東京大学 THE UNIVERSITY OF TOKYO

2020.02.21版

①です。PDは、Public Domainの略。前半はSingle-cell RNA-seqデータもbulk RNA-seqデータと特性に違いはないんじゃないかという内容。後半は3群間比較や生存曲線の描画に関する内容。2月10, 14, and 21日の計3回分。

ゲノム医学（第1-3回）

[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/

①

資料はコチラ

①ググって、②私のホームページの、③講義、のところにPDFがあります（のでメモなどをとる必要はありません）。

東大　門田

門田　幸二のホームページ

名前　門田　幸二（かどた　こうじ）

所属　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
東京大学 微生物科学イノベーション連携研究機構

身分　准教授

研究分野　バイオインフォマティクス（トランスクリプトーム解析）

- 研究テーマ (last modified 2018/03/01)
- 原著論文 (last modified 2018/03/08)
- 総説・解説記事・翻訳など (last modified 2017/11/13)
- 略歴 (last modified 2018/04/09)
- 講義 (last modified 2018/04/06)
- 講演・講習会など (last modified 2018/05/04) NEW
- 外部研究資金 (last modified 2018/04/06)
- その他 (last modified 2018/05/17) NEW
- リンク集 (last modified 2018/05/11) NEW

研究テーマ

トランスクリプトーム解析手法の開発。本ユニットでは、様々なトランスクリプトームデータの解析や新規解析手法の開発を通じて、農学生命科学への応用を目指します。「数式を並べ立てた難解な方法を凌駕する"シンプルな方法"の開発」をモットーとしています。これまでの主な研究成果を三つのカテゴリーで分けていますが、いずれも「トランスクリプトーム解析」でひとまとめにできます。また、実験系の方でも気軽に研究成果を利用可能なように「(Rで)マイクロアレイデータ解析」と「(Rで)塩基配列解析」上にも下記開発手法中の一部について、その利用法を記述しています。

資料はコチラ

①ググって、②私のホームページの、③講義、のところにPDFがあります（のでメモなどをとる必要はありません）。④このあたりです。

講義 ③

- 東大・院農・アグリバイオ「バイオスタティスティクス基礎論」(2006-2008年度、分担)
- 東大・院農・アグリバイオ「農学生命情報科学実習I」(2005-2008年度、分担)
- 東大・院農・アグリバイオ「機能ゲノム学」（2005-2008年度は分担、2014-2019年度）
- 東大・院農・アグリバイオ「バイオインフォマティクス基礎実習」(2004-2008年度、分担)
- 東大・院農・アグリバイオ「プロテオーム情報学」（2009年度、分担）
- 東大・院農・アグリバイオ「バイオインフォマティクスリテラシーII」（2009年度、分担）
- 東大・院農・アグリバイオ「ゲノム情報解析基礎」（2010-2018年度は分担、2019年度）
- 東大・院農・アグリバイオ「オーム情報解析」（2010-2013年度、分担）
- 東大・院農・アグリバイオ「農学生命情報科学特論I」（2010-2014年度は分担、2015-2016年度、2018-2019年度）
- 東大・院農・アグリバイオ「農学生命情報科学特論II」（2016年度）
- 東大・院農・アグリバイオ「農学生命情報科学特論III」（2011, 2013年度、分担）
- 東大・院農「情報生命工学」（1コマ; 2003, 2005, 2009年度）
- 東大・農学部「生物情報工学」（2コマ; 2005-2007年度）
- 東大・農学部「生物情報科学」（1コマ; 2008-2015年度）
- 東大・農学部「生物情報科学I」（1コマ; 2016-2019年度）
- 東大・農学部展開科目「バイオインフォマティクス」（2016-2017年度、分担）
- バイオインフォマティクス人材育成講座 スタンダードコース「バイオインフォマティクス 次世代シーケンサー編」（4コマ; 2011年度; 於沖縄工業高等専門学校(沖縄); 2011.10.15）
- 琉球大学・農学部「食品機能科学特別講義I」（3コマ; 2012年度; H24年度バイオインフォマティクス・スタンダードコースの一環; 2012.09.06;「講義資料」;「課題」）
- 奈良先端科学技術大学院大学(NAIST)・バイオサイエンス研究科「ゲノム機能解析特論」（2013年度; NAIST植物グローバル教育プロジェクト・平成25年度ワークショップの一環; 2013.06.06;「ゲノム・トランスクリプトームの各種解析をRで行う」）
- 奈良先端科学技術大学院大学(NAIST)・バイオサイエンス研究科「ゲノム機能解析特論」（2014年度; NAIST植物グローバル教育プロジェクト・平成26年度ワークショップの一環; 2014.06.12;「(Rで)塩基配列解析の利用法：GC含量計算から発現変動解析まで」）
- 横浜市立大学・大学院医学研究科「ゲノム医学」：第1回(2019.01.25), 第2回(2019.02.22), 第3回(2019.03.15)
- 横浜市立大学・大学院医学研究科「ゲノム医学」：第1回(2020.02.10), 第2回(2020.02.14), 第3回(2020.02.21) ④

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

昨年度の計3回の講義内容です。

昨年度の講義内容

- 第1回（2019年1月25日）：バイオインフォ概論とRの基礎

  バイオインフォマティクスを学ぶ上でのRの位置づけや、基本的な利用法に関する本当に極初級者向けの解説

- 第2回（2019年2月22日）：Rパッケージの話

  Rを利用する際によく聞くパッケージというものの概念的な話や、どのようにして利用したいパッケージを見つけ出すかなどの話。

- 第3回（2019年3月15日）：RNA-seq発現解析

  ガン vs. 正常などの状態の異なるグループ間でのクラスタリングや発現変動解析を行う実例や結果の解釈についての解説。Rを覚える時間がないヒトでもウェブツールを利用して同様の解析ができる話など。

昨年度の第1回の講義内容です。

Contents（2019年1月）

- アグリバイオの歴史（一兵卒の視点から）
- アグリバイオの教育プログラム
- 2つのウェブページ
  - （Rで）マイクロアレイデータ解析
  - （Rで）塩基配列解析
- バイオインフォマティクスとNGS（次世代シークエンサ）
  - テクノロジーの栄枯盛衰～マイクロアレイからRNA-seq（NGS）へ～
  - NGSハンズオン講習会（H26～29）
  - 日本乳酸菌学会誌のNGS連載（H26～）
- バイオインフォマティクスとR
  - バイオインフォマティクス・スキル標準
  - Rの基本的な利用法

昨年度の第2回の講義内容です。

Contents（2019年2月）

- R本体とRパッケージの関係
- Rパッケージ：BioconductorとCRAN
- Rの基本的な利用法のおさらい：Biostringsを用いた翻訳配列の取得
  - 入力ファイルのダウンロード、Rの起動を作業ディレクトリの変更
  - コピペ実行、コードの解説
- 複数の例題を実行して理解を深める
- 「（Rで）塩基配列解析」の興味ある項目の例題をこなす
  - Biostrings, BSgenome
- Bioconductorで興味あるパッケージを探す

昨年度の第3回の講義内容です。

Contents（2019年3月）

- トランスクリプトーム（RNA-seq）解析の原理、データ解析戦略のイメージ
- 公共カウントデータの取得（recount2）
- サンプル間クラスタリング
  - 手元のデータを実行、結果の解釈
  - グループ間の分離度を客観的に示すスコア
- TCC-GUI
  - ファイルのアップロード、グループラベル情報の付与、平均シルエットスコア（AS値）
  - 探索的解析（Exploratory Analysis）：階層的クラスタリングやPCAなど
  - 発現変動解析（TCC Computation）
- すぐにDisconnectedとなる問題とその対策
  - RStudioのインストール
  - TCC-GUIローカル版の起動

Contents（2019年8月）

2019年8月28日に「バイオインフォマティクス解析集中トレーニングコース」の一環?!としてお話しさせていただいた内容です。

- 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

Contents（2019年8月）

2019年10月にいただいたアンケート結果では、①「3群間比較が難しかった」、「後半が難しかった」、「時間が短かった」というコメントが散見されました。

- 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 ①
- （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

Contents（2019年8月）

- 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
- 2群間比較：実データ、TCC（反復増やすとDEG増える）
- ② 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
- ③ Single-cell RNA-seq（scRNA-seq）
- バイオインフォマティクス実習

2019年10月にいただいたアンケート結果では、①「3群間比較が難しかった」、「後半が難しかった」、「時間が短かった」というコメントが散見されました。また、②bulk RNA-seqは古いので、③scRNA-seqの最新の話を知りたかったというコメントもありました。

Contents（2019年8月）

- 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
- 2群間比較：実データ、TCC（反復増やすと
- ② 他グループによる性能評価論文（TCCが
- 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
- ③ Single-cell RNA-seq（scRNA-seq）
- バイオインフォマティクス実習

2019年10月にいただいたアンケート結果では、①「3群間比較が難しかった」、「後半が難しかった」、「時間が短かった」というコメントが散見されました。また、②bulk RNA-seqは古いので、③scRNA-seqの最新の話を知りたかったというコメントもありました。しかし、③scRNA-seqデータ解析の要素技術の多くは、②bulk RNA-seq由来のもの。知見もまた然り。

Contents（2019年8月）

- 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
- 2群間比較：実データ、TCC（反復増やすと
- ② 他グループによる性能評価論文（TCCが非
- TCCで3群間比較、baySeqも組み合わせて
- （Rで）塩基配列解析
- ③ Single-cell RNA-seq（scRNA-seq）
- バイオインフォマティクス実習

2019年10月にいただいたアンケート結果では、①「3群間比較が難しかった」、「後半が難しかった」、「時間が短かった」というコメントが散見されました。また、②bulk RNA-seqは古いので、③scRNA-seqの最新の話を知りたかったというコメントもありました。しかし、③scRNA-seqデータ解析の要素技術の多くは、②bulk RNA-seq由来のもの。知見もまた然り。プログラミング言語についても、「敷居が高いが計算が速いコンパイラ方式のC言語 vs. 敷居が低いが計算が遅いインタプリタ方式のPerl言語」みたいな昔話と同様?!、「敷居が高いが計算が速いPython vs. 閾値が低いが計算が遅いR」みたいな感じで、どっちがよい?的な話もあるようです。

Contents（2019年8月）

- 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
- 2群間比較：実データ、TCC（反復増やすと
- ② 他グループによる性能評価論文（TCCが非
- TCCで3群間比較、baySeqも組み合わせて
- （Rで）塩基配列解析
- ③ Single-cell RNA-seq（scRNA-seq）
- バイオインフォマティクス実習

2019年10月にいただいたアンケート結果では、①「3群間比較が難しかった」、「後半が難しかった」、「時間が短かった」というコメントが散見されました。また、②bulk RNA-seqは古いので、③scRNA-seqの最新の話を知りたかったというコメントもありました。しかし、③scRNA-seqデータ解析の要素技術の多くは、②bulk RNA-seq由来のもの。知見もまた然り。プログラミング言語についても、「敷居が高いが計算が速いコンパイラ方式のC言語 vs. 敷居が低いが計算が遅いインタプリタ方式のPerl言語」みたいな昔話と同様?!、「敷居が高いが計算が速いPython vs. 閾値が低いが計算が遅いR」みたいな感じで、どっちがよい?的な話もあるようです。今Pythonが流行っている理由は、深層学習が流行っているから。「深く学ぶ → 沢山計算する → 高速計算可能なPython」という流れだと思います。つまり深層学習でなければRで十分。

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

DECENT論文1

scRNA-seq発現変動解析用Rパッケージ DECENTの論文。

Bioinformatics. 2019 Dec 15;35(24):5155-5162. doi: 10.1093/bioinformatics/btz453.

DECENT: differential expression with capture efficiency adjustmeNT for single-cell RNA-seq data.

Ye C[1,2,3], Speed TP[1,4], Salim A[1,5,6].

Author information

Abstract

MOTIVATION: Dropout is a common phenomenon in single-cell RNA-seq (scRNA-seq) data, and when left unaddressed it affects the validity of the statistical analyses. Despite this, few current methods for differential expression (DE) analysis of scRNA-seq data explicitly model the process that gives rise to the dropout events. We develop DECENT, a method for DE analysis of scRNA-seq data that explicitly and accurately models the molecule capture process in scRNA-seq experiments.

