

細胞種ごとの 遺伝子発現変動を推定する Rパッケージ omicwas

竹内史比古

国立国際医療研究センター(NCGM)研究所

2025年3月19日

Bio"Pack"athon

<http://www.fumihiko.takeuchi.name>

自己紹介

- 竹内史比古（ふみひこ）
- 居住地
 - 高校までは京都
 - 大学からは主に東京
 - 現在は東京とメルボルン
- 現職
 - 国立国際医療研究センター(NCGM)研究所
生命情報科学研究部
テニユアトラック部長
 - Baker Heart and Diabetes Institute
Head of Bioinformatics
- 趣味
 - ジョギング
 - 統計学

研究歴

	病原体の ゲノム解析	生活習慣病 の遺伝的 要因解明	生活習慣病 の病態解明
2001 東京大学・情報科学専攻 PhD			
2000 NCGM・医療生態学研究部	↓	↓	
2004 順天堂大学・感染制御科学	↓		
2006 Sanger Institute		↓	
2008 NCGM・遺伝子研究部		↓	
2011 感染研・病原体ゲノム解析研究センター	↓		
2015 NCGM・遺伝子研究部		↓	↓
2023 Baker Heart and Diabetes Institute		↓	↓

要旨

【背景】 病変組織と正常組織を比較して、遺伝子発現変動や、CpGサイトのメチル化変動が解析されている。

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

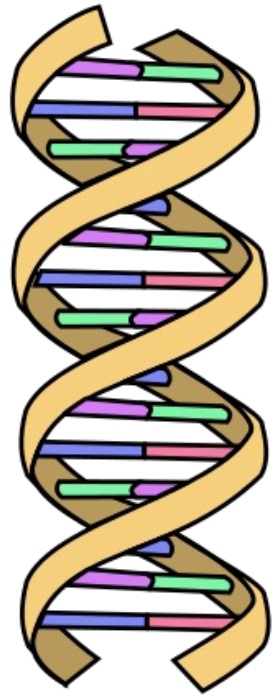
【方法】 従来法は線形回帰。非線形リッジ回帰によるプログラムを開発した。シミュレーションと実データで検証。

【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**細胞種ごとの遺伝子発現やDNAメチル化変動**を推定できた。

