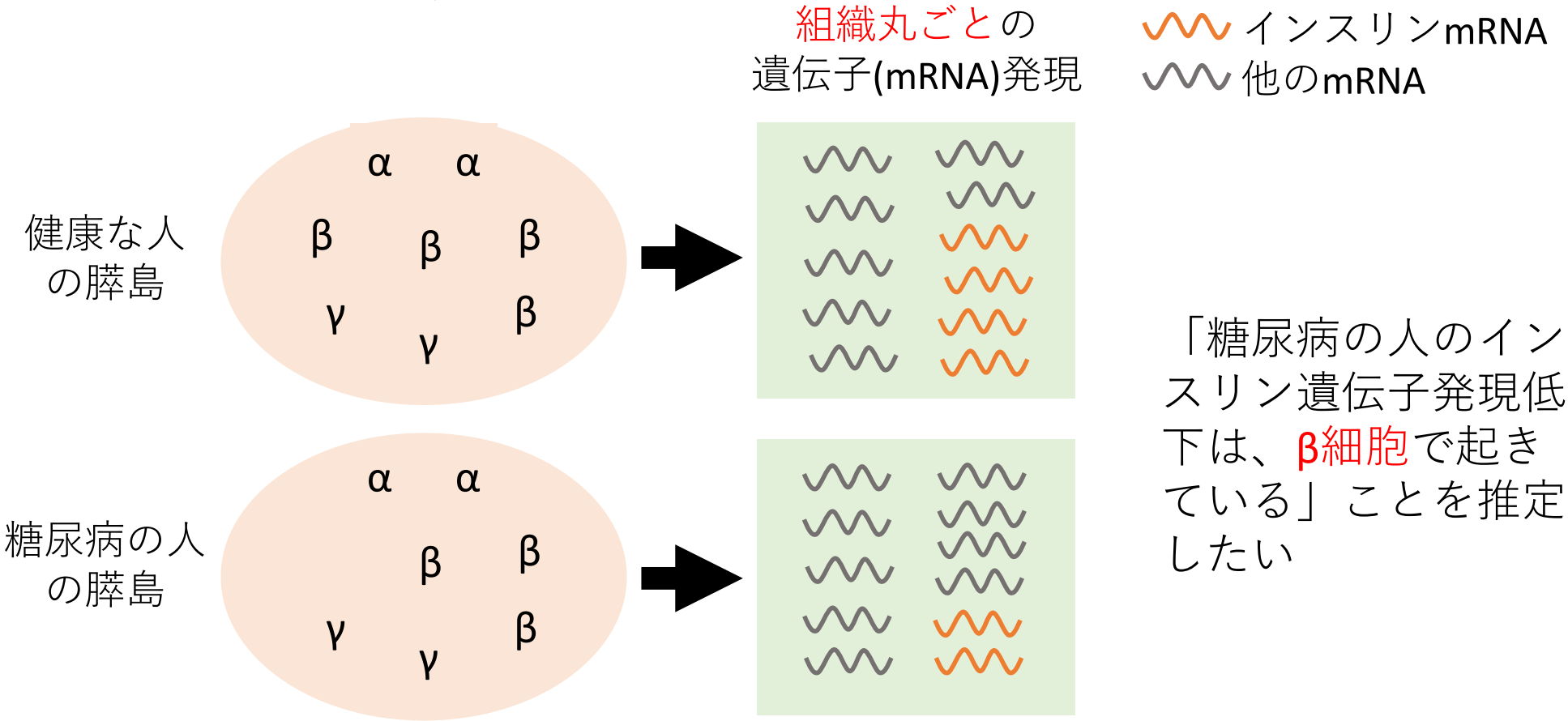
組織：膵臓ランゲルハンス島
細胞種：α細胞, β細胞, γ細胞, …



- インスリンはβ細胞のみで発現する
- 糖尿病の人のβ細胞では、インスリンの発現が低下

簡便に測定できること：
組織丸ごとの遺伝子発現変動

- 細胞種ごとの遺伝子発現を推定したい
  - 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



「糖尿病の人のインスリン遺伝子発現低下は、β細胞で起きている」ことを推定したい

従来の定式化（線形回帰）

- 以後、1つの遺伝子に注目
- 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
  - 組織での遺伝子発現 Y_i

