

トランスクリプトームのバイオインフォマ ティクス解析：発現変動解析とその周辺

¹東京大学・大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
²東京大学・微生物科学イノベーション連携研究機構
門田幸二(かどた こうじ)
kadota@iu.a.u-tokyo.ac.jp
<http://www.iu.a.u-tokyo.ac.jp/~kadota/>

Contents

- 自己紹介と東大アグリバイオの紹介
- トランスクリプトーム解析、発現解析、発現変動解析、実験デザイン
- 2群間比較: 実データ、TCC(反復増やすとDEG増える)
- 他グループによる性能評価論文(TCCが非推奨となる場合も!)
- TCCで3群間比較、baySeqも組み合わせて発現パターンまで得る
- (Rで)塩基配列解析
- Single-cell RNA-seq(scRNA-seq)
- バイオインフォマティクス実習

マイペースな理論屋

昔(1999-2011年)はマイクロアレイ、最近(2011以降)はRNA-seqデータを取り扱っています。一貫して、**トランスクリプトーム解析手法の開発**に取り組んでおり、**理論屋**の部類に属します。

略歴

- 2002年3月 東京大学・大学院農学生命科学研究科 博士課程修了
- 2002年4月 産業技術総合研究所・生命情報科学研究センター
- 2003年11月 放射線医学総合研究所・先端遺伝子発現研究センター
- 2005年2月~ 東京大学・大学院農学生命科学研究科

①博士課程修了後、②3年弱の修行期間を経て、
③現在の所属先で**一兵卒**になって早14年。

自己紹介

略歴

- 2002年3月 東京大学・大学院農学生命科学研究科 博士課程修了 ①
- 2002年4月 産業技術総合研究所・生命情報科学研究センター
- 2003年11月 放射線医学総合研究所・先端遺伝子発現研究センター } ②
- 2005年2月～ 東京大学・大学院農学生命科学研究科 ③

自己紹介

③現在の所属先は、当時あちこちで行われていた
バイオインフォ人材養成プログラムの一つとして、
④清水謙多郎 教授を中心に2004年に設立されま
した。

略歴

- 2002年3月 東京大学・大学院農学生命科学研究科 博士課程修了 ①
- 2002年4月 産業技術総合研究所・生命情報科学研究センター
- 2003年11月 放射線医学総合研究所・先端遺伝子発現研究センター } ②
- 2005年2月～ 東京大学・大学院農学生命科学研究科 ③
アグリバイオインフォマティクス人材養成プログラム(2004/10-2009/3)
アグリバイオインフォマティクス教育研究プログラム(2009/4～現在)



<http://www.bi.a.u-tokyo.ac.jp/~shimizu/>

①所属先の正式名称は発足当時と異なっており、
ややこしいので、②アグリバイオと称しています。

アグリバイオ

略歴

- 2002年3月 東京大学・大学院農学生命科学研究科 博士課程修了
- 2002年4月 産業技術総合研究所・生命情報科学研究センター
- 2003年11月 放射線医学総合研究所・先端遺伝子発現研究センター
- 2005年2月～ 東京大学・大学院農学生命科学研究科

アグリバイオ インフォマティクス人材養成プログラム (2004/10-2009/3)

アグリバイオ インフォマティクス教育研究プログラム (2009/4～現在)



アグリバイオ

①アグリバイオでググると、②一応上位にランクイン。15年の歴史がありますので、まったく無名というわけではない…はず。

アグリバイオ - Google 検索

https://www.google.com/search?hl=ja&source=hp&ei=m2ZCXf...

Google

アグリバイオ

すべて ニュース 地図 画像 ショッピング もっと見る 設定 ツ

約 1,850,000 件 (0.37 秒)

アグリバイオインフォマティクス教育研究ユニット: ホーム
www.iu.a.u-tokyo.ac.jp/

お知らせ - 受講に関する更新情報. 講義日当日に受講登録をする場合は16:30までに事務局までお越しください。UTASで履修登録をする際に、講義日程が重複していないにも関わらずエラーが出る場合は、どちらか一方をUTASで履修登録し、大学院教務課に (...)

受講生の方へ · 事務局 · 本学の大学院生以外の方 · システム生物学概論

各講義のページ | アグリバイオインフォマティクス教育研究ユニット
www.iu.a.u-tokyo.ac.jp/ · [教育プログラム](#)

カテゴリー, 科目名, 実施ターム · 単位, 実施 曜日, 基礎, 生物配列解析基礎, 生命科学のためのデータベースの利用と基本的な解析手法について講義します。配列データベースや機能データベースの使用法を紹介するとともに、ホモロジー検索、モチーフ解析、Perl ...

お問い合わせ | アグリバイオインフォマティクス教育研究ユニット
www.iu.a.u-tokyo.ac.jp/main_contact.html

アグリバイオインフォマティクス教育研究ユニット事務局 〒113-8657 東京都文京区弥生1-1-1 東京大学大学院農学生命科学研究科 電子メール: . 事務局へのアクセス (農学部弥生キャンパス地図) . 事務局は、農学部2号館地下1階 (14-2号室) にあります。

アグリバイオ | 北隆館WEBサイト
hokuryukan-ns.co.jp/cms/book_category/x01/

アグリバイオ 2019年4月号. ケミカル/バイオロジーの農業生産向上への応用. Application to the

東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
Agricultural Bioinformatics Research Unit

受講生の方へ 研究者の方へ

ホーム
本ユニットについて
メンバー
教育プログラム
研究フォーラム
イベント
お問い合わせ
リンク

ようこそ!!
アグリバイオインフォマティクス
教育研究ユニットへ

バイオインフォマティクスの実践的基礎教育から、関連した農学生命科学研究と研究支援、本分野の社会連携、国際拠点の形成を目指して

お知らせ - 受講に関する更新情報

- 講義日当日に受講登録をする場合は16:30までに事務局までお越しください。
- UTASで履修登録をする際に、講義日程が重複していないにも関わらずエラーが出る場合は、どちらか一方をUTASで履修登録し、大学院教務課に (ホームページ) 追加登録申請書を出してください。 **NEW!!**
- 2019年度受講生募集要項はこちら(PDF)です。
- 受講に関するお問い合わせはこちらをご確認ください。
Q & A(本学の大学院生の方)
Q & A(本学の大学院生以外の方)
- 成績証明書発行を希望される方は申請用紙「Word形式、PDF形式」に必要な事項を記入し、事務局までご連絡ください。

東京大学
The University of Tokyo

メンバー

①代表者は研究科長。ほぼ専任教員の中のトップは、②寺田 透先生です。寺田先生が、受講ガイダンス、サーバ管理、無線LANなどインフラ関係の大変な業務を担当してくださっていますm(_ _)m。

The screenshot shows a web browser window displaying the website for the Agricultural Bioinformatics Research Unit at the University of Tokyo. The page title is 'アグリバイオインフォマティクス教育研究ユニット' (Agricultural Bioinformatics Education Research Unit). The navigation menu includes 'ホーム', '本ユニットについて', 'メンバー', '教育プログラム', '研究フォーラム', 'イベント', 'お問い合わせ', and 'リンク'. The 'メンバー' (Members) section is expanded, showing a list of roles: 'プログラム代表者', '専任教員', '兼任教員', '講師', and '事務職員'. Under '教育プログラムメンバー (2019年度版)', there are two entries:

役職	氏名	所属	担当
プログラム代表者	堤 伸浩 / TSUTSUMI Nobuhiro	(大学院農学生命科学研究科長)	プログラム代表者
専任教員	寺田 透 / TERADA Tohru	(大学院情報学環・学際情報学府 / 准教授)	研究活動：構造バイオインフォマティクス・分子シミュレーションを用いたタンパク質の機能発現メカニズムの解明

Red arrows with numbers 1 and 2 point to the portraits of Nobuhiro Tsutsumi and Tohru Terada, respectively.

メンバー

①他の専任教員として、大森 良弘先生が昨年8月より加入。植物の研究をされています。②多くの皆様のご協力によって、アグリバイオの教育プログラムが成り立っています。

メンバー | アグリバイオインフォマティクス × +

← → ↻ 🏠 ⓘ 保護されていない通信 | www.iu.a.u-tokyo.ac.jp/main_member.html ☆ 🗑️ 👤 ⋮

	<p>大森良弘 / OMORI Yoshihiro (大学院農学生命科学研究科 / 特任准教授)</p> <p>研究活動：フィールドインフォマティクス。イオノミクスを介したフィールド環境における植物成長制御ネットワークの解明、植物成長予測、ならびに植物栄養診断技術の開発</p>
兼任教員 (運営・講義)	<ul style="list-style-type: none">■ 清水謙多郎 (東大・農 / 教授)■ 岸野洋久 (東大・農 / 教授)■ 野尻秀昭 (東大・農 / 教授)■ 永田宏次 (東大・農 / 准教授)
兼任教員 (講義)	<ul style="list-style-type: none">■ 岩田洋佳 (東大・農 / 准教授)
非常勤講師 (講義)	<ul style="list-style-type: none">■ 北田修一 (東京海洋大学 / 名誉教授)■ 麻生川 稔 (日本電気株式会社 / 主席技術主幹)■ 有田正規 (国立遺伝学研究所 / 教授)■ 大島研郎 (法政大学 / 教授)■ 井澤 毅 (東大・農・生産・環境生物学専攻 / 教授)■ 郭 威 (東大・農・附属生態調和農学機構 / 特任助教)■ 中道礼一郎 (資源研究センター / 主任研究員)■ 戸田陽介 (名古屋大学 / 招聘教員)■ 市橋泰範 (理化学研究所 / チームリーダー)
研究員	<ul style="list-style-type: none">■ 阿部 紘一 (特任研究員)■ 根上 樹 (特任研究員)
事務職員	<ul style="list-style-type: none">■ 寺田朋子 (学術支援職員)■ 三浦 文 (学術支援職員)

東京大学大学院農学生命科学研究科 〒113-8657 東京都文京区弥生1-1-1
Copyright © アグリバイオインフォマティクス教育研究ユニット



メンバー

①他の専任教員として、大森 良弘先生が昨年8月より加入。植物の研究をされています。②多くの皆様のご協力によって、アグリバイオの教育プログラムが成り立っています。アグリバイオは5年前に外部予算が切れて研究科予算で運営されています。厳しい予算状況の中、人件費の確保など毎年心労の多い雑務を③清水謙多郎先生が担当してくださっています。

メンバー | アグリバイオインフォマティクス × +
保護されていない通信 | www.iu.a.u-tokyo.ac.jp/main_member.html

	 <p>大森良弘 / OMORI Yoshihiro (大学院農学生命科学研究科 / 特任)</p> <p>研究活動：フィールドインフォマティクスを介したフィールド環境における植物成長制御ネットワークの解明、植物成長予測、ならびに植物栄養診断技術の開発</p>
兼任教員 (運営・講義)	<ul style="list-style-type: none"> ■ 清水謙多郎 (東大・農 / 教授) ■ 岸野洋久 (東大・農 / 教授) ■ 野尻秀昭 (東大・農 / 教授) ■ 永田宏次 (東大・農 / 准教授)
兼任教員 (講義)	<ul style="list-style-type: none"> ■ 岩田洋佳 (東大・農 / 准教授)
非常勤講師 (講義)	<ul style="list-style-type: none"> ■ 北田修一 (東京海洋大学 / 名誉教授) ■ 麻生川 稔 (日本電気株式会社 / 主席技術主幹) ■ 有田正規 (国立遺伝学研究所 / 教授) ■ 大島研郎 (法政大学 / 教授) ■ 井澤 毅 (東大・農・生産・環境生物学専攻 / 教授) ■ 郭 威 (東大・農・附属生態調和農学機構 / 特任助教) ■ 中道礼一郎 (資源研究センター / 主任研究員) ■ 戸田陽介 (名古屋大学 / 招聘教員) ■ 市橋泰範 (理化学研究所 / チームリーダー)
研究員	<ul style="list-style-type: none"> ■ 阿部 紘一 (特任研究員) ■ 根上 樹 (特任研究員)
事務職員	<ul style="list-style-type: none"> ■ 寺田 朋子 (学術支援職員) ■ 三浦 文 (学術支援職員)



東京大学大学院農学生命科学研究科 〒113-8657 東京都文京区弥生1-1-1
Copyright © アグリバイオインフォマティクス教育研究ユニット

教育プログラム

アグリバイオの①教育プログラムは、②大きく3つのカテゴリーに分けられ…

東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
Agricultural Bioinformatics Research Unit

受講生の方へ 研究者の方へ

ホーム > 教育プログラム

教育プログラム

▼ プログラム概要 ▼ 講義について ▼ 受講について ▼ 各講義のページ
▼ スケジュール

プログラム概要

本プログラムで開講する講義科目は、大きく3つのカテゴリー（基礎、方法論、先端トピックス）に分けられます。カテゴリーと各講義の関係については、[各講義のページ](#)をご覧ください。

カテゴリー	目的
基礎	主にバイオインフォマティクスを利用した研究経験のない方を対象としています。生命科学のための各種データベースの利用法やバイオインフォマティクスを利用した様々なツールの利用法、統計の基礎を学ぶことができます。
方法論	「基礎」の科目を土台として、様々な実験手法（トランスクリプトーム解析法、質量分析法など）や計算機的手法（パターン認識や機械学習、統計モデルやモデル選択、分子シミュレーション法、データ正規化や多重比較問題への対処）について解説します。
先端トピックス	企業や大学の研究者が、それぞれの最先端の研究課題について講義・実習を行います。ここでは、バイオインフォマティクスの実際の活用例に触れることで、個々の研究課題へのフィードバックを目指します。また、農学生命情報科学特別演習では、本プログラム教員による研究指導を受けることができます。

教育プログラム

アグリバイオの①教育プログラムは、②大きく3つのカテゴリーに分けられ…様々な講義科目があります。

東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
Agricultural Bioinformatics Research Unit

ホーム > 教育プログラム

教育プログラム

- ▼ プログラム概要
- ▼ 講義について
- ▼ 受講生の方へ
- ▼ スケジュール

プログラム概要

本プログラムで開講する講義科目は、大きく3つ(ス)に分けられます。カテゴリーと各講義の関係

カテゴリー	目的
基礎	主にバイオインフォマティクスを学ぶ。生命科学のための各種データを利用した様々なツールの利用法。
方法論	「基礎」の科目を土台として、様々な方法、質量分析法など)や計算機的手法やモデル選択、分子シミュレーション)について解説します。
先端トピックス	企業や大学の研究者が、それぞれ研究しています。ここでは、バイオインフォマティクスで、個々の研究課題へのフィードバックを目指します。また、農学生命情報科学特別演習では、本プログラム教員による研究指導を受けることができます。

先端トピックス

セミナー・討論形式 研究指導

農学生命情報科学特別演習

農学生命情報科学特論 I

農学生命情報科学特論 II

農学生命情報科学特論 III

農学生命情報科学特論 IV

方法論

講義・実習を一体化

生物配列統計学 システム生物学概論 知識情報処理論

オーム情報解析 機能ゲノム学 分子モデリングと分子シミュレーション

フィールドインフォマティクス

基礎

講義・実習を一体化

ゲノム情報解析基礎 構造バイオインフォマティクス基礎

生物配列解析基礎 バイオスタティスティクス基礎論

教育プログラム

アグリバイオの①教育プログラムは、②大きく3つのカテゴリーに分けられ…様々な講義科目があります。フリーソフトウェアRを使う講義科目が多いのが特徴(といわれる)。これは設立当初の全体方針によります。

教育プログラム | アグリバイオインフォマ × +
 ← → ↻ 🏠 ⓘ 保護されていない通信 | www.iu.a.u-tokyo.ac.jp/main_education.html



東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
 Agricultural Bioinformatics Research Unit

+ サイトマップ + English

🌿 受講生の方へ 🌿 研究者の方へ

- + ホーム
- + 本ユニットについて
- + メンバー
- + 教育プログラム **①**
- + 研究フォーラム
- + イベント
- + お問い合わせ
- + リンク

ホーム > 教育プログラム

教育プログラム

▼ プログラム概要 ▼ 講義について ▼ 受講に
 ▼ スケジュール

プログラム概要

本プログラムで開講する講義科目は、大きく3つ(ス)に分けられます。カテゴリーと各講義の関係

カテゴリー	目的
基礎	主にバイオインフォマティクスを学ぶ。生命科学のための各種データを利用した様々なツールの利用法。
方法論	「基礎」の科目を土台として、様々な方法、質量分析法など)や計算機的手法やモデル選択、分子シミュレーション)について解説します。
先端トピックス	企業や大学の研究者が、それぞれに携わっています。ここでは、バイオインフォマティクスで、個々の研究課題へのフィードバックを目指します。また、農学生命情報科学特別演習では、本プログラム教員による研究指導を受けることができます。

先端トピックス

セミナー・
 討論形式
 研究指導

農学生命情報科学特別演習

農学生命情報科学特論 I
 農学生命情報科学特論 II
 農学生命情報科学特論 III
 農学生命情報科学特論 IV

方法論

講義・実習を
 一体化

生物配列統計学 システム生物学概論 知識情報処理論
 オーム情報解析 機能ゲノム学 分子モデリングと分子シミュレーション
 フィールドインフォマティクス

基礎

講義・実習を
 一体化

ゲノム情報解析基礎 構造バイオインフォマティクス基礎
 生物配列解析基礎 バイオスタティスティクス基礎論

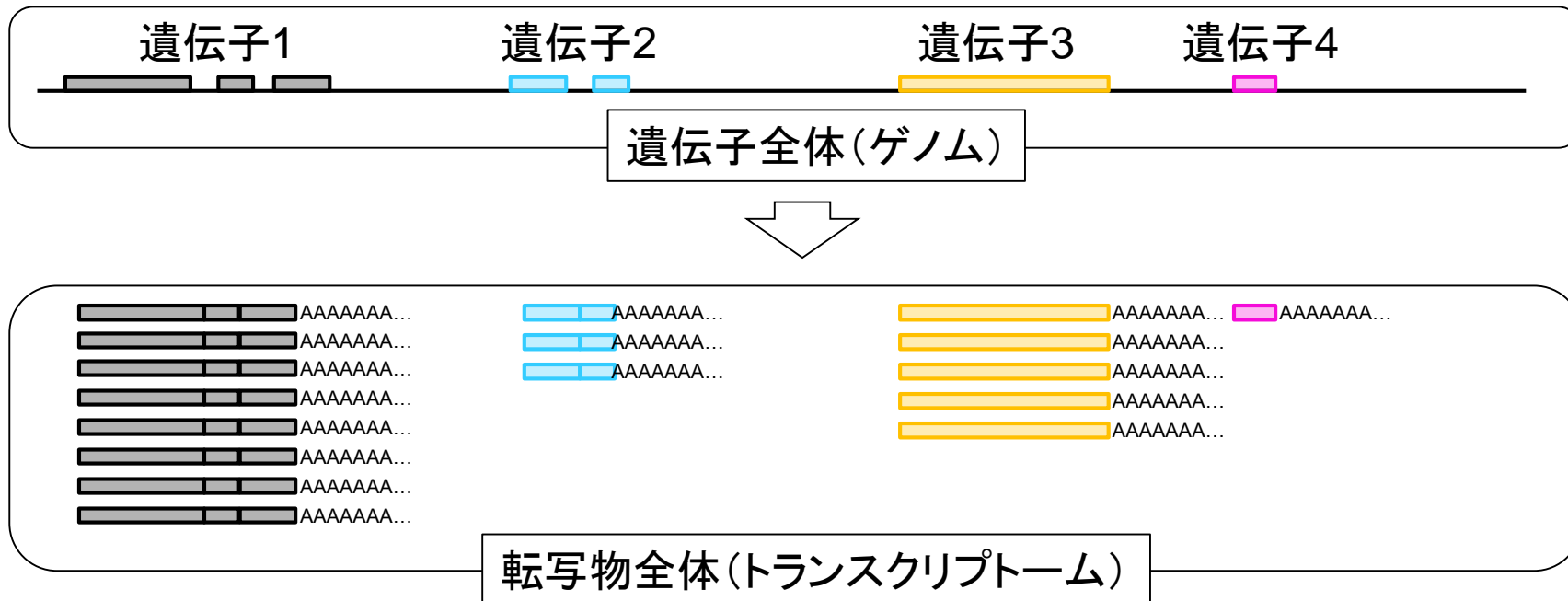


Contents

- 自己紹介と東大アグリバイオの紹介
- トランスクリプトーム解析、発現解析、発現変動解析、実験デザイン
- 2群間比較: 実データ、TCC(反復増やすとDEG増える)
- 他グループによる性能評価論文(TCCが非推奨となる場合も!)
- TCCで3群間比較、baySeqも組み合わせて発現パターンまで得る
- (Rで)塩基配列解析
- Single-cell RNA-seq(scRNA-seq)
- バイオインフォマティクス実習

トランスクリプトーム解析

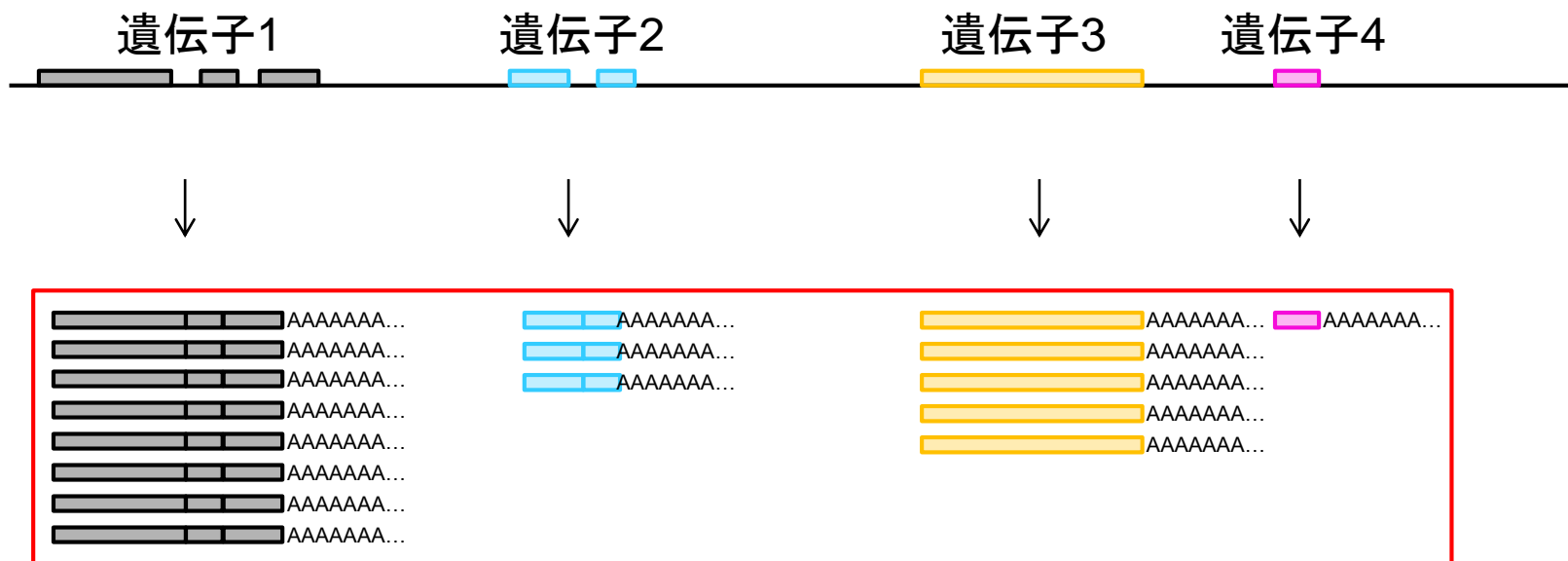
調べたいサンプル内で働いているRNAの種類(塩基配列)や量(発現量)を調べるのがトランスクリプトーム解析。



- ・遺伝子1は沢山転写されている(発現している)
- ・遺伝子4はごくわずかしか転写されてない
- ・...