RESULTS: We show that DECENT demonstrates improved DE performance over existing DE methods that do not explicitly model dropout. This improvement is consistently observed across several public scRNA-seq datasets generated using different technological platforms. The gain in improvement is especially large when the capture process is overdispersed. DECENT maintains type I error well while achieving better sensitivity. Its performance without spike-ins is almost as good as when spike-ins are used to calibrate the capture model.

AVAILABILITY AND IMPLEMENTATION: The method is implemented as a publicly available R package available from https://github.com/cz-ye/DECENT.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics online.

PMID: 31197307 PMCID: PMC6954660 DOI: 10.1093/bioinformatics/btz453

DECENT論文2

scRNA-seq発現変動解析用Rパッケージ DECENTの論文。①比較的最近Publishされたものです。②のURLから利用可能とあるが、私の経験上③githubに置いてあるプログラムは、使い方の説明が不親切。

Bioinformatics. 2019 Dec 15;):5155-5162. doi: 10.1093/bioinformatics/btz453.

DECENT: differential expression with capture efficiency adjustr single-cell RNA-seq data.

Ye C[1,2,3], Speed TP[1,4], Salim A[1,5,6].

Author information

Abstract

MOTIVATION: Dropout is a common phenomenon in single-cell RNA-seq (scRNA-seq) data, and when left unaddressed it affects the validity of the statistical analyses. Despite this, few current methods for differential expression (DE) analysis of scRNA-seq data explicitly model the process that gives rise to the dropout events. We develop DECENT, a method for DE analysis of scRNA-seq data that explicitly and accurately models the molecule capture process in scRNA-seq experiments.

RESULTS: We show that DECENT demonstrates improved DE performance over existing DE methods that do not explicitly model dropout. This improvement is consistently observed across several public scRNA-seq datasets generated using different technological platforms. The gain in improvement is especially large when the capture process is overdispersed. DECENT maintains type I error well while achieving better sensitivity. Its performance without spike-ins is almost as good as when spike-ins are used to calibrate the capture model.

AVAILABILITY AND IMPLEMENTATION: The method is implemented as a publicly available R package available from https://github.com/cz-ye/DECENT.

SUPPLEMENTARY ATION: Supplement ata are available at Bioinformatics online.

PMID: 31197307 PMCID: PMC6954660 DOI: 10.1093/bioinformatics/btz453

DECENT論文3

PDFファイルだと、①5160ページの右下あたりの段落を抜粋。

Bioinformatics. 2019 Dec 15;35(24):5155-5162. doi: 10.1093/bioinformatics/btz453.

DECENT: differential expression with capture efficiency adjustmeNT for single-cell RNA-seq data.

Ye C[1,2,3], Speed TP[1,4], Salim A[1,5,6].

Author information

①

Abstract

MOTIVATION: Dropo
unaddressed it affec
expression (DE) ana
We develop DECEN
molecule capture pro

RESULTS: We show
not explicitly model c
datasets generated u
the capture process
performance without

AVAILABILITY AND
available from https:/

SUPPLEMENTARY I

PMID: 31197307 PMCI

The benchmarking so far was based on two group comparisons. DECENT performs statistical tests under the under the well-established GLM framework and can readily accommodate more complex experimental designs. The Soumillon *et al.*'s data are a time course experiment, with three time points involved in adipose stem cell differentiation. This allowed us to have a glance at how different DE methods perform on more complex UMI-based scRNA-seq experiments beyond two-group comparisons. We tested the hypothesis that expression of a gene is constant across the three time points. Except for SCDE, which is designed only for two group comparison, and TASC, which requires spike-ins, other methods were compared in this setting. The reference genuine DEGs across the three time points were also derived from the matching bulk experiments. DECENT again outperformed all other methods with an even more pronounced advantage (Supplementary Fig. S11).

DECENT論文4

PDFファイルだと、①5160ページの右下あたりの段落を抜粋。このscRNA-seq論文でも行われているように、大抵の性能評価は②2群間比較で行われる。

Bioinformatics. 2019 Dec 15;35(24):5155-5162. doi: 10.1093/bioinformatics/btz453.

DECENT: differential expression with capture efficiency adjustmeNT for single-cell RNA-seq data.

Ye C[1,2,3], Speed TP[1,4], Salim A[1,5,6].

Author information

②

Abstract

MOTIVATION: Dropo
unaddressed it affect
expression (DE) ana
We develop DECEN
molecule capture pro

RESULTS: We show
not explicitly model c
datasets generated u
the capture process
performance without

AVAILABILITY AND
available from https:/

SUPPLEMENTARY I

PMID: 31197307 PMCI

The benchmarking so far was based on two group comparisons. DECENT performs statistical tests under the under the well-established GLM framework and can readily accommodate more complex experimental designs. The Soumillon *et al.*'s data are a time course experiment, with three time points involved in adipose stem cell differentiation. This allowed us to have a glance at how different DE methods perform on more complex UMI-based scRNA-seq experiments beyond two-group comparisons. We tested the hypothesis that expression of a gene is constant across the three time points. Except for SCDE, which is designed only for two group comparison, and TASC, which requires spike-ins, other methods were compared in this setting. The reference genuine DEGs across the three time points were also derived from the matching bulk experiments. DECENT again outperformed all other methods with an even more pronounced advantage (Supplementary Fig. S11).

DECENT論文5

PDFファイルだと、①5160ページの右下あたりの段落を抜粋。このscRNA-seq論文でも行われているように、大抵の性能評価は②2群間比較で行われる。③Rパッケージ DECENTは、一般化線形モデル（GLM）を用いた検定を行う。

Bioinformatics. 2019 Dec 15;35(24):5155-5162. doi: 10.1093/bioinformatics/btz453.

DECENT: differential expression with capture efficiency adjustr
single-cell RNA-seq data.

Ye C[1,2,3], Speed TP[1,4], Salim A[1,5,6].

Author information

Abstract

MOTIVATION: Dropo
unaddressed it affect
expression (DE) ana
We develop DECEN
molecule capture pro

RESULTS: We show
not explicitly model c
datasets generated u
the capture process
performance without

AVAILABILITY AND
available from https:/

SUPPLEMENTARY I

PMID: 31197307 PMCI

The benchmarking so far was based on two group comparisons. DECENT performs statistical tests under the under the well-established GLM framework and can readily accommodate more complex experimental designs. The Soumillon *et al.*'s data are a time course experiment, with three time points involved in adipose stem cell differentiation. This allowed us to have a glance at how different DE methods perform on more complex UMI-based scRNA-seq experiments beyond two-group comparisons. We tested the hypothesis that expression of a gene is constant across the three time points. Except for SCDE, which is designed only for two group comparison, and TASC, which requires spike-ins, other methods were compared in this setting. The reference genuine DEGs across the three time points were also derived from the matching bulk experiments. DECENT again outperformed all other methods with an even more pronounced advantage (Supplementary Fig. S11).

③

DECENT論文6

Bioinformatics. 2019 Dec 15;35(24):5155-5162. doi: 10.1093/bioinformatics/btz453.

DECENT: differential expression with capture efficiency adjustr
single-cell RNA-seq data.

Ye C[1,2,3], Speed TP[1,4], Salim A[1,5,6].

Author information

Abstract

MOTIVATION: Dropo
unaddressed it affect
expression (DE) ana
We develop DECEN
molecule capture pro

RESULTS: We show
not explicitly model c
datasets generated u
the capture process
performance without

AVAILABILITY AND
available from https:/

SUPPLEMENTARY I

PMID: 31197307 PMCI

The benchmarking so far was based on two group
performs statistical tests under the under the well-
and can readily accommodate more complex expe
Soumillon *et al.*'s data are a time course experime
involved in adipose stem cell differentiation. This
how different DE methods perform on more complex UMI-based scRNA-seq experiments beyond two-group comparisons. We tested the hypothesis that expression of a gene is constant across the three time points. Except for SCDE, which is designed only for two group comparison, and TASC④hich requires spike-ins, other methods were compared in this setting. The reference genuine DEGs across the three time points were also derived from the matching bulk experiments. DECENT again outperformed all other methods with an even more pronounced advantage (Supplementary Fig. S11).

PDFファイルだと、①5160ページの右下あたりの段落を抜粋。このscRNA-seq論文でも行われているように、大抵の性能評価は②2群間比較で行われる。③Rパッケージ DECENTは、一般化線形モデル（GLM）を用いた検定を行う。GLMの意味するところは、例えば3群間比較（G1 vs. G2 vs. G3）のような複雑な実験デザインにも対応しているということ。④この場合の帰無仮説は「G1 = G2 = G3」としている。従って、検定結果のp値が低い遺伝子がどの群間で発現変動しているのかは不明。

DECENT論文7

PDFファイルだと、①5159ページの左下あたり。

Bioinformatics. 2019 Dec 15;35(24):5155-5162. doi: 10.1093/bioinformatics/btz453.

DECENT: differential expression with capture efficiency adjustmeNT for single-cell RNA-seq data.

Ye C[1,2,3], Speed TP[1,4], Salim A[1,5,6].

Author information

①

Abstract

MOTIVATION: Dropo
unaddressed it affec
expression (DE) ana
We develop DECEN
molecule capture pro

RESULTS: We show
not explicitly model c
datasets generated u
the capture process
performance without

AVAILABILITY AND
available from https:/

SUPPLEMENTARY I

PMID: 31197307 PMCI

3.1 Benchmarking using simulated data

We simulated 20 datasets, each consisting of 500 cells belonging to 2 cell types (224 cells from cell type 1 versus 276 cells from cell type 2) with 3000 endogenous genes and 50 spike-ins. The observed count were generated under the DECENT model using parameters estimated from Tung's dataset (see Supplementary Materials for details). In each dataset, we set ~10% of the genes to be DEGs. Figure 2 shows that DECENT estimates gene-specific pre-dropout proportion of zeroes and variance, as well as the actual pre-dropout counts unbiasedly. Figure 3 shows that DECENT's performance in detecting DEGs also appear to be competitive when compared with existing methods, namely SCDE (Kharchenko *et al.*, 2014), MAST (Finak *et al.*, 2015), Monocle2 (Qiu *et al.*, 2017; Trapnell *et al.*, 2014), ZINB-WaVE adjusted edgeR (Van den Berge *et al.*, 2018) and edgeR (McCarthy *et al.*, 2012). Over the 20 datasets, the mean(SD) of the partial area under the receiver operating characteristic (pAUROC) for DECENT is 0.708(0.001), followed by MAST with 0.687(0.001) (see

DECENT論文8

PDFファイルだと、①5159ページの左下あたり。②シミュレーションデータの解析では、全遺伝子のうち10%までの発現変動遺伝子（DEGs）しか想定していない。

Bioinformatics. 2019 Dec 15;35(24):5155-5162. doi: 10.1093/bioinformatics/btz453.

DECENT: differential expression with capture efficiency adjustmeNT for single-cell RNA-seq data.

Ye C[1,2,3], Speed TP[1,4], Salim A[1,5,6].

Author information

①

Abstract

MOTIVATION: Dropo
unaddressed it affect
expression (DE) ana
We develop DECEN
molecule capture pro

RESULTS: We show
not explicitly model d
datasets generated u
the capture process
performance without

AVAILABILITY AND
available from https:/

SUPPLEMENTARY I

PMID: 31197307 PMCI

3.1 Benchmarking using simulated data

We simulated 20 datasets, each consisting of 500 cells belonging to 2 cell types (224 cells from cell type 1 versus 276 cells from cell type 2) with 3000 endogenous genes and 50 spike-ins. The observed count were generated under the DECENT model using parameters estimated from Tung's dataset (see Supplementary Materials for details). In each dataset, we set ~10% of the genes to be DEGs. Figure 2 shows that DECENT estimates gene-specific pre-dropout proportion of zeroes and variance, as well as the actual pre-dropout counts unbiasedly. Figure 3 shows that DECENT's performance in detecting DEGs also appear to be competitive when compared with existing methods, namely SCDE (Kharchenko *et al.*, 2014), MAST (Finak *et al.*, 2015), Monocle2 (Qiu *et al.*, 2017; Trapnell *et al.*, 2014), ZINB-WaVE adjusted edgeR (Van den Berge *et al.*, 2018) and edgeR (McCarthy *et al.*, 2012). Over the 20 datasets, the mean(SD) of the partial area under the receiver operating characteristic (pAUROC) for DECENT is 0.708(0.001), followed by MAST with 0.687(0.001) (see

②

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

NBID論文1

scRNA-seq発現変動解析用Rパッケージ NBIDの論文。

Genome Biol. 2018 May 31;19(1):70. doi: 10.1186/s13059-018-1438-9.