BMC Bioinformatics 22:141 <https://rdcu.be/chh6N>

CRAN omicwas <https://github.com/fumi-github/omicwas>



- 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

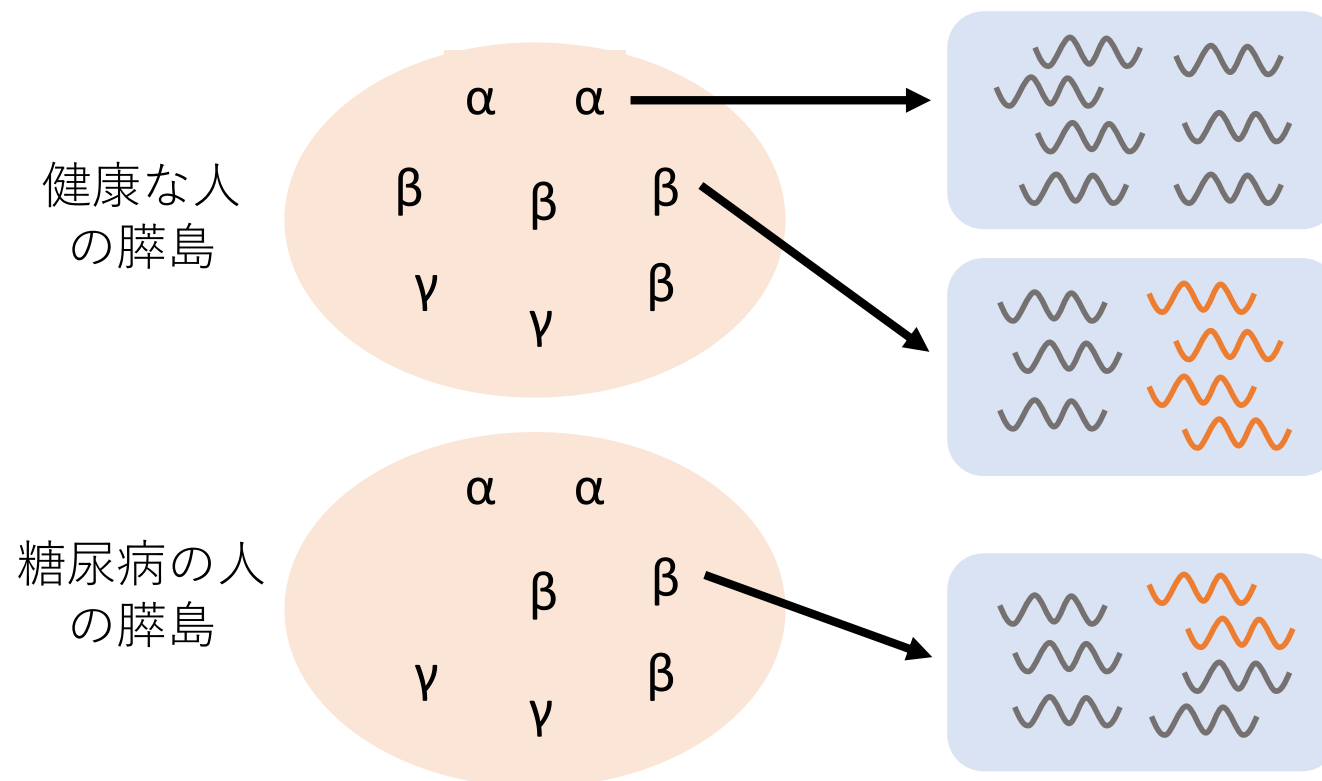
- 病気で、特定の細胞種の遺伝子発現が変動する

組織：膵臓ランゲルハンス島

細胞種： α 細胞, β 細胞, γ 細胞, ...