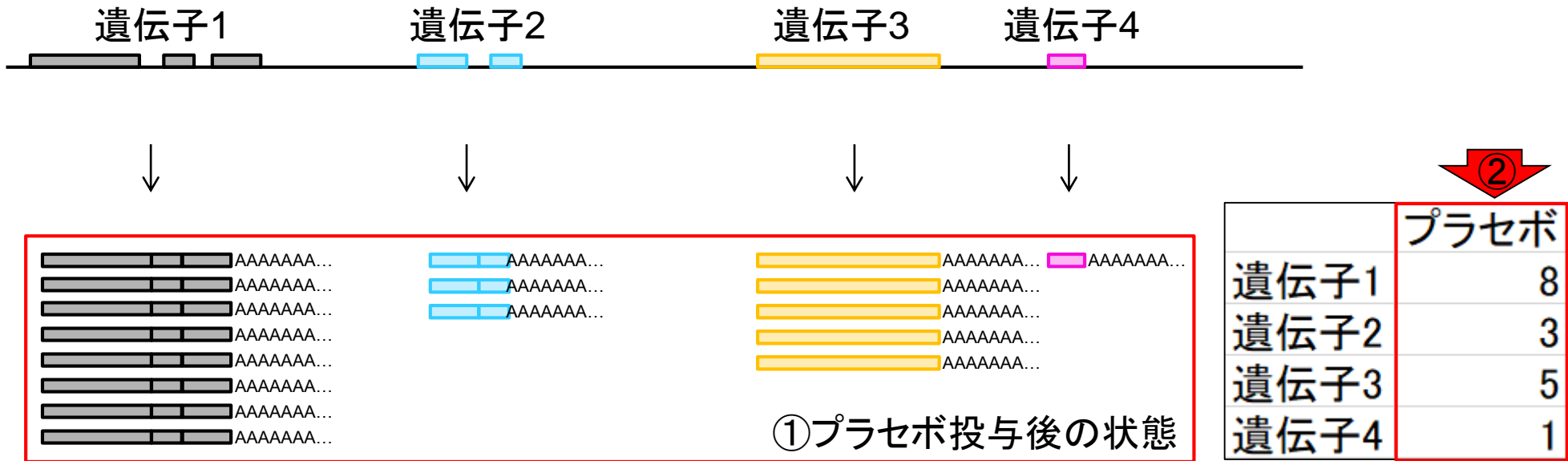
発現解析

調べたいサンプル内で働いているRNAの種類(塩基配列)や量(発現量)を調べるのがトランスクリプトーム解析。発現量に特化した解析を**発現解析**という。



発現解析

赤枠を①プラセボ投与後の状態だとすると、1つの実験(ある患者さんの癌サンプル)の発現データ取得後の結果として、②で示すような数値ベクトルが得られる。



発現解析

赤枠を①プラセボ投与後の状態だとすると、1つの実験(ある患者さんの癌サンプル)の発現データ取得後の結果として、②で示すような数値ベクトルが得られる。通常は何かと比較して違いを見たいので



	② プラセボ
遺伝子1	8
遺伝子2	3
遺伝子3	5
遺伝子4	1

①プラセボ投与後の状態

プラセボ vs. 薬剤A

赤枠を①プラセボ投与後の状態だとすると、1つの実験(ある患者さんの癌サンプル)の発現データ取得後の結果として、②で示すような数値ベクトルが得られる。通常は何かと比較して違いを見たいので…例えば③薬剤A投与後の発現データを取得します。



	プラセボ
遺伝子1	8
遺伝子2	3
遺伝子3	5
遺伝子4	1



	薬剤A
遺伝子1	7
遺伝子2	15
遺伝子3	5
遺伝子4	6

プラセボ vs. 薬剤A

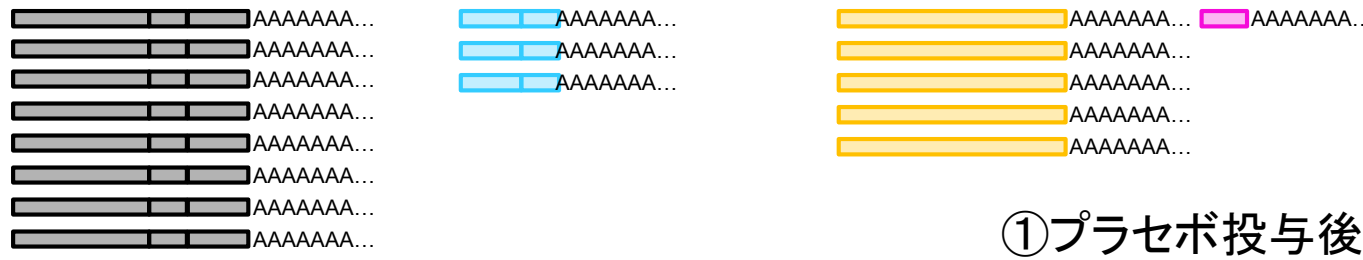
これは、①プラセボと②薬剤Aの数値ベクトル同士を比較するための最もシンプルな実験デザイン。目的は、薬剤Aに反応して発現が変化する遺伝子(発現変動遺伝子)の同定。

遺伝子1

遺伝子2

遺伝子3

遺伝子4



	① プラセボ	② 薬剤A
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	6

発現変動解析

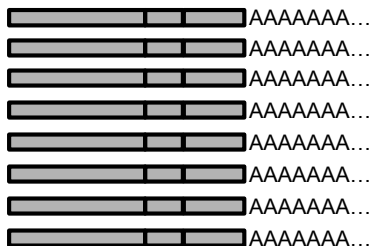
これは、①プラセボと②薬剤Aの数値ベクトル同士を比較するための最もシンプルな実験デザイン。目的は、薬剤Aに反応して発現が変化する遺伝子(発現変動遺伝子)の同定。赤枠がデータ解析を行う際の入力データ。この段階で、ベクトルではなく行列となる。この業界で遺伝子発現行列と呼ばれるものに相当します。

遺伝子1

遺伝子2

遺伝子3

遺



①プラセボ投与後



②薬剤A投与後

	プラセボ	薬剤A
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	6

発現変動解析の①入力と②出力の概念図。③log比などの解析結果が得られる。

発現変動解析の入出力

①入力

	プラセボ	薬剤A
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	6



②出力

	プラセボ	薬剤A	log比	p値
遺伝子1	8	7	-0.2	
遺伝子2	3	15	2.32	
遺伝子3	5	5	0	
遺伝子4	1	6	2.58	



発現変動解析の入出力

発現変動解析の①入力と②出力の概念図。③log比などの解析結果が得られる。例えば、④は $\log_2(6/1) = 2.584963$ として得られます。最も発現変動の度合いが高いのは、log比で評価すると遺伝子4になります。

①入力

	プラセボ	薬剤A
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	6



②出力

	プラセボ	薬剤A	log比	p値
遺伝子1	8	7	-0.2	
遺伝子2	3	15	2.32	
遺伝子3	5	5	0	
遺伝子4	1	6	2.58	



発現変動解析の入出力

発現変動解析の①入力と②出力の概念図。③log比などの解析結果が得られる。例えば、④は $\log_2(6/1) = 2.584963$ として得られます。最も発現変動の度合いが高いのは、log比で評価すると遺伝子4になります。統計的な検定手法が適用できる場合は、出力結果として⑤p値とそれに関連した指標も得ることができます。

①入力

	プラセボ	薬剤A
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	6



②出力

	プラセボ	薬剤A	log比	p値
遺伝子1	8	7	-0.2	
遺伝子2	3	15	2.32	
遺伝子3	5	5	0	
遺伝子4	1	6	2.58	



Contents

- 自己紹介と東大アグリバイオの紹介
- トランスクリプトーム解析、発現解析、発現変動解析、**実験デザイン**
- 2群間比較: 実データ、TCC(反復増やすとDEG増える)
- 他グループによる性能評価論文(TCCが非推奨となる場合も!)
- TCCで3群間比較、baySeqも組み合わせて発現パターンまで得る
- (Rで)塩基配列解析
- Single-cell RNA-seq(scRNA-seq)
- バイオインフォマティクス実習

実験デザイン

通常は、①プラセボと②薬剤Aの発現変動解析を1サンプルずつのデータで行うことはしない。

	① プラセボ	② 薬剤A
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	6



②出力

	プラセボ	薬剤A	log比	p値
遺伝子1	8	7	-0.2	
遺伝子2	3	15	2.32	
遺伝子3	5	5	0	
遺伝子4	1	6	2.58	

実験デザイン

通常は、①プラセボと②薬剤Aの発現変動解析を1サンプルずつのデータで行うことはしない。自分は「プラセボ vs. 薬剤A」の比較をしているつもりでも、実際には③「女性 vs. 男性」、④肥満度の違い、⑤ウイルス感染の有無の影響などを調べているだけかもしれないからです。

	⑤ 感染アリ 低肥満度 ③ 女性	感染ナシ 高肥満度 男性	④
	プラセボ	薬剤A	
遺伝子1	8	7	→
遺伝子2	3	15	
遺伝子3	5	5	
遺伝子4	1	6	

②出力

	プラセボ	薬剤A	log比	p値
遺伝子1	8	7	-0.2	
遺伝子2	3	15	2.32	
遺伝子3	5	5	0	
遺伝子4	1	6	2.58	

2群間比較

通常は、同一条件の反復データを取得して、「①プラセボ群 vs. ②薬剤A投与群」のようなグループ(群)間での比較を行います。これは3反復の例。

	①プラセボ群			②薬剤A投与群		
	患者1	患者2	患者3	患者4	患者5	患者6
遺伝子1	8	7	8	7	8	8
遺伝子2	3	4	2	15	16	14
遺伝子3	5	6	6	5	5	6
遺伝子4	1	7	3	6	2	4

2群間比較

通常は、同一条件の反復データを取得して、「①プラセボ群 vs. ②薬剤A投与群」のようなグループ(群)間での比較を行います。これは3反復の例。こうすることで、

	①プラセボ群			②薬剤A投与群		
	患者1	患者2	患者3	患者4	患者5	患者6
遺伝子1	8	7	8	7	8	8
遺伝子2	3	4	2	15	16	14
遺伝子3	5	6	6	5	5	6
遺伝子4	1	7	3	6	2	4

2群間比較

通常は、同一条件の反復データを取得して、「①プラセボ群 vs. ②薬剤A投与群」のようなグループ(群)間での比較を行います。これは3反復の例。こうすることで、**反復なしデータ**で③最も発現変動していた遺伝子4が、

	①プラセボ群		②薬剤A投与群		log比	p値
	患者1		患者4			
遺伝子1	8		7		-0.2	
遺伝子2	3		15		2.32	
遺伝子3	5		5		0	
遺伝子4	1		6		2.58	③

2群間比較

通常は、同一条件の反復データを取得して、「①プラセボ群 vs. ②薬剤A投与群」のようなグループ(群)間での比較を行います。これは3反復の例。こうすることで、反復なしデータで③最も発現変動していた遺伝子4が、**反復あり**にすると実はそうでもなかったといったことがわかります。

	①プラセボ群			②薬剤A投与群			log比	p値
	患者1	患者2	患者3	患者4	患者5	患者6		
遺伝子1	8	7	8	7	8	8	0	
遺伝子2	3	4	2	15	16	14	2.32	
遺伝子3	5	6	6	5	5	6	-0.1	
遺伝子4	1	7	3	6	2	4	0.13	③

2群間比較

通常は、同一条件の反復データを取得して、「①プラセボ群 vs. ②薬剤A投与群」のようなグループ(群)間での比較を行います。これは3反復の例。こうすることで、反復なしデータで③最も発現変動していた遺伝子4が、反復ありにすると実はそうでもなかったといったことがわかります。また、**反復ありデータにすることで、一般的な発現変動解析手法を適用でき、④p値などの結果を得られます。**若干不正確

①プラセボ群

②薬剤A投

	患者1	患者2	患者3	患者4	患者5	患者6	log比	p値
遺伝子1	8	7	8	7	8	8	0	
遺伝子2	3	4	2	15	16	14	2.32	
遺伝子3	5	6	6	5	5	6	-0.1	
遺伝子4	1	7	3	6	2	4	0.13	



Contents

- 自己紹介と東大アグリバイオの紹介
- トランスクリプトーム解析、発現解析、発現変動解析、実験デザイン
- 2群間比較: 実データ、TCC(反復増やすとDEG増える)
- 他グループによる性能評価論文(TCCが非推奨となる場合も!)
- TCCで3群間比較、baySeqも組み合わせて発現パターンまで得る
- (Rで)塩基配列解析
- Single-cell RNA-seq(scRNA-seq)
- バイオインフォマティクス実習

2群間比較用の実データ

2013年と古いですが、2群間比較(3反復 vs. 3反復)用のRNA-seq解析論文。

[Genome Res.](#) 2013 Oct;23(10):1563-79. doi: 10.1101/gr.154872.113. Epub 2013 Jul 26.

Sumoylation at chromatin governs coordinated repression of a transcriptional program essential for cell growth and proliferation.

[Neyret-Kahn H¹](#), [Benhamed M](#), [Ye T](#), [Le Gras S](#), [Cossec JC](#), [Lapaquette P](#), [Bischof O](#), [Ouspenskaia M](#), [Dasso M](#), [Seeler J](#), [Davidson I](#), [Dejean A](#).

Author information

Abstract

Despite numerous studies on specific sumoylated transcriptional regulators, the global role of SUMO on chromatin in relation to transcription regulation remains largely unknown. Here, we determined the genome-wide localization of SUMO1 and SUMO2/3, as well as of UBC9 (encoded by UBE2I) and PIAS4 (encoded by PIAS4), two markers for active sumoylation, along with Pol II and histone marks in proliferating versus senescent human fibroblasts together with gene expression profiling. We found that, whereas SUMO alone is widely distributed over the genome with strong association at active promoters, active sumoylation occurs most prominently at promoters of histone and protein biogenesis genes, as well as Pol I rRNAs and Pol III tRNAs. Remarkably, these four classes of genes are up-regulated by inhibition of sumoylation, indicating that SUMO normally acts to restrain their expression. In line with this finding, sumoylation-deficient cells show an increase in both cell size and global protein levels. Strikingly, we found that in senescent cells, the SUMO machinery is selectively retained at histone and tRNA gene clusters, whereas it is massively released from all other unique chromatin regions. These data, which reveal the highly dynamic nature of the SUMO landscape, suggest that maintenance of a repressive environment at histone and tRNA loci is a hallmark of the senescent state. The approach taken in our study thus permitted the identification of a common biological output and uncovered hitherto unknown functions for active sumoylation at chromatin as a key mechanism that, in dynamically marking chromatin by a simple modifier, orchestrates concerted transcriptional regulation of a network of genes essential for cell growth and proliferation.

PMID: 23893515 PMCID: [PMC3787255](#) DOI: [10.1101/gr.154872.113](#)

2群間比較用の実データ

2013年と古いですが、2群間比較(3反復 vs. 3反復)用のRNA-seq解析論文。
Proliferative vs. Ras-induced senescent human primary fibroblasts の比較をしているようです。

	Pro群			Ras群		
	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
ENSG000000000419	282	354	208	165	301	209
ENSG000000000457	201	254	183	166	296	148
ENSG000000000460	114	112	101	55	81	59
ENSG000000000938	0	0	0	2	2	1
ENSG000000000971	747	914	605	252	414	147
...						

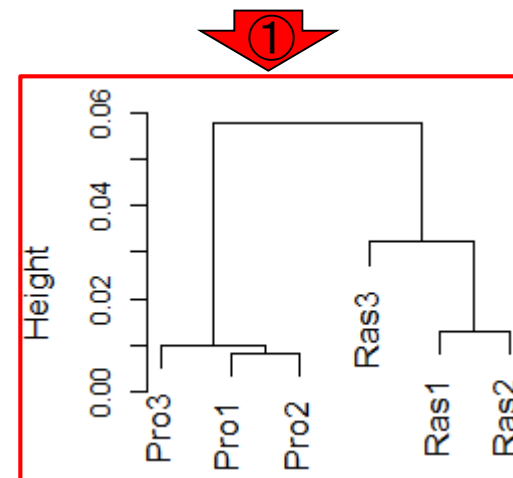
60,234 genes

クラスタリング結果

サンプル間クラスタリング結果。①このように群ごとに明瞭に分かれている場合は、発現変動遺伝子(DEG)が沢山得られることが期待されます。

	Pro群			Ras群		
	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
ENSG000000000419	282	354	208	165	301	209
ENSG000000000457	201	254	183	166	296	148
ENSG000000000460	114	112	101	55	81	59
ENSG000000000938	0	0	0	2	2	1
ENSG000000000971	747	914	605	252	414	147
...						

60,234 genes

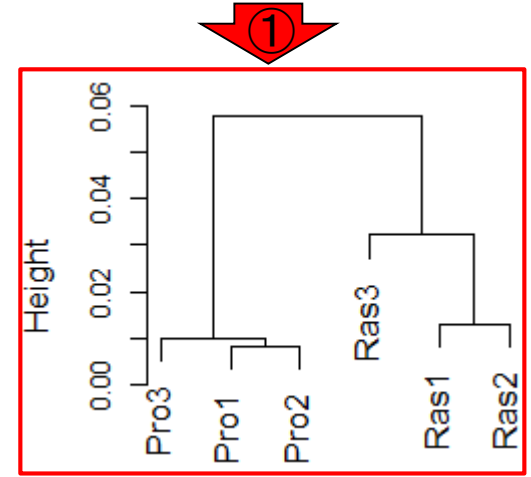


平均シルエットスコア

サンプル間クラスタリング結果。①このように群ごとに明瞭に分かれている場合は、発現変動遺伝子(DEG)が沢山得られることが期待されます。詳細はすっ飛ばしますが、Pro群 vs. Ras群の平均シルエットスコア(AS値)は0.69。0よりも大きく最大値(1)に近い値なので、AS値からもDEGの多さが予想されるデータ。

	Pro群			Ras群		
	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
ENSG000000000419	282	354	208	165	301	209
ENSG000000000457	201	254	183	166	296	148
ENSG000000000460	114	112	101	55	81	59
ENSG000000000938	0	0	0	2	2	1
ENSG000000000971	747	914	605	252	414	147
...						

60,234 genes



Contents

- 自己紹介と東大アグリバイオの紹介
- トランスクリプトーム解析、発現解析、発現変動解析、実験デザイン
- 2群間比較: 実データ、TCC(反復増やすとDEG増える)
- 他グループによる性能評価論文(TCCが非推奨となる場合も!)
- TCCで3群間比較、baySeqも組み合わせて発現パターンまで得る
- (Rで)塩基配列解析
- Single-cell RNA-seq(scRNA-seq)
- バイオインフォマティクス実習

BMC Bioinformatics. 2013 Jul 9;14:219. doi: 10.1186/1471-2105-14-219.

TCC: an R package for comparing tag count data with robust normalization strategies.

Sun J¹, Nishiyama T, Shimizu K, Kadota K.

Author information

Abstract

BACKGROUND: Differential expression analysis based on "next-generation" sequencing technologies is a fundamental means of studying RNA expression. We recently developed a multi-step normalization method (called TbT) for two-group RNA-seq data with replicates and demonstrated that the statistical methods available in four R packages (edgeR, DESeq, baySeq, and NBPSeg) together with TbT can produce a well-ranked gene list in which true differentially expressed genes (DEGs) are top-ranked and non-DEGs are bottom ranked. However, the advantages of the current TbT method come at the cost of a huge computation time. Moreover, the R packages did not have normalization methods based on such a multi-step strategy.

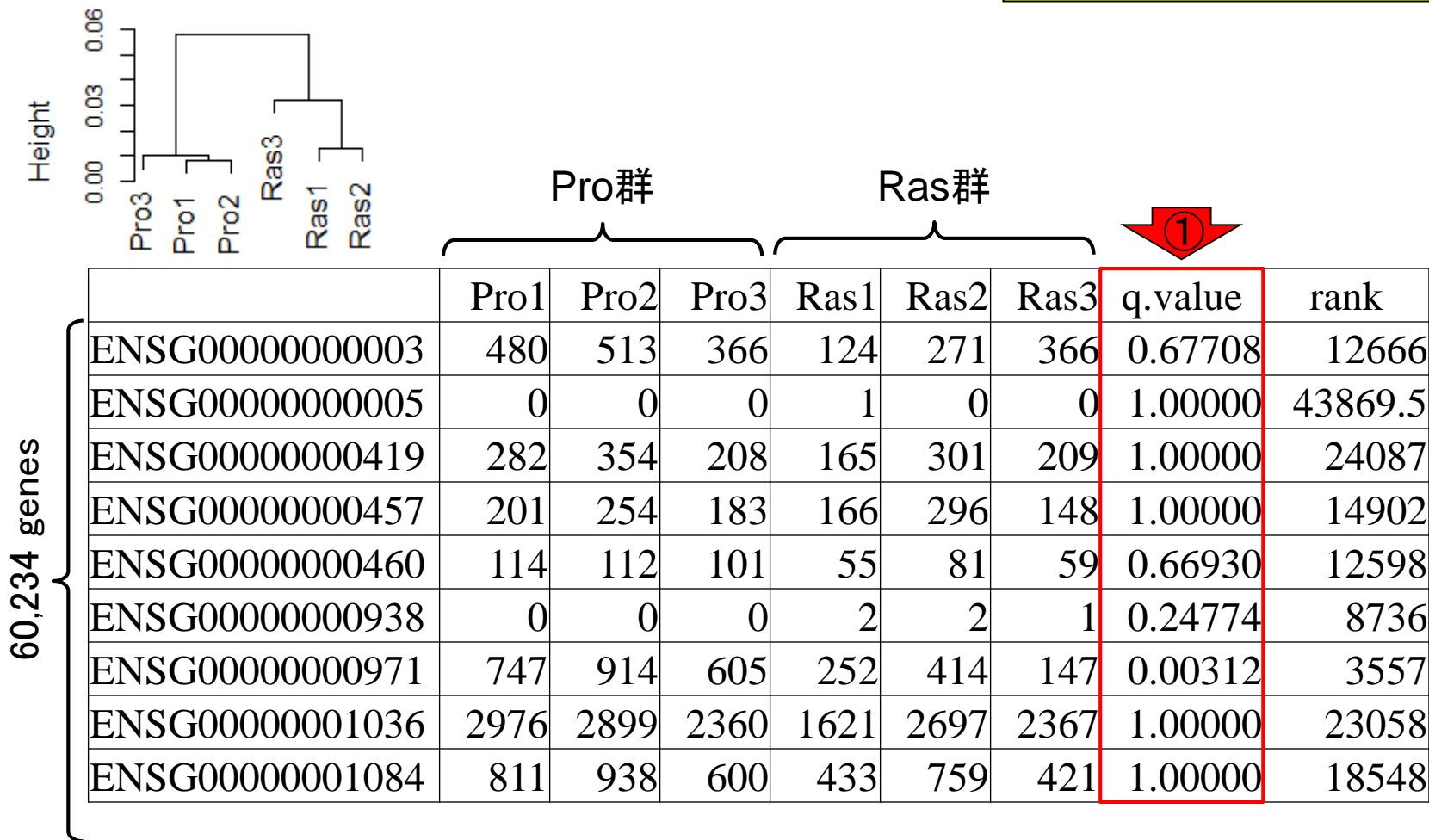
RESULTS: TCC (an acronym for Tag Count Comparison) is an R package that provides a series of functions for differential expression analysis of tag count data. The package incorporates multi-step normalization methods, whose strategy is to remove potential DEGs before performing the data normalization. The normalization function based on this DEG elimination strategy (DEGES) includes (i) the original TbT method based on DEGES for two-group data with or without replicates, (ii) much faster methods for two-group data with or without replicates, and (iii) methods for multi-group comparison. TCC provides a simple unified interface to perform such analyses with combinations of functions provided by edgeR, DESeq, and baySeq. Additionally, a function for generating simulation data under various conditions and alternative DEGES procedures consisting of functions in the existing packages are provided. Bioinformatics scientists can use TCC to evaluate their methods, and biologists familiar with other R packages can easily learn what is done in TCC.

CONCLUSION: DEGES in TCC is essential for accurate normalization of tag count data, especially when up- and down-regulated DEGs in one of the samples are extremely biased in their number. TCC is useful for analyzing tag count data in various scenarios ranging from unbiased to extremely biased differential expression. TCC is available at <http://www.iu.a.u-tokyo.ac.jp/~kadota/TCC/> and will appear in Bioconductor (<http://bioconductor.org/>) from ver. 2.13.