UMI-count modeling and differential expression analysis for single-cell RNA sequencing.

Chen W[1], Li Y[2], Easton J[1], Finkelstein D[1], Wu G[1], Chen X[3].

⊕ Author information

Abstract

Read counting and unique molecular identifier (UMI) counting are the principal gene expression quantification schemes used in single-cell RNA-sequencing (scRNA-seq) analysis. By using multiple scRNA-seq datasets, we reveal distinct distribution differences between these schemes and conclude that the negative binomial model is a good approximation for UMI counts, even in heterogeneous populations. We further propose a novel differential expression analysis algorithm based on a negative binomial model with independent dispersions in each group (NBID). Our results show that this properly controls the FDR and achieves better power for UMI counts when compared to other recently developed packages for scRNA-seq analysis.

KEYWORDS: Differential expression analysis; Negative binomial; Unique molecular identifier

PMID: 29855333 PMCID: PMC5984373 DOI: 10.1186/s13059-018-1438-9

NBID論文2

scRNA-seq発現変動解析用Rパッケージ NBIDの論文。bulk RNA-seqではリードカウントのみだったが、①scRNA-seqでは、「リードカウント」と「UMIカウント」の2種類の数値行列作成手段がある。

Genome Biol. 2018 May 31;19(1):70. doi: 10.1186/s13059-018-1438-9.

UMI-count modeling and differential expression anal single-cell RNA sequencing.

Chen W[1], Li Y[2], Easton J[1], Finkelstein D[1], Wu G[1], Chen X[3].

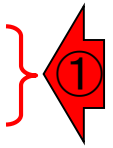
⊕ Author information

Abstract

Read counting and unique molecular identifier (UMI) counting are the principal gene expression quantification schemes used in single-cell RNA-sequencing (scRNA-seq) analysis. By using multiple scRNA-seq datasets, we reveal distinct distribution differences between these schemes and conclude that the negative binomial model is a good approximation for UMI counts, even in heterogeneous populations. We further propose a novel differential expression analysis algorithm based on a negative binomial model with independent dispersions in each group (NBID). Our results show that this properly controls the FDR and achieves better power for UMI counts when compared to other recently developed packages for scRNA-seq analysis.

KEYWORDS: Differential expression analysis; Negative binomial; Unique molecular identifier

PMID: 29855333 PMCID: PMC5984373 DOI: 10.1186/s13059-018-1438-9



NBID論文3

scRNA-seq発現変動解析用Rパッケージ NBIDの論文。bulk RNA-seqではリードカウントのみだったが、①scRNA-seqでは、「リードカウント」と「UMIカウント」の2種類の数値行列作成手段がある。② scRNA-seqのUMIカウントデータは、負の二項分布（negative binomial model）に従う。これは、bulk RNA-seqのカウントデータと同じだということ。

Genome Biol. 2018 May 31;19(1):70. doi: 10.1186/s13059-018-1438-9.

UMI-count modeling and differential expression anal single-cell RNA sequencing.

Chen W[1], Li Y[2], Easton J[1], Finkelstein D[1], Wu G[1], Chen X[3].

Author information

Abstract

Read counting and unique molecular identifier (UMI) counting are the principal gene expression quantification schemes used in single-cell RNA-sequencing (scRNA-seq) analysis. By using multiple scRNA-seq datasets, we reveal distinct distribution differences between these schemes and conclude that the negative binomial model is a good approximation for UMI counts, even in heterogeneous populations. We further propose a novel differential expression analysis algorithm based on a negative binomial model with independent dispersions in each group (NBID). Our results show that this properly controls the FDR and achieves better power for UMI counts when compared to other recently developed packages for scRNA-seq analysis.

KEYWORDS: Differential expression analysis; Negative binomial; Unique molecular identifier

PMID: 29855333 PMCID: PMC5984373 DOI: 10.1186/s13059-018-1438-9

NBID論文4

PDFファイルだと、①13ページの右上あたりを抜粋。Discussionの最後のほう。

single-cell analysis. Even though only pairwise analyses were considered in the current study, the general form of NBID allows multiple groups to be tested simultaneously, as in the generalized linear model framework.

Differential expression analysis can be used to reveal differences among samples run on separate lanes or plates. However, it should be pointed out that batch effects are expected to overlap with biological differences in this setting. It is better to account for batch effects by proper experimental design, such as by including multiple biological replicates for each group. Another typical application of differential expression analysis is to identify potential biomarkers for inferred cell subpopulations. One potential caveat in this setting is that cells are usually clustered from the same data. Therefore, *p* values or FDR values derived from the differential expression analysis might be overly optimistic. However, the result is still useful for prioritizing potential biomarkers for further validation.

NBID論文5

single-cell analysis. Even though only pairwise analyses were co
the current study, the general form of NBID allows multiple grou
tested simultaneously, as in the generalized linear model framework.

Differential expression analysis can be used to reveal differences among samples run on separate lanes or plates. However, it should be pointed out that batch effects are expected to overlap with biological differences in this setting. It is better to account for batch effects by proper experimental design, such as by including multiple biological replicates for each group. Another typical application of differential expression analysis is to identify potential biomarkers for inferred cell subpopulations. One potential caveat in this setting is that cells are usually clustered from the same data. Therefore, *p* values or FDR values derived from the differential expression analysis might be overly optimistic. However, the result is still useful for prioritizing potential biomarkers for further validation.

論文中では2群間比較しか考慮していないが、NBID自体はGLMの枠組みで3群以上の比較も可能とのこと。しかしこのパッケージもマニュアルが不親切で使う気になれない（個人の感想です）。

NBID論文6

single-cell analysis. Even though only pairwise analyses were co
the current study, the general form of NBID allows multiple grou
tested simultaneously, as in the generalized linear model framew

Differential expression analysis can be used to reveal differences
samples run on separate lanes or plates. However, it should be po
batch effects are expected to overlap with biological differences i
It is better to account for batch effects by proper experimental de
by including multiple biological replicates for each group. Anoth
application of differential expression analysis is to identify poten
biomarkers for inferred cell subpopulations. One potential caveat
setting is that cells are usually clustered from the same data. Ther
values or FDR values derived from the differential expression an
be overly optimistic. However, the result is still useful for prioriti
potential biomarkers for further validation.

論文中では2群間比較しか考慮していないが、NBID自体はGLMの枠組みで3群以上の比較も可能とのこと。しかしこのパッケージもマニュアルが不親切で使う気になれない（個人の感想です）。これはNBID法というより、scRNA-seqの一般的な解析手順に起因する注意喚起。scRNA-seqは、まず探索的な解析（クラスタリング）により似た発現パターンを示す細胞をグループ化する。この段階ですでに「同一グループ内の細胞間の全体的な類似度が高いことが保証されている」ので、適当なp値やFDR閾値を満たす発現変動遺伝子（DEG）が多数得られてしまうので注意してね、ということです。ある程度心得のあるヒトなら当たり前だよね、という話ですが念のため。

Content（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

sc性能評価論文1

scRNA-seqのベストプラクティスに関する比較的最近の論文。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analysis pipelines.

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing (scRNA-seq) methods has created a large variety of experimental and computational pipelines for which best practices have not yet been established. Here, we use simulations based on five scRNA-seq library protocols in combination with nine realistic differential expression (DE) setups to systematically evaluate three mapping, four imputation, seven normalisation and four differential expression testing approaches resulting in ~3000 pipelines, allowing us to also assess interactions among pipeline steps. We find that choices of normalisation and library preparation protocols have the biggest impact on scRNA-seq analyses. Specifically, we find that library preparation determines the ability to detect symmetric expression differences, while normalisation dominates pipeline performance in asymmetric DE-setups. Finally, we illustrate the importance of informed choices by showing that a good scRNA-seq pipeline can have the same impact on detecting a biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-019-12266-7

sc性能評価論文2

scRNA-seqのベストプラクティスに関する比較的最近の論文。ライブラリ調整法からスタートして、マッピング、データ補完、正規化、発現変動までの計3,000パイプラインの性能を評価している。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analy

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing (scRNA-seq) methods has created a large variety of experimental and computational pipelines for which best practices have not yet been established. Here, we use simulations based on five scRNA-seq library protocols in combination with nine realistic differential expression (DE) setups to systematically evaluate three mapping, four imputation, seven normalisation and four differential expression testing approaches resulting in ~3000 pipelines, allowing us to also assess interactions among pipeline steps. We find that choices of normalisation and library preparation protocols have the biggest impact on scRNA-seq analyses. Specifically, we find that library preparation determines the ability to detect symmetric expression differences, while normalisation dominates pipeline performance in asymmetric DE-setups. Finally, we illustrate the importance of informed choices by showing that a good scRNA-seq pipeline can have the same impact on detecting a biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-019-12266-7

sc性能評価論文3

scRNA-seqのベストプラクティスに関する比較的最近の論文。ライブラリ調整法からスタートして、マッピング、データ補完、正規化、発現変動までの計3,000パイプラインの性能を評価している。結論としては、正規化とライブラリ調整でどの方法を選択するかが重要なポイント。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analy

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing (scRNA-seq) methods has created a large variety of experimental and computational pipelines for which best practices have not yet been established. Here, we use simulations based on five scRNA-seq library protocols in combination with nine realistic differential expression (DE) setups to systematically evaluate three mapping, four imputation, seven normalisation and four differential expression testing approaches resulting in ~3000 pipelines, allowing us to also assess interactions among pipeline steps. We find that choices of normalisation and library preparation protocols have the biggest impact on scRNA-seq analyses. Specifically, we find that library preparation determines the ability to detect symmetric expression differences, while normalisation dominates pipeline performance in asymmetric DE-setups. Finally, we illustrate the importance of informed choices by showing that a good scRNA-seq pipeline can have the same impact on detecting a biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-019-12266-7

sc性能評価論文4

scRNA-seqのベストプラクティスに関する比較的最近の論文。ライブラリ調整法からスタートして、マッピング、データ補完、正規化、発現変動までの計3,000パイプラインの性能を評価している。結論としては、正規化とライブラリ調整でどの方法を選択するかが重要なポイント。この論文も2群間比較のみで評価していて、symmetricな発現変動を同定する際にはライブラリ調整が重要で、asymmetricな発現変動を同定する際には正規化が重要。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analy

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing (scRNA-seq) method
large variety of experimental and computational pipelines for which best prac
yet been established. Here, we use simulations based on five scRNA-seq lib
combination with nine realistic differential expression (DE) setups to systema
three mapping, four imputation, seven normalisation and four differential expression testing approaches resulting in ~3000 pipelines, allowing us to also assess interactions among pipeline steps. We find that choices of normalisation and library preparation protocols have the biggest impact on scRNA-seq analyses. Specifically, we find that library preparation determines the ability to detect symmetric expression differences, while normalisation dominates pipeline performance in asymmetric DE-setups. Finally, we illustrate the importance of informed choices by showing that a good scRNA-seq pipeline can have the same impact on detecting a biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-019-12266-7