- 推定したいパラメータ
  - 遺伝子発現のベースレベル α_h
  - X_i の効果 β_h
  - C_i の効果 γ

- 細胞種 h での遺伝子発現

$$\alpha_h + \beta_h X_i$$

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

- **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**

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

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

従来法の問題点、本研究の提案

- 遺伝子発現比較はlog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（次スライド）

➔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 λ

- 細胞種 h での遺伝子発現

$$\alpha_h + \beta_h X_i$$

- 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 ⇄ logスケール

正規化 $f = \log$

逆 $g = \exp$

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

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

この遺伝子のmRNAの細胞種 h での割合

➔ 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
- 血液中の白血球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] λ_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

- λ ↑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

遺伝子発現のベースレベル $\boldsymbol{\alpha}$
>> X_i の効果 $\boldsymbol{\beta}$
C_i の効果 $\boldsymbol{\gamma}$

- 非線形モデル

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

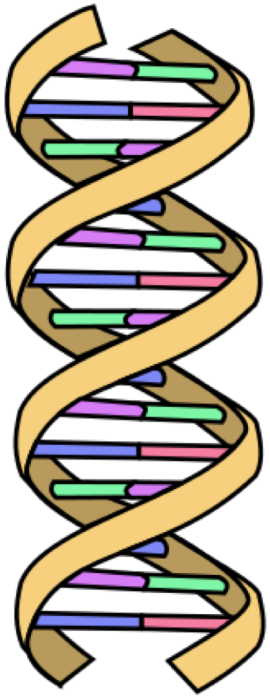
$$f(Y_i) = \mu_i + \varepsilon_i$$

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

- リッジ回帰

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

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 λ 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



- 対象とする問題
- 従来法の問題点
- 提案法

- シミュレーションによる検証
- 実データでの検証

比較検討した検定アルゴリズム

- 周辺モデル
  - 周辺モデル
  - csSAM.monovariate
- 非線形回帰・リッジ回帰（本研究）
  - log.ridge
  - log
  - ridge
- 全体モデル
  - 全体モデル
  - $f = g$ = identity
  - TOAST
  - csSAM.lm
- 周辺+全体のハイブリッドモデル（本研究）
  - 周辺モデルでFDR<0.05かつ
全体モデルでP<0.05かつ同じ向きの効果