PMID: 23837715 PMCID: PMC3716788 DOI: 10.1186/1471-2105-14-219

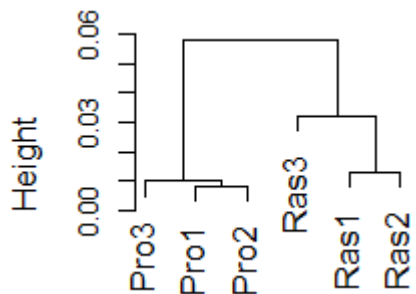
TCC実行結果の一部

TCC実行結果の一部として、①q.value (q-value)とrank (順位情報)を表示。q-valueは、adjusted p-valueとも呼ばれる。



TCC実行結果の一部

TCC実行結果の一部として、①q.value (q-value)とrank (順位情報)を表示。q-valueは、adjusted p-valueとも呼ばれる。発現変動順にソートした結果。②上位6個は、いずれもRas群で高発現パターンの遺伝子であることがわかる。



Pro群 Ras群

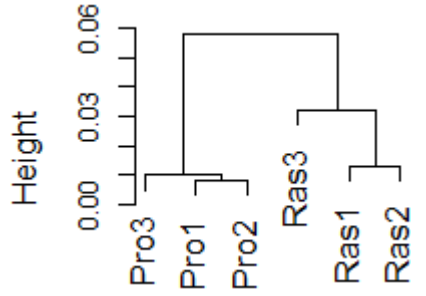
60,234 genes

	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	q.value	rank
ENSG00000240386	0	0	0	4398	6094	7683	0.00000	1
ENSG00000128564	18	27	19	2038	2657	2138	0.00000	2
ENSG00000188064	9	7	10	1027	1362	1264	0.00000	3
ENSG00000101188	7	6	11	1054	1518	1050	0.00000	4
ENSG00000145107	5	5	2	470	742	501	0.00000	5
ENSG00000243742	84	63	52	2072	3185	2657	0.00000	6
ENSG00000163431	4342	3927	4153	50	85	41	0.00000	7
ENSG00000204291	1420	1497	1329	16	30	18	0.00000	8
ENSG00000181634	127	198	68	9606	#####	#####	0.00000	9



False Discovery Rate

偽陽性率10% (10% false positive rate; 10% FPR)を満たす遺伝子数は、p-value < 0.10で得られる。同様に、偽発見率10% (10% false discovery rate; 10% FDR)を満たす遺伝子数は、q-value < 0.10で得られる。



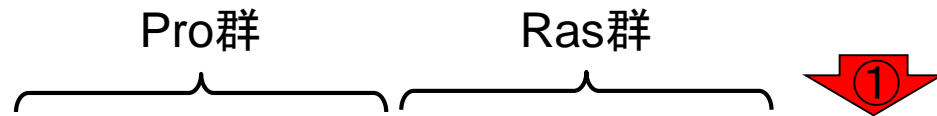
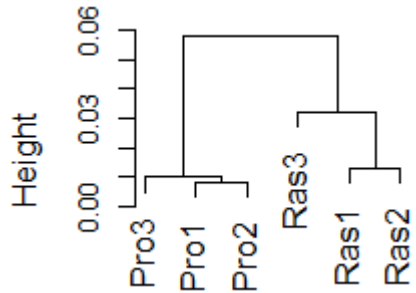
60,234 genes

	Pro群			Ras群				
	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	q.value	rank
ENSG00000006715	978	1035	705	378	620	401	0.09980	6754
ENSG00000205060	1207	1369	1052	528	877	412	0.09980	6756
ENSG00000271075	2	6	1	0	0	0	0.09980	6755
ENSG00000150753	4292	4142	3305	3297	5334	4732	0.09986	6757
ENSG00000150907	25	21	20	6	9	9	0.09987	6758
ENSG00000233247	319	338	249	226	506	364	0.09998	6759
ENSG00000226261	4	8	1	0	1	0	0.10001	6760
ENSG00000136859	2558	2190	2370	929	1327	1360	0.10006	6762
ENSG00000164414	349	416	274	115	208	195	0.10006	6763



False Discovery Rate

10% FDRを満たす遺伝子数は6,759個。これは「許容する偽物(non-DEG)混入割合」に相当し、例えば6,759個中675.9個が理論上偽物だということ。



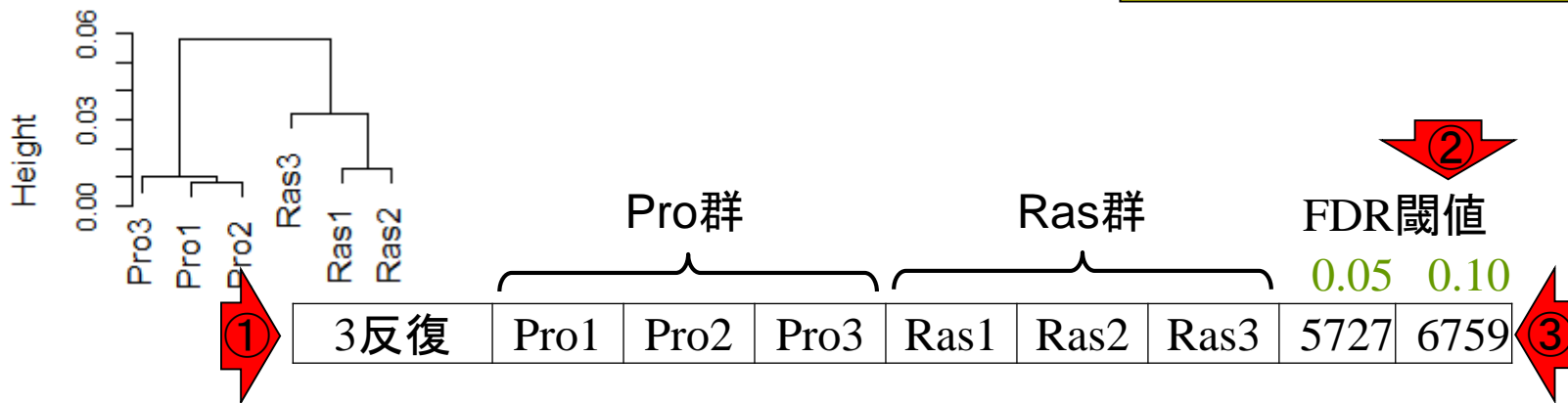
60,234 genes

	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	q.value	rank
ENSG00000006715	978	1035	705	378	620	401	0.09980	6754
ENSG00000205060	1207	1369	1052	528	877	412	0.09980	6756
ENSG00000271075	2	6	1	0	0	0	0.09980	6755
ENSG00000150753	4292	4142	3305	3297	5334	4732	0.09986	6757
ENSG00000150907	25	21	20	6	9	9	0.09987	6758
ENSG00000233247	319	338	249	226	506	364	0.09998	6759
ENSG00000226261	4	8	1	0	1	0	0.10001	6760
ENSG00000136859	2558	2190	2370	929	1327	1360	0.10006	6762
ENSG00000164414	349	416	274	115	208	195	0.10006	6763



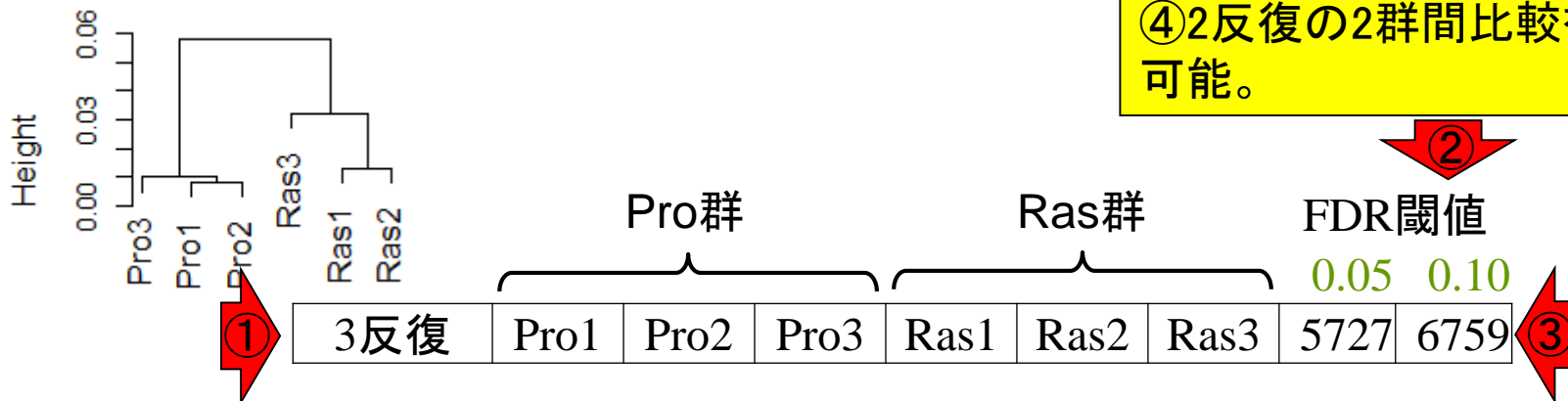
3反復で2群間比較

①3反復の2群間比較の結果として、②10% FDRを満たす遺伝子数は、③6,759個であった。



2反復で2群間比較

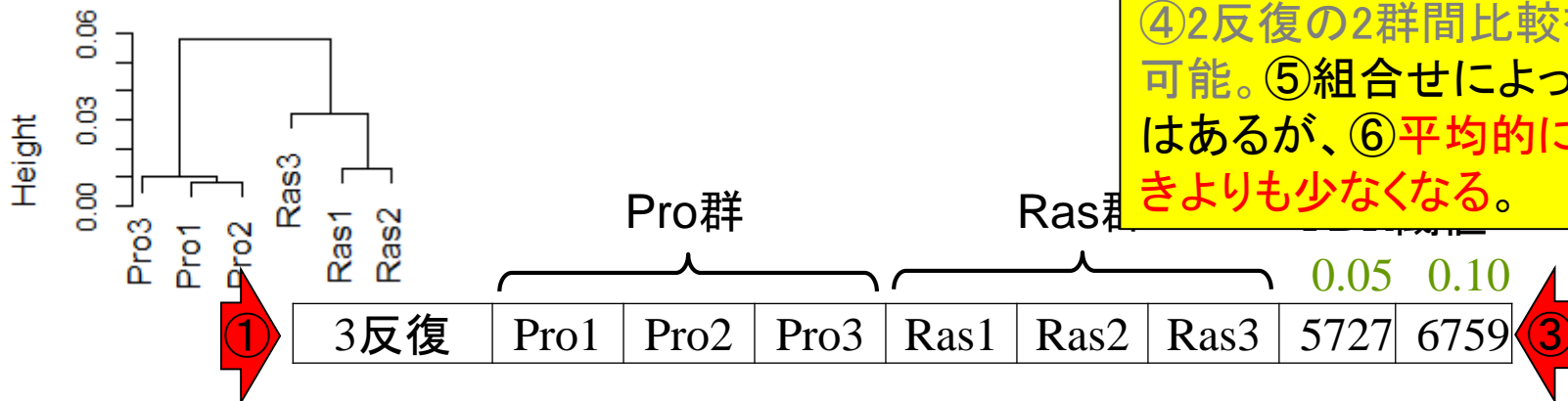
①3反復の2群間比較の結果として、②10% FDRを満たす遺伝子数は、③6,759個であった。元は3反復のデータなので、④2反復の2群間比較を9通り行うことが可能。



2反復	Pro1	Pro2		Ras1	Ras2		8086	9026
2反復	Pro1	Pro2		Ras1		Ras3	3550	4371
2反復	Pro1	Pro2			Ras2	Ras3	3282	4059
2反復	Pro1		Pro3	Ras1	Ras2		7739	8578
2反復	Pro1		Pro3	Ras1		Ras3	3330	3986
2反復	Pro1		Pro3		Ras2	Ras3	3186	3889
2反復		Pro2	Pro3	Ras1	Ras2		6545	7444
2反復		Pro2	Pro3	Ras1		Ras3	3210	3883
2反復		Pro2	Pro3		Ras2	Ras3	3120	3821

2反復で2群間比較

①3反復の2群間比較の結果として、②10% FDRを満たす遺伝子数は、③6,759個であった。元は3反復のデータなので、④2反復の2群間比較を9通り行うことが可能。⑤組合せによって結果にバラつきはあるが、⑥平均的には、③3反復のときよりも少なくなる。



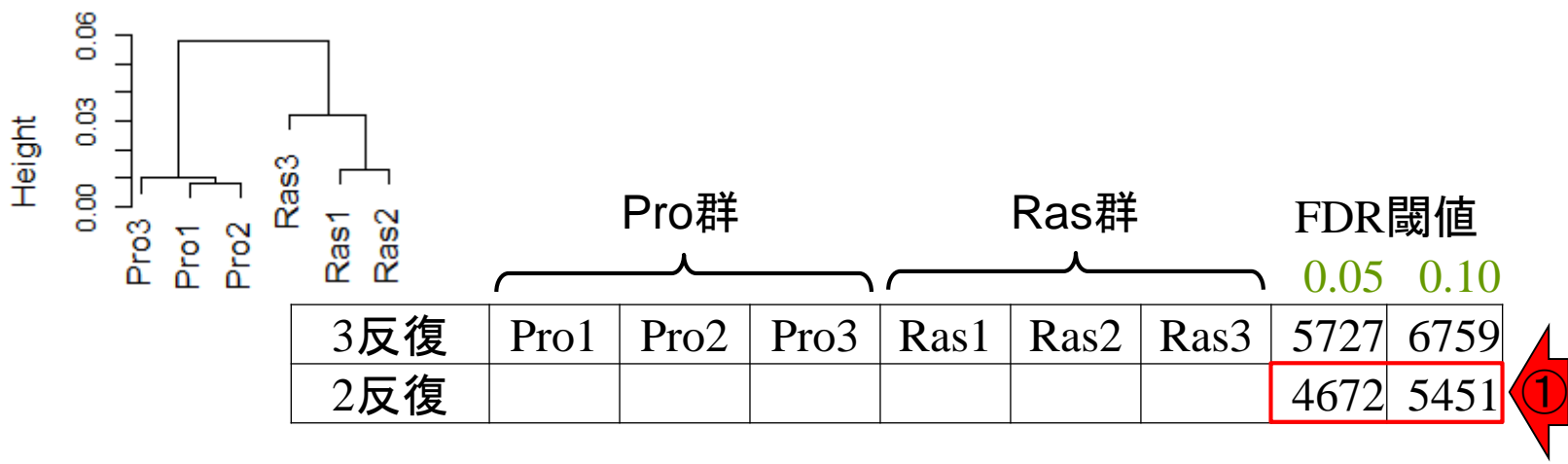
④	2反復	Pro1	Pro2		Ras1	Ras2		8086	9026
	2反復	Pro1	Pro2		Ras1		Ras3	3550	4371
	2反復	Pro1	Pro2			Ras2	Ras3	3282	4059
	2反復	Pro1		Pro3	Ras1	Ras2		7739	8578
	2反復	Pro1		Pro3	Ras1		Ras3	3330	3986
	2反復	Pro1		Pro3		Ras2	Ras3	3186	3889
	2反復		Pro2	Pro3	Ras1	Ras2		6545	7444
	2反復		Pro2	Pro3	Ras1		Ras3	3210	3883
	2反復		Pro2	Pro3		Ras2	Ras3	3120	3821
	平均						4672	5451	

⑤

⑥

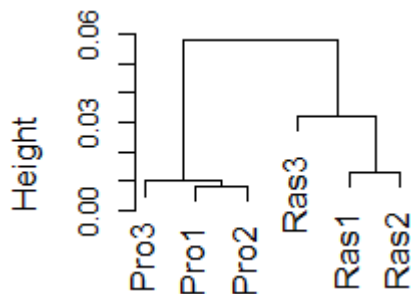
①反復数が減ると(3→2)、FDR閾値を満たす遺伝子数も減る。

ここまでのまとめ



反復なしで2群間比較

①反復数が減ると(3→2)、FDR閾値を満たす遺伝子数も減る。②反復なしにすると大幅に減る。



	Pro群			Ras群			FDR閾値	
							0.05	0.10
3反復	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	5727	6759
2反復							4672	5451

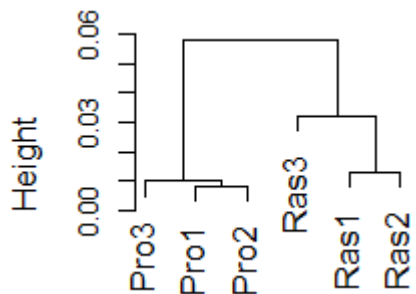
Pro1			Ras1			116	135
Pro1				Ras2		165	217
Pro1					Ras3	2	4
	Pro2		Ras1			77	102
	Pro2			Ras2		137	170
	Pro2				Ras3	1	3
		Pro3	Ras1			120	161
		Pro3		Ras2		143	185
		Pro3			Ras3	1	4

平均 85 109

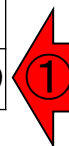


反復増やすとDEG増える

「①反復なしデータで実行するとほとんどDEGが得られなくなるんですけど、やり方が間違ってますか?」という質問をときどき受けます。このような結果になるのは、少なくとも私の中では常識です。

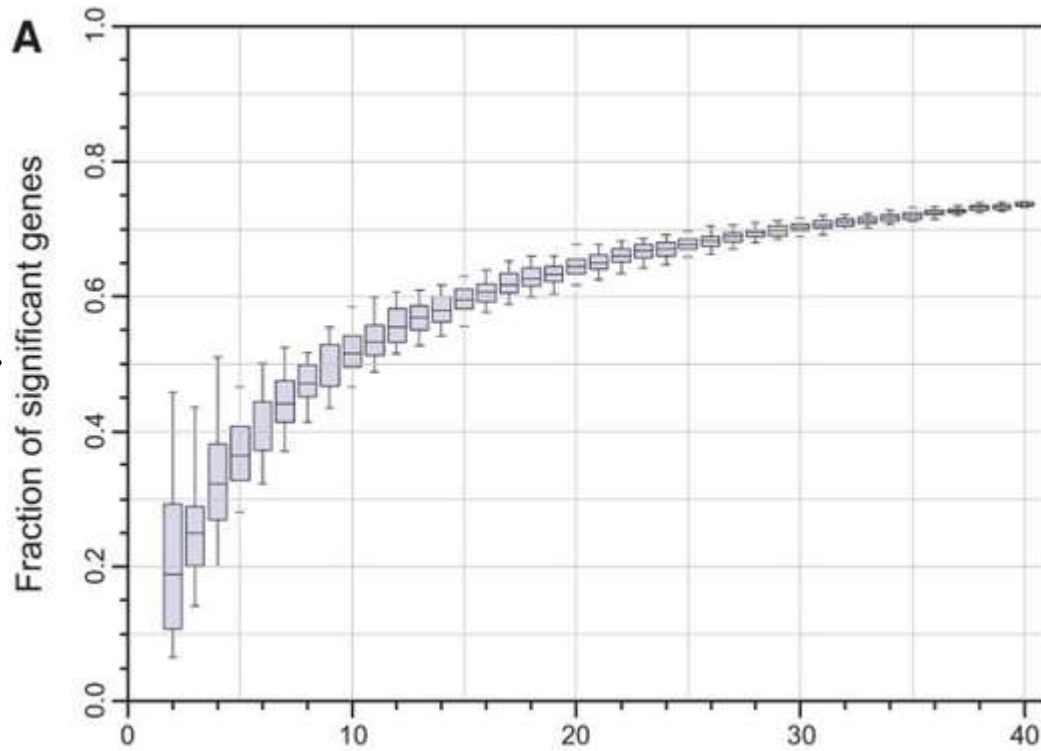


	Pro群			Ras群			FDR閾値	
							0.05	0.10
3反復	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	5727	6759
2反復							4672	5451
反復なし							85	109



反復増やすとDEG増える

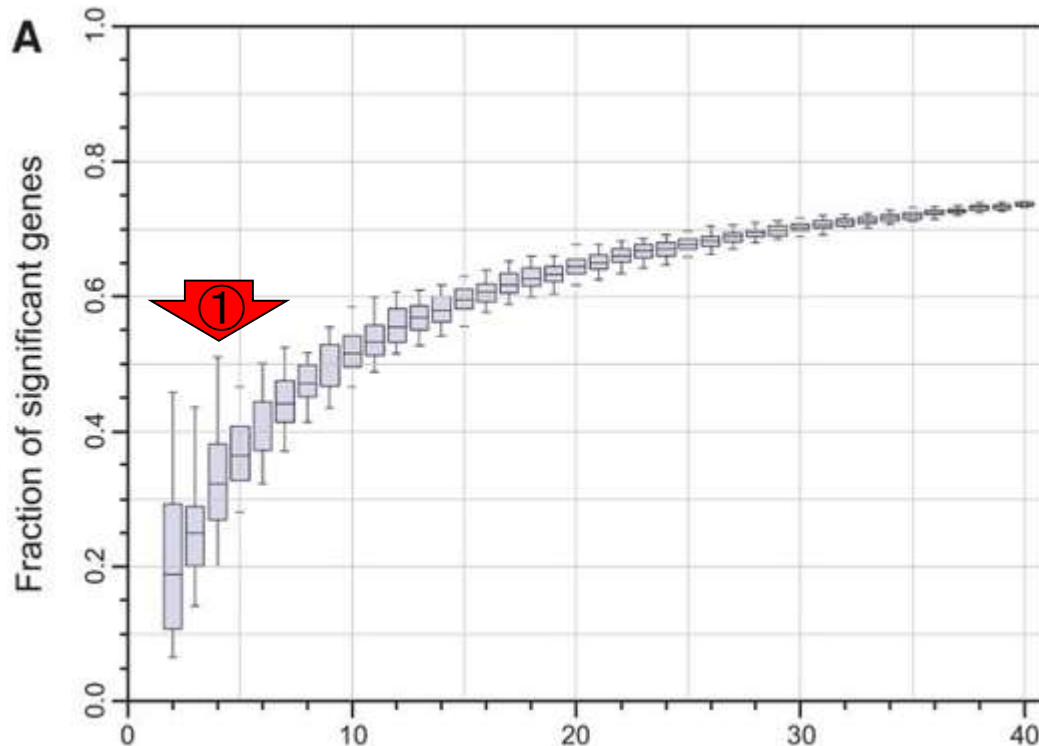
①の論文のFig. 1A。②横軸は反復数で、
③縦軸は全遺伝子に占めるDEGの割合。
これは2群間比較用で、各群につき42反
復もあるデータです。



① Schurch et al., *RNA*, **22**: 839–851, 2016

© 2016 Schurch et al.; Published by Cold Spring
Harbor Laboratory Press for the RNA Society

反復増やすとDEG増える



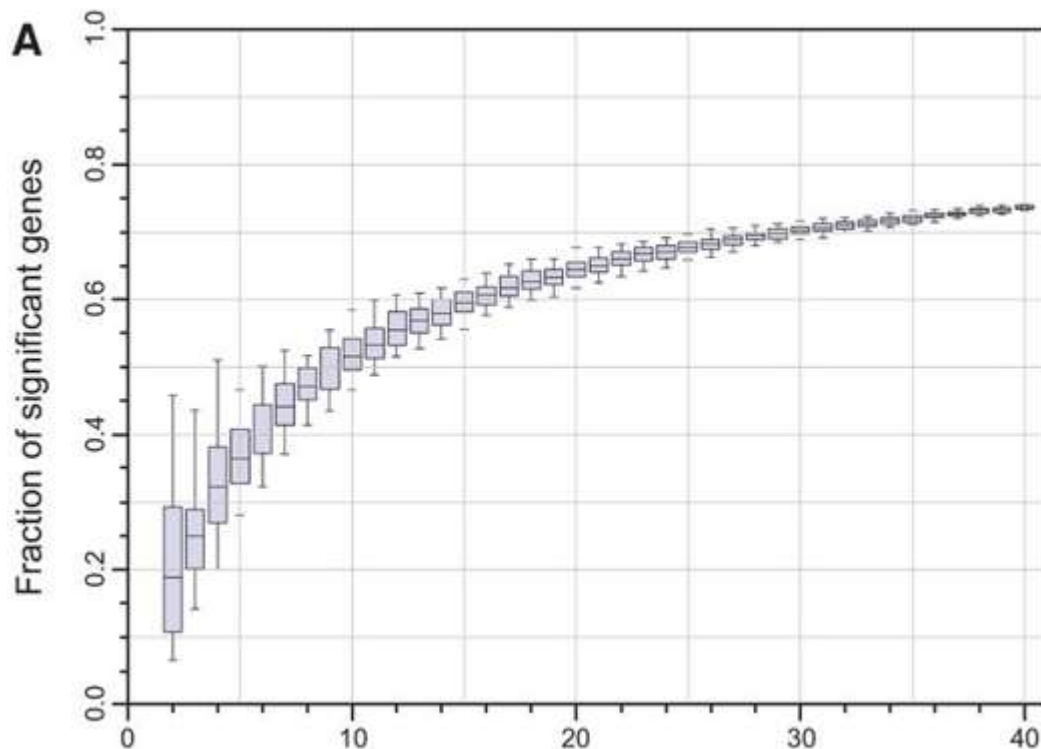
box plotになっている理由は、ランダムサンプリングを行っているから。例えば①は4反復分をランダムにサンプリングして、DEGの割合(P_{DEG})を算出する作業を何度も繰り返した結果。全体的に反復数が多いほど結果が安定することがわかる。そして、反復数が多いほど P_{DEG} の値が大きくなり、②一定値(約0.74)に近づいていることもわかる。これは、edgeRというRパッケージの解析結果

Schurch et al., *RNA*, **22**: 839–851, 2016

© 2016 Schurch et al.; Published by Cold Spring Harbor Laboratory Press for the RNA Society

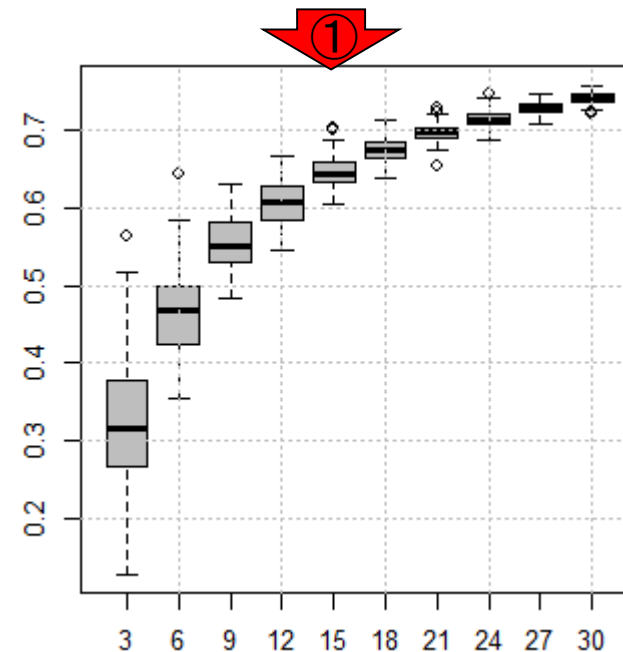
反復増やすとDEG増える

①はTCCで同じデータを解析した結果。
 ②の論文のAdditional file 3aです。傾向は全く同じですね。何か比較解析を行う際には、反復数を揃えて実行した結果に基づいて考察するのが基本。



Schurch et al., *RNA*, **22**: 839–851, 2016

© 2016 Schurch et al.; Published by Cold Spring Harbor Laboratory Press for the RNA Society



Zhao et al., *Biol. Proc. Online*, **20**: 5, 2018

Contents

- 自己紹介と東大アグリバイオの紹介
- トランスクリプトーム解析、発現解析、発現変動解析、実験デザイン
- 2群間比較: 実データ、TCC(反復増やすとDEG増える)
- 他グループによる性能評価論文(TCCが非推奨となる場合も!)
- TCCで3群間比較、baySeqも組み合わせて発現パターンまで得る
- (Rで)塩基配列解析
- Single-cell RNA-seq(scRNA-seq)
- バイオインフォマティクス実習

性能評価論文

2018年の論文で、TCC(に実装されているDEGES正規化法)と他の正規化法の比較がなされています。

[Brief Bioinform.](#) 2018 Sep 28;19(5):776-792. doi: 10.1093/bib/bbx008.

Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions.

[Evans C](#)¹, [Hardin J](#)², [Stoebel DM](#)³.

+ Author information

Abstract

RNA-Seq is a widely used method for studying the behavior of genes under different biological conditions. An essential step in an RNA-Seq study is normalization, in which raw data are adjusted to account for factors that prevent direct comparison of expression measures. Errors in normalization can have a significant impact on downstream analysis, such as inflated false positives in differential expression analysis. An underemphasized feature of normalization is the assumptions on which the methods rely and how the validity of these assumptions can have a substantial impact on the performance of the methods. In this article, we explain how assumptions provide the link between raw RNA-Seq read counts and meaningful measures of gene expression. We examine normalization methods from the perspective of their assumptions, as an understanding of methodological assumptions is necessary for choosing methods appropriate for the data at hand. Furthermore, we discuss why normalization methods perform poorly when their assumptions are violated and how this causes problems in subsequent analysis. To analyze a biological experiment, researchers must select a normalization method with assumptions that are met and that produces a meaningful measure of expression for the given experiment.