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

symmetric（対称）1

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analy

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing (scRNA-seq) method
large variety of experimental and computational pipelines for which best prac
yet been established. Here, we use simulations based on five scRNA-seq lib
combination with nine realistic differential expression (DE) setups to systema
three mapping, four imputation, seven normalisation and four differential exp
approaches resulting in ~3000 pipelines, allowing us to also assess interactions among pipeline steps. We find that choices of normalisation and library preparation protocols have the biggest impact on scRNA-seq analyses. Specifically, we find that library preparation determines the ability to detect symmetric expression differences, while normalisation dominates pipeline performance in asymmetric DE-setups. Finally, we illustrate the importance of informed choices by showing that a good scRNA-seq pipeline can have the same impact on detecting a biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-019-12266-7

scRNA-seqのベストプラクティスに関する比較的最近の論文。ライブラリ調整法からスタートして、マッピング、データ補完、正規化、発現変動までの計3,000パイプラインの性能を評価している。結論としては、正規化とライブラリ調整でどの方法を選択するかが重要なポイント。この論文も2群間比較のみで評価していて、symmetricな発現変動を同定する際にはライブラリ調整が重要で、asymmetricな発現変動を同定する際には正規化が重要。「2群間でsymmetricな発現変動」をいくつか例示します。

symmetric（対称）2

10遺伝子×6サンプル（or 細胞）からなる、リードカウントデータ行列またはUMIカウントデータ行列が手元にあり…

	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

symmetric（対称）3

10遺伝子×6サンプル（or 細胞）からなる、リードカウントデータ行列またはUMIカウントデータ行列が手元にあり、G1という状態のグループ（群）とG2という状態のグループ（群）の2群間比較を行いたい場合で考えます。

	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

symmetric（対称）4

このデータは、①発現変動遺伝子（DEG）を含む割合が20%（P_{DEG} = Effect size = 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

① P_{DEG} or Effect size

②

symmetric（対称）5

このデータは、①発現変動遺伝子（DEG）を含む割合が20%（P_{DEG} = Effect size = 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

symmetric（対称）6

このデータは、①発現変動遺伝子（DEG）を含む割合が20%（P_{DEG} = Effect size = 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）。このようなDEG数がG1群とG2群で同程度という状態をsymmetricと表現しています。

	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

①

symmetric（対称）7

このデータは、①発現変動遺伝子（DEG）を含む割合が20%（P_{DEG} = Effect size = 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）。このようなDEG数がG1群とG2群で同程度という状態をsymmetricと表現しています。さきほどの性能評価論文では、DEGの割合のことをEffect sizeと表現しており、この場合はEffect size = 20%となります。

	G1群			G2群		
	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G
gene1	14	110	73	0	0	
gene2	1	0	11	12	539	546
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

symmetric（対称）8

このようなデータの場合は、Effect size = 40%となります。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）なので、もちろんsymmetricなデータです。

	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

①

symmetric（対称）9

このようなデータの場合は、Effect size = 60%となります。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）なので、もちろんsymmetricなデータです。

	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

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

asymmetric（非対称）1

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analy

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing (scRNA-seq) method
large variety of experimental and computational pipelines for which best prac
yet been established. Here, we use simulations based on five scRNA-seq lib
combination with nine realistic differential expression (DE) setups to systema
three mapping, four imputation, seven normalisation and four differential exp
approaches resulting in ~3000 pipelines, allowing us to also assess interactions among pipeline steps. We find that choices of normalisation and library preparation protocols have the biggest impact on scRNA-seq analyses. Specifically, we find that library preparation determines the ability to detect symmetric expression differences, while normalisation dominates pipeline performance in asymmetric DE-setups. Finally, we illustrate the importance of informed choices by showing that a good scRNA-seq pipeline can have the same impact on detecting a biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-019-12266-7

scRNA-seqのベストプラクティスに関する比較的最近の論文。ライブラリ調整法からスタートして、マッピング、データ補完、正規化、発現変動までの計3,000パイプラインの性能を評価している。結論としては、正規化とライブラリ調整でどの方法を選択するかが重要なポイント。この論文も2群間比較のみで評価していて、symmetricな発現変動を同定する際にはライブラリ調整が重要で、asymmetricな発現変動を同定する際には正規化が重要。「2群間でasymmetricな発現変動」をいくつか例示します。

asymmetric（非対称）2

①DEGを含む割合が20%（P_{DEG} = Effect size = 0.2）で、②残りはnon-DEG。

	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

①

②

asymmetric（非対称）3

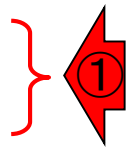
①DEGを含む割合が20%（P_{DEG} = Effect size = 0.2）で、②残りはnon-DEG。①全てのDEGがG1群で高発現のような状態をasymmetricといいます。

	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

asymmetric（非対称）4

①DEGを含む割合が20%（P_{DEG} = Effect size = 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状態をasymmetricといいます。さきほどの性能評価論文では、特に全てのDEGがどちらかの群に偏っている場合をcompletely asymmetricと表現しています。

	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



asymmetric（非対称）5

このようなデータの場合は、Effect size = 40%となります。①全てのDEGがG1群で高発現なので、completely asymmetric なデータです。

	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

asymmetric（非対称）6

このようなデータの場合は、Effect size = 60%となります。①全てのDEGがG1群で高発現なので、completely asymmetric なデータです。

	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

①

asymmetric（非対称）7

このようなデータの場合は、Effect size = 60%となります。①全てのDEGがG1群で高発現というではないので、asymmetricなデータです。

	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

①

②

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

sc性能評価論文再訪1

①scRNA-seqのベストプラクティスに関する比較的最近の論文の、②Fig. 1。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analysis pipelines.

①

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

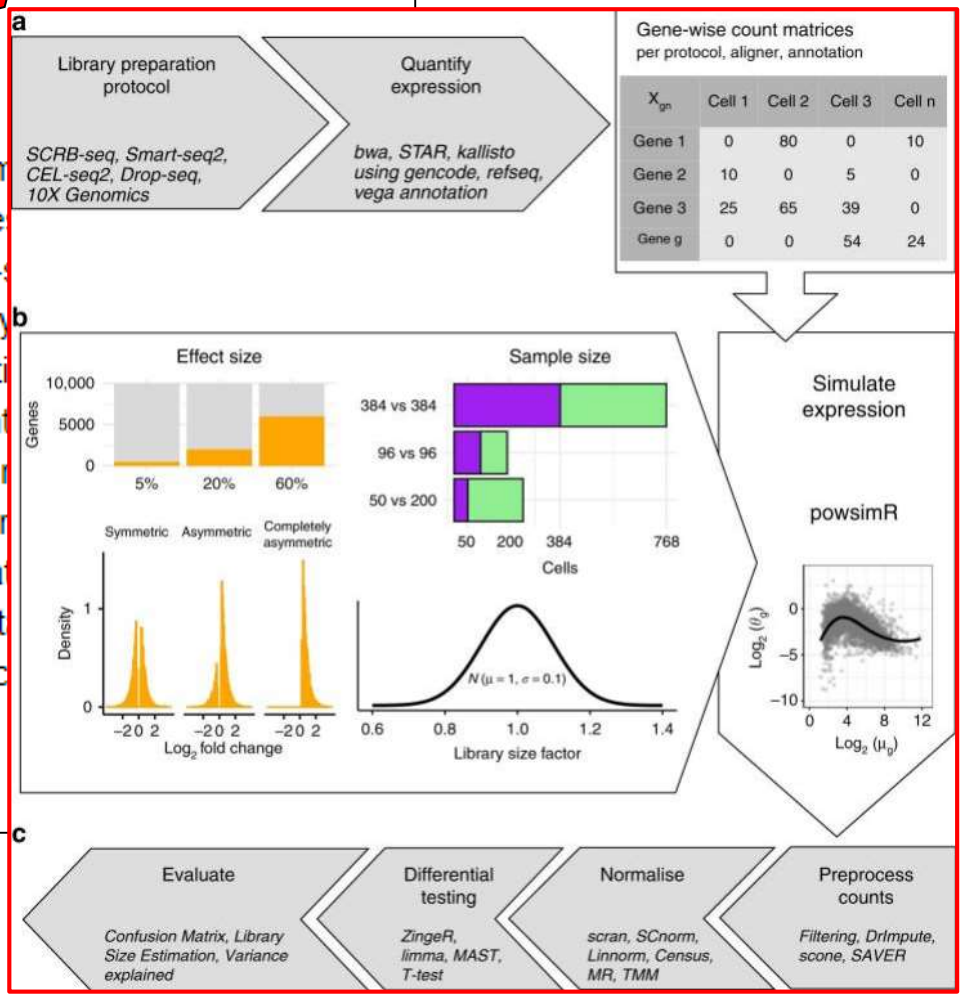
Author information

②

Abstract

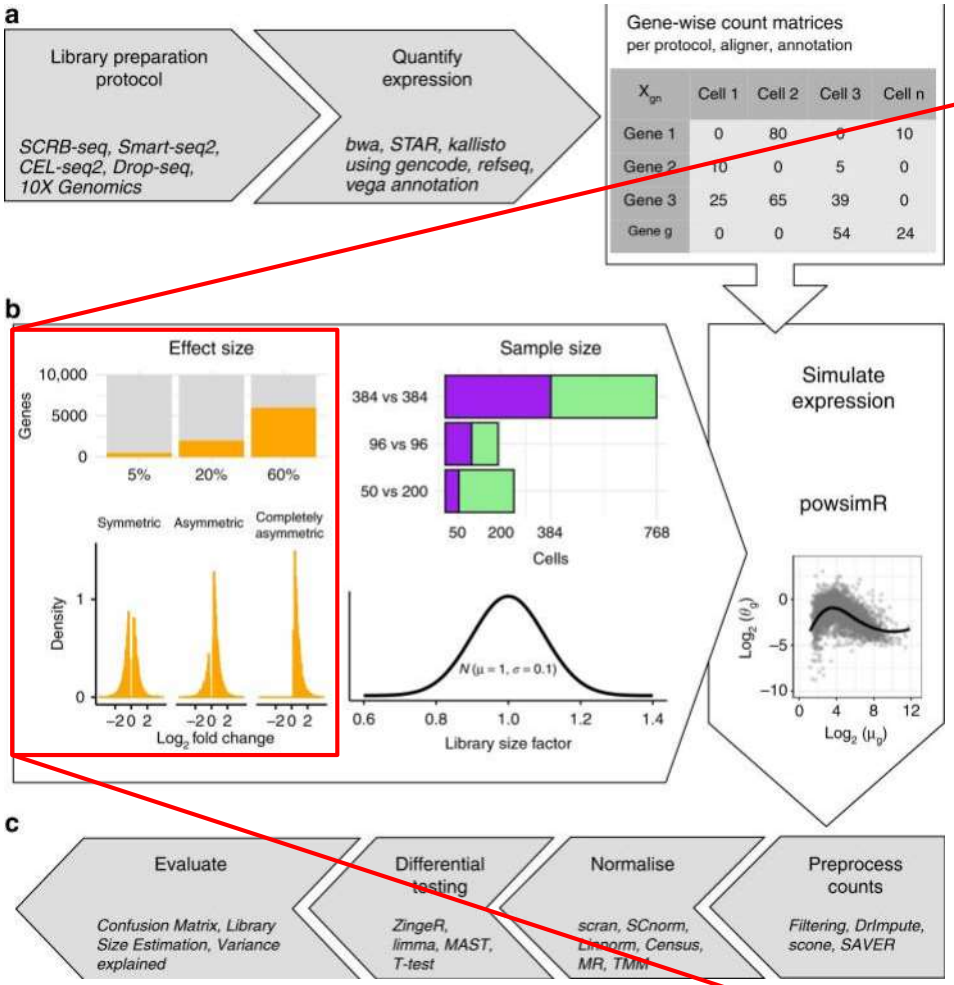
The recent rapid spread of single cell RNA sequencing (scRNA-seq) m
large variety of experimental and computational pipelines for which be
yet been established. Here, we use simulations based on five scRNA-
combination with nine realistic differential expression (DE) setups to sy
three mapping, four imputation, seven normalisation and four differenti
approaches resulting in ~3000 pipelines, allowing us to also assess int
pipeline steps. We find that choices of normalisation and library prepar
biggest impact on scRNA-seq analyses. Specifically, we find that librar
the ability to detect symmetric expression differences, while normalisat
performance in asymmetric DE-setups. Finally, we illustrate the import
by showing that a good scRNA-seq pipeline can have the same impac
biological signal as quadrupling the sample size.