細胞の
遺伝子(mRNA)発現

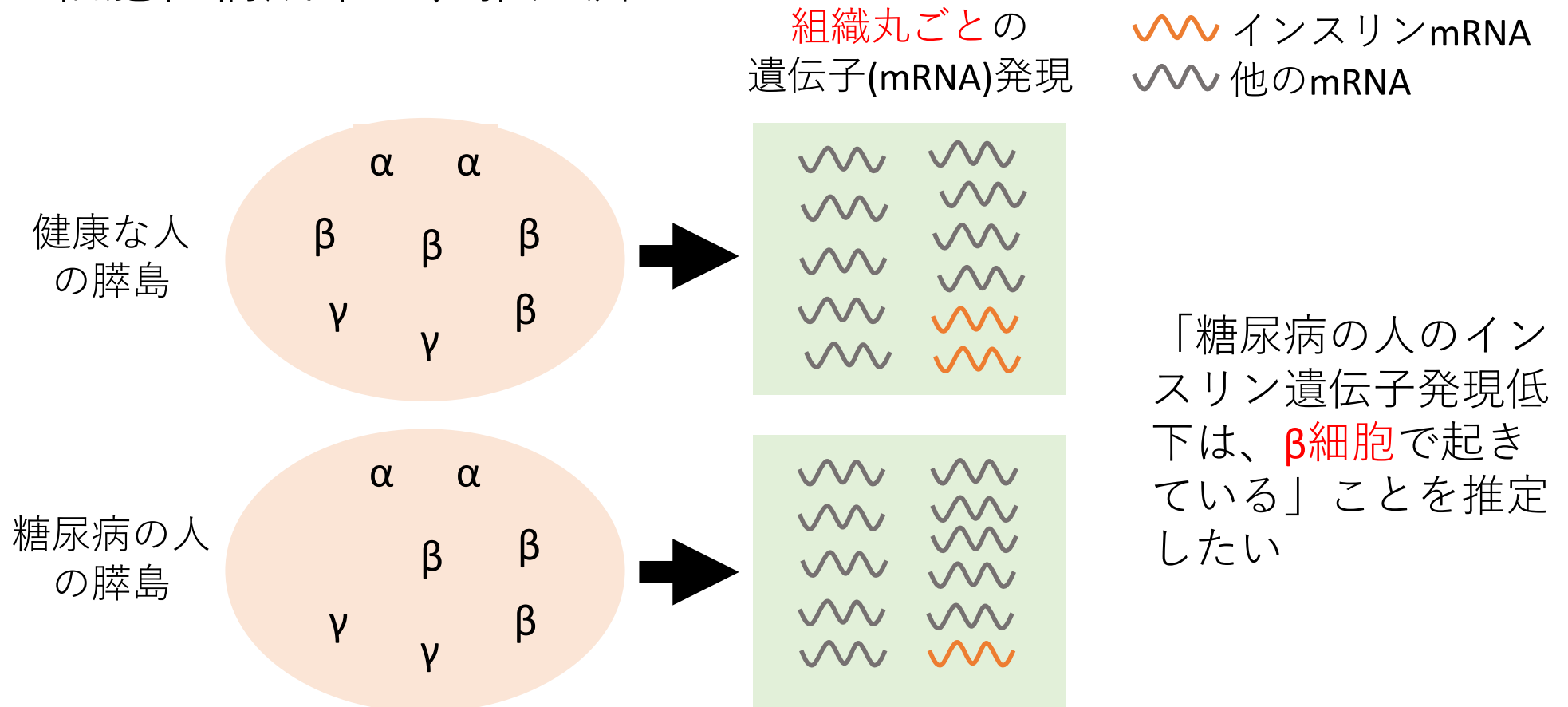
~~~~~ インスリンmRNA  
~~~~~ 他のmRNA



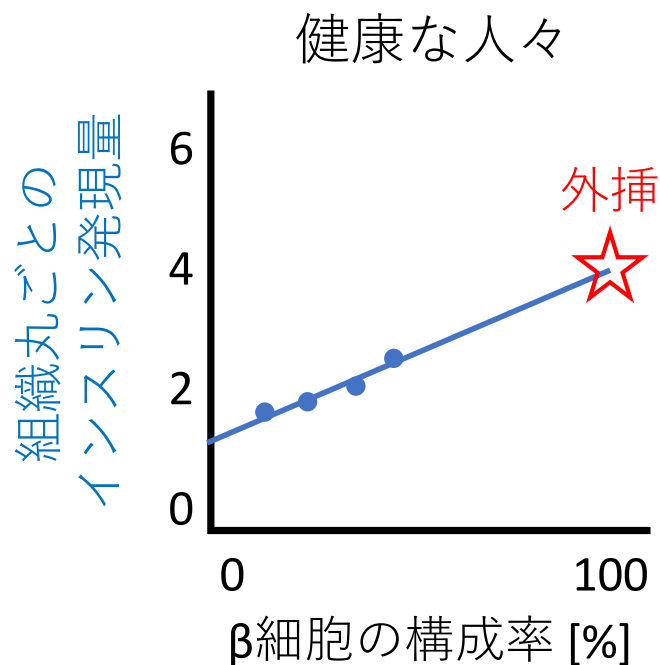
- インスリンは β 細胞のみで発現する
- 糖尿病の人の β 細胞では、インスリンの発現が低下

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

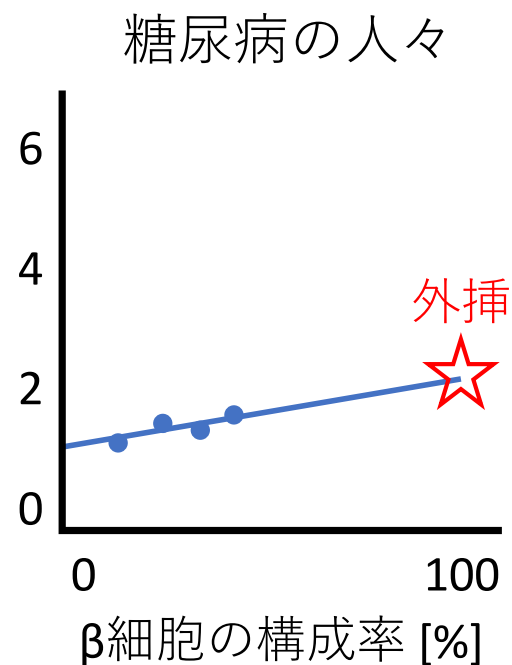


統計的推定の考え方



↓推定

β 細胞の
インスリン発現量 = 4



↓推定

β 細胞の
インスリン発現量 = 2

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_{hi}
 - 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