PMID: 28334202 PMCID: [PMC6171491](#) DOI: [10.1093/bib/bbx008](#)

性能評価論文の結論1

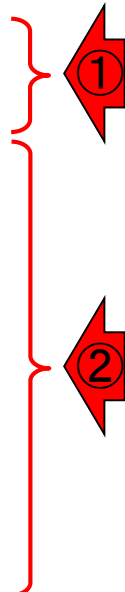
TCCとedgeR (or DESeq2)の性能は、
DEGの偏りがない(unbiased)場合は互角。
。

	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene1	14	110	73	0	0	10
gene2	1	0	11	12	539	346
gene3	8	19	6	11	12	14
gene4	8	6	3	5	1	52
gene5	22	16	7	1	8	0
gene6	436	696	543	774	808	835
gene7	10	0	11	9	0	8
gene8	10	5	5	27	20	1
gene9	101	71	13	49	63	63
gene10	1	2	2	0	0	0

性能評価論文の結論1

TCCとedgeR (or DESeq2)の性能は、DEGの偏りが無い(unbiased)場合は互角。このデータは、①DEGを含む割合が20% ($P_{\text{DEG}} = 0.2$)。②残りはnon-DEG。

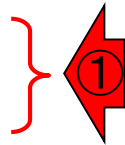
	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene1	14	110	73	0	0	10
gene2	1	0	11	12	539	346
gene3	8	19	6	11	12	14
gene4	8	6	3	5	1	52
gene5	22	16	7	1	8	0
gene6	436	696	543	774	808	835
gene7	10	0	11	9	0	8
gene8	10	5	5	27	20	1
gene9	101	71	13	49	63	63
gene10	1	2	2	0	0	0



性能評価論文の結論1

TCCとedgeR (or DESeq2)の性能は、DEGの偏りが無い(unbiased)場合は互角。このデータは、①DEGを含む割合が20% ($P_{\text{DEG}} = 0.2$)。②残りはnon-DEG。①DEGの半分はG1群で高発現(gene1)、残りの半分はG2群で高発現(gene2)。

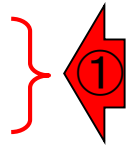
	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene1	14	110	73	0	0	10
gene2	1	0	11	12	539	346
gene3	8	19	6	11	12	14
gene4	8	6	3	5	1	52
gene5	22	16	7	1	8	0
gene6	436	696	543	774	808	835
gene7	10	0	11	9	0	8
gene8	10	5	5	27	20	1
gene9	101	71	13	49	63	63
gene10	1	2	2	0	0	0



性能評価論文の結論1

TCCとedgeR (or DESeq2)の性能は、DEGの偏りが無い(unbiased)場合は互角。このデータは、①DEGを含む割合が20% ($P_{DEG} = 0.2$)。②残りはnon-DEG。①DEGの半分はG1群で高発現(gene1)、残りの半分はG2群で高発現(gene2)。この例のように、DEGの発現変動パターンに偏りが無い場合の性能は互角。

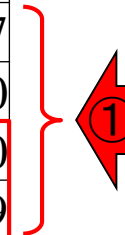
	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene1	14	110	73	0	0	10
gene2	1	0	11	12	539	346
gene3	8	19	6	11	12	14
gene4	8	6	3	5	1	52
gene5	22	16	7	1	8	0
gene6	436	696	543	774	808	835
gene7	10	0	11	9	0	8
gene8	10	5	5	27	20	1
gene9	101	71	13	49	63	63
gene10	1	2	2	0	0	0



性能評価論文の結論1

DEGの発現変動パターンに偏りが無い場合の性能は互角なので、DEGの割合は無関係。例えば、①DEGを含む割合が40% ($P_{\text{DEG}} = 0.4$) だったとしても、DEGの半分はG1群で高発現 (gene1 and 2)、残りの半分はG2群で高発現 (gene3 and 4) であれば性能は互角。

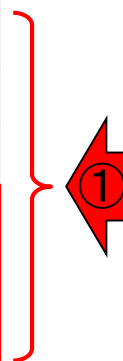
	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene1	14	110	73	0	0	7
gene2	29	0	263	1	25	0
gene3	8	19	6	193	227	150
gene4	33	6	3	443	41	139
gene5	22	16	7	2	17	16
gene6	436	696	543	594	520	681
gene7	10	0	11	5	1	8
gene8	10	5	5	35	100	5
gene9	101	71	13	35	26	73
gene10	1	2	2	0	1	0



性能評価論文の結論1

DEGの発現変動パターンに偏りがない場合の性能は互角なので、DEGの割合は無関係。例えば、①DEGを含む割合が60% ($P_{\text{DEG}} = 0.6$) だったとしても、DEGの半分はG1群で高発現 (gene1, 2, and 3)、残りの半分はG2群で高発現 (gene4, 5, and 6) であれば性能は互角。

	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene1	14	110	73	10	8	7
gene2	29	0	263	5	2	0
gene3	184	188	168	8	12	3
gene4	10	36	19	216	41	558
gene5	24	3	24	23	278	308
gene6	784	503	677	11160	10965	13551
gene7	10	0	11	7	1	8
gene8	10	5	3	9	100	5
gene9	101	71	44	94	26	73
gene10	1	2	5	0	1	0



性能評価論文の結論2

①DEGを含む割合が20% ($P_{\text{DEG}} = 0.2$)で、
②残りはnon-DEG。①全てのDEGがG1群で高発現のような、**DEGの発現変動パターンに偏りがある**場合は性能差が出る。
この例のようなDEGを含む割合が20%程度だとTCCの性能は高い。

	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene1	87	306	646	13	26	10
gene2	233	76	201	8	13	5
gene3	4	1	0	2	4	0
gene4	221	106	56	360	124	129
gene5	150	170	154	190	149	140
gene6	16	35	23	22	12	12
gene7	6	0	1	4	3	4
gene8	1	3	2	1	3	5
gene9	5	5	1	0	8	1
gene10	89	37	95	126	61	41



性能評価論文の結論2

①DEGを含む割合が40% ($P_{\text{DEG}} = 0.4$)で、
②残りはnon-DEG。①全てのDEGがG1群で高発現のような、**DEGの発現変動パターンに偏りがある**場合は性能差が出る。
この例のようなDEGを含む割合が40%程度でもTCCの性能は高い。

	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene1	87	306	21	8	18	3
gene2	233	76	360	12	5	34
gene3	69	42	144	5	1	10
gene4	3949	1408	2600	263	162	392
gene5	178	153	125	138	166	127
gene6	15	21	22	39	28	31
gene7	6	0	2	0	0	1
gene8	1	16	5	1	3	7
gene9	5	3	2	13	8	10
gene10	89	52	258	256	70	242

①

②

性能評価論文の結論2

①DEGを含む割合が60% ($P_{DEG} = 0.6$)で、
 ②残りはnon-DEG。①全てのDEGがG1群で高発現のような、**DEGの発現変動パターンに偏りがある**場合は性能差が出る。
 この例のようなDEGを含む割合が60%程度でもTCCの性能は高い。

	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene1	87	77	16	5	26	10
gene2	233	169	21	4	13	5
gene3	69	0	94	5	4	0
gene4	3949	5787	9658	144	124	129
gene5	3361	3211	2707	150	149	140
gene6	332	348	251	22	12	12
gene7	1	8	2	4	3	4
gene8	3	0	0	1	3	5
gene9	1	5	2	0	8	1
gene10	125	175	265	126	61	41



性能評価論文の結論2

①DEGを含む割合が60% ($P_{\text{DEG}} = 0.6$)で、
 ②残りはnon-DEG。①DEGの5/6がG1群
 で高発現のような、**DEGの発現変動パター**
ーンに偏りがある場合も性能差が出る。
 この例のようなDEGを含む割合が60%程
 度でもTCCの性能は高い。


	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene1	87	306	21	8	18	3
gene2	233	76	360	12	5	34
gene3	69	42	144	5	1	10
gene4	3949	1408	2600	263	162	392
gene5	3361	3009	2477	138	166	127
gene6	15	21	22	721	476	684
gene7	6	0	2	0	0	1
gene8	1	16	5	1	3	7
gene9	5	3	2	13	8	10
gene10	89	52	258	256	70	242



性能評価論文の結論3

但し、①DEGを含む割合が80% ($P_{\text{DEG}} = 0.8$)で、②残りはnon-DEG。①DEGの6/8がG1群で高発現のような、**DEGがほとんどで偏りがある場合にはTCCの性能は急激に落ちる(ワーストレベル)。**

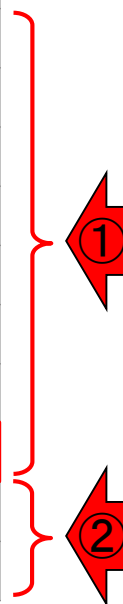
	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene1	87	77	16	5	14	2
gene2	233	169	21	4	17	1
gene3	69	0	94	5	1	0
gene4	3949	5787	9658	144	147	87
gene5	3361	3211	2707	150	123	136
gene6	332	348	251	22	24	32
gene7	1	8	2	57	41	30
gene8	3	0	0	117	59	49
gene9	1	5	2	0	5	1
gene10	125	175	265	62	148	419



性能評価論文の結論3

但し、①DEGを含む割合が80% ($P_{DEG} = 0.8$)で、②残りはnon-DEG。①DEGの7/8がG1群で高発現のような、DEGがほとんどで偏りがある場合にはTCCの性能は急激に落ちる(ワーストレベル)。

	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene1	87	56	382	8	2	1
gene2	233	24	551	12	14	0
gene3	69	55	9	5	1	0
gene4	3949	5958	2299	263	171	93
gene5	3361	2717	3113	138	154	139
gene6	332	870	475	39	20	24
gene7	28	27	128	0	4	0
gene8	0	5	2	18	116	20
gene9	3	1	6	1	3	0
gene10	104	143	258	101	134	530



性能評価論文のまとめ

性能評価にTCC(正確にはDEGES正規化法)を含めているだけなので、原著論文ではTCCを主語とした書き方にはなっていない。

- 1. DEGの偏りが無い(unbiased)場合は互角。DEGの割合は無関係。
- DEGに偏りがある場合は性能差が出る
 2. DEGの割合が60%程度以下の場合は、TCCの性能は高い(ほぼパーフェクト)
 3. DEGの割合が70%程度以上の場合は、TCCの性能は低い(ワーストレベル)

私は、①他グループの性能評価論文で、
②の事実を初めて知りました。

性能評価論文のまとめ

- 1. DEGの偏りが無い(unbiased)場合は互角。DEGの割合は無関係。
- DEGに偏りがある場合は性能差が出る
 - 2. DEGの割合が60%程度以下の場合は、TCCの性能は高い(ほぼパーフェクト)
 - 3. DEGの割合が60%程度以上の場合は、TCCの性能は低い(ワーストレベル)

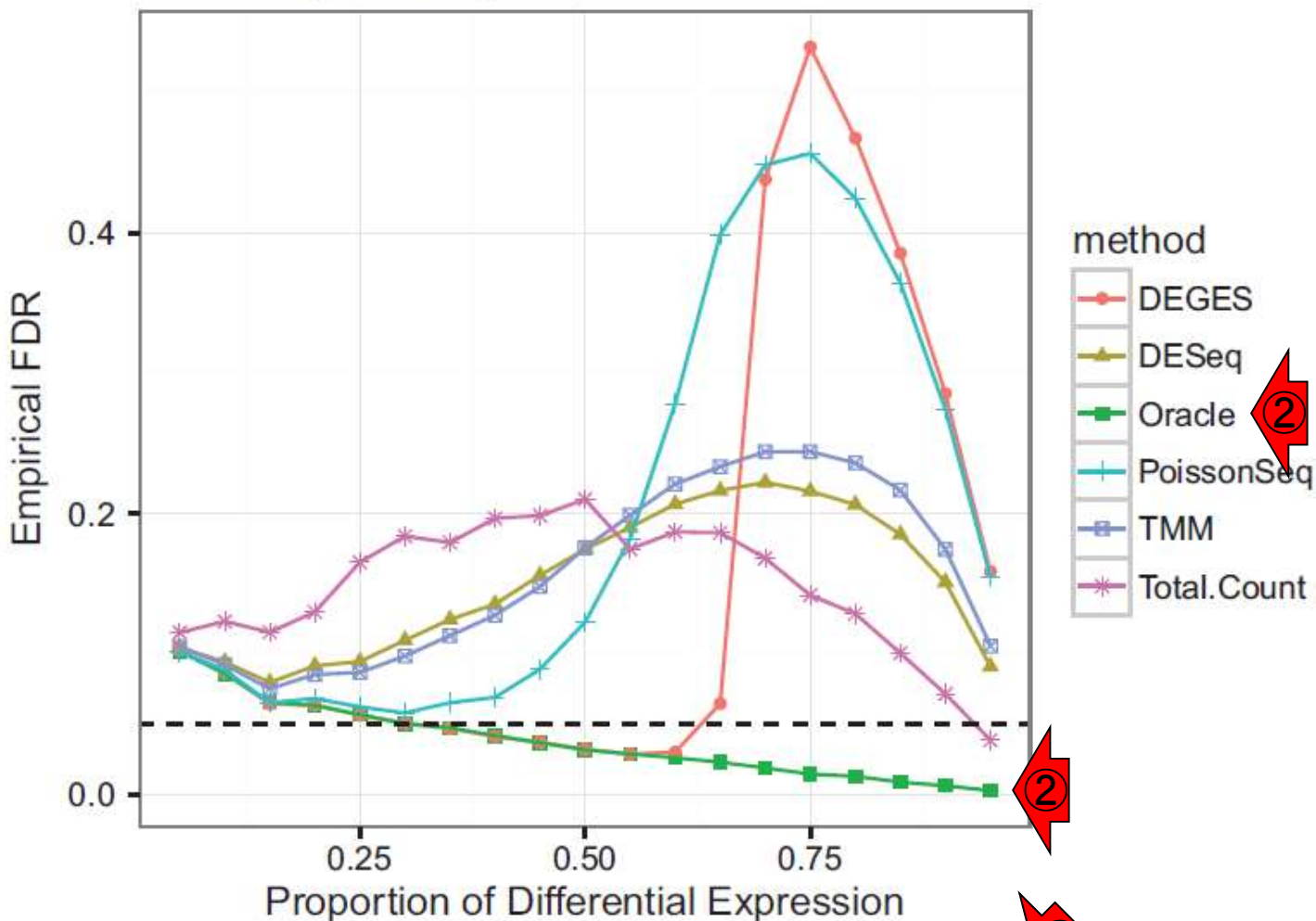
②

①

性能評価論文のFig. 8

eFDR by proportion of DE
asymmetry, different mRNA/cell

①他グループの性能評価論文の実際の図。横軸がDEGの割合。縦軸の値が②Oracle(神託;理想値と解釈すればよい)に近ければよい。

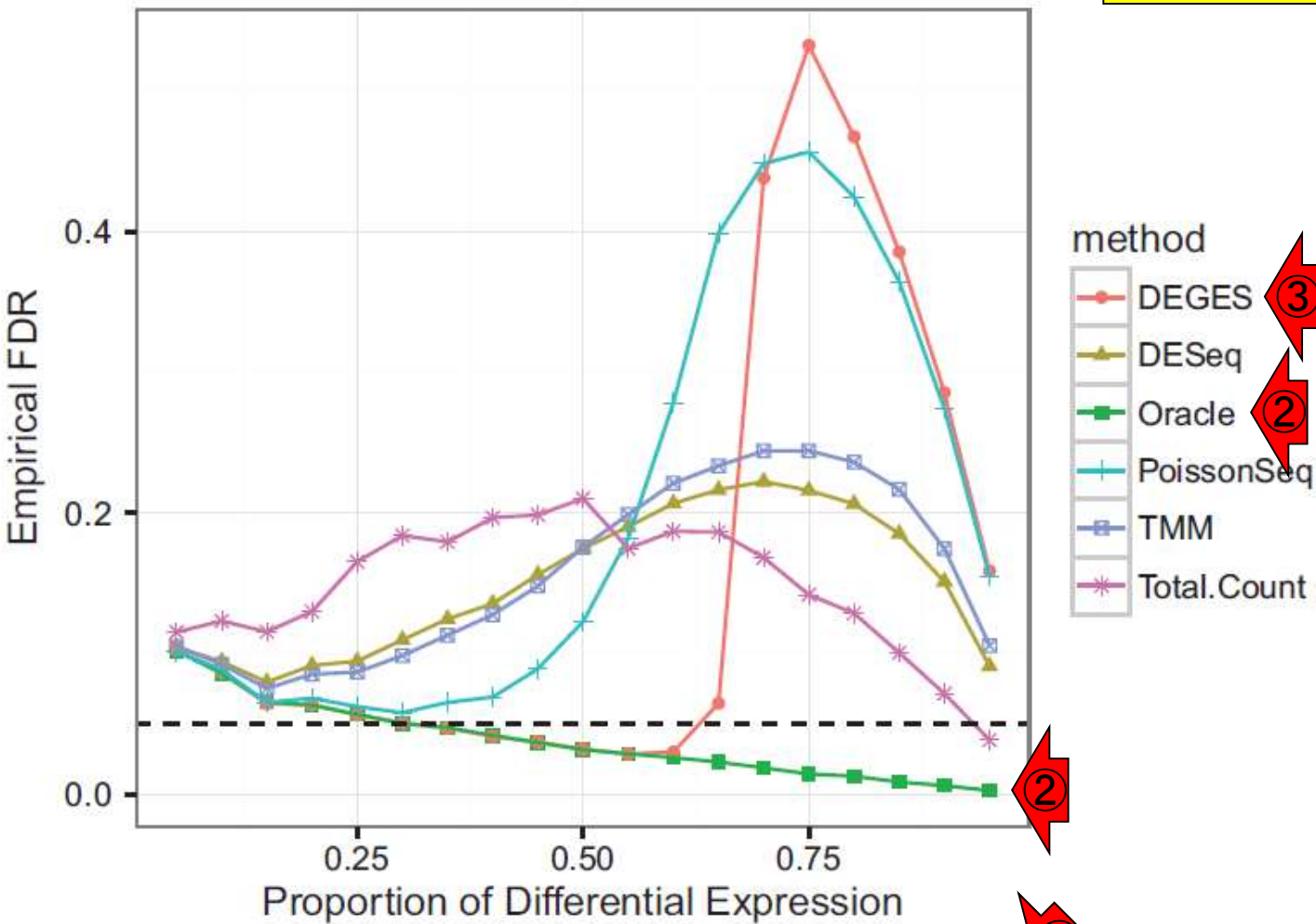


正規化法性能評価論文のFig. 8の右上の図

性能評価論文のFig. 8

eFDR by proportion of DE
asymmetry, different mRNA/cell

①他グループの性能評価論文の実際の図。横軸がDEGの割合。縦軸の値が②Oracle(神託;理想値と解釈すればよい)に近ければよい。③DEGESがTCCのこと。

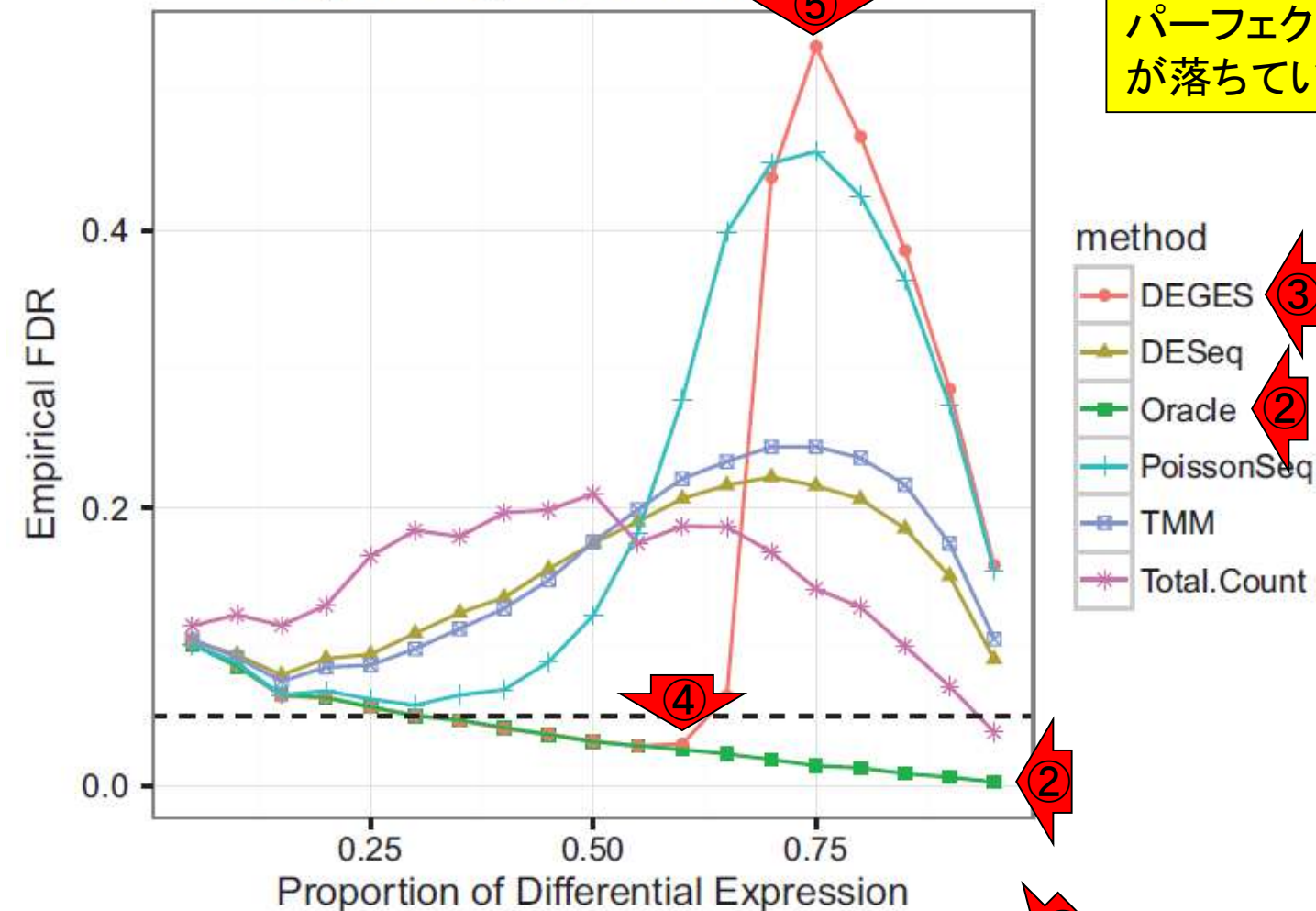


正規化法性能評価論文のFig. 8の右上の図

性能評価論文のFig. 8

eFDR by proportion of DE
asymmetry, different mRNA/cell

①他グループの性能評価論文の実際の図。横軸がDEGの割合。縦軸の値が②Oracle(神託;理想値と解釈すればよい)に近ければよい。③DEGESがTCCのこと。④DEGの割合が60%くらいまではほぼパーフェクトだが、その後は一気に性能が落ちていき、⑤75%以降はワースト。