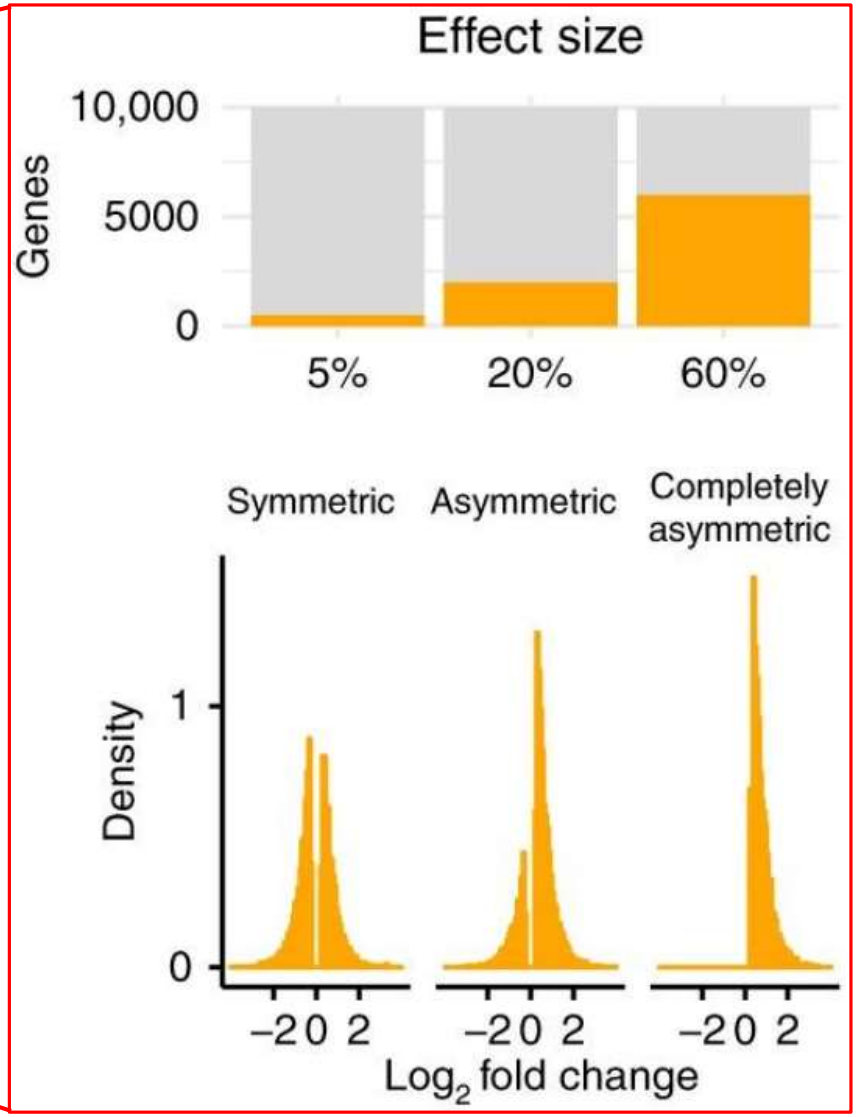
PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-019-12266-7



sc性能評価論文再訪2

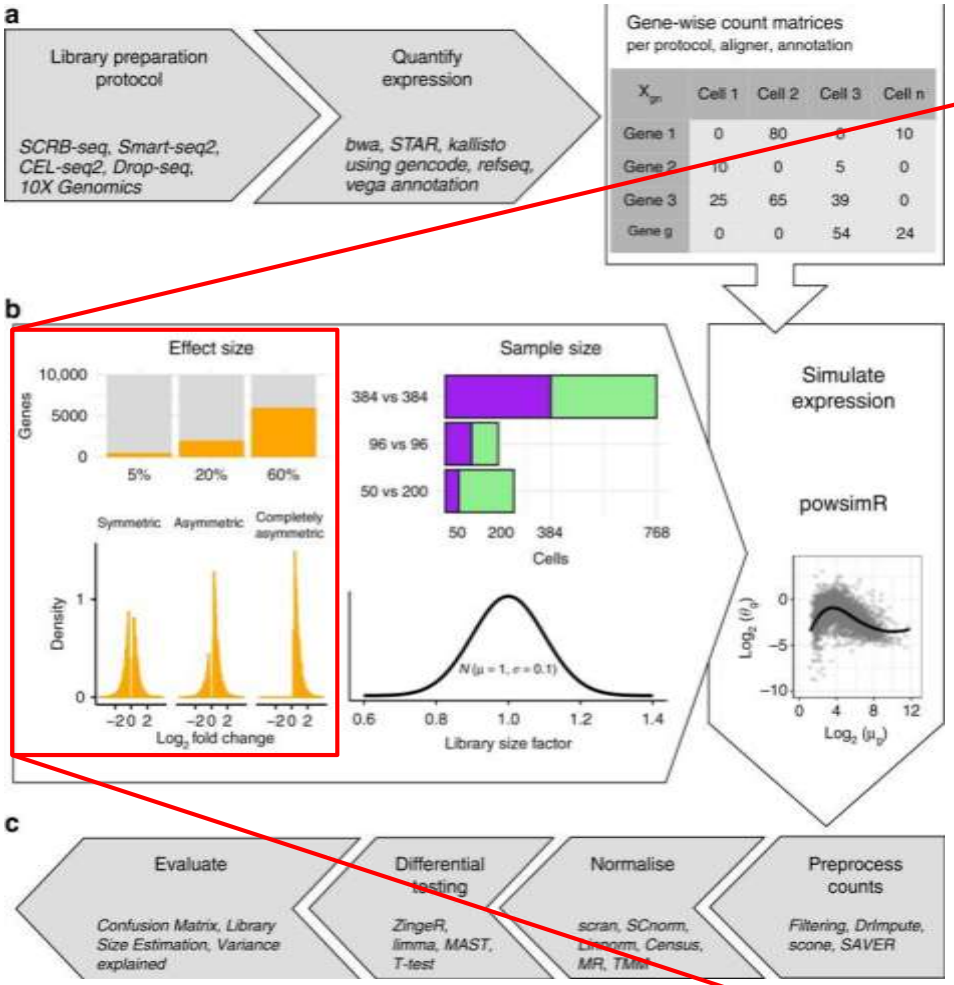
①scRNA-seqのベストプラクティスに関する比較的最近の論文の、②Fig. 1。赤枠はシミュレーション条件に関する説明。

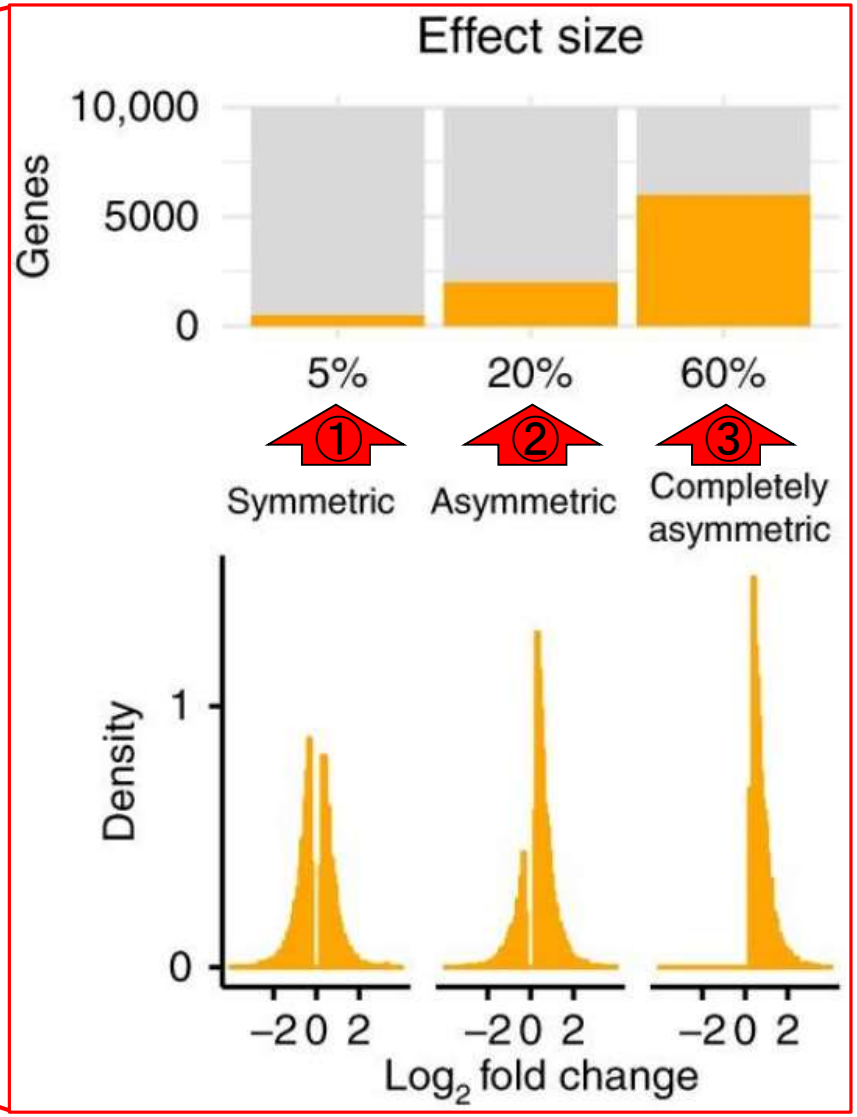




SC性能評価論文再訪3

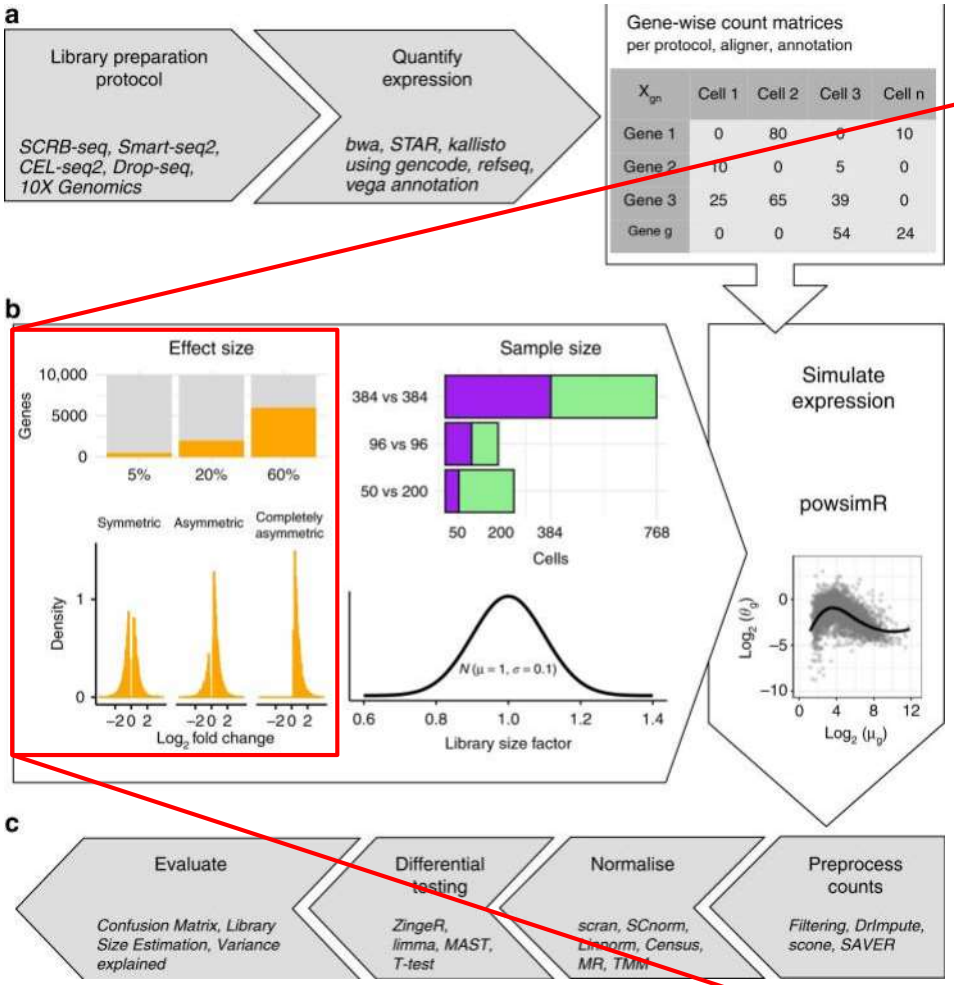
全10,000遺伝子中、DEGの割合（Effect size）が、①5%（= 500個）、②20%（2,000個）、③60%（6,000個）という3条件で実行。