$$\alpha_h + \beta_h X_i$$

↓ $\sum_h W_{h,i} \times$

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

- 細胞種 h での遺伝子発現

$$\alpha_h + \beta_h X_i$$

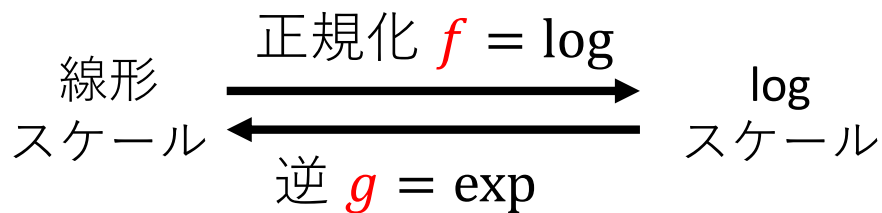
- 線形モデル



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

- 非線形モデル

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



- 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

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

正則化パラメータ λ

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

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

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

ほぼ定数 $\cdot X_i$

- 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

| | 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 |

| 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, $\text{MSE}[\hat{\boldsymbol{\beta}}(\lambda)]$ 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 $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = \boldsymbol{U} \boldsymbol{D} \boldsymbol{V}^T$
 - $\text{MSE}[\hat{\boldsymbol{\beta}}(\lambda)] = \|\text{Bias}[\hat{\boldsymbol{\beta}}(\lambda)]\|^2 + \text{tr}(\text{Var}[\hat{\boldsymbol{\beta}}(\lambda)])$
 $= \sum_{m=1}^M \left(\frac{\lambda}{d_m^2 + \lambda} \right)^2 (\boldsymbol{v}_m^T \boldsymbol{\beta})^2 + \left(\frac{d_m^2}{d_m^2 + \lambda} \right)^2 \left(\frac{\sigma^2}{d_m^2} \right)$
 - 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

- $\lambda \uparrow$ に伴い、
Bias $\uparrow$
Var $\downarrow$
• 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 α
>> X_i の効果 β
 C_i の効果 γ

- 非線形モデル

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

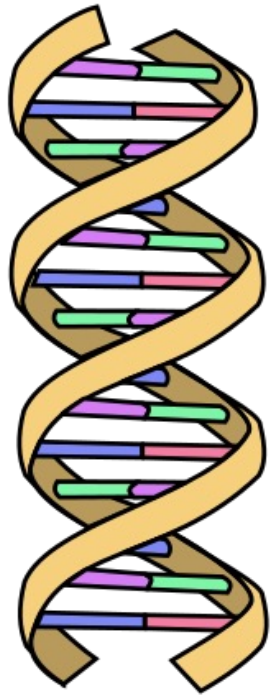
$$f(Y_i) = \mu_i + \varepsilon_i$$

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

- リッジ回帰

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

1. 最小2乗回帰（1回目）
 - $\beta = \gamma = \mathbf{0}$ として $\hat{\alpha}(0)$ を推定
 - $\hat{\sigma}^2$ を推定
2. 最小2乗回帰（2回目）
 - $\alpha = \hat{\alpha}(0)$ として $\hat{\beta}(0), \hat{\gamma}(0)$ を推定
 - 正則化パラメータ λ を決定
3. リッジ回帰
 - $\alpha = \hat{\alpha}(0)$ として $\hat{\beta}(\lambda), \hat{\gamma}(\lambda)$ 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
 - $f = g = \text{identity}$
 - 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