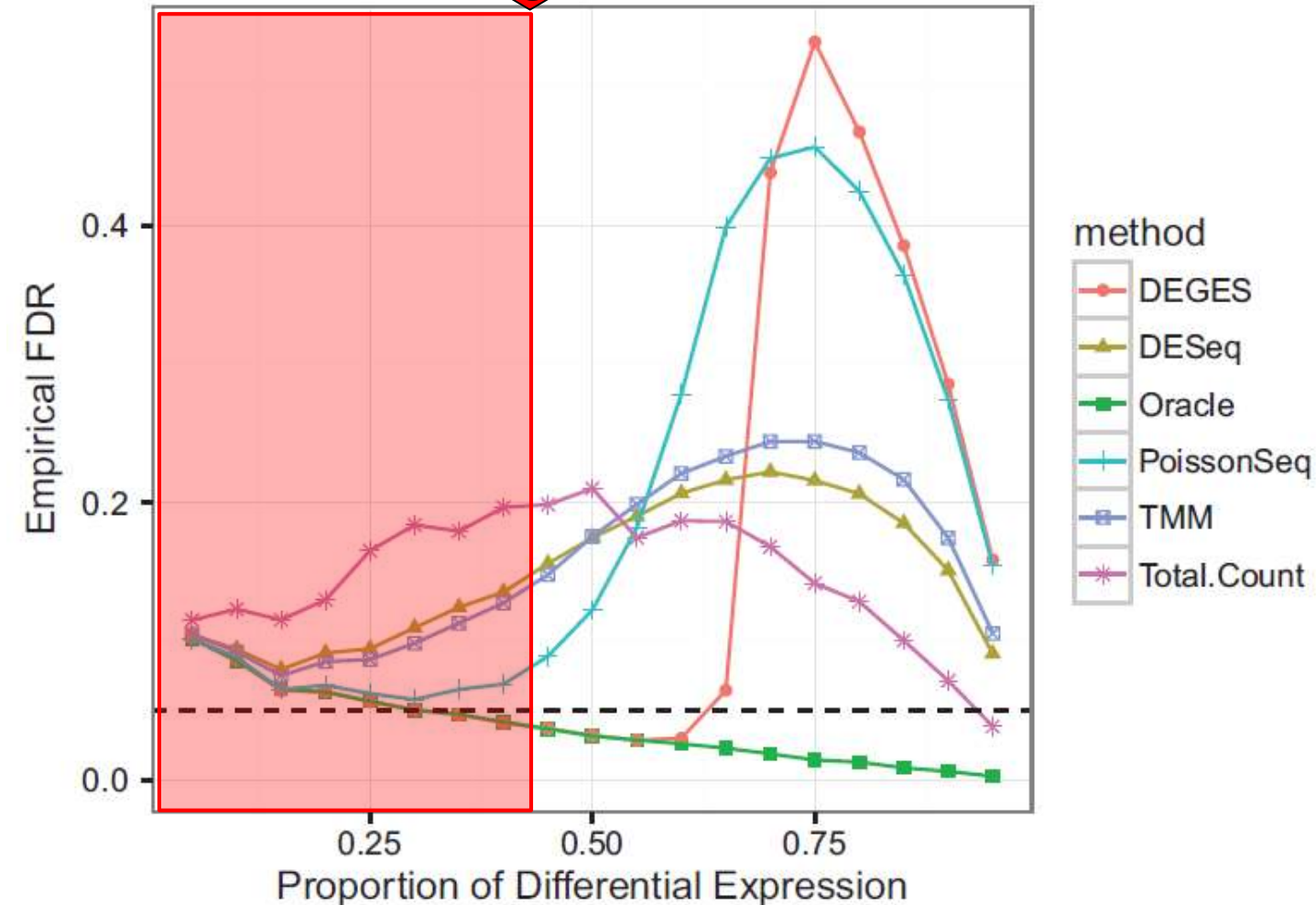


正規化法性能評価論文のFig. 8の右上の図

ノーフリーランチ定理

eFDR by proportion of DE
asymmetry different mRNA/cell

私の常識の範囲では、発現変動解析で
想定されるDEGの割合は、①せいぜい
40%程度まで。

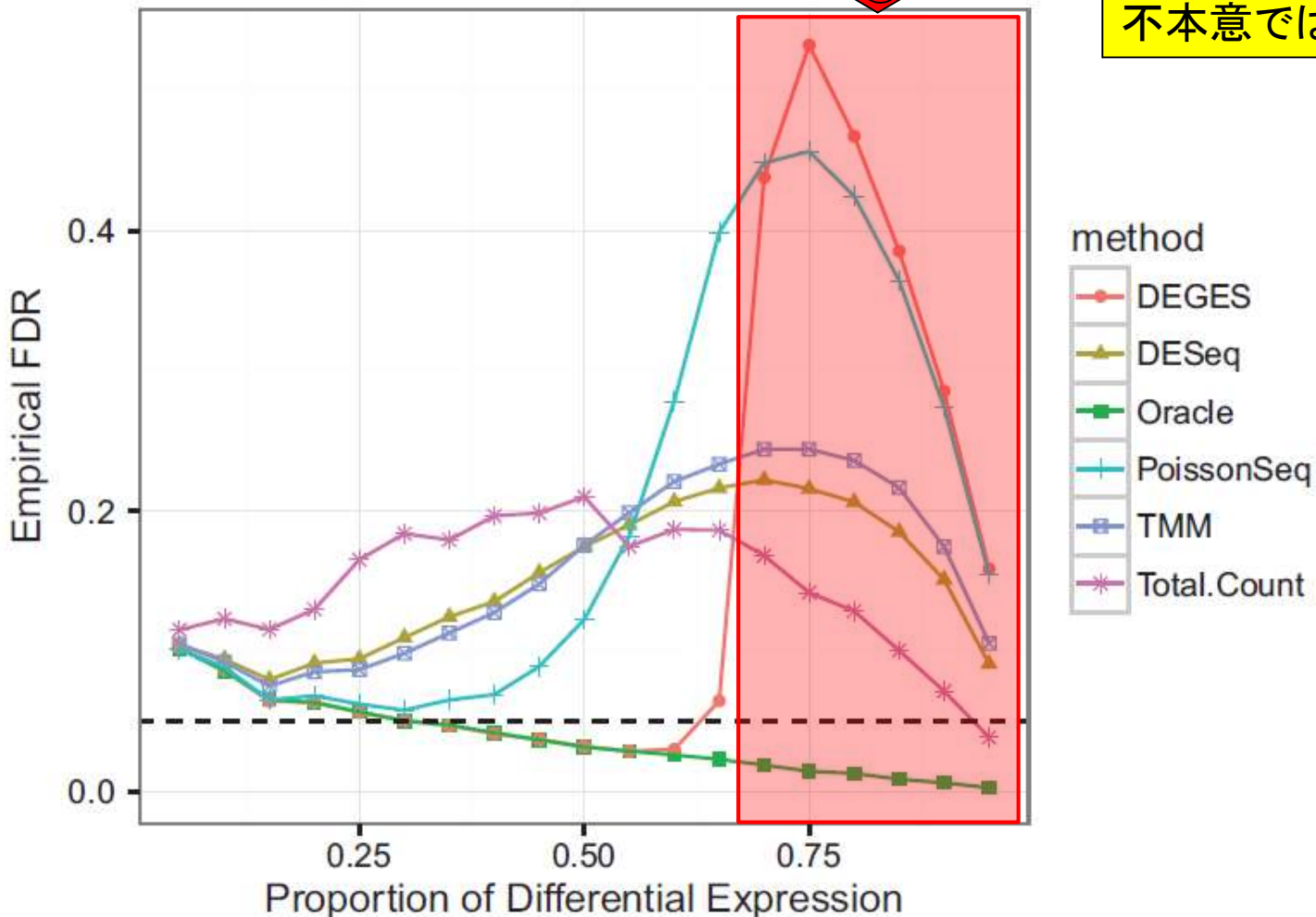


ノーフリーランチ定理

eFDR by proportion of DE
asymmetry, different mRNA/ #



私の常識の範囲では、発現変動解析で想定されるDEGの割合は、①せいぜい40%程度まで。それゆえ、②60%以上がDEGで、しかも片方の群に偏っているような条件まで含めて議論されるのは若干不本意ではある。

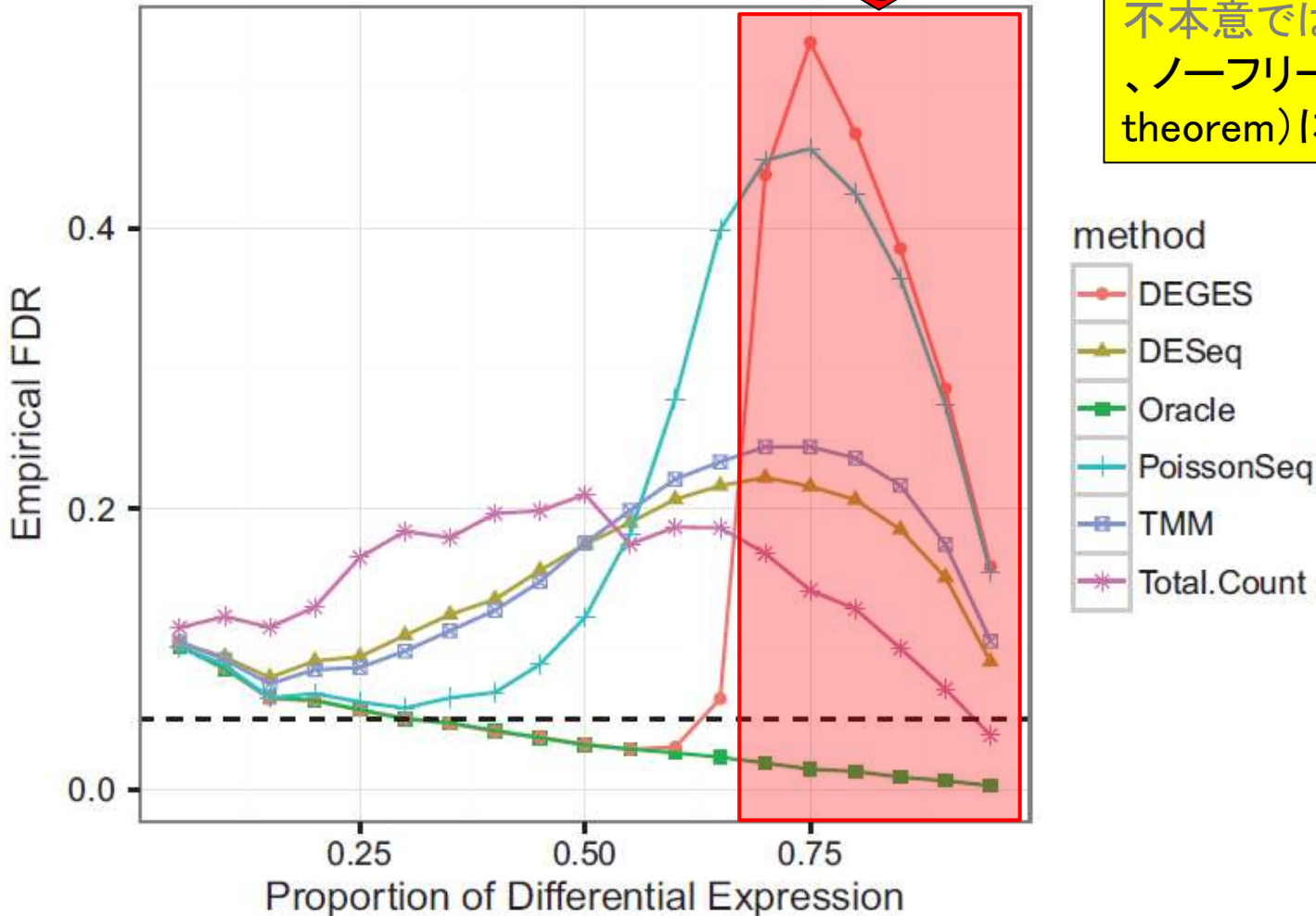


ノーフリーランチ定理

eFDR by proportion of DE
asymmetry, different mRNA/ #



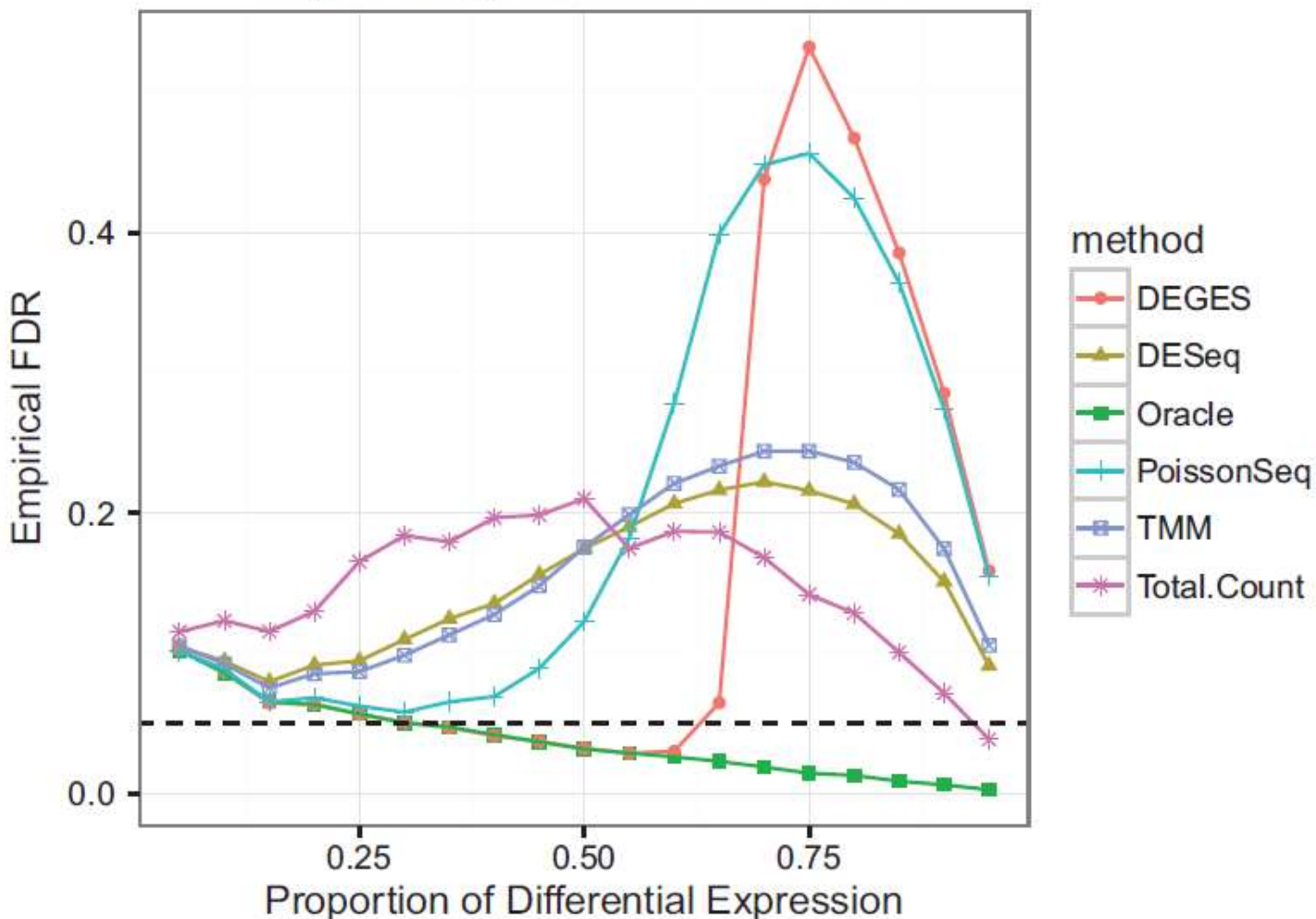
私の常識の範囲では、発現変動解析で想定されるDEGの割合は、①せいぜい40%程度まで。それゆえ、②60%以上がDEGで、しかも片方の群に偏っているような条件まで含めて議論されるのは若干不本意ではある。しかしこの結果自体は、ノーフリーランチ定理(no-free-lunch theorem)に合致するものであり妥当。



ノーフリーランチ定理

eFDR by proportion of DE
asymmetry, different mRNA/cell

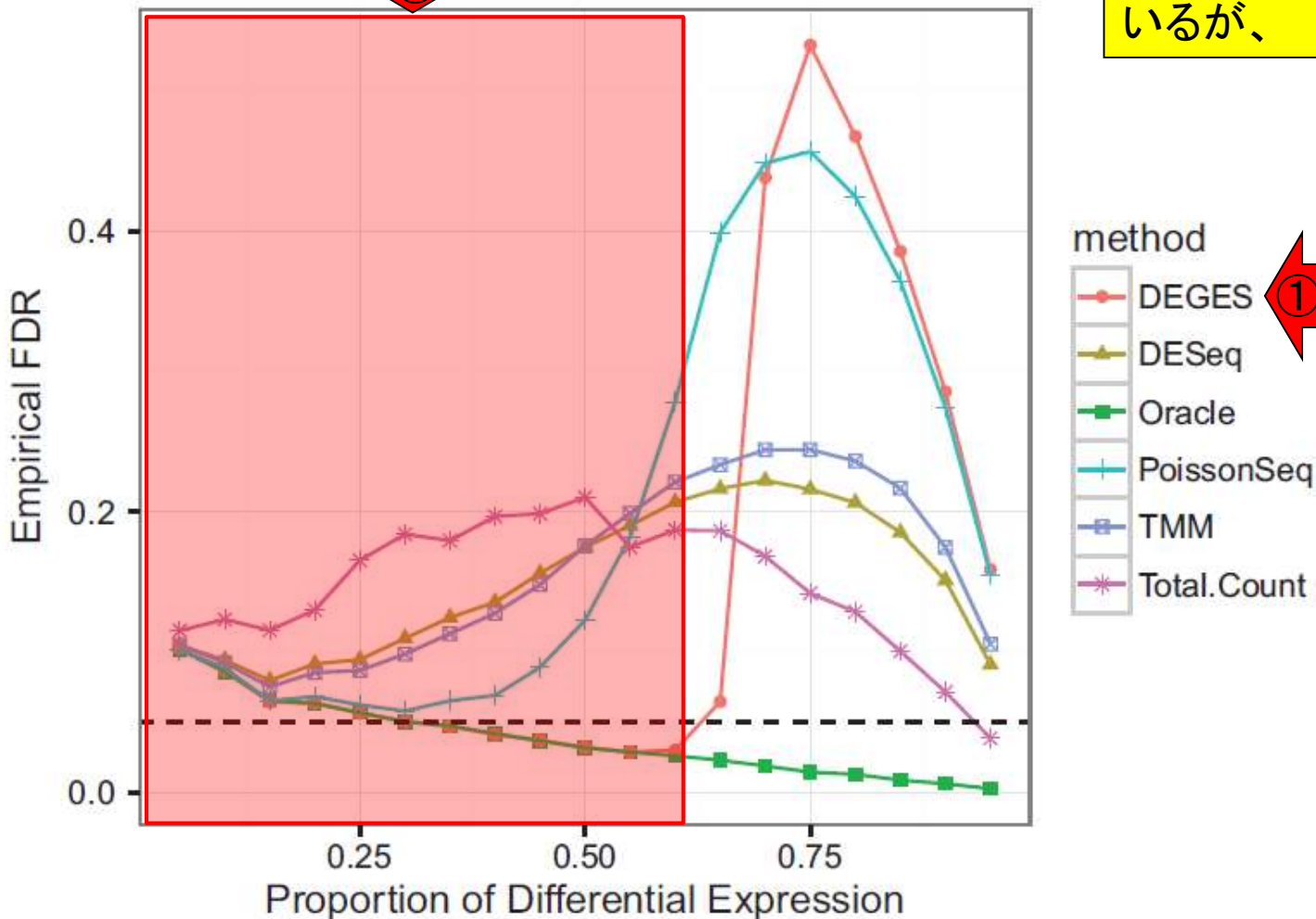
ノーフリーランチ定理をこの図で説明すると、「全ての横軸の範囲(DEGの割合)において性能のよい方法は存在しない」ということ。



ノーフリーランチ定理

eFDR by proportion of DE
asymmetry, different mRNA/cell

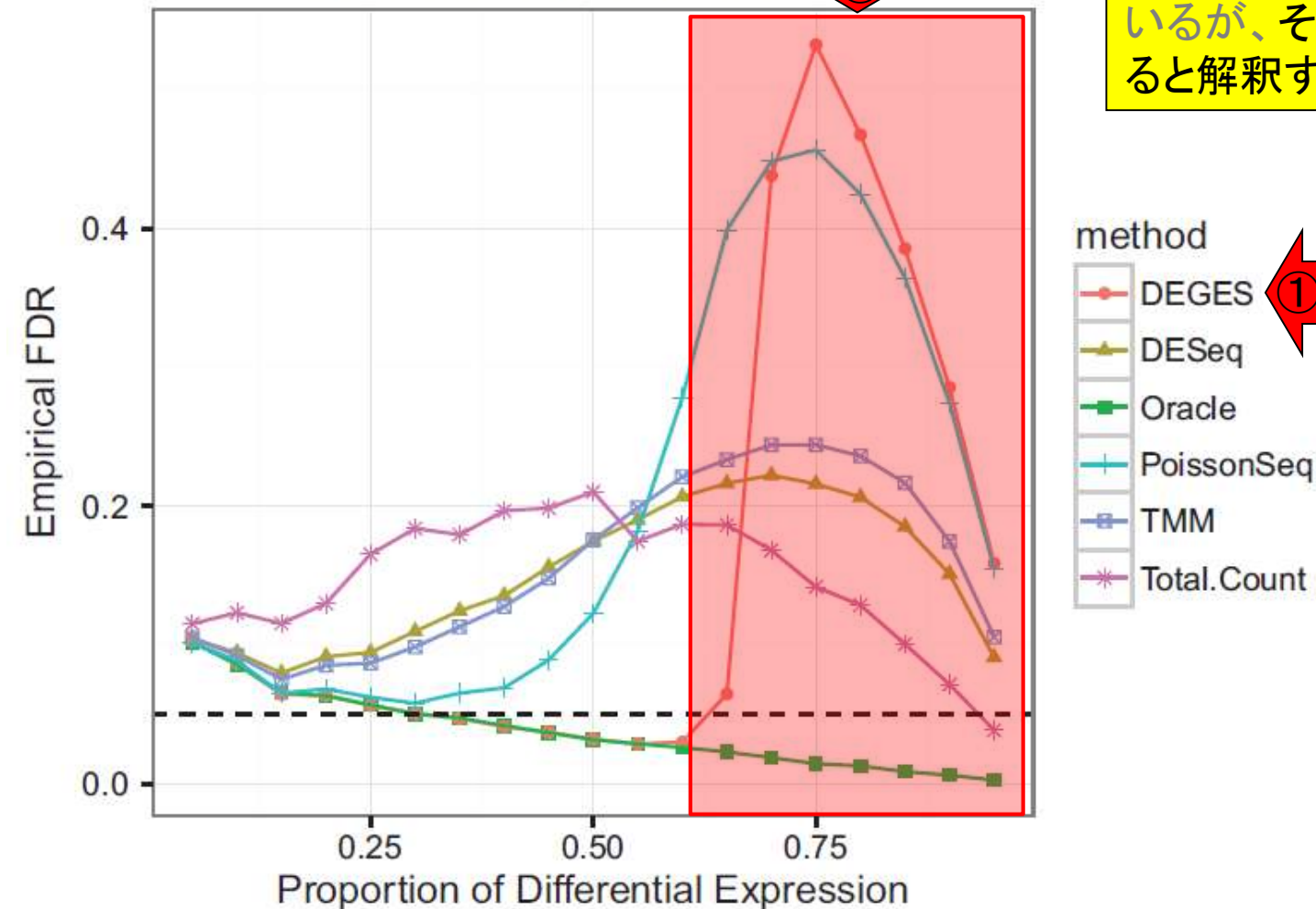
ノーフリーランチ定理をこの図で説明すると、「全ての横軸の範囲(DEGの割合)において性能のよい方法は存在しない」ということ。①DEGES(TCCのこと)は、②の範囲でほぼパーフェクトな性能を示しているが、



ノーフリーランチ定理

eFDR by proportion of DE
asymmetry, different mRNA

ノーフリーランチ定理をこの図で説明すると、「全ての横軸の範囲(DEGの割合)において性能のよい方法は存在しない」ということ。①DEGES(TCCのこと)は、②の範囲でほぼパーフェクトな性能を示しているが、その代償を③の範囲で払っていると解釈すればよい。

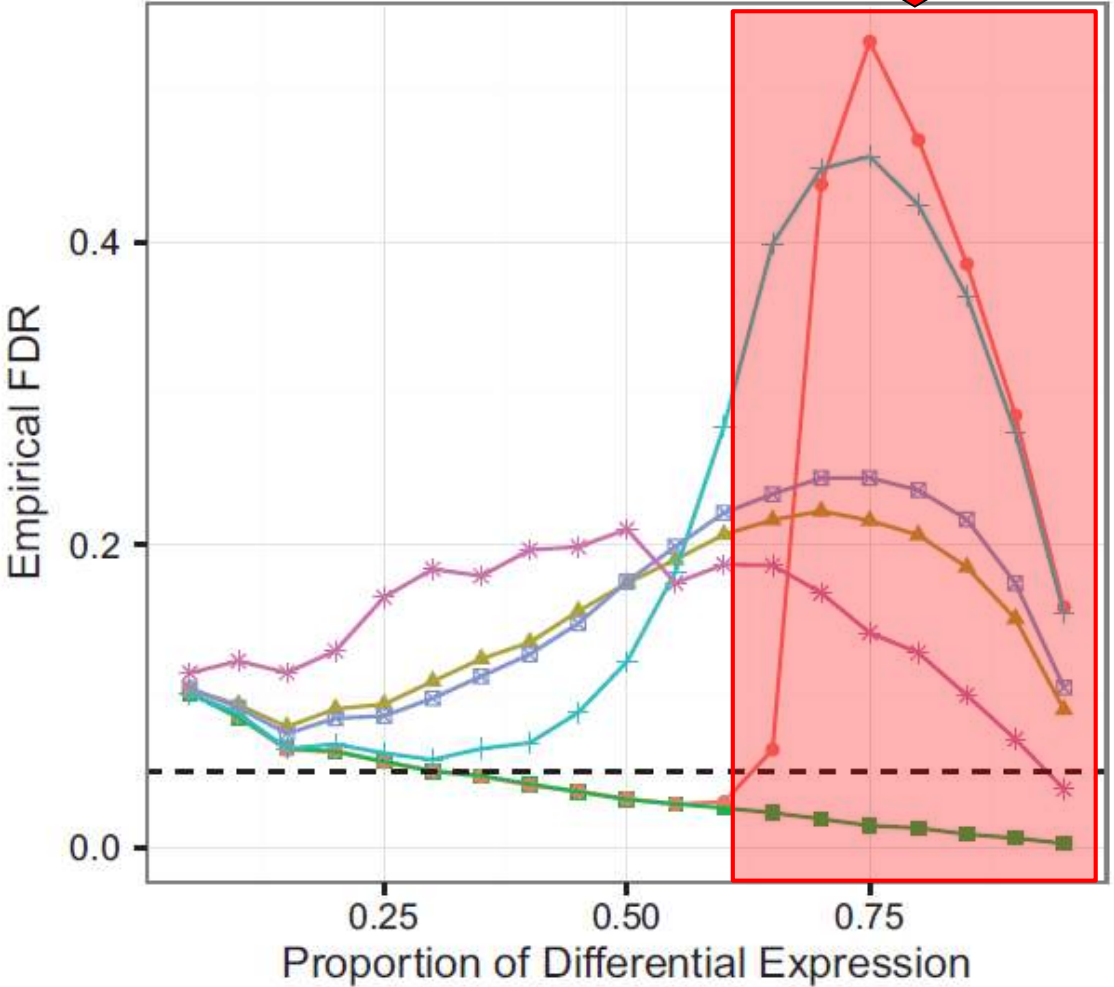


ノーフリーランチ定理

eFDR by proportion of DE
asymmetry, different mRNA

③

ノーフリーランチ定理をこの図で説明すると、「全ての横軸の範囲(DEGの割合)において性能のよい方法は存在しない」ということ。①DEGES(TCCのこと)は、②の範囲でほぼパーフェクトな性能を示しているが、その代償を③の範囲で払っていると解釈すればよい。実用上は、③DEGの割合が非常に高いTCC実行結果が得られたら、別パッケージの利用を推奨。



- method
- DEGES ①
 - DESeq
 - Oracle
 - PoissonSeq
 - TMM
 - Total.Count



Contents

- 自己紹介と東大アグリバイオの紹介
- トランスクリプトーム解析、発現解析、発現変動解析、実験デザイン
- 2群間比較: 実データ、TCC(反復増やすとDEG増える)
- 他グループによる性能評価論文(TCCが非推奨となる場合も!)
- TCCで3群間比較、baySeqも組み合わせて発現パターンまで得る
- (Rで)塩基配列解析
- Single-cell RNA-seq(scRNA-seq)
- バイオインフォマティクス実習

TCCで3群間比較

	A群			B群			C群		
	A1	A2	A3	B1	B2	B3	C1	C2	C2
gene_1	691	364	869	21	96	89	41	81	69
gene_2	11	83	125	7	0	1	1	4	7
gene_3	24	8	8	0	0	4	4	2	5
gene_4	34	5	9	0	0	0	0	4	0
gene_5	16	30	13	0	1	3	2	1	1
gene_6	0	0	2	0	0	0	0	0	1
gene_7	0	21	9	0	3	0	2	0	0
gene_8	639	472	462	54	55	31	16	39	37
gene_9	14	59	44	21	8	3	0	4	2

...