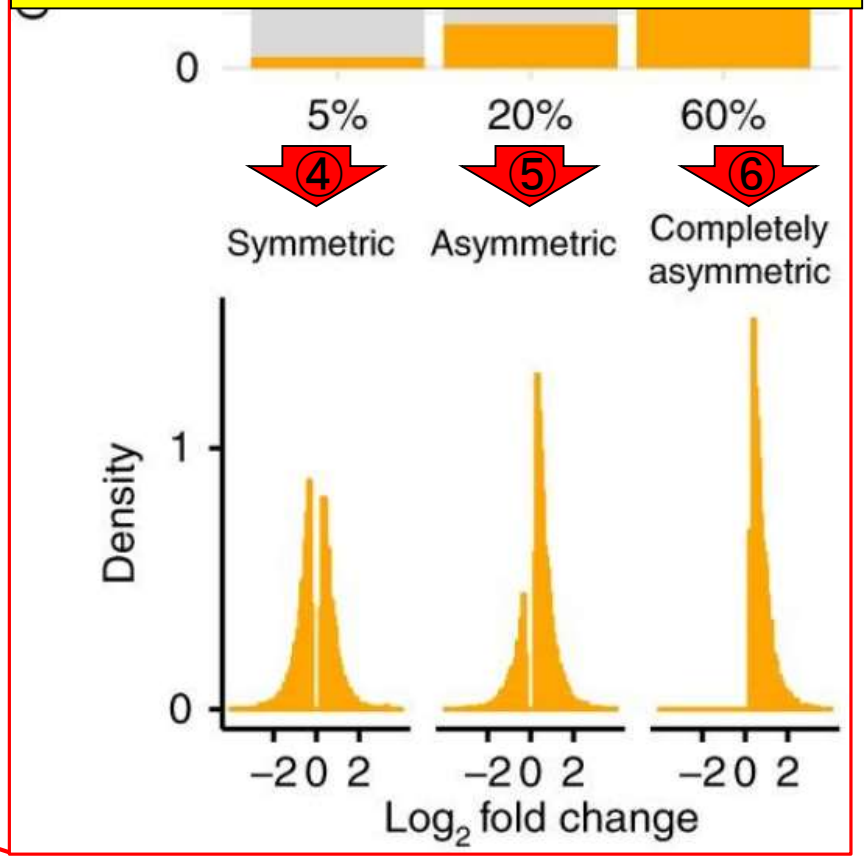




sc性能評価論文再訪4

全10,000遺伝子中、DEGの割合（Effect size）が、①5%（= 500個）、②20%（2,000個）、③60%（6,000個）という3条件で実行。DEG数がG1群とG2群で、④同程度という状態をsymmetric、⑤どちらかの群に偏っている状態をasymmetric、そして⑥完全に偏っている状態をcompletely asymmetricと表現しています。





sc性能評価論文再訪5

①scRNA-seqのベストプラクティスに関する比較的最近の論文の、②Introductionの一部抜粋。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analysis pipelines.

①

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

②

Abstract

The recent rapid spread of single cell RNA sequencing
large variety of experimental and computational pipelir
yet been established. Here, we use simulations based
combination with nine realistic differential expression (
three mapping, four imputation, seven normalisation ar
approaches resulting in ~3000 pipelines, allowing us to
pipeline steps. We find that choices of normalisation ar
biggest impact on scRNA-seq analyses. Specifically, w
the ability to detect symmetric expression differences,
performance in asymmetric DE-setups. Finally, we illus
by showing that a good scRNA-seq pipeline can have
biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

sc性能評価論文再訪6

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

①scRNA-seqのベストプラクティスに関する比較的最近の論文の、②Introductionの一部抜粋。旧来の発現変動解析はsymmetricデータを仮定している。

sc性能評価論文再訪7

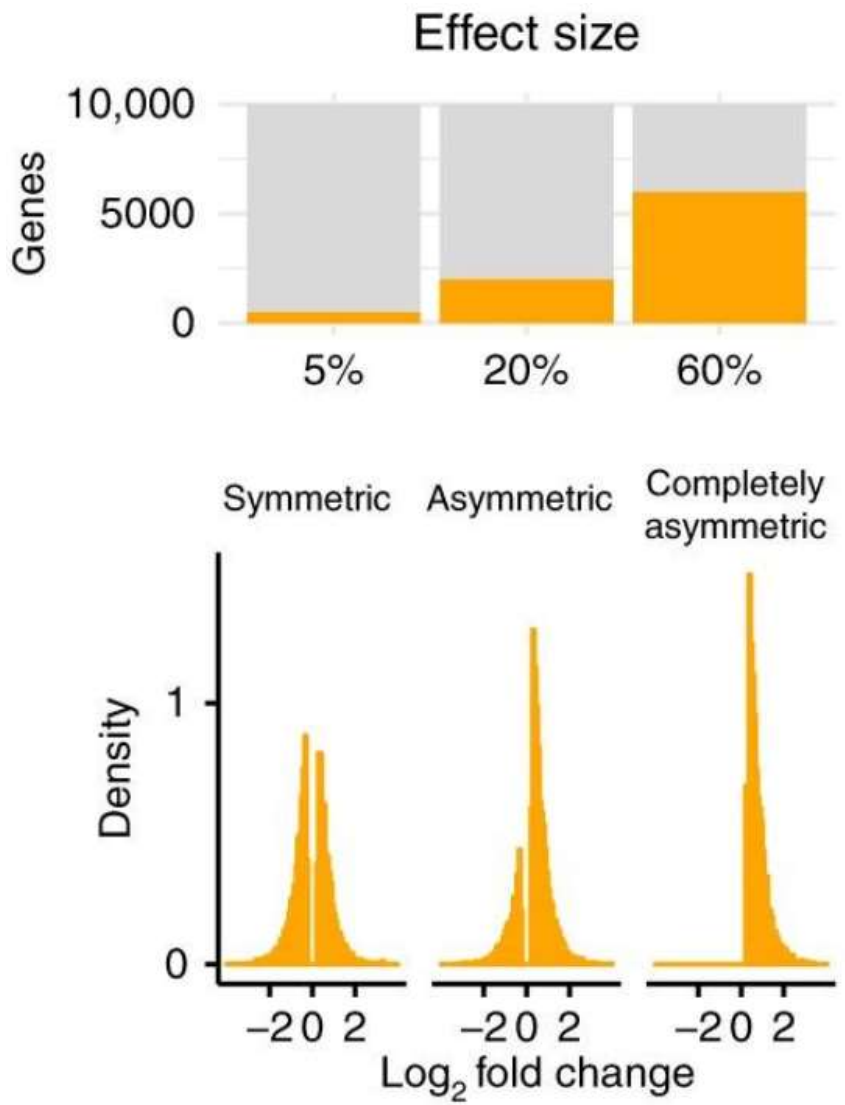
One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

①scRNA-seqのベストプラクティスに関する比較的最近の論文の、②Introductionの一部抜粋。旧来の発現変動解析はsymmetricデータを仮定している。DEGの割合が少なくnonDEGが大多数（つまりEffect sizeが小さい）、または特定の群で高発現なものと低発現なものが同程度（つまりsymmetric）だということを意味する。

sc性能評価論文再訪8

赤下線で言いたいことは、旧来の発現変動解析は…

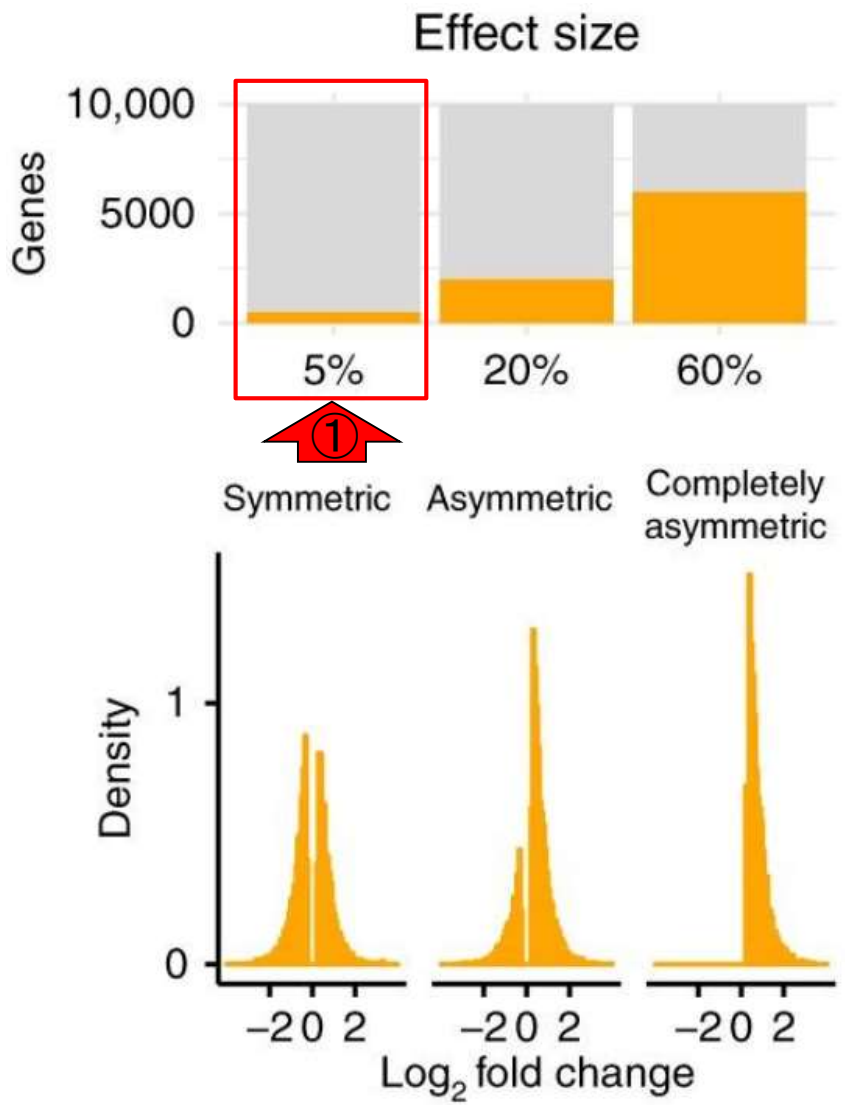
One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.