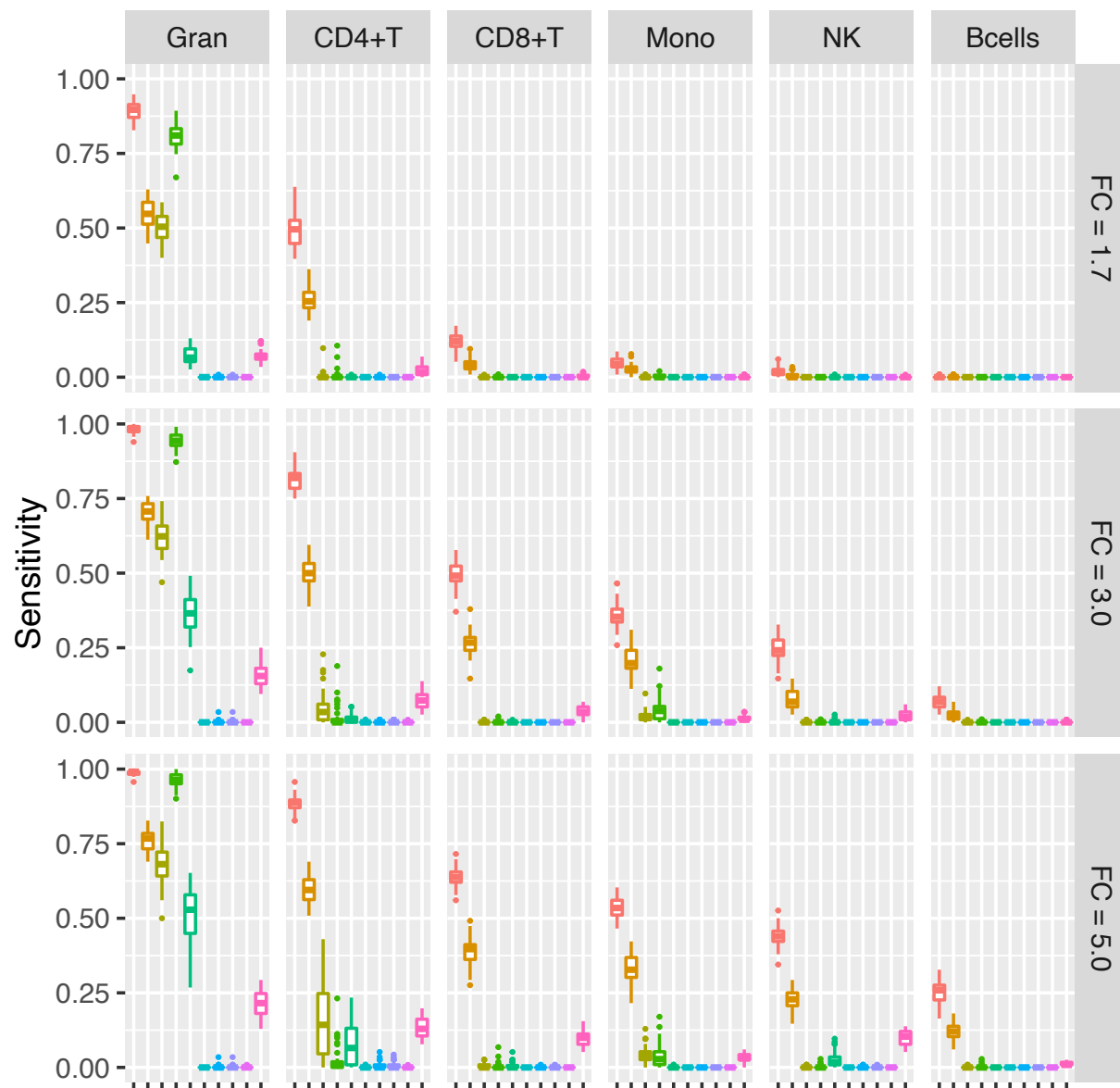
[方法] 実データに基づくシミュレーション

- 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、 $2.5\% \div 6 = 58$ 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