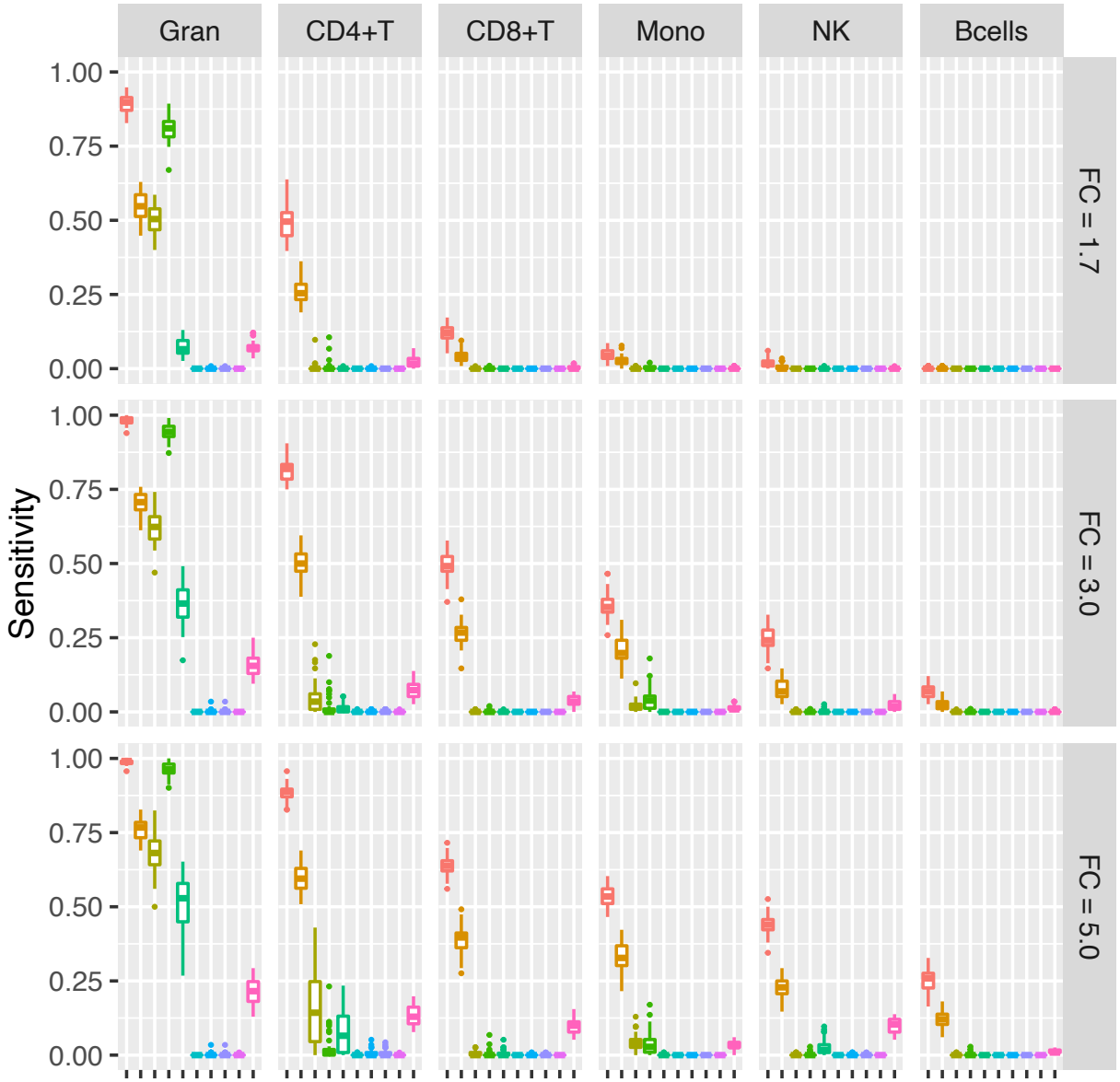
- 有意水準
  - 数万個の遺伝子の多重検定補整が必要
  - False discovery rate (FDR) <0.05

[方法] 実データに基づくシミュレーション

- GTExプロジェクトの全血RNA解読
  - 389検体
  - 14,038遺伝子
- 白血球6種の構成率 W_{hi} はDeconCellプログラムで推定
  - 顆粒球、 $CD4^+T$、 $CD8^+T$、単球、 NK、 B細胞
- 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
- 遺伝子を無作為に分類
  - [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発現上昇
    - 各細胞種につき、2.5% ÷ 6 = 58遺伝子が発現上昇
  - [2.5%] 単一細胞種が罹患者で発現低下
  - 発現変動は1.7倍、3倍、5倍のいずれかに固定
  - 各人各細胞種の発現量をランダム生成
    - 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(FC)➔高
高割合の細胞種➔高