sc性能評価論文再訪9

赤下線で言いたいことは、旧来の発現変動解析は、①Effect sizeが小さく（データに占めるDEG数は多くない）、

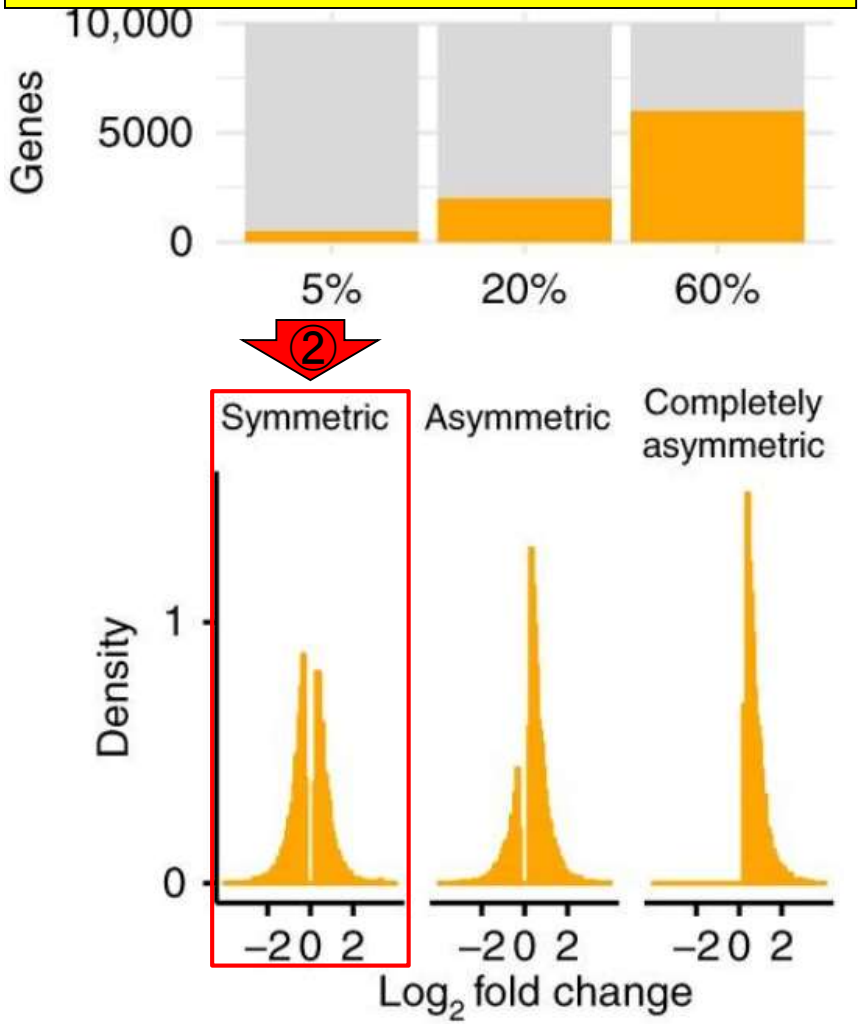
One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.



sc性能評価論文再訪10

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

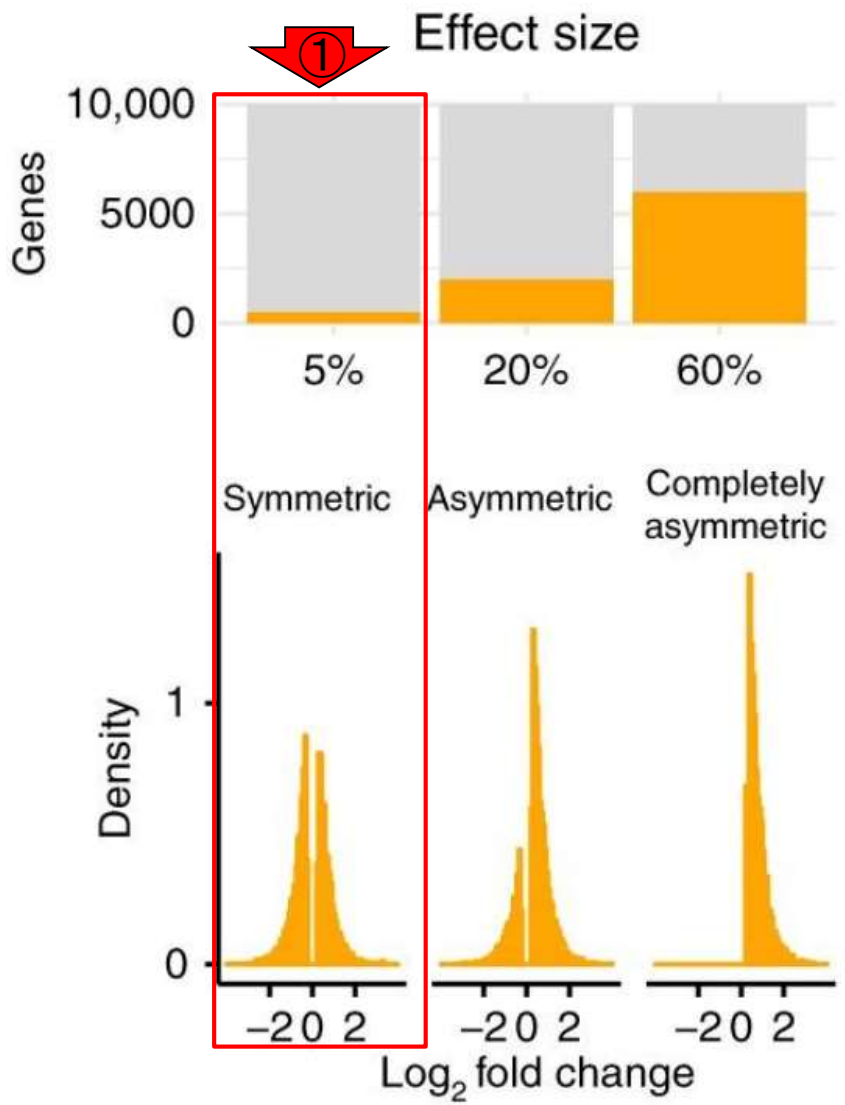
赤下線で言いたいことは、旧来の発現変動解析は、①Effect sizeが小さく（データに占めるDEG数は多くない）、②特定の群で高発現なものと低発現なものが同程度（つまりsymmetric）ということ。



sc性能評価論文再訪11

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

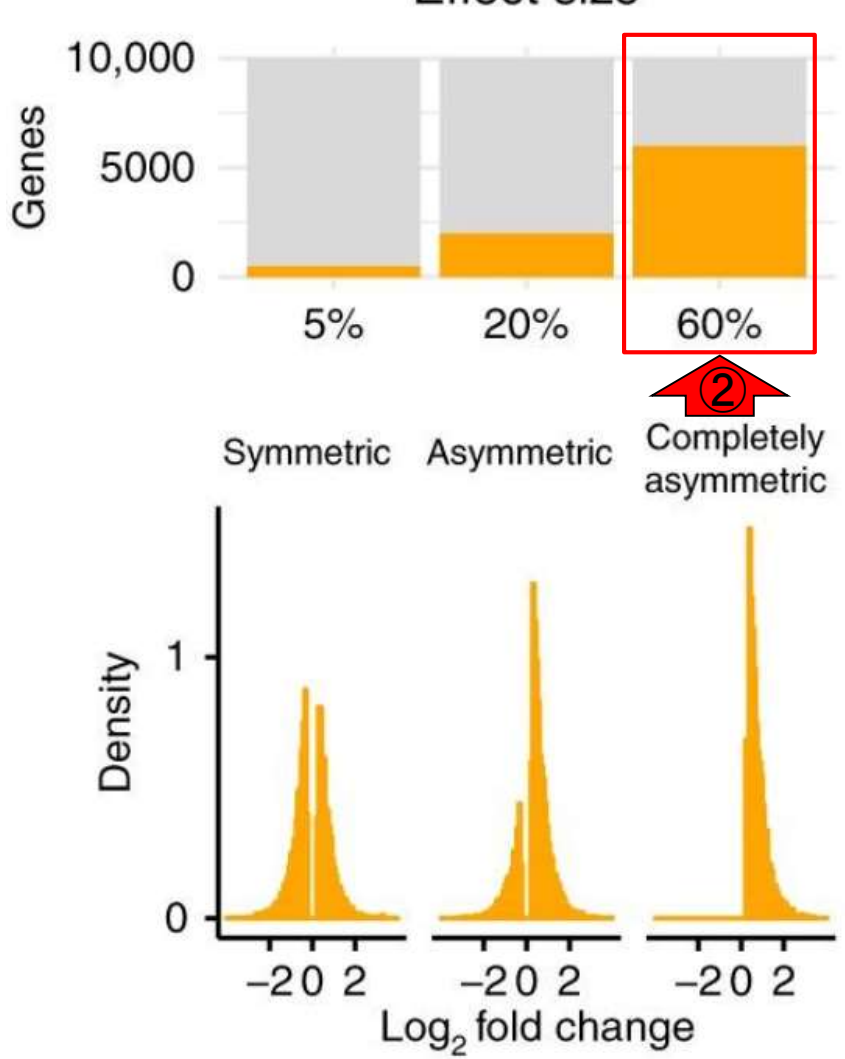
多様なシングルセルを解析する時代においては、①赤枠のような仮定はもはや時代遅れだ。



sc性能評価論文再訪12

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

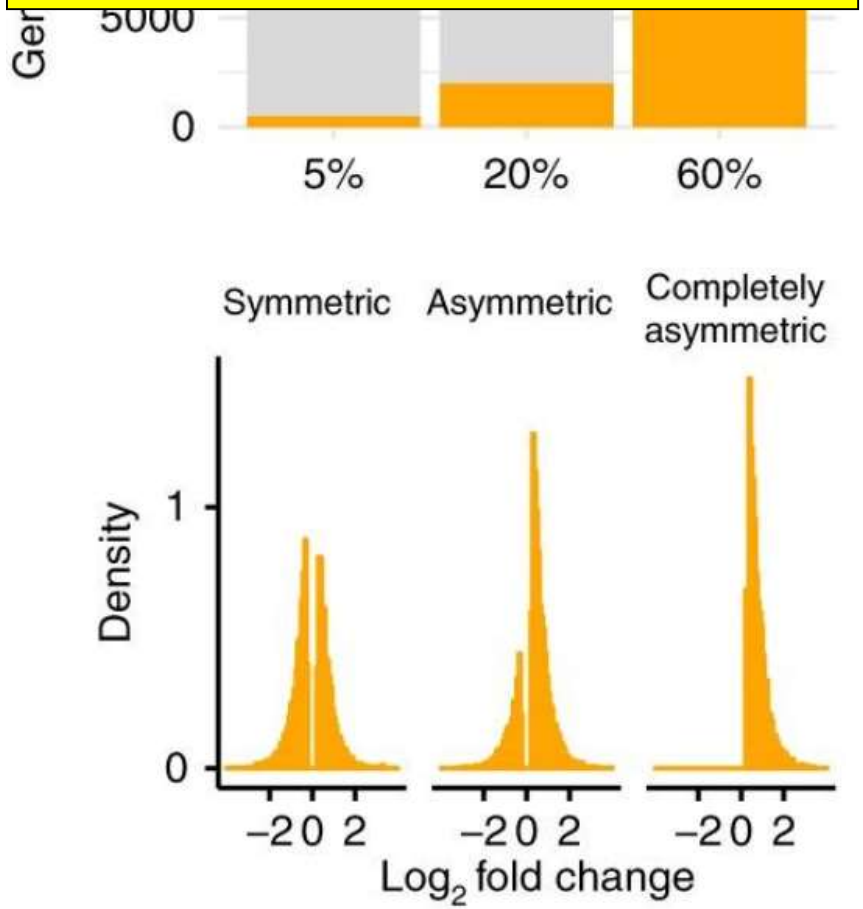
多様なシングルセルを解析する時代においては、①赤枠のような仮定はもはや時代遅れだ。②こういうscRNA-seqデータも実際にはある。



sc性能評価論文再訪13

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

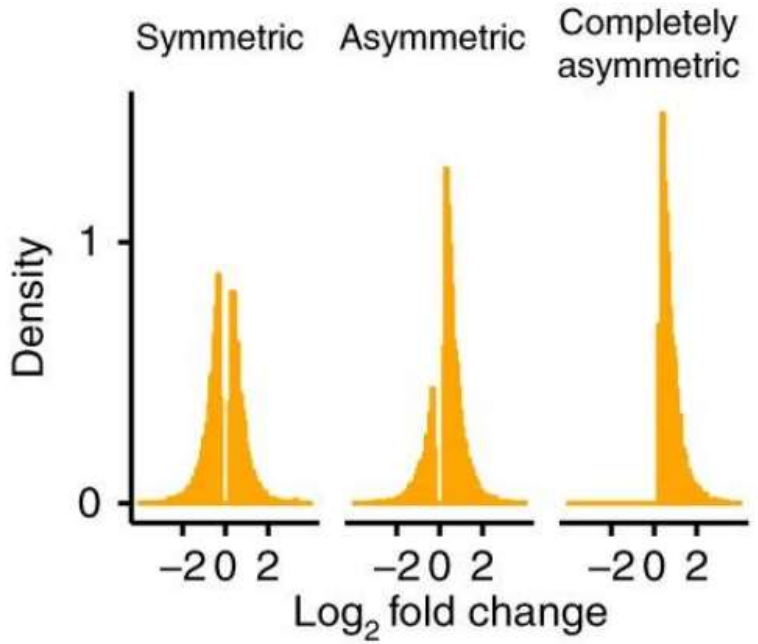
多様なシングルセルを解析する時代においては、①赤枠のような仮定はもはや時代遅れだ。②こういうscRNA-seqデータも実際にはある。このような非対称性（asymmetry）の問題の有無が、scRNA-seqとbulk RNA-seqとの特徴の違いであり、これまで対処されてこなかった。



sc性能評価論文再訪14

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

多様なシングルセルを解析する時代においては、①赤枠のような仮定はもはや時代遅れだ。②こういうscRNA-seqデータも実際にはある。このような非対称性（asymmetry）の問題の有無が、scRNA-seqとbulk RNA-seqとの特徴の違いであり、これまで対処されてこなかった。それゆえ、（2019年10月に出た）本研究では、2群間比較で様々なシミュレーション条件で性能評価を行った。



sc性能評価論文再訪15

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with s① to large differences in mRNA content including the entire spectrum of possible DE-settings.