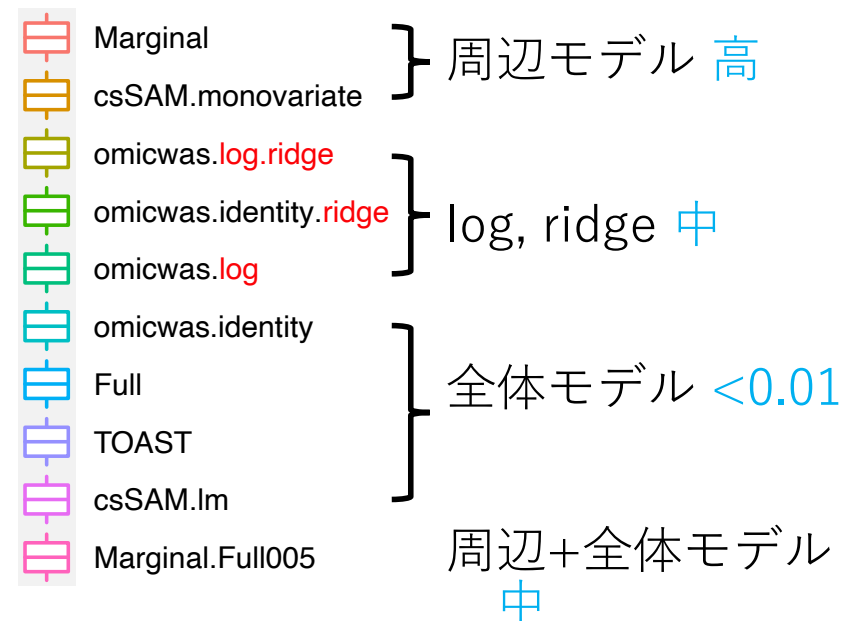
感度 Sensitivity

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

強い効果(FC) → 高
高割合の細胞種 → 高

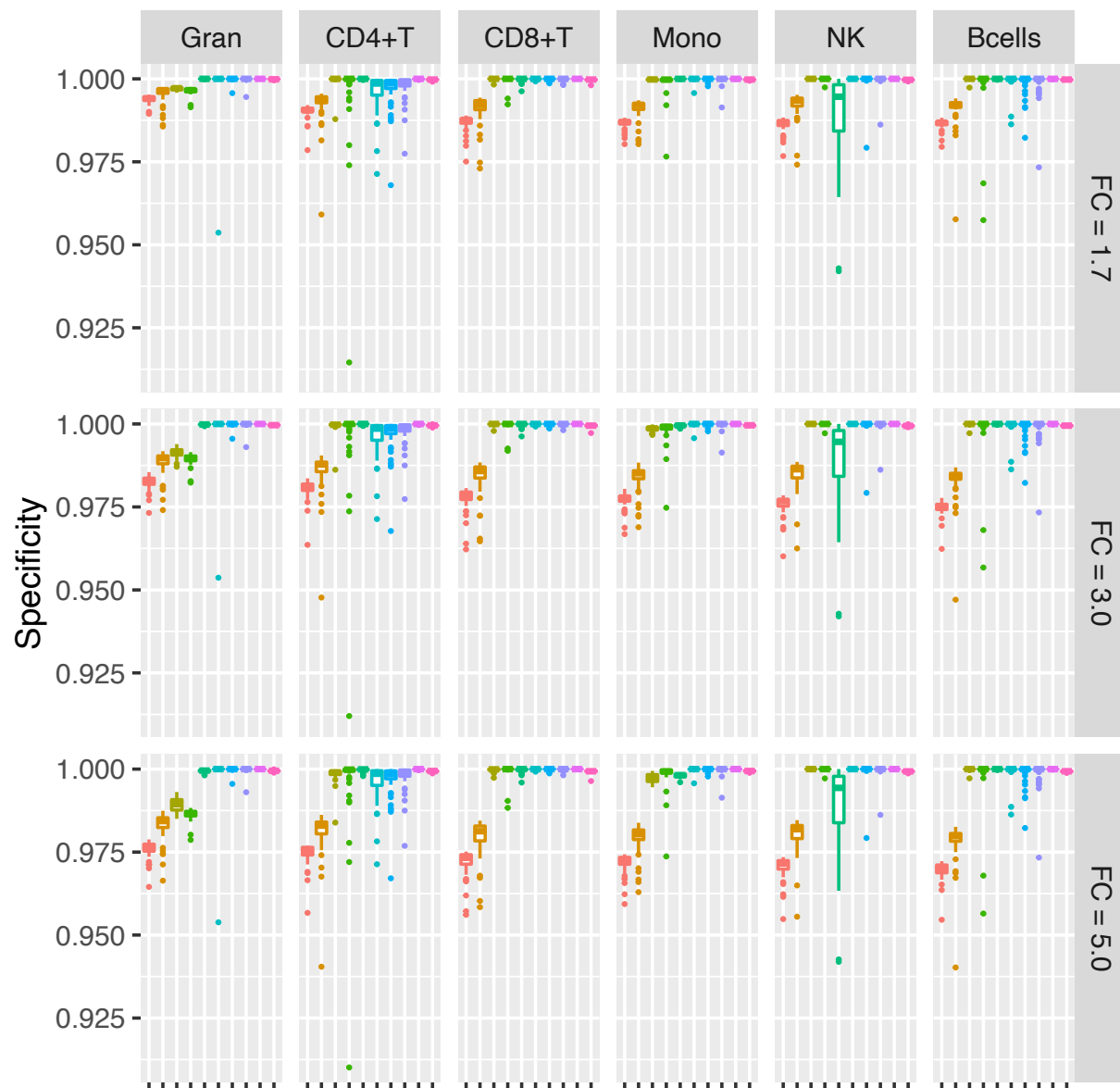


Algorithm



特異度 Specificity

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



どれも平均 >0.96

Algorithm

- Marginal
- csSAM.monovariate
- omicwas.log.ridge
- omicwas.identity.ridge
- omicwas.log
- omicwas.identity
- Full
- TOAST
- csSAM.lm
- Marginal.Full005