Algorithm

- Marginal
- csSAM.monovariate

周辺モデル 高

- omicwas.log.ridge
- omicwas.identity.ridge
- omicwas.log

log, ridge 中

- omicwas.identity
- Full
- TOAST
- csSAM.lm

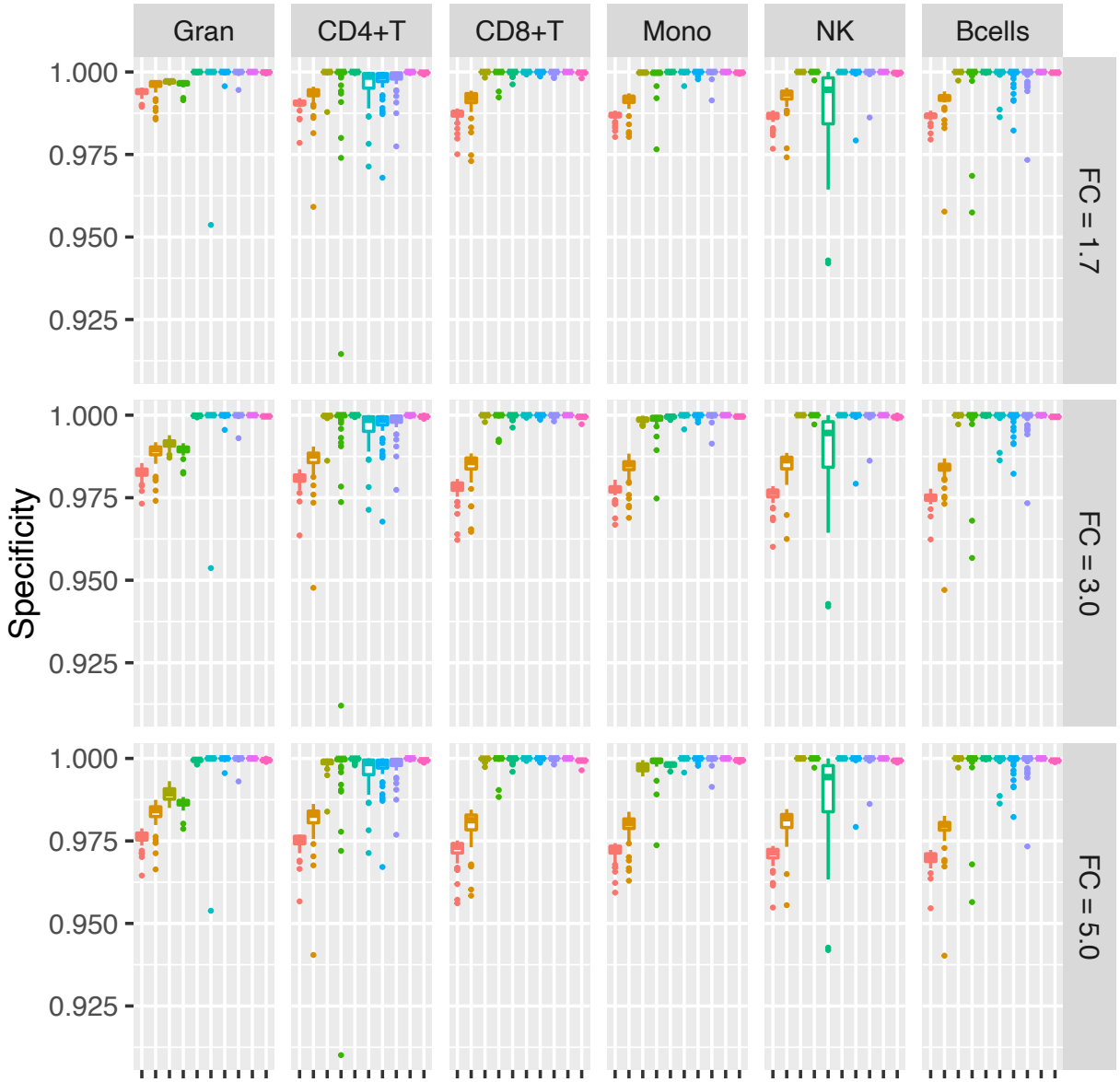
全体モデル <0.01

- Marginal.Full005

周辺+全体モデル 中

特異度 Specificity

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



どれも平均>0.96

Algorithm

- Marginal
- csSAM.monovariate

周辺モデル >0.96

- omicwas.log.ridge
- omicwas.identity.ridge
- omicwas.log

log, ridge >0.98

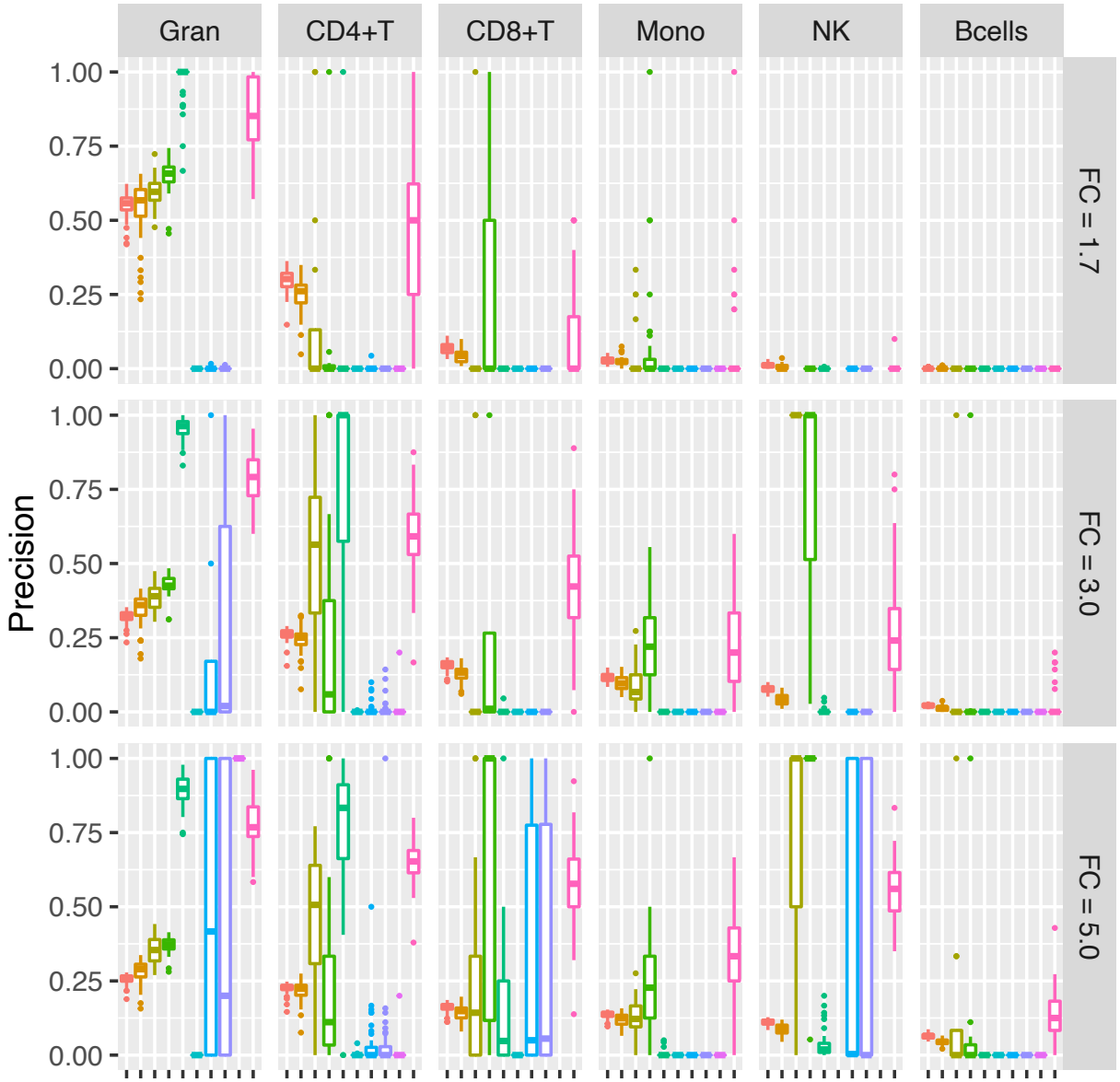
- omicwas.identity
- Full
- TOAST
- csSAM.lm

全体モデル >0.99

- Marginal.Full005

周辺+全体モデル >0.99

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



Algorithm

- Marginal
- csSAM.monovariate

周辺モデル
0～0.55

- omicwas.log.ridge
- omicwas.identity.ridge