TCCで3群間比較

A群 vs. B群 vs. C群のようなデータの場合、TCCは、ANOVAのような「どこかの群間で発現変動している順にランキングできる結果」しか返しません。

	A群			B群			C群			q.value
	A1	A2	A3	B1	B2	B3	C1	C2	C2	
gene_1295	5415	5290	4941	315	419	397	310	328	328	5.56E-33
gene_1087	1859	2013	1375	106	81	66	65	108	96	5.56E-33
gene_554	7882	7549	8641	531	383	611	289	324	472	5.56E-33
gene_1676	732	571	891	#####	#####	#####	868	1016	1274	8.65E-33
gene_48	830	906	729	39	40	34	65	53	81	8.65E-33
gene_879	804	647	713	32	33	32	61	61	37	1.04E-32
gene_2335	118	112	135	1430	1440	1395	82	99	107	4.83E-30
gene_2692	180	171	169	149	136	188	2671	1772	2437	6.46E-30
gene_1138	1393	493	706	20	22	9	36	50	55	1.19E-29

...

TCCで3群間比較

A群 vs. B群 vs. C群のようなデータの場合、TCCは、ANOVAのような「どこかの群間で発現変動している順にランキングできる結果」しか返しません。例えば、①第1～3位はA群で高発現パターン、

	A群			B群			C群			q.value
	A1	A2	A3	B1	B2	B3	C1	C2	C2	
gene_1295	5415	5290	4941	315	419	397	310	328	328	5.56E-33
gene_1087	1859	2013	1375	106	81	66	65	108	96	5.56E-33
gene_554	7882	7549	8641	531	383	611	289	324	472	5.56E-33
gene_1676	732	571	891	####	####	####	868	1016	1274	8.65E-33
gene_48	830	906	729	39	40	34	65	53	81	8.65E-33
gene_879	804	647	713	32	33	32	61	61	37	1.04E-32
gene_2335	118	112	135	1430	1440	1395	82	99	107	4.83E-30
gene_2692	180	171	169	149	136	188	2671	1772	2437	6.46E-30
gene_1138	1393	493	706	20	22	9	36	50	55	1.19E-29



TCCで3群間比較

A群 vs. B群 vs. C群のようなデータの場合、TCCは、ANOVAのような「どこかの群間で発現変動している順にランキングできる結果」しか返しません。例えば、①第1～3位はA群で高発現パターン、②第4位はB群で高発現パターン、

	A群			B群			C群			q.value
	A1	A2	A3	B1	B2	B3	C1	C2	C2	
gene_1295	5415	5290	4941	315	419	397	310	328	328	5.56E-33
gene_1087	1859	2013	1375	106	81	66	65	108	96	5.56E-33
gene_554	7882	7549	8641	531	383	611	289	324	472	5.56E-33
gene_1676	732	571	891	####	####	####	868	1016	1274	8.65E-33
gene_48	830	906	729	39	40	34	65	53	81	8.65E-33
gene_879	804	647	713	32	33	32	61	61	37	1.04E-32
gene_2335	118	112	135	1430	1440	1395	82	99	107	4.83E-30
gene_2692	180	171	169	149	136	188	2671	1772	2437	6.46E-30
gene_1138	1393	493	706	20	22	9	36	50	55	1.19E-29



TCCで3群間比較

A群 vs. B群 vs. C群のようなデータの場合、TCCは、ANOVAのような「どこかの群間で発現変動している順にランキングできる結果」しか返しません。例えば、①第1～3位はA群で高発現パターン、②第4位はB群で高発現パターン、③その他はこんな感じ。

	A群			B群			C群			q.value	
	A1	A2	A3	B1	B2	B3	C1	C2	C2		
gene_1295	5415	5290	4941	315	419	397	310	328	328	5.56E-33	①
gene_1087	1859	2013	1375	106	81	66	65	108	96	5.56E-33	
gene_554	7882	7549	8641	531	383	611	289	324	472	5.56E-33	
gene_1676	732	571	891	####	####	####	868	1016	1274	8.65E-33	②
gene_48	830	906	729	39	40	34	65	53	81	8.65E-33	③
gene_879	804	647	713	32	33	32	61	61	37	1.04E-32	
gene_2335	118	112	135	1430	1440	1395	82	99	107	4.83E-30	
gene_2692	180	171	169	149	136	188	2671	1772	2437	6.46E-30	
gene_1138	1393	493	706	20	22	9	36	50	55	1.19E-29	

発現パターン情報も...

TCCは、①～③のような発現パターンを自動的に同定する機能を提供していないが...

	A群			B群			C群			q.value	
	A1	A2	A3	B1	B2	B3	C1	C2	C2		
gene_1295	5415	5290	4941	315	419	397	310	328	328	5.56E-33	①
gene_1087	1859	2013	1375	106	81	66	65	108	96	5.56E-33	
gene_554	7882	7549	8641	531	383	611	289	324	472	5.56E-33	
gene_1676	732	571	891	####	####	####	868	1016	1274	8.65E-33	②
gene_48	830	906	729	39	40	34	65	53	81	8.65E-33	③
gene_879	804	647	713	32	33	32	61	61	37	1.04E-32	
gene_2335	118	112	135	1430	1440	1395	82	99	107	4.83E-30	
gene_2692	180	171	169	149	136	188	2671	1772	2437	6.46E-30	
gene_1138	1393	493	706	20	22	9	36	50	55	1.19E-29	

発現パターン情報も...

TCCは、①～③のような発現パターンを自動的に同定する機能を提供していないが、④のような発現パターン分類結果も欲しい!

	A群			B群			C群			q.value	orderings
	A1	A2	A3	B1	B2	B3	C1	C2	C2		
gene_1295	5415	5290	4941	315	419	397	310	328	328	5.56E-33	A>other
gene_1087	1859	2013	1375	106	81	66	65	108	96	5.56E-33	A>other
gene_554	7882	7549	8641	531	383	611	289	324	472	5.56E-33	A>other
gene_1676	732	571	891	####	####	####	868	1016	1274	8.65E-33	B>other
gene_48	830	906	729	39	40	34	65	53	81	8.65E-33	A>other
gene_879	804	647	713	32	33	32	61	61	37	1.04E-32	A>other
gene_2335	118	112	135	1430	1440	1395	82	99	107	4.83E-30	B>other
gene_2692	180	171	169	149	136	188	2671	1772	2437	6.46E-30	C>other
gene_1138	1393	493	706	20	22	9	36	50	55	1.19E-29	A>other



お約束は事後検定だが

一般的によく行われる手順は、事後検定 (post-hoc test)。例えば、3通りの2群間比較 (A vs. B, A vs. C, and B vs. C) を行い、その結果に基づいて④のような結論を導くことは理論上は可能だが、現実には結構面倒。

	A群			B群			C群			q.value	orderings
	A1	A2	A3	B1	B2	B3	C1	C2	C2		
gene_1295	5415	5290	4941	315	419	397	310	328	328	5.56E-33	A>other
gene_1087	1859	2013	1375	106	81	66	65	108	96	5.56E-33	A>other
gene_554	7882	7549	8641	531	383	611	289	324	472	5.56E-33	A>other
gene_1676	732	571	891	#####	#####	#####	868	1016	1274	8.65E-33	B>other
gene_48	830	906	729	39	40	34	65	53	81	8.65E-33	A>other
gene_879	804	647	713	32	33	32	61	61	37	1.04E-32	A>other
gene_2335	118	112	135	1430	1440	1395	82	99	107	4.83E-30	B>other
gene_2692	180	171	169	149	136	188	2671	1772	2437	6.46E-30	C>other
gene_1138	1393	493	706	20	22	9	36	50	55	1.19E-29	A>other



Contents

- 自己紹介と東大アグリバイオの紹介
- トランスクリプトーム解析、発現解析、発現変動解析、実験デザイン
- 2群間比較: 実データ、TCC(反復増やすとDEG増える)
- 他グループによる性能評価論文(TCCが非推奨となる場合も!)
- TCCで3群間比較、baySeqも組み合わせて発現パターンまで得る
- (Rで)塩基配列解析
- Single-cell RNA-seq(scRNA-seq)
- バイオインフォマティクス実習

Osabe法

3群間比較で発現パターン分類まで行うための推奨解析パイプライン提唱論文。筆頭著者のOsabe氏の名前を冠して、Osabe法と勝手に命名。

Bioinform Biol Insights. 2019 Jul 8;13:1177932219860817. doi: 10.1177/1177932219860817. eCollection 2019.

Accurate Classification of Differential Expression Patterns in a Bayesian Framework With Robust Normalization for Multi-Group RNA-Seq Count Data.

Osabe T¹, Shimizu K^{1,2}, Kadota K^{1,2}.

Author information

Abstract

Empirical Bayes is a choice framework for differential expression (DE) analysis for multi-group RNA-seq count data. Its characteristic ability to compute posterior probabilities for predefined expression patterns allows users to assign the pattern with the highest value to the gene under consideration. However, current Bayesian methods such as baySeq and EBSeq can be improved, especially with respect to normalization. Two R packages (baySeq and EBSeq) with their default normalization settings and with other normalization methods (MRN and TCC) were compared using three-group simulation data and real count data. Our findings were as follows: (1) the Bayesian methods coupled with TCC normalization performed comparably or better than those with the default normalization settings under various simulation scenarios, (2) default DE pipelines provided in TCC that implements a generalized linear model framework was still superior to the Bayesian methods with TCC normalization when overall degree of DE was evaluated, and (3) baySeq with TCC was robust against different choices of possible expression patterns. In practice, we recommend using the default DE pipeline provided in TCC for obtaining overall gene ranking and then using the baySeq with TCC normalization for assigning the most plausible expression patterns to individual genes.

KEYWORDS: RNA-seq; differential expression analysis; empirical Bayes; expression patterns; normalization

PMID: 31312083 PMCID: [PMC6614939](#) DOI: [10.1177/1177932219860817](#)

Osabe法を一言でいえば...

	A1	A2	A3	B1	B2	B3	C1	C2	C2	q.value	orderings
gene_1295	5415	5290	4941	315	419	397	310	328	328	5.56E-33	A>other
gene_1087	1859	2013	1375	106	81	66	65	108	96	5.56E-33	A>other
gene_554	7882	7549	8641	531	383	611	289	324	472	5.56E-33	A>other
gene_1676	732	571	891	####	####	####	868	1016	1274	8.65E-33	B>other
gene_48	830	906	729	39	40	34	65	53	81	8.65E-33	A>other
gene_879	804	647	713	32	33	32	61	61	37	1.04E-32	A>other
gene_2335	118	112	135	1430	1440	1395	82	99	107	4.83E-30	B>other
gene_2692	180	171	169	149	136	188	2671	1772	2437	6.46E-30	C>other
gene_1138	1393	493	706	20	22	9	36	50	55	1.19E-29	A>other

Osabe法を一言でいえば

①が入力で、②のような結果を得るもの。
 ③は従来のTCCで得られる結果と同じ。
 ④発現パターン分類結果を独立に計算して付加したのがOsabe法。

	A1	A2	A3	B1	B2	B3	C1	C2	C2	q.value	orderings
gene_1295	5415	5290	4941	315	419	397	310	328	328	5.56E-33	A>other
gene_1087	1859	2013	1375	106	81	66	65	108	96	5.56E-33	A>other
gene_554	7882	7549	8641	531	383	611	289	324	472	5.56E-33	A>other
gene_1676	732	571	891	####	####	####	868	1016	1274	8.65E-33	B>other
gene_48	830	906	729	39	40	34	65	53	81	8.65E-33	A>other
gene_879	804	647	713	32	33	32	61	61	37	1.04E-32	A>other
gene_2335	118	112	135	1430	1440	1395	82	99	107	4.83E-30	B>other
gene_2692	180	171	169	149	136	188	2671	1772	2437	6.46E-30	C>other
gene_1138	1393	493	706	20	22	9	36	50	55	1.19E-29	A>other



TCC正規化+baySeq

①が入力で、②のような結果を得るもの。
 ③は従来のTCCで得られる結果と同じ。
 ④発現パターン分類結果を独立に計算して付加したのがOsabe法。④はTCCで得られた頑健な正規化係数を、baySeqという経験ベイズ系の発現変動解析用Rパッケージと組み合わせたものです。



	A1	A2	A3	B1	B2	B3	C1	C2	C2	q.value	orderings
gene_1295	5415	5290	4941	315	419	397	310	328	328	5.56E-33	A>other
gene_1087	1859	2013	1375	106	81	66	65	108	96	5.56E-33	A>other
gene_554	7882	7549	8641	531	383	611	289	324	472	5.56E-33	A>other
gene_1676	732	571	891	####	####	####	868	1016	1274	8.65E-33	B>other
gene_48	830	906	729	39	40	34	65	53	81	8.65E-33	A>other
gene_879	804	647	713	32	33	32	61	61	37	1.04E-32	A>other
gene_2335	118	112	135	1430	1440	1395	82	99	107	4.83E-30	B>other
gene_2692	180	171	169	149	136	188	2671	1772	2437	6.46E-30	C>other
gene_1138	1393	493	706	20	22	9	36	50	55	1.19E-29	A>other

TCC正規化+baySeq

baySeq実行時に、①計5つの発現パターンを考慮し、パターンごとの当てはまり度合い(事後確率)を調べています。そして、最も当てはまりのよいパターンを…



	A1	A2	A3	B1	B2	B3	C1	C2	C2	nonDEG	DEG_A	DEG_B	DEG_C	DEGall
gene_1295	5415	5290	4941	315	419	397	310	328	328	0.000	0.999	0.000	0.000	0.001
gene_1087	1859	2013	1375	106	81	66	65	108	96	0.000	0.996	0.000	0.000	0.004
gene_554	7882	7549	8641	531	383	611	289	324	472	0.000	0.997	0.000	0.000	0.003
gene_1676	732	571	891	#####	#####	#####	868	1016	1274	0.000	0.000	0.993	0.000	0.007
gene_48	830	906	729	39	40	34	65	53	81	0.000	0.845	0.000	0.000	0.155
gene_879	804	647	713	32	33	32	61	61	37	0.000	0.917	0.000	0.000	0.083
gene_2335	118	112	135	1430	1440	1395	82	99	107	0.000	0.000	0.975	0.000	0.025
gene_2692	180	171	169	149	136	188	2671	1772	2437	0.000	0.000	0.000	0.995	0.005
gene_1138	1393	493	706	20	22	9	36	50	55	0.000	0.613	0.002	0.001	0.384

TCC正規化+baySeq

baySeq実行時に、①計5つの発現パターンを考慮し、パターンごとの当てはまり度合い(事後確率)を調べています。そして、最も当てはまりのよいパターンを、こんな感じで同定し…



	A1	A2	A3	B1	B2	B3	C1	C2	C2	nonDEG	DEG_A	DEG_B	DEG_C	DEGall
gene_1295	5415	5290	4941	315	419	397	310	328	328	0.000	0.999	0.000	0.000	0.001
gene_1087	1859	2013	1375	106	81	66	65	108	96	0.000	0.996	0.000	0.000	0.004
gene_554	7882	7549	8641	531	383	611	289	324	472	0.000	0.997	0.000	0.000	0.003
gene_1676	732	571	891	####	####	####	868	1016	1274	0.000	0.000	0.993	0.000	0.007
gene_48	830	906	729	39	40	34	65	53	81	0.000	0.845	0.000	0.000	0.155
gene_879	804	647	713	32	33	32	61	61	37	0.000	0.917	0.000	0.000	0.083
gene_2335	118	112	135	1430	1440	1395	82	99	107	0.000	0.000	0.975	0.000	0.025
gene_2692	180	171	169	149	136	188	2671	1772	2437	0.000	0.000	0.000	0.995	0.005
gene_1138	1393	493	706	20	22	9	36	50	55	0.000	0.613	0.002	0.001	0.384

TCC正規化+baySeq

baySeq実行時に、①計5つの発現パターンを考慮し、パターンごとの当てはまり度合い(事後確率)を調べています。そして、最も当てはまりのよいパターンを、こんな感じで同定し、②こんな感じで出力しています。

	A1	A2	A3	B1	B2	B3	C1	C2	C2	nonDEG	DEG_A	DEG_B	DEG_C	DEGall	pattern
gene_1295	5415	5290	4941	315	419	397	310	328	328	0.000	0.999	0.000	0.000	0.001	DEG_A
gene_1087	1859	2013	1375	106	81	66	65	108	96	0.000	0.996	0.000	0.000	0.004	DEG_A
gene_554	7882	7549	8641	531	383	611	289	324	472	0.000	0.997	0.000	0.000	0.003	DEG_A
gene_1676	732	571	891	#####	#####	#####	868	1016	1274	0.000	0.000	0.993	0.000	0.007	DEG_B
gene_48	830	906	729	39	40	34	65	53	81	0.000	0.845	0.000	0.000	0.155	DEG_A
gene_879	804	647	713	32	33	32	61	61	37	0.000	0.917	0.000	0.000	0.083	DEG_A
gene_2335	118	112	135	1430	1440	1395	82	99	107	0.000	0.000	0.975	0.000	0.025	DEG_B
gene_2692	180	171	169	149	136	188	2671	1772	2437	0.000	0.000	0.000	0.995	0.005	DEG_C
gene_1138	1393	493	706	20	22	9	36	50	55	0.000	0.613	0.002	0.001	0.384	DEG_A

TCC正規化+baySeq

しかし、例えば①は単にA群で発現変動しているということを意味しているにすぎず、A群で高発現なのか低発現なのかまでは示していません。

	A1	A2	A3	B1	B2	B3	C1	C2	C2	nonDEG	DEG_A	DEG_B	DEG_C	DEGall	pattern
gene_1295	5415	5290	4941	315	419	397	310	328	328	0.000	0.999	0.000	0.000	0.001	DEG_A
gene_1087	1859	2013	1375	106	81	66	65	108	96	0.000	0.996	0.000	0.000	0.004	DEG_A
gene_554	7882	7549	8641	531	383	611	289	324	472	0.000	0.997	0.000	0.000	0.003	DEG_A
gene_1676	732	571	891	#####	#####	#####	868	1016	1274	0.000	0.000	0.993	0.000	0.007	DEG_B
gene_48	830	906	729	39	40	34	65	53	81	0.000	0.845	0.000	0.000	0.155	DEG_A
gene_879	804	647	713	32	33	32	61	61	37	0.000	0.917	0.000	0.000	0.083	DEG_A
gene_2335	118	112	135	1430	1440	1395	82	99	107	0.000	0.000	0.975	0.000	0.025	DEG_B
gene_2692	180	171	169	149	136	188	2671	1772	2437	0.000	0.000	0.000	0.995	0.005	DEG_C
gene_1138	1393	493	706	20	22	9	36	50	55	0.000	0.613	0.002	0.001	0.384	DEG_A

①

TCC正規化+baySeq

しかし、例えば①は単にA群で発現変動しているということを意味しているにすぎず、A群で高発現なのか低発現なのかまでは示していません。baySeqは②の大小関係の情報まで出力してくれるので…



	A1	A2	A3	B1	B2	B3	C1	C2	C2	nonDEG	DEG_A	DEG_B	DEG_C	DEGall	pattern	orderings
gene_1295	5415	5290	4941	315	419	397	310	328	328	0.000	0.999	0.000	0.000	0.001	DEG_A	A>other
gene_1087	1859	2013	1375	106	81	66	65	108	96	0.000	0.996	0.000	0.000	0.004	DEG_A	A>other
gene_554	7882	7549	8641	531	383	611	289	324	472	0.000	0.997	0.000	0.000	0.003	DEG_A	A>other
gene_1676	732	571	891	#####	#####	#####	868	1016	1274	0.000	0.000	0.993	0.000	0.007	DEG_B	B>other
gene_48	830	906	729	39	40	34	65	53	81	0.000	0.845	0.000	0.000	0.155	DEG_A	A>other
gene_879	804	647	713	32	33	32	61	61	37	0.000	0.917	0.000	0.000	0.083	DEG_A	A>other
gene_2335	118	112	135	1430	1440	1395	82	99	107	0.000	0.000	0.975	0.000	0.025	DEG_B	B>other
gene_2692	180	171	169	149	136	188	2671	1772	2437	0.000	0.000	0.000	0.995	0.005	DEG_C	C>other
gene_1138	1393	493	706	20	22	9	36	50	55	0.000	0.613	0.002	0.001	0.384	DEG_A	A>other



TCC正規化+baySeq

しかし、例えば①は単にA群で発現変動しているということを意味しているにすぎず、A群で高発現なのか低発現なのかまでは示していません。baySeqは②の大小関係の情報まで出力してくれるので、どのパターンがいくつあったのかをより詳細に知ることができます。



	A1	A2	A3	B1	B2	B3	C1	C2	C2	nonDEG	DEG_A	DEG_B	DEG_C	DEGall	pattern	orderings
gene_1295	5415	5290	4941	315	419	397	310	328	328	0.000	0.999	0.000	0.000	0.001	DEG_A	A>other
gene_1087	1859	2013	1375											0.004	DEG_A	A>other
gene_554	7882	7549	8641											0.003	DEG_A	A>other
gene_1676	732	571	891	#										0.007	DEG_B	B>other
gene_48	830	906	729											0.155	DEG_A	A>other
gene_879	804	647	713											0.083	DEG_A	A>other
gene_2335	118	112	135	1										0.025	DEG_B	B>other
gene_2692	180	171	169											0.005	DEG_C	C>other
gene_1138	1393	493	706											0.384	DEG_A	A>other

```

R Console
> table(out$MAP)

  DEG_A  DEG_B  DEG_C  DEGall  nonDEG
1293    725    475     14    7493

> table(orderings)
orderings
      A>B>C  A>C>B  A>other  B>C>A  B>other
7493         3         4    1245         2         689
C>A>B  C>B>A  C>other  other>A  other>B  other>C
      1         4        458        48        36        17

> |
    
```



DEGallの説明



	A1	A2	A3	B1	B2	B3	C1	C2	C2	nonDEG	DEG_A	DEG_B	DEG_C	DEGall
gene_1295	5415	5290	4941	315	419	397	310	328	328	0.000	0.999	0.000	0.000	0.001
gene_1087	1859	2013	1375	106	81	66	65	108	96	0.000	0.996	0.000	0.000	0.004
gene_554	7882	7549	8641	531	383	611	289	324	472	0.000	0.997	0.000	0.000	0.003
gene_1676	732	571	891	####	####	####	868	1016	1274	0.000	0.000	0.993	0.000	0.007
gene_48	830	906	729	39	40	34	65	53	81	0.000	0.845	0.000	0.000	0.155
gene_879	804	647	713	32	33	32	61	61	37	0.000	0.917	0.000	0.000	0.083
gene_2335	118	112	135	1430	1440	1395	82	99	107	0.000	0.000	0.975	0.000	0.025
gene_2692	180	171	169	149	136	188	2671	1772	2437	0.000	0.000	0.000	0.995	0.005
gene_1138	1393	493	706	20	22	9	36	50	55	0.000	0.613	0.002	0.001	0.384

DEGallの説明

ちなみに、①DEGallは、全ての群間で発現変動しているパターン。②の事後確率がまあまあ高い理由は、この発現パターン(A >> C > B)を見れば納得できる。

	A1	A2	A3	B1	B2	B3	C1	C2	C2	nonDEG	DEG_A	DEG_B	DEG_C	DEGall
gene_1295	5415	5290	4941	315	419	397	310	328	328	0.000	0.999	0.000	0.000	0.001
gene_1087	1859	2013	1375	106	81	66	65	108	96	0.000	0.996	0.000	0.000	0.004
gene_554	7882	7549	8641	531	383	611	289	324	472	0.000	0.997	0.000	0.000	0.003
gene_1676	732	571	891	####	####	####	868	1016	1274	0.000	0.000	0.993	0.000	0.007
gene_48	830	906	729	39	40	34	65	53	81	0.000	0.845	0.000	0.000	0.155
gene_879	804	647	713	32	33	32	61	61	37	0.000	0.917	0.000	0.000	0.083
gene_2335	118	112	135	1430	1440	1395	82	99	107	0.000	0.000	0.975	0.000	0.025
gene_2692	180	171	169	149	136	188	2671	1772	2437	0.000	0.000	0.000	0.995	0.005
gene_1138	1393	493	706	20	22	9	36	50	55	0.000	0.613	0.002	0.001	0.384



Contents

- 自己紹介と東大アグリバイオの紹介
- トランスクリプトーム解析、発現解析、発現変動解析、実験デザイン
- 2群間比較: 実データ、TCC(反復増やすとDEG増える)
- 他グループによる性能評価論文(TCCが非推奨となる場合も!)
- TCCで3群間比較、baySeqも組み合わせて発現パターンまで得る
- (Rで)塩基配列解析
- Single-cell RNA-seq(scRNA-seq)
- バイオインフォマティクス実習

Osabe法の利用手段

3群間比較で発現パターン分類まで行うための推奨解析パイプライン(Osabe法)の利用手段。

Bioinform Biol Insights. 2019 Jul 8;13:1177932219860817. doi: 10.1177/1177932219860817. eCollection 2019.