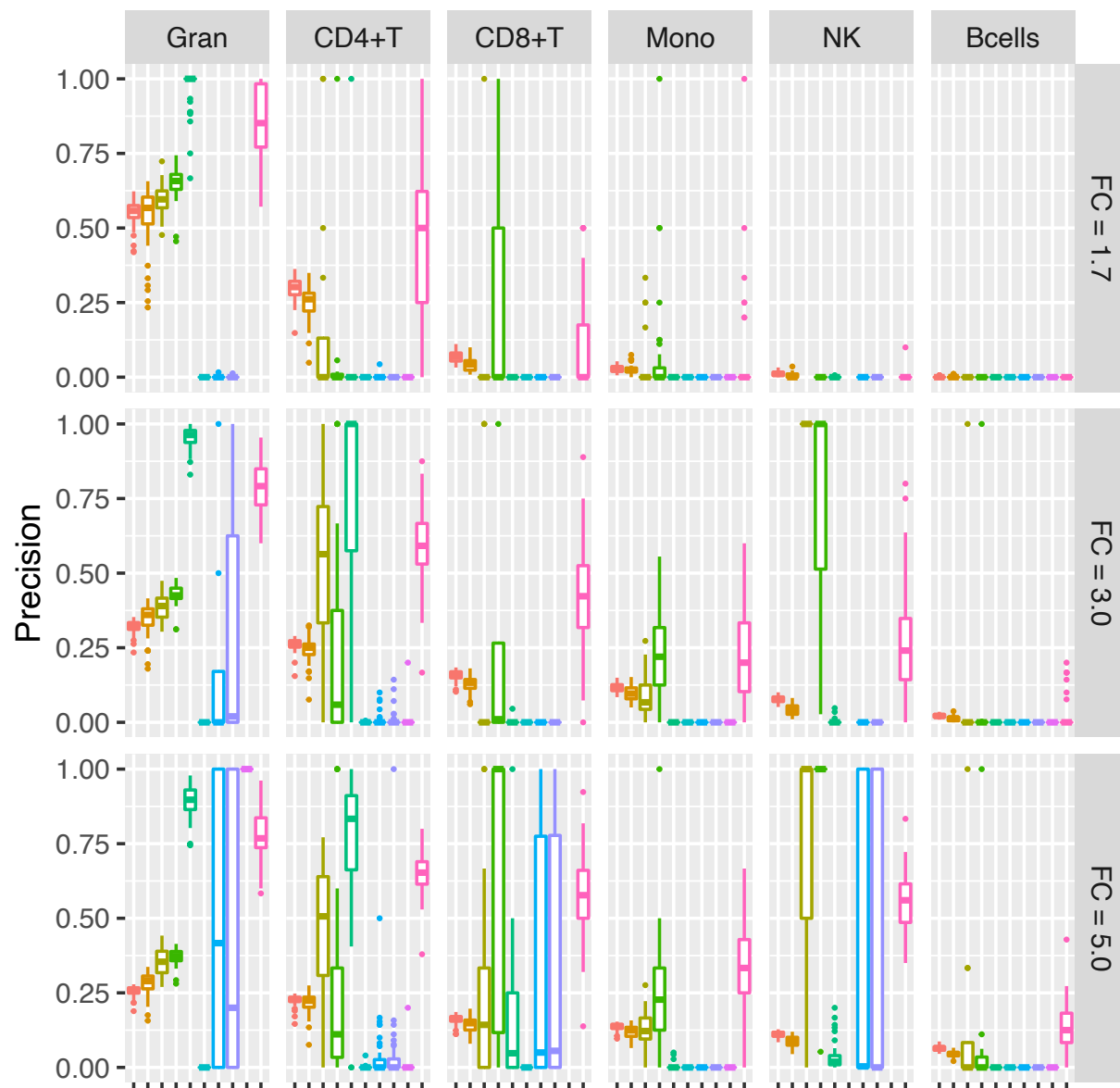
} 周辺モデル >0.96

} log, ridge >0.98

} 全体モデル >0.99

周辺+全体モデル
 >0.99

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



Algorithm

- Marginal
 - csSAM.monovariate
 - omicwas.log.ridge
 - omicwas.identity.ridge
 - omicwas.log
 - omicwas.identity
 - Full
 - TOAST
 - csSAM.lm
 - Marginal.FullI005
- } 周辺モデル
 0~0.55
- } log, ridge
 0~1
- } 全体モデル
 除外
- } 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

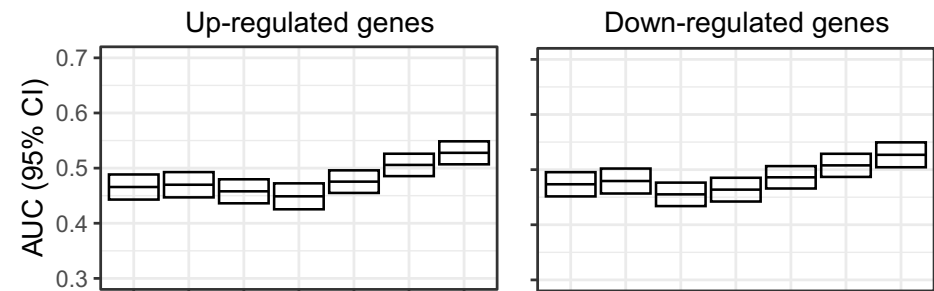
- ROC曲線のAUCで評価
- 標準誤差はjackknife法で推定
 - 染色体上の位置で並べて、100分割