- omicwas.log

log, ridge
0～1

- omicwas.identity
- Full
- TOAST
- csSAM.lm

全体モデル
除外

- Marginal.Full005

周辺+全体モデル
0～0.84

周辺モデルがベスト　2/18

log, ridgeがベスト　13/18

周辺+全体モデルがベスト　3/18

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 csSAM.monovariate	高	>0.96	2/18でベスト	△
log.ridge ridge log	中	>0.98	13/18でベスト	△
全体 identity TOAST csSAM.lm	<0.01	>0.99		×
周辺+全体	中	>0.99	3/18でベスト	△

- 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
- log, ridge
  - 周辺モデルと全体モデルの中間
  - 3者で甲乙つけられず
- 周辺+全体のハイブリッドモデル
  - 周辺モデルでFDR<0.05かつ ➔感度
  - 全体モデルでP<0.05 ➔特異度

[方法] 実データでの検証

- 年齢と関連する細胞種特異的遺伝子発現変動
- 学習データ
  - GTExプロジェクトの全血RNA解読
  - 389検体
  - Z値を計算
- 検証データ
  - GSE56047 (Reynolds et al. 2014, PMID: 25404168)
  - 分離した単球: 1202検体
  - 分離したCD4$^+$T: 214検体
    - 検体少ない、変動小さい
  - P<0.05を正解とする