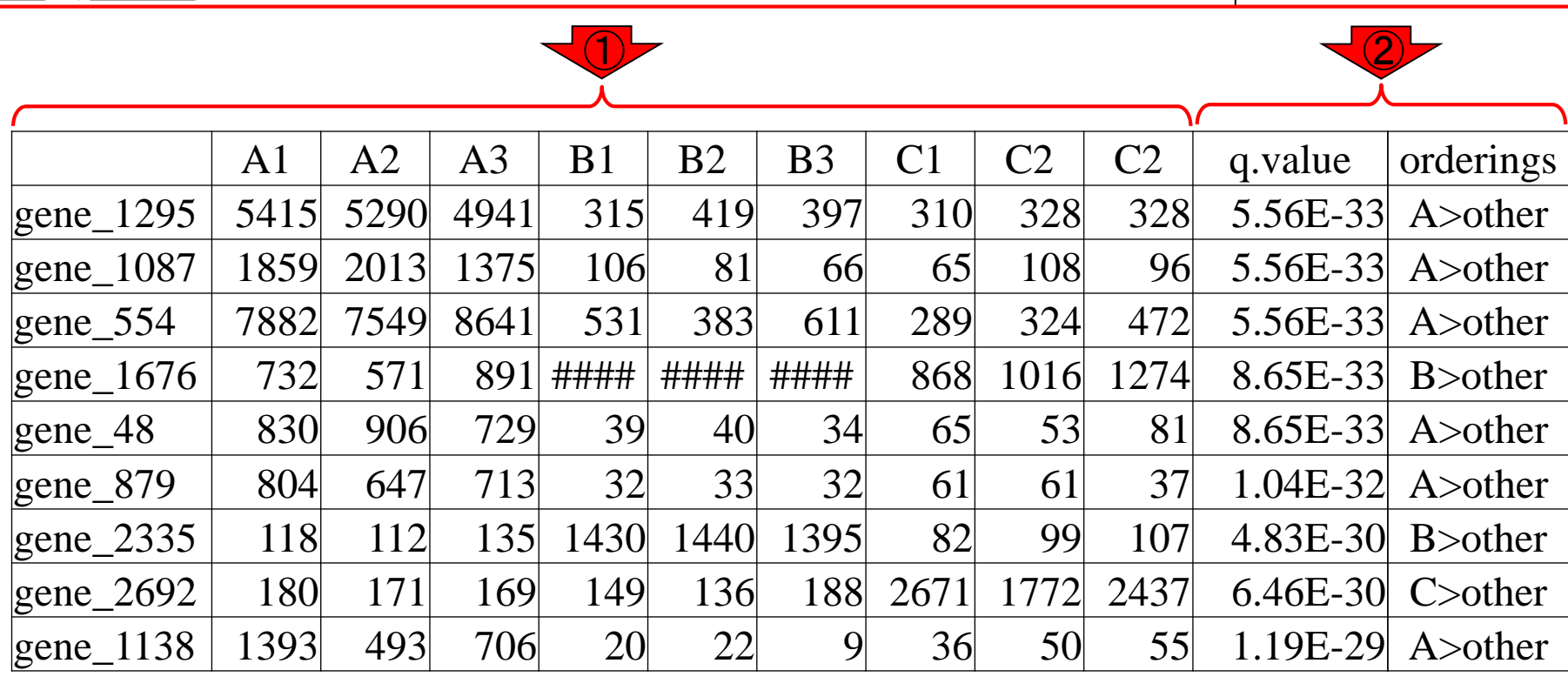
Accurate Classification of Differential Expression Patterns in a Bayesian Framework With Robust Normalization for Multi-Group RNA-Seq Count Data.

Osabe T¹, Shimizu K^{1,2}, Kadota K^{1,2}.

Author info

Abstract

Empirical Bayes data. Its character to assign the p methods such packages (bay (MRN and TC follows: (1) the with the default TCC that impl TCC normaliz different choic provided in TC assigning the



	A1	A2	A3	B1	B2	B3	C1	C2	C2	q.value	orderings
gene_1295	5415	5290	4941	315	419	397	310	328	328	5.56E-33	A>other
gene_1087	1859	2013	1375	106	81	66	65	108	96	5.56E-33	A>other
gene_554	7882	7549	8641	531	383	611	289	324	472	5.56E-33	A>other
gene_1676	732	571	891	#####	#####	#####	868	1016	1274	8.65E-33	B>other
gene_48	830	906	729	39	40	34	65	53	81	8.65E-33	A>other
gene_879	804	647	713	32	33	32	61	61	37	1.04E-32	A>other
gene_2335	118	112	135	1430	1440	1395	82	99	107	4.83E-30	B>other
gene_2692	180	171	169	149	136	188	2671	1772	2437	6.46E-30	C>other
gene_1138	1393	493	706	20	22	9	36	50	55	1.19E-29	A>other

KEYWORDS: RNA-seq; differential expression analysis; empirical Bayes; expression patterns; normalization

PMID: 31312083 PMCID: [PMC6614939](#) DOI: [10.1177/1177932219860817](#)

(Rで)塩基配列解析



(Rで)塩基配列解析 (last modified 2019/07/19, since 2010)

このウェブページのR関連部分は、[インストール | についての推奨手順 \(Windows2018.11.15版と Macintosh2018.11.27版\)](#)に従ってフリーソフトRと必要なパッケージをインストール済みであるという前提で記述しています。初心者の方は[基本的な利用法 \(Windows2019.03.12版と Macintosh2019.03.12版\)](#)で自習してください。2018年7月に[\(Rで\)塩基配列解析の一部 \(講習会・書籍・学会誌など\)](#)を切り分けて[サブページ](#)に移行しました。(2018/07/18)

What's new? ([過去のお知らせはこちら](#))

- 「解析 | 発現変動 | 3群間 | 対応なし | 複製なし | [TCC\(Sun_2013\)](#)」、および「解析 | 発現変動 | 2群間 | 対応なし | 複製なし | [TCC\(Sun_2013\)](#)」で内部的に利用していたオプションを "[トップページへ](#)" "deseq" に切り替えました。理由はDESeq2を使うとエラーが出るようになったからです(山本裕二郎)

(Rで)塩基配列解析

①ココで提供。沢山項目がありますが、
②ここでOsabe法を利用できます。

(Rで)塩基配列解析



(last modified

このウェブ
Macintosh2

提で記述し
で自習して
てサブペー

What's ne

• 「解析 |
群間 | 対
ら"desec

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html

- 解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 基礎 | [edgeR\(Robinson 2010\)](#) (last modified 2015/02/03)
- 解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 応用 | [TCC\(Sun 2013\)](#) (last modified 2016/05/31)推奨
- 解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 応用 | Blekhmanデータ | [TCC\(Sun 2013\)](#) (last modified 2018/06/18)
- 解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 応用 | [TCC+baySeq\(Osabe 2019\)](#) (last modified 2019/07/17)推奨 **NEW**
- 解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 応用 | [TCC+EBSeq\(Osabe 2019\)](#) (last modified 2019/07/10) **NEW**
- 解析 | 発現変動 | 3群間 | 対応なし | 複製なし | [DESeq2\(Love 2014\)](#) (last modified 2016/06/01)
- 解析 | 発現変動 | 3群間 | 対応なし | 複製なし | [TCC\(Sun 2013\)](#) (last modified 2019/07/11)推奨 **NEW**
- 解析 | 発現変動 | 5群間 | 対応なし | 複製あり | [TCC\(Sun 2013\)](#) (last modified 2015/11/05)推奨
- [解析 | 発現変動 | 時系列 | について](#) (last modified 2019/05/31)
- 解析 | 発現変動 | 時系列 | [maSigPro\(Nueda 2014\)](#) (last modified 2015/08/16)
- 解析 | 発現変動 | 時系列 | [Bayesian model-based clustering\(Nascimento 2012\)](#) (last modified 2012/09/10)
- [解析 | 発現変動 | exon/isoform | について](#) (last modified 2018/04/12)
- 解析 | 発現変動 | exon/isoform | [DEXseq\(Anders 2012\)](#) (last modified 2014/06/23)
- [解析 | 機能解析 | について](#) (last modified 2018/06/24)
- [解析 | 機能解析 | GMTファイル取得 | について](#) (last modified 2018/07/17)
- 解析 | 機能解析 | GMTファイル取得 | [EGSEAdata\(Alhamdoosh 2017\)](#) (last modified 2018/06/27)

[トップページへ](#)

Contents

- 自己紹介と東大アグリバイオの紹介
- トランスクリプトーム解析、発現解析、発現変動解析、実験デザイン
- 2群間比較: 実データ、TCC(反復増やすとDEG増える)
- 他グループによる性能評価論文(TCCが非推奨となる場合も!)
- TCCで3群間比較、baySeqも組み合わせて発現パターンまで得る
- (Rで)塩基配列解析
- Single-cell RNA-seq(scRNA-seq)
- バイオインフォマティクス実習

(Rで)塩基配列解析

最近は①や②でsingle-cell RNA-seq (scRNA-seq)に関する情報も提供しています。例えば、①でscRNA-seqの前処理の必要性を述べています。

(Rで)塩基配列解析

x +

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html

- 解析 | 前処理 | ID変換 | [Ensembl Gene ID中のバージョン情報を除去](#) (last modified 2018/08/08)
- 解析 | 前処理 | ID変換 | Ensembl Gene ID --> gene symbols | [基礎](#) (last modified 2018/08/15)
- 解析 | 前処理 | ID変換 | Ensembl Gene ID --> gene symbols | [RangedSummarizedExperiment](#) (last modified 2018/08/15)
- 解析 | 前処理 | [scRNA-seq](#) | [①](#) [について](#) (last modified 2019/06/26)
- 解析 | [クラスタリング](#) | [RNA-seq](#) | [②](#) [について](#) (last modified 2019/04/04)
- 解析 | [クラスタリング](#) | [RNA-seq](#) | サンプル間 | [hdust](#) (last modified 2015/02/26)
- 解析 | [クラスタリング](#) | [RNA-seq](#) | サンプル間 | [TCC\(Sun_2013\)](#) (last modified 2018/08/06)
- 解析 | [クラスタリング](#) | [RNA-seq](#) | 遺伝子間(基礎) | [MBCluster.Seq\(Si_2014\)](#) (last modified 2018/09/23)
- 解析 | [クラスタリング](#) | [RNA-seq](#) | 遺伝子間(応用) | [TCC正規化\(Sun_2013\)+MBCluster.Seq\(Si_2014\)](#) (last modified 2016/05/30)
- 解析 | [クラスタリング](#) | [scRNA-seq](#) | [②](#) [について](#) (last modified 2019/06/27)
- 解析 | [外れサンプル検出](#) | [について](#) (last modified 2019/03/28)
- 解析 | [発現変動](#) | [について](#)(2013年頃の記載事項で記念に残しているだけ) (last modified 2014/07/10)
- 解析 | [発現変動](#) | [について](#) (last modified 2019/05/24)
- 解析 | [発現変動](#) | [2群間](#) | [対応なし](#) | [について](#) (last modified 2016/10/07)
- 解析 | [発現変動](#) | [2群間](#) | [対応なし](#) | [複製あり](#) | [DESeq2\(Love_2014\)](#) (last modified 2015/11/15)
- 解析 | [発現変動](#) | [2群間](#) | [対応なし](#) | [複製あり](#) | [TCC\(Sun_2013\)](#) (last modified 2015/07/07)推奨

[トップページへ](#)

scRNA-seqと前処理

ざっくり言うと、scRNA-seqの発現行列データには、発現していないことを意味する0という数値が多く含まれる。これがサンプル間クラスタリングなどの解析結果に悪影響を与えるので、前処理が必要だということ。

(Rで)塩基配列解析

x

+

← → ↻ 🏠 ⓘ 保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r

解析 | 前処理 | scRNA-seq | について ①

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、多くの遺伝子のカウントがゼロ(0)という特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現している遺伝子が異なるため本当に発現していない」という生物学的な理由(biological cause)によるもの。そしてもう1つは「本当は発現しているんだけども捉えられていない」という技術的な理由(technical cause)によるものです。特に、後者の技術的な理由でゼロカウントになることを「ドロップアウト(dropout)」と言います([Zappia et al., Genome Biol., 2017](#))。データのほとんどがゼロでそれが検出されたりされなかったりするデータ(遺伝子)が多いのが特徴ですが、これがデータ解析時に悪さをするので前処理(preprocessing)が必要なのです。

しかし、一言で前処理と言っても、実際には多様な処理が含まれます。例えば、転写物レベルのカウントデータを遺伝子レベルにつぶす(collapseする)処理や、外部コントロールとして用いるspike-in転写物のカウント数がやたらと多い細胞のデータ(実験上のミスや細胞が死んでいるなどの理由による)の除去、ごく少数の遺伝子のみのカウント数とそのライブラリの総カウント数の占めるようなlow-complexity librariesの存在確認、低発現遺伝子(low-abundance genes)やドロップアウト率の高い遺伝子(genes with high dropout rate)のフィルタリングなどが挙げられます([McCarthy et al., Bioinformatics, 2017](#))。この他にも、反復データをうまく利用して真の発現レベル(true expression levels)を推定するプログラム(inferとかimputationとかdenoisingとかrecoverなどの単語を含むものが該当します)もこの範疇に含めてよいと思います。

R用:

- [scater](#)(通常用) : [McCarthy et al., Bioinformatics, 2017](#)
- [netSmooth](#)(imputation用) : [Ronen and Akalin, F1000Res., 2018](#)
- [scImpute](#)(imputation用) : [Li and Li, Nat Commun., 2018](#)
- [DrImpute](#)(imputation用) : [Gong et al., BMC Bioinformatics, 2018](#)

[トップページへ](#)

scRNA-seqと前処理

多数の方法が提案されている。評価基準のほとんどが、データの視覚化の良し悪しに関するものとなっている。

(Rで塩基配列解析

保護されていない通信

low-complexity librariesの存在は、高い遺伝子(genes with high dropout)の発現レベル(expression levels)を推定するプログラムを包含するのが該当しますもこの範疇

R用:

- [scater](#)(通常用) : [McCarthy et al.](#)
- [netSmooth](#)(imputation用) : [R](#)
- [scImpute](#)(imputation用) : [Li a](#)
- [DrImpute](#)(imputation用) : [Go](#)
- [SAVER](#)(imputation用) : [Huang](#)
- [MAGIC](#)(imputation用) : [van D](#)
- [LSImpute](#)(imputation用; RのS
- [DoubletFinder](#)(doublets同定用)

R以外:

- [BISCUIT](#)(imputation用; プロク
- [kNN-smoothing](#)(imputation用)
- [MAGIC](#)(imputation用) : [van D](#)
- [AutoImpute](#)(imputation用) : [T](#)

(Rで塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.ht...

R以外:

- [BISCUIT](#)(imputation用; プログラムは非公開?!): [Azizi et al., Genom Comput Biol., 2017](#)
- [kNN-smoothing](#)(imputation用) : [Wagner et al., bioRxiv, 2017](#)
- [MAGIC](#)(imputation用) : [van Dijk et al., Cell, 2018](#)
- [AutoImpute](#)(imputation用) : [Talwar et al., Sci Rep., 2018](#)
- [scVI](#)(通常用) : [Lopez et al., Nat Methods, 2018](#)
- [VASC](#)(imputation用) : [Wang and Gu, Genomics Proteomics Bioinformatics, 2018](#)
- [scScope](#)(imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#)(imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらでも変わる的な) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

[トップページへ](#)

scRNA-seqと前処理

多数の方法が提案されている。評価基準のほとんどが、データの視覚化の良し悪しに関するものとなっている。理由はシンプル。scRNA-seqは、未知の細胞種 (cell types) を発見する探索的な解析 (exploratory analysis) が中心だから。若干語弊あり。ある細胞種から別の細胞種に変化する軌跡を追うtrajectory解析などもある。

(Rで塩基配列解析

← → ↻ 🏠 ⓘ 保護されていない通信

low-complexity librariesの存在は、高い遺伝子 (genes with high dropout) (Bioinformatics, 2017)。この他にも、expression levels) を推定するプログラムを含むものが該当します) もこの範疇

R用:

- [scater](#) (通常用) : [McCarthy et al.](#)
- [netSmooth](#) (imputation用) : [R](#)
- [scImpute](#) (imputation用) : [Li a](#)
- [DrImpute](#) (imputation用) : [Go](#)
- [SAVER](#) (imputation用) : [Huang](#)
- [MAGIC](#) (imputation用) : [van D](#)
- [LSImpute](#) (imputation用; RのS
- [DoubletFinder](#) (doublets同定用)

R以外:

- [BISCUIT](#) (imputation用; プロク
- [kNN-smoothing](#) (imputation用)
- [MAGIC](#) (imputation用) : [van D](#)
- [AutoImpute](#) (imputation用) : [T](#)

(Rで塩基配列解析

← → ↻ 🏠 ⓘ 保護されていない通信

R以外:

- [BISCUIT](#) (imputation用; プログラ
- [kNN-smoothing](#) (imputation用) : [Wagner et al., bioRxiv, 2017](#)
- [MAGIC](#) (imputation用) : [van Dijk et al., Cell, 2018](#)
- [AutoImpute](#) (imputation用) : [Talwar et al., Sci Rep., 2018](#)
- [scVI](#) (通常用) : [Lopez et al., Nat Methods, 2018](#)
- [VASC](#) (imputation用) : [Wang and Gu, Genomics Proteomics Bioinformatics, 2018](#)
- [scScope](#) (imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#) (imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#) (imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#) (imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#) (doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- 手法比較 (SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- 手法比較 (Variational Autoencoder (VAE) 系はパラメータチューニング次第でいくらでも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

[トップページへ](#)

クラスタリングが重要

それゆえ、①前処理だけでなく、②クラスタリングもscRNA-seqデータ解析の主要部分となっている。

(Rで塩基配列解析

x +

← → ↻ 🏠 ⓘ 保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html

- 解析 | 前処理 | ID変換 | [Ensembl Gene ID中のバージョン情報を除去](#) (last modified 2018/08/08)
 - 解析 | 前処理 | ID変換 | Ensembl Gene ID --> gene symbols | [基礎](#) (last modified 2018/08/15)
 - 解析 | 前処理 | ID変換 | Ensembl Gene ID --> gene symbols | [RangedSummarizedExperiment](#) (last modified 2018/08/15)
 - [解析 | 前処理 | scRNA-seq | ① について](#) (last modified 2019/06/26)
 - [解析 | クラスタリング | RNA-seq | ② について](#) (last modified 2019/04/04)
 - 解析 | クラスタリング | RNA-seq | サンプル間 | [hdust](#) (last modified 2015/02/26)
 - 解析 | クラスタリング | RNA-seq | サンプル間 | [TCC\(Sun_2013\)](#) (last modified 2018/08/06)
 - 解析 | クラスタリング | RNA-seq | 遺伝子間(基礎) | [MBCluster.Seq\(Si_2014\)](#) (last modified 2018/09/23)
 - 解析 | クラスタリング | RNA-seq | 遺伝子間(応用) | [TCC正規化\(Sun_2013\)+MBCluster.Seq\(Si_2014\)](#) (last modified 2016/05/30)
 - [解析 | クラスタリング | scRNA-seq | ② について](#) (last modified 2019/06/27)
 - [解析 | 外れサンプル検出 | について](#) (last modified 2019/03/28)
 - [解析 | 発現変動 | について\(2013年頃の記載事項で記念に残しているだけ\)](#) (last modified 2014/07/10)
 - [解析 | 発現変動 | について](#) (last modified 2019/05/24)
 - [解析 | 発現変動 | 2群間 | 対応なし | について](#) (last modified 2016/10/07)
 - 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | [DESeq2\(Love_2014\)](#) (last modified 2015/11/15)
 - 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | [TCC\(Sun_2013\)](#) (last modified 2015/07/07)推奨
- [トップページへ](#)

クラスタリングが重要

それゆえ、①前処理だけでなく、②クラスタリングもscRNA-seqデータ解析の主要部分となっている。PCAやt-SNEなどを利用して2次元平面上にプロットし、既知の同一細胞種が同じクラスターに含まれるかなどが評価される。

解析 | クラスタリング | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)データ用です。 [scRNA-tools](#)はscRNA-seqデータ解析全般のツールのデータベースですが、この中にクラスタリングプログラムも含まれています。(scRNA-seqとの対比として)昔ながらのbulk RNA-seqでは、どのサンプルがどの群に属するかが既知なので 群間で発現が異なる遺伝子(Differentially Expressed Genes; DEG)の検出がcommon taskでした。しかしscRNA-seqでは、通常どのサンプルがどの群に属するかが不明なので、探索的な解析(Exploratory Analysis)が中心となります。つまりクラスタリングが重要だということです。そのため、多くのscRNA-seq用のクラスタリングプログラムは どのサンプルがどの群に属するかを割り当てることにフォーカスしています([Zappia et al., Genome Biol., 2017](#))。このやり方はサンプルが細胞の状態が変化しない成熟細胞(mature cells)の場合に有効です。しかし発生段階(developmental stages)では、幹細胞(stem cells)が成熟細胞へと分化します。よって、特定の群への割り当てというのは適切ではなく、ある細胞(one cell type)が別の種類の細胞へと連続的に変化していく軌跡(continuous trajectory)を並べる(ordering)ようなプログラムを利用する必要があります。bulk RNA-seqでも経時変化に特化したものと通常のクラスタリングに分けられるようなものだと思えば納得しやすいでしょう。また、クラスタリングの際に重要なのは、効果的な次元削減(dimensionality reduction)です。例えば、主成分分析(PCA)だと数万遺伝子(つまり数万次元)のデータを2次元や3次元に削減した状態でデータを表示させています。このPCA (Rのprcomp関数を実行することと同じ)もまた、次元削減法の1つと捉えることができます。scRNA-seqデータに特化したのが、t-SNE, ZIFA, CIDRのような以下にリストアップされているものたちです。

R用:

- [PHATE\(通常用\)](#) : [Moon et al., bioRxiv, 2018](#)
- [Monocle\(trajjectory用\)](#) : [Trapnell et al., Nat Biotechnol., 2014](#)
- [Seurat\(通常用\)](#) : [Satija et al., Nat Biotechnol., 2015](#)

[トップページへ](#)

クラスタリングが重要

それゆえ、①前処理だけでなく、②クラスタリングもscRNA-seqデータ解析の主要部分となっている。PCAやt-SNEなどを利用して2次元平面上にプロットし、既知の同一細胞種が同じクラスターに含まれるかなどが評価される。①Reviewや②ガイドラインの論文もいくつかあります。

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#al

解析 | クラスタリング | scRNA-seq | について

single-cell RNA-seq (scRNA-seq) データベースですが、この中にクラスタリングは含まれていません。bulk RNA-seqでは、どのサンプルにどの遺伝子が発現しているか (Expressed Genes; DEG) の検出は比較的容易ですが、どの細胞に属するかが不明なので、探索的クラスタリングが必要だということです。そのため、多様な細胞状態を割り当てることにフォーカスしたクラスタリングアルゴリズムが細胞の状態が変化しない成熟段階 (mature stages) では、幹細胞 (stem cells) ではなく、ある細胞 (one cell type) を識別する (ordering) ようなプログラムが有用です。通常のクラスタリングに分けられないような細胞群を識別するために、重要なのは、効果的な次元削減 (dimensionality reduction) (つまり数万次元) のデータを2次元や3次元に削減することと同じ) もまた、t-SNE, ZIFA, CIDR のような

R用:

- PHATE(通常用) : [Moon et al., Nat Rev Genet., 2019](#)
- Monocle(trajjectory用) : [Chen et al., Nat Commun., 2019](#)
- Seurat(通常用) : [Satija et al., Nat Rev Genet., 2019](#)

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#abo...

- PAGA(両方) : [Wolf et al., Genome Biol., 2019](#)
- STREAM(trajjectory用) : [Chen et al., Nat Commun., 2019](#)
- VPAC(?用) : [Chen et al., BMC Bioinformatics, 2019](#)

Review、ガイドライン、パイプライン系:

- Granatum(パイプライン) : [Zhu et al., Genome Med., 2017](#)
- Review : [Menon V, Brief Funct Genomics., 2018](#)
- scRNA-tools(ツールのデータベース) : [Zappia et al., PLoS Comput Biol., 2018](#)
- 手法比較(Seuratがよい。ribosomal genesのような高発現遺伝子はクラスタリング時に誇張された影響を与えているかも?) : [Freytag et al., F1000Res., 2018](#)
- DuoClustering2018(ベンチマーク) : [Duò et al., F1000Res., 2018](#)
- ① Review : [Kiselev et al., Nat Rev Genet., 2019](#)
- Review(scRNA-seq全般の話; クラスタリングは一部のみ) : [Chen et al., Front Genet., 2019](#)
- ② ガイドライン(best practices) : [Luecken and Theis, Mol Syst Biol., 2019](#)

[トップページへ](#)

前処理法と評価基準

①前処理法(imputation法)を、②visualization以外の様々な評価基準で比較した論文もある。この論文では、例えば③DCA(オートエンコーダを用いてノイズを除去する方法)は、本来補正すべきでないものに対しても補正する傾向にある、などということが述べられている。

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#al

解析 | 前処理 | scRNA-seq | について ①

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、いう特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現している遺伝子が異なるため本当に発現していない」という生物学的な理由(biological cause)によるもの。そしてもう1つは「本当は発現しているんだけども捉えられていない」という技術的な理由(technical cause)によるもの。これを「ドロップアウト(dropout)」と言います。ドロップアウトは、遺伝子発現が検出されたりされなかったりするデータ(遺伝子発現データの前処理(preprocessing)が必要なのです。

しかし、一言で前処理と言っても、実際には、レベルにつぶす(collapseする)処理や、外部コタ(実験上のミスや細胞が死んでいるなどの理由)でカウント数の占めるようなlow-complexity libraryの生成、高発現遺伝子(genes with high dropout rate)の除去(2017)。その他にも、反復データをうまく利用してimputationとかdenoisingとかrecoverなどの

R用:

- [scater](#)(通常用) : [McCarthy et al., Bioinformatics, 2016](#)
- [netSmooth](#)(imputation用) : [Ronen et al., Nat Commun., 2019](#)
- [scImpute](#)(imputation用) : [Li and Li, Nat Commun., 2018](#)
- [DrImpute](#)(imputation用) : [Gong et al., Nat Commun., 2019](#)

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_s...