| 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 | |

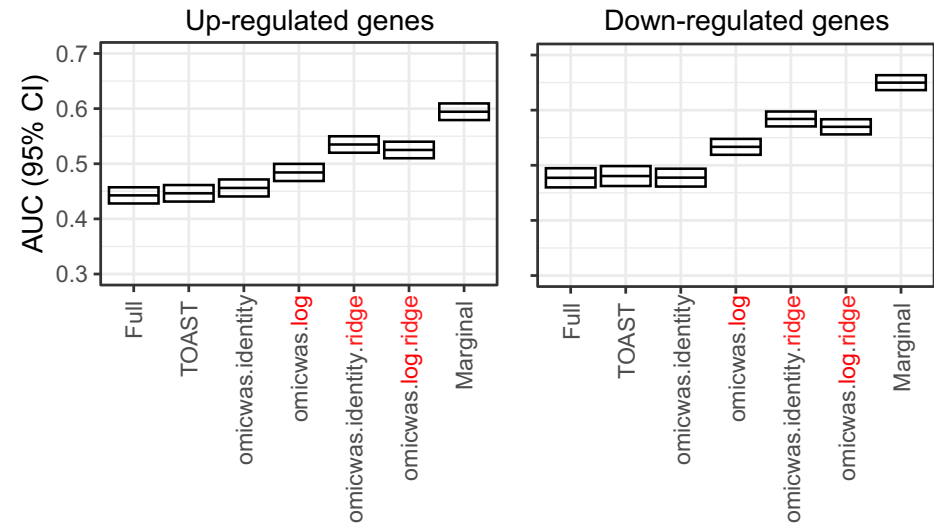
実データでの検証結果

- 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 $AUC \approx 0.5$ 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

(C) Age association in CD4⁺ T cells



(D) Age association in Monocytes



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>
- BMC Bioinformatics 22:141
 - <https://rdcu.be/chh6N>

Rパッケージomicwasの開発 (1/2)

- 動機
 - Rを十数年使ってきて、作る側もやりたくなった
 - 初めてパッケージ開発
- 教材
 - Hadley Wickham and Jennifer Bryan, R Packages
<https://r-pkgs.org/>
 - Cheatsheet
<https://rstudio.github.io/cheatsheets/html/package-development.html>
- 環境
 - MacBook Pro
 - RStudio
 - devtools
 - testthat
 - git, GitHub

Rパッケージomicwasの開発 (2/2)

- テストは大事
 - 早めにテストを書いておくと、コードを間違えたときに気づける
 - Vignetteの例と使い回す
- 機械的チェックは全てパスしないと(mis-spelled words以外)、CRANのボランティアは見てくれない
- 2019/10/16 v0.1
 - リッジ回帰
 - 学会発表、GitHub公開
- 2020/3/3 v0.3.1
 - CRAN公開
- 2020/6/8 v0.7.0
 - 非線形回帰を追加
 - CRAN公開

2020/1/20

If there are references describing the methods in your package, please add these in the description field of your DESCRIPTION file in the form

authors (year) <doi:...>

authors (year) <arXiv:...>

authors (year, ISBN:...)

or if those are not available: <https:...>

with no space after 'doi:', 'arXiv:', 'https:' and angle brackets for auto-linking.

(If you want to add a title as well please put it in quotes: "Title")

`%dontrun{}` should only be used if the example really cannot be executed (e.g. because of missing additional software, missing API keys, ...) by the user. That's why wrapping examples in `%dontrun{}` adds the comment ("`# Not run:`") as a warning for the user.

Does not seem necessary.

Please unwrap the examples if they are executable in < 5 sec, or create additionally small toy examples to allow automatic testing.

(You could also replace `%dontrun{}` with `%donttest` if it takes longer than 5 sec to be executed, but it would be preferable to have automatic checks for functions.)

Please do not set a seed to a specific number within a function.

e.g.: `set.seed(123)` in `omicwas.R`

Please ensure that your functions do not write by default or in your examples/vignettes/tests in the user's home filesystem (including the package directory and `getwd()`). That is not allowed by CRAN policies. Please only write/save files if the user has specified a directory in the function themselves. Therefore please omit any default path = `getwd()` in writing functions.

In your examples/vignettes/tests you can write to `tempdir()`.

2020/2/1

Please add small
executable examples in
your Rd-files to enable
automatic testing.

2020/2/19

Thanks,
on its way to CRAN.