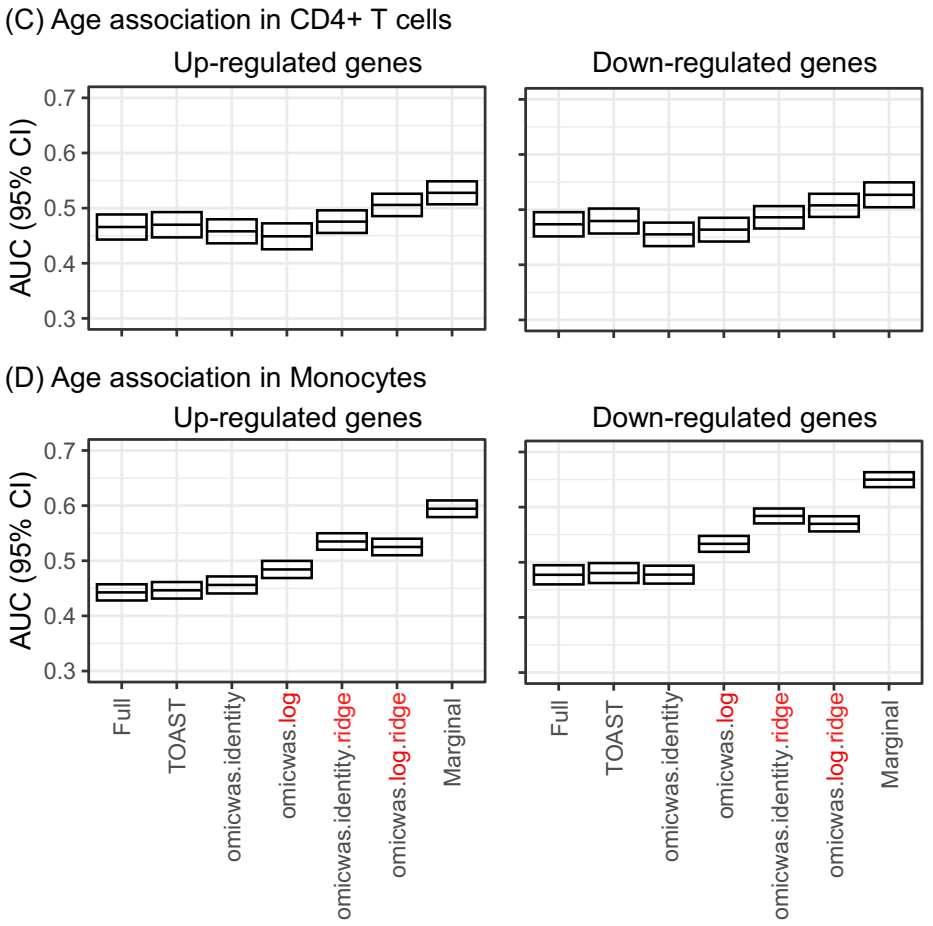
- ROC曲線のAUCで評価
- 標準誤差はjackknife法で推定
  - 染色体上の位置で並べて、100分割

遺伝子数		加齢に伴うCD4+Tでの発現変動			
		上昇	無し	低下	
加齢に伴う単球での発現変動	上昇	178	1762	141	2081
	無し	458	8029	535	9022
	低下	57	1761	166	1984
		693	11552	842	

実データでの検証結果

- 年齢と関連する細胞種特異的遺伝子発現変動

- CD4⁺T
  - どのアルゴリズムもAUC≒0.5でシグナル無し
  - そもそも正解のシグナルが小さい
- 単球
  - 周辺モデル
    ＞ log, ridge
    ＞ 全体モデル



周辺+全体モデルは未計算

まとめ

- 細胞種ごとの遺伝子発現変動を検出する方法を提案した
- リッジ回帰により、多重共線性に対処した
- 非線形回帰により、logスケールをモデル化した。これは多重共線性も和らげる
- 感度・陽性的中率のどちらも悪くはない
- 他を凌駕するモデルは未だなく、改良余地あり

- Rパッケージomicwas
  - CRAN, https://github.com/fumi-github/omicwas
- bioRxivプレプリント（今月中に更新予定）
  - https://dx.doi.org/10.1101/2020.06.18.158758