- [scScope](#)(imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#)(imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらかでも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

[トップページへ](#)

前処理法と評価基準

①この論文では、(全データに対して smoothing をかける②DCAなどと違って) ゼロカウントデータに対してのみ smoothing をかける③SAVERがよい(例: Fig. 1A)という結論だが... そうこうしているうちに④SAVER-Xという次期バージョンも出ている。

(Rで塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#al

解析 | 前処理 | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、多くの遺伝子のカウントがゼロ(0)という特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現している遺伝子が異なるため本当に発現していない」という生物学的な理由(biological cause)によるもの。そしてもう1つは「本当は発現しているんだけども捉えられていない」という技術的な理由(technical cause)によるもの。これを「ドロップアウト(dropout)」と言います。ドロップアウトは、遺伝子の発現が検出されたりされなかったりするデータ (遺伝子発現データの前処理(preprocessing)が必要なのです。

しかし、一言で前処理と言っても、実際には、ノイズを除去する処理や、外部コ変数(実験上のミスや細胞が死んでいるなどの理由)による変動を補正する処理、低複雑性ライブラリ(Low-complexity library)の除去、高発現遺伝子(genes with high dropout rate)の除去(2017)。その他にも、反復データをうまく利用してimputationとかdenoisingとかrecoverなどの

R用:

- [scater](#)(通常用) : [McCarthy et al., Bioinformatics, 2016](#)
- [netSmooth](#)(imputation用) : [Ronen et al., Nat Commun., 2019](#)
- [scImpute](#)(imputation用) : [Li and Li, Nat Commun., 2018](#)
- [DrImpute](#)(imputation用) : [Gong et al., Nat Commun., 2019](#)

(Rで塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_s...

- [scScope](#)(imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#)(imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- ① 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- ② 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらでも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

[トップページへ](#)

前処理法と評価基準

①この論文の中では、Splatterというプログラムを用いてシミュレーションデータを生成し、発現変動遺伝子の検出精度を調べている。結論としては、imputation(前処理)を行わないraw count dataが最もよかった(Fig. 2E)。Imputationを行ったデータの中では、②SAVERが最もよいパフォーマンスだった。

(Rで塩基配列解析)

x +

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#al

解析 | 前処理 | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、いう特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現している遺伝子が異なるため本来に発現していない」という生物学的な理由(biological cause)によるもの、そしてもう1つは「本当は発現しているんだけども捉えられていない」という技術的な理由(technical cause)によるもの、この2つを「ドロップアウト(dropout)」と言います。ドロップアウトは、遺伝子発現量がゼロに設定されたりされなかったりするデータ(遺伝子発現量がゼロ)を「ドロップアウト(dropout)」と言います。ドロップアウトを除去する必要があるのです。

しかし、一言で前処理と言っても、実際には、レベルにつぶす(collapseする)処理や、外部ノイズ(実験上のミスや細胞が死んでいるなどの理由)によるカウント数の占めるようなlow-complexity libraryの除去、高発現遺伝子(genes with high dropout rate)の除去(2017)。その他にも、反復データをうまく利用してimputationとかdenoisingとかrecoverなどの処理があります。

R用:

- [scater](#)(通常用) : [McCarthy et al., Bioinformatics, 2016](#)
- [netSmooth](#)(imputation用) : [Ronen et al., Nat Commun., 2018](#)
- [scImpute](#)(imputation用) : [Li and Li, Nat Commun., 2018](#)
- [DrImpute](#)(imputation用) : [Gong et al., Nat Commun., 2018](#)

(Rで塩基配列解析)

x +

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_s...

- [scScope](#)(imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#)(imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- ① 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- ② 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらかも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

[トップページへ](#)

前処理法と評価基準

①この論文のFigs. 1 and 2の結果は、負の二項分布モデル(NB model)から生成したシミュレーションデータ由来。そして②SAVERはNBモデルを仮定した model-based method。それゆえ、SAVERがよい性能を示すのは当たり前といえば当たり前。もちろん、①のpage 5の右側の真ん中あたりでもきっちり書かれているが、AbstractをFig.だけをざっと眺めただけではわからないところなので読者の力量が問われるところ。

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#al

解析 | 前処理 | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、いう特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現していない」という生物学的な理由(biological cause)によるもの。そしてもう1つは「検出されていない」という技術的な理由(technical cause)によるもの。これを「ドロップアウト(dropout)」と言います。ドロップアウトは、遺伝子発現が検出されなかったりするデータ (遺伝子発現データの前処理)が必要なのです。

しかし、一言で前処理と言っても、実際には、ノイズを除去する処理や、外部コ変数(実験上のミスや細胞が死んでいるなどの理由)による変動の占めるようなlow-complexity libraryの補正(高次元データの次元削減)の低い遺伝子(genes with high dropout rate)の補正(2017)。その他にも、反復データをうまく利用してimputationとかdenoisingとかrecoverなどの

R用:

- [scater](#)(通常用) : [McCarthy et al., Bioinformatics, 2016](#)
- [netSmooth](#)(imputation用) : [Ronen et al., Nat Commun., 2019](#)
- [scImpute](#)(imputation用) : [Li and Li, Nat Commun., 2018](#)
- [DrImpute](#)(imputation用) : [Gong et al., Nat Commun., 2019](#)

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#al

- [scScope](#)(imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#)(imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- ① 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- ② 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらかでも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

[トップページへ](#)

シミュレーションデータ

①この論文では、シミュレーションの枠組みでSplatterプログラムが利用されているが、scRNA-seqデータの特徴をうまく捉えられているかについては疑問が残る (page7の左下あたりにも記載あり)。

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#about_analysis_prepro...

解析 | 前処理 | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、多くの遺伝子のカウントがゼロ(0)という特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現している遺伝子が異なるため本当に発現していない」という生物学的な理由(biological cause)によるもの。そしてもう1つは「本当は発現しているんだけども捉えられていない」という技術的な理由(technical cause)によるもの。これを「ドロップアウト(dropout)」と言います。ドロップアウトは、データ(遺伝子発現)が欠けたりされなかったりするデータ(遺伝子発現)を「ドロップアウト(dropout)」と言います。ドロップアウトを補正するための前処理(preprocessing)が必要なのです。

しかし、一言で前処理と言っても、実際には、レベルにつぶす(collapseする)処理や、外部コタ(実験上のミスや細胞が死んでいるなどの理由)でカウント数の占めるようなlow-complexity libraryの生成(高次元遺伝子)の除去(genes with high dropout rate) (2017)。この他にも、反復データをうまく利用してimputationとかdenoisingとかrecoverなどの

R用:

- [scater](#)(通常用) : [McCarthy et al., Biostatistics, 2016](#)
- [netSmooth](#)(imputation用) : [Ronen et al., Nat Commun., 2019](#)
- [scImpute](#)(imputation用) : [Li and Li, Nat Commun., 2018](#)
- [DrImpute](#)(imputation用) : [Gong et al., Nat Commun., 2019](#)

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_s...

- [scScope](#)(imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#)(imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- ① 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらかも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

[トップページへ](#)

シミュレーションデータ

実際、①から辿れる、②の論文のFig. 1を眺めると、③Splatterよりも、④powsimRパッケージを用いてシミュレーションデータを生成するほうがよさそうな印象を受ける。

(Rで)塩基配列解析

x +

← → ↻ 🏠 ⓘ 保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#

- カウント情報取得 | シミュレーションデータ | RNA-seq | Biological rep. | 2群間 | 基礎 | [TCC\(Sun 2013\)](#) (last modified 2018/07/22)
- カウント情報取得 | シミュレーションデータ | RNA-seq | Biological rep. | 2群間 | 応用 | [TCC\(Sun 2013\)](#) (last modified 2018/07/22)
- カウント情報取得 | シミュレーションデータ | RNA-seq | Biological rep. | 3群間 | 基礎 | [TCC\(Sun 2013\)](#) (last modified 2018/07/22)
- [カウント情報取得 | シミュレーションデータ | scRNA-seq | について](#) ① modified 2019/04/12
- カウント情報取得 | シミュレーションデータ | scRNA-seq | 基礎(同一細胞群) | [Splatter\(Zappia 2017\)](#) (last modified 2019/04/11)
- カウント情報取得 | シミュレーションデータ | scRNA-seq | 応用 | [powsimR\(Vieth et al. 2017\)](#) (last modified 2019/04/11)
- カウント情報取得 | シミュレーションデータ | scRNA-seq | 評価系 | [countsimQC\(Soneson and Robinson 2018\)](#) (last modified 2019/04/11)
- [配列長とカウント数の関係](#) (last modified 2019/04/11)
- [正規化 | について](#) (last modified 2019/04/11)
- 正規化 | 基礎 | [RPK or CPK \(配列長\)](#) (last modified 2019/04/11)
- 正規化 | 基礎 | [RPM or CPM \(総リ\)](#) (last modified 2019/04/11)
- 正規化 | 基礎 | [RPKM\(Mortazavi 2003\)](#) (last modified 2019/04/11)
- 正規化 | 基礎 | [TPM\(Li 2010\)](#) (last modified 2019/04/11)
- 正規化 | サンプル内 | [EDASeq\(Risso et al. 2014\)](#) (last modified 2019/04/11)

(Rで)塩基配列解析

x +

← → ↻ 🏠 ⓘ 保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#ab... ☆ 🧑

カウント情報取得 | シミュレーションデータ | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)用のシミュレーションデータを作成するものです。

Rパッケージ:

- [BASiCS](#) : [Vallejos et al., PLoS Comput Biol., 2015](#)
- [scDD](#) : [Korthauer et al., Genome Biol., 2016](#)
- [Splatter](#) : [Zappia et al., Genome Biol., 2017](#) ③
- [powsimR](#) : [Vieth et al., Bioinformatics, 2017](#) ④
- [countsimQC](#)(作成というより評価系) : [Soneson and Robinson, Bioinformatics, 2018](#) ②

R以外:

- [PROSSTT](#) : [Papadopoulos et al., Bioinformatics, 2019](#)

[トップページへ](#)

オートエンコーダ系の話

(Rで塩基配列解析)

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#about_analysis_prepro...

解析 | 前処理 | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、多くの遺伝子のカウントがゼロ(0)という特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現している遺伝子が異なるため本当に発現していない」という生物学的な理由(biological cause)によるもの。そしてもう1つは「本当は発現しているんだけども捉えられていない」という技術的な理由(technical cause)によるもの。この技術的な理由を「ドロップアウト(dropout)」と言います。ドロップアウトは、遺伝子の発現が検出されなかったりされなかったりするデータ (遺伝子発現データの前処理(preprocessing)が必要なのです。

しかし、一言で前処理と言っても、実際には、ドロップアウトを補正する処理や、外部ノイズ(実験上のミスや細胞が死んでいるなどの理由)によるカウント数の占めるようなlow-complexity libraryの除去、高発現遺伝子(genes with high dropout rate)の除去(2017)。この他にも、反復データをうまく利用してimputationとかdenoisingとかrecoverなどの処理が行われます。

R用:

- [scater](#)(通常用) : [McCarthy et al., Bioinformatics, 2016](#)
- [netSmooth](#)(imputation用) : [Ronen et al., Nat Commun., 2019](#)
- [scImpute](#)(imputation用) : [Li and Li, Nat Commun., 2018](#)
- [DrImpute](#)(imputation用) : [Gong et al., Nat Commun., 2019](#)

(Rで塩基配列解析)

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_s...

- [scScope](#)(imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#)(imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらかでも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

[トップページへ](#)

オートエンコーダ系の話

①DCAは、②オートエンコーダに基づく方法(autoencoder-based method)。オートエンコーダは、元々はニューラルネットワークを使用した次元削減のためのアルゴリズムとして提案されたもの。

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#about_analysis_prepro...

解析 | 前処理 | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、多くの遺伝子のカウントがゼロ(0)という特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現している遺伝子が異なるため本当に発現していない」という生物学的な理由(biological cause)によるもの。そしてもう1つは「本当は発現しているんだけども捉えられていない」という技術的な理由(technical cause)によるもの。これを「ドロップアウト(dropout)」と言います。ドロップアウトは、データが欠けたりされなかったりするデータ(遺伝子)を「補完(imputation)」する必要があります。

しかし、一言で前処理と言っても、実際にはレベルにつぶす(collapseする)処理や、外部コタ(実験上のミスや細胞が死んでいるなどの理由)でカウント数の占めるようなlow-complexity libraryの除去や、高発現遺伝子(genes with high dropout rate)の除去(2017)。この他にも、反復データをうまく利用してimputationとかdenoisingとかrecoverなどの処理が行われます。

R用:

- [scater](#)(通常用) : [McCarthy et al., Bioinformatics, 2016](#)
- [netSmooth](#)(imputation用) : [Ronen et al., BMC Bioinformatics, 2016](#)
- [scImpute](#)(imputation用) : [Li and Li, Nature Methods, 2018](#)
- [DrImpute](#)(imputation用) : [Gong et al., Nature Methods, 2018](#)

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_s...

- [scScope](#)(imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#)(imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらかでも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

[トップページへ](#)

オートエンコーダ系の話

①最近ではノイズ除去目的でも利用されている。それをscRNA-seq用にパラメータチューニングしたのが、②DCAという理解でよい。

(Rで塩基配列解析)

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#about_analysis_prepro...

解析 | 前処理 | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、多くの遺伝子のカウントがゼロ(0)という特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現している遺伝子が異なるため本当に発現していない」という生物学的な理由(biological cause)によるもの。そしてもう1つは「本当は発現しているんだけども捉えられていない」という技術的な理由(technical cause)によるもの。これを「ドロップアウト(dropout)」と言います。ドロップアウトは、検出されたりされなかったりするデータ(遺伝子発現量)を「補完(imputation)」が必要なのです。

しかし、一言で前処理と言っても、実際には、ドロップアウトを補完する処理や、外部ノイズ(実験上のミスや細胞が死んでいるなどの理由)による低複雑性ライブラリ(low-complexity library)の除去、高発現遺伝子(genes with high dropout rate)の除去(2017)。この他にも、反復データをうまく利用してimputationとかdenoisingとかrecoverなどの処理が行われます。

R用:

- [scater](#)(通常用) : [McCarthy et al., Bioinformatics, 2016](#)
- [netSmooth](#)(imputation用) : [Ronen et al., Nat Commun., 2019](#)
- [scImpute](#)(imputation用) : [Li and Li, Nat Commun., 2018](#)
- [DrImpute](#)(imputation用) : [Gong et al., Nat Commun., 2019](#)

(Rで塩基配列解析)

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_s...

- [scScope](#)(imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#)(imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらかでも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

[トップページへ](#)

オートエンコーダ系の話

オートエンコーダは、ざっくり言ってEncoderとDecoderの2つの部分に分かれる。Encoderのところでデータの圧縮(compression)が行われる。このときにランダムノイズなどの本質的でない情報(non-essential sources of variation)がふるい落とされるので、結果的にノイズ除去を行っていることになる。次に、Decoderのところでデータの再構築(reconstruction)が行われる。再構築とは、入力と同じ遺伝子数(次元数)からなるノイズ除去後の出力結果(denoised output)を得ることに相当する。

(Rで)塩基配列解析

x +

← → ↻ 🏠 ⓘ 保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#a

解析 | 前処理 | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、いう特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現していない」という生物学的な理由(biological cause)によるもの。そしてもう1つは「検出されていない」という技術的な理由(technical cause)によるもの。これを「ドロップアウト(dropout)」と言います。ドロップアウトは、遺伝子発現が検出されなかったり、あるいは検出されなかったりするデータ(遺伝子発現データの前処理)が必要なのです。

しかし、一言で前処理と言っても、実際には、レベルにつぶす(collapseする)処理や、外部コタ(実験上のミスや細胞が死んでいるなどの理由)でカウント数の占めるようなlow-complexity libraryの除去、高発現遺伝子(genes with high dropout rate)の除去(2017)。この他にも、反復データをうまく利用してimputationとかdenoisingとかrecoverなど

R用:

- [scater](#)(通常用) : [McCarthy et al., Bi](#)
- [netSmooth](#)(imputation用) : [Ronen](#)
- [scImpute](#)(imputation用) : [Li and Li](#)
- [DrImpute](#)(imputation用) : [Gong et](#)

(Rで)塩基配列解析

x +

← → ↻ 🏠 ⓘ 保護されていない

- [scScope](#)(imputation用) : [De](#)
- [SAVER-X](#)(imputation用) : [W](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらかでも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

[トップページへ](#)

オートエンコーダ系の話

①の論文中 (page 6の左下あたり) で②を引用して述べていることとして、オートエンコーダはパフォーマンスに大きな影響を与える多くのパラメータを含む。それは、③でも述べているようにパラメータチューニング次第で精度が大きく変わりうるということ。

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#al

解析 | 前処理 | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、多くの遺伝子のカウントがゼロ(0)という特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現している遺伝子が異なるため本当に発現していない」という生物学的な理由(biological cause)によるもの。そしてもう1つは「本当は発現しているんだけども捉えられていない」という技術的な理由(technical cause)によるもの。これを「ドロップアウト(dropout)」と言います。ドロップアウトは、遺伝子の発現が検出されたりされなかったりするデータ (遺伝子発現データの前処理(preprocessing)が必要なのです。

しかし、一言で前処理と言っても、実際には、ノイズを除去する処理や、外部コトバ(実験上のミスや細胞が死んでいるなどの理由)による低複雑性ライブラリ(low-complexity library)の除去、高発現遺伝子(genes with high dropout rate)の除去(2017)。この他にも、反復データをうまく利用してimputationとかdenoisingとかrecoverなどの処理が行われます。

R用:

- [scater](#)(通常用) : [McCarthy et al., Bioinformatics, 2016](#)
- [netSmooth](#)(imputation用) : [Ronen et al., Nat Commun., 2019](#)
- [scImpute](#)(imputation用) : [Li and Li, Nat Commun., 2018](#)
- [DrImpute](#)(imputation用) : [Gong et al., Nat Commun., 2019](#)

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_s...

- [scScope](#)(imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#)(imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

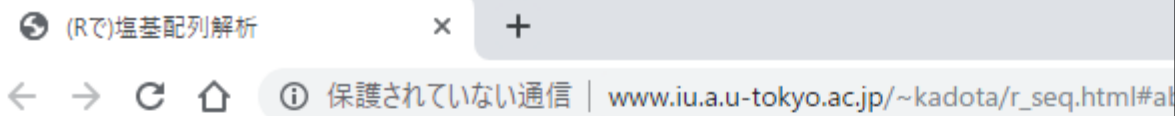
Review、ガイドライン、パイプライン系:

- ① 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- ② 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらかでも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

③ [トップページへ](#)

オートエンコーダ系の話

①の論文中 (page 6の左下あたり) で②を引用して述べていることとして、オートエンコーダはパフォーマンスに大きな影響を与える多くのパラメータを含む。それは、③でも述べているようにパラメータチューニング次第で精度が大きく変わってしまうということ。感覚的には、やろうと思えば都合のよいシミュレーション条件のみで性能評価して「俺の方法が一番」と言えるということ。



解析 | 前処理 | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、いう特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現していない」という生物学的な理由(biological cause)によるもの。そしてもう1つは「検出されていない」という技術的な理由(technical cause)によるもの。これを「ドロップアウト(dropout)」と言います。ドロップアウトは、遺伝子発現が検出されたりされなかったりするデータ (遺伝子発現データの前処理 (preprocessing)が必要なのです。

しかし、一言で前処理と言っても、実際には、ノイズを除去する処理や、外部コ変数(実験上のミスや細胞が死んでいるなどの理由)による変動の占めるようなlow-complexity libraryの補正、高発現遺伝子(genes with high dropout rate)の補正(2017)。その他にも、反復データをうまく利用してimputationとかdenoisingとかrecoverなどの処理が行われます。

R用:

- [scater](#)(通常用) : [McCarthy et al., Bioinformatics, 2016](#)
- [netSmooth](#)(imputation用) : [Ronen et al., Nat Commun., 2019](#)
- [scImpute](#)(imputation用) : [Li and Li, Nat Commun., 2018](#)
- [DrImpute](#)(imputation用) : [Gong et al., Nat Commun., 2019](#)



- [scScope](#)(imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#)(imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- ① 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- ② 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらかでも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

[トップページへ](#)

オートエンコーダ系の話

例えば、評価用データセットで、my methodのみパラメータチューニングしておく。そして他の方法はデフォルトで実行すると、自分の方法が有利になります。
①のpage 7で、②ノーフリーランチ定理(no free lunch theorem)と絡めて述べられています。

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#al

解析 | 前処理 | scRNA-seq | について

single-cell RNA-seq (scRNA-seq)のカウントデータは「疎(sparse)」です。つまり、多くの遺伝子のカウントがゼロ(0)という特徴を持ちます。これには2つの理由があり、1つは「細胞のタイプごとに発現している遺伝子が異なるため本当に発現していない」という生物学的な理由(biological cause)によるもの。そしてもう1つは「本当は発現しているんだけども捉えられていない」という技術的な理由(technical cause)によるもの。これを「ドロップアウト(dropout)」と言います。ドロップアウトは、遺伝子の発現が検出されたりされなかったりするデータ (遺伝子発現データの前処理(preprocessing)が必要なのです。

しかし、一言で前処理と言っても、実際には、ノイズを除去する処理や、外部コスタ(実験上のミスや細胞が死んでいるなどの理由)による低複雑性ライブラリ(low-complexity library)の除去、高発現遺伝子(genes with high dropout rate)の除去(2017)。その他にも、反復データをうまく利用してimputationとかdenoisingとかrecoverなどの処理が行われます。

R用:

- [scater](#)(通常用) : [McCarthy et al., Bioinformatics, 2016](#)
- [netSmooth](#)(imputation用) : [Ronen et al., Nature Methods, 2016](#)
- [scImpute](#)(imputation用) : [Li and Li, Nature Methods, 2018](#)
- [DrImpute](#)(imputation用) : [Gong et al., Nature Methods, 2018](#)

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_s...

- [scScope](#)(imputation用) : [Deng et al., bioRxiv, 2018](#)
- [SAVER-X](#)(imputation用) : [Wang et al., bioRxiv, 2018](#)
- [DCA](#)(imputation用) : [Eraslan et al., Nat Commun., 2019](#)
- [McImpute](#)(imputation用) : [Mongia et al., Front Genet., 2019](#)
- [Scrublet](#)(doublets同定用) : [Wolock et al., Cell Syst., 2019](#)

Review、ガイドライン、パイプライン系:

- 手法比較(SAVERが一番無難という結論) : [Andrews and Hemberg, F1000Res., 2018](#)
- 手法比較(Variational Autoencoder(VAE)系はパラメータチューニング次第でいくらかも変わる) : [Hu and Greene, Pac Symp Biocomput., 2019](#)

① [トップページへ](#)

②

Contents

- 自己紹介と東大アグリバイオの紹介
- トランスクリプトーム解析、発現解析、発現変動解析、実験デザイン
- 2群間比較: 実データ、TCC(反復増やすとDEG増える)
- 他グループによる性能評価論文(TCCが非推奨となる場合も!)
- TCCで3群間比較、baySeqも組み合わせて発現パターンまで得る
- (Rで)塩基配列解析
- Single-cell RNA-seq(scRNA-seq)
- バイオインフォマティクス実習

バイオインフォ実習

今日の話は、①中林先生と藩先生らによって行われているバイオインフォマテイクス実習の、②に相当する部分です。scRNA-seqの内容については、③9月5日に実施予定。

先端医科学研究センター バイオインフォマテイクス解析室

2019年度 バイオインフォマテイクス実習



開催内容

1人1台のパソコンを使用して大規模データの解析手法について学びます。自身のノートパソコンを会場に持ち込んで作業を行うことも可能です。

- 会場 横浜市立大学福浦キャンパス 看護棟 4階 M402情報処理室
- 定員 30名
- 講師 中林 潤 (横浜市立大学 先端医科学研究センター バイオインフォマテイクス 准教授)
藩 龍馬 (横浜市立大学 医学部 免疫学 助教)

体験実習	2019年 4月11日 (木) 17:30-18:30 (体験実習は60分)	実習概要説明、PC操作説明
第1回	2019年 5月9日 (木) 17:30-19:00	RNA-seqデータ解析 シーケンスデータのマッピング、 カウント、可視化
第2回	2019年 7月4日 (木) 17:30-19:00	RNA-seqデータ解析 発現変動遺伝子同定、 遺伝子の機能推定
第3回	2019年 9月5日 (木) 17:30-19:00	scRNA-seqデータ解析 データ処理、細胞のクラスタ化
第4回	2019年 11月14日 (木) 17:30-19:00	ATAC-seqデータ解析 マッピング、モチーフ解析
第5回	2020年 1月9日 (木) 17:30-19:00	Whole exome解析 マッピング、SNP解析

体験実習は、
申込不要・無料です



Bioinformatic