多様なシングルセルを解析する時代においては、①赤枠のような仮定はもはや時代遅れだ。②こういうscRNA-seqデータも実際にはある。このような非対称性（asymmetry）の問題の有無が、scRNA-seqとbulk RNA-seqとの特徴の違いであり、これまで対処されてこなかった。それゆえ、（2019年10月に出た）本研究では、2群間比較で様々なシミュレーション条件で性能評価を行った。しかしながら、①Thereforeにかかる前の、前提が間違っている。

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

DEGES論文1

（当時はbulkという表現はなかったが）①
2012年の時点でbulk RNA-seq用のasymmetryに対応した正規化法に関する論文が出ている。

Algorithms Mol Biol. 2012 Apr 5;7(1):5. doi: 10.1186/1748-7188-7-5.

A normalization strategy for comparing tag count data.

Kadota K[1], Nishiyama T, Shimizu K.

Author information

Abstract

BACKGROUND: High-throughput sequencing, such as ribonucleic acid sequencing (RNA-seq) and chromatin immunoprecipitation sequencing (ChIP-seq) analyses, enables various features of organisms to be compared through tag counts. Recent studies have demonstrated that the normalization step for RNA-seq data is critical for a more accurate subsequent analysis of differential gene expression. Development of a more robust normalization method is desirable for identifying the true difference in tag count data.

RESULTS: We describe a strategy for normalizing tag count data, focusing on RNA-seq. The key concept is to remove data assigned as potential differentially expressed genes (DEGs) before calculating the normalization factor. Several R packages for identifying DEGs are currently available, and each package uses its own normalization method and gene ranking algorithm. We compared a total of eight package combinations: four R packages (edgeR, DESeq, baySeq, and NBPSeq) with their default normalization settings and with our normalization strategy. Many synthetic datasets under various scenarios were evaluated on the basis of the area under the curve (AUC) as a measure for both sensitivity and specificity. We found that packages using our strategy in the data normalization step overall performed well. This result was also observed for a real experimental dataset.

CONCLUSION: Our results showed that the elimination of potential DEGs is essential for more accurate normalization of RNA-seq data. The concept of this normalization strategy can widely be applied to other types of tag count data and to microarray data.

PMID: 22475125 PMCID: PMC3341196 DOI: 10.1186/1748-7188-7-5

DEGES論文2

（当時はbulkという表現はなかったが）①2012年の時点でbulk RNA-seq用のasymmetryに対応した正規化法に関する論文が出ている。②Fig. 1aを反時計回りに90度回転させたもの。

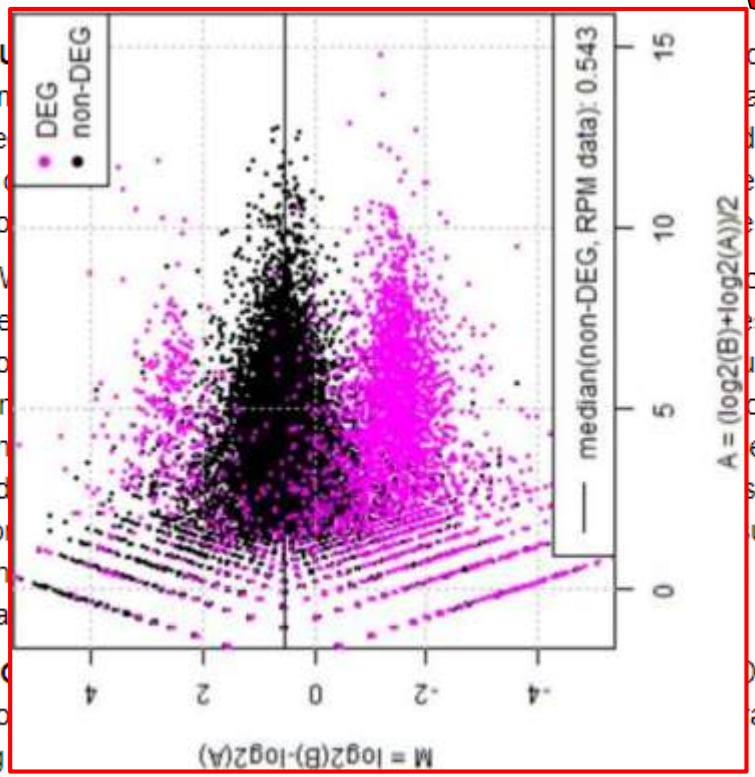
Algorithms Mol Biol. 2012 Apr 5;7(1):5. doi: 10.1186/1748-7188-7-5.

A normalization strategy for comparing tag count data.

Kadota K[1], Nishiyama T, Shimizu K.

Author information

②



Abstract

BACKGROUcid sequencing (RNA-seq) and chromatin in ables various features of organisms to be compare that the normalization step for RNA-seq data is ential gene expression. Development of a more ro true difference in tag count data.

RESULTS: V cusing on RNA-seq. The key concept is to remove es (DEGs) before calculating the normalizatio urrently available, and each package uses its own compared a total of eight package combination eq) with their default normalization settings and s under various scenarios were evaluated o ure for both sensitivity and specificity. We found th step overall performed well. This result was a

CONCLUSIOEGs is essential for more accurate normalizatio ategy can widely be applied to other types of tag

PMID: 22475125 PMCID: PMC3341196 DOI: 10.1186/1748-7188-7-5

DEGES論文3

（当時はbulkという表現はなかったが）①2012年の時点でbulk RNA-seq用のasymmetryに対応した正規化法に関する論文が出ている。②Fig. 1aを反時計回りに90度回転させたもの。③よりも④のほうにDEGが偏っていることがわかる。これがasymmetryと同じ意味。

Algorithms Mol Biol. 2012 Apr 5;7(1):5. doi: 10.1186/1748-7188-7-5.

A normalization strategy for comparing tag count data.

Kadota K[1], Nishiyama T, Shimizu K.

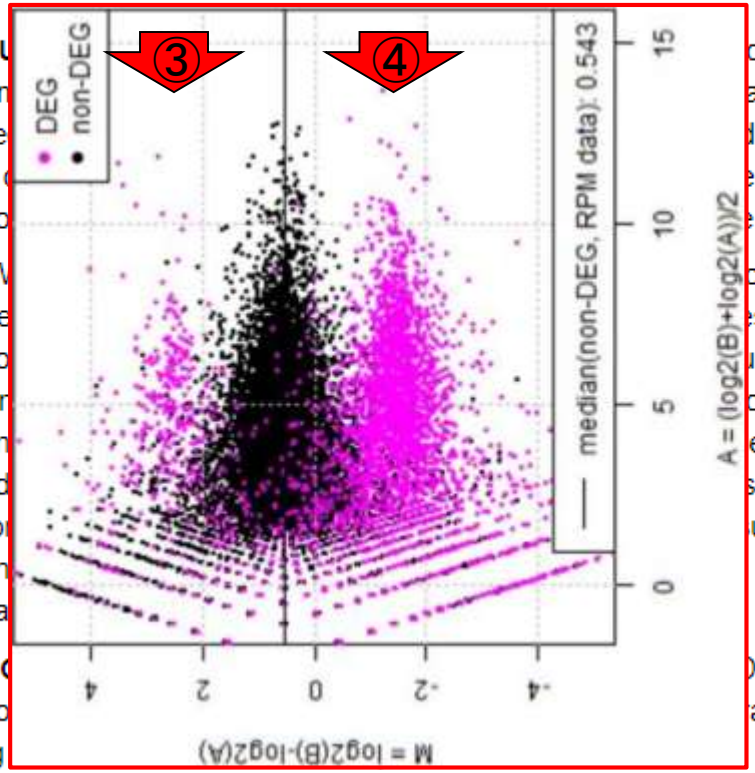
Author information

Abstract

BACKGROUcid sequencing (RNA-seq) and chromatin in ables various features of organisms to be compare that the normalization step for RNA-seq data is ential gene expression. Development of a more ro true difference in tag count data.

RESULTS: V cusing on RNA-seq. The key concept is to remove es (DEGs) before calculating the normalizatio urrently available, and each package uses its own compared a total of eight package combination eq) with their default normalization settings and s under various scenarios were evaluated o ure for both sensitivity and specificity. We found th step overall performed well. This result was a

CONCLUSIOEGs is essential for more accurate normalizatio ategy can widely be applied to other types of tag



PMID: 22475125 PMCID: PMC3341196 DOI: 10.1186/1748-7188-7-5

DEGES論文4

Algorithms Mol Biol. 2012 Apr 5;7(1):5. doi: 10.1186/1748-7188-7-5.

A normalization strategy for comparing tag count data.

Kadota K[1], Nishiyama T, Shimizu K.

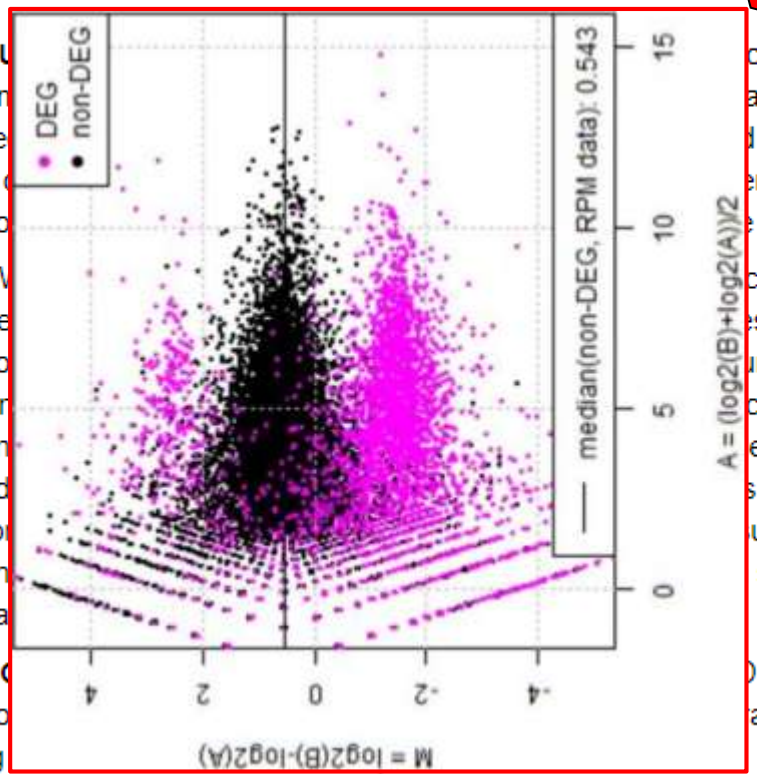
Author information

Abstract

BACKGROUcid sequencing (RNA-s
chromatin in ables various features
be compare d that the normalization
seq data is ential gene expression.
of a more ro e true difference in tag

RESULTS: V cusing on RNA-seq. Th
is to remove s (DEGs) before calcul
normalizatio urrently available, and e
uses its own compared a total of eig
combination eq) with their default n
settings and s under various scenari
evaluated o ure for both sensitivity
We found th step overall performed
result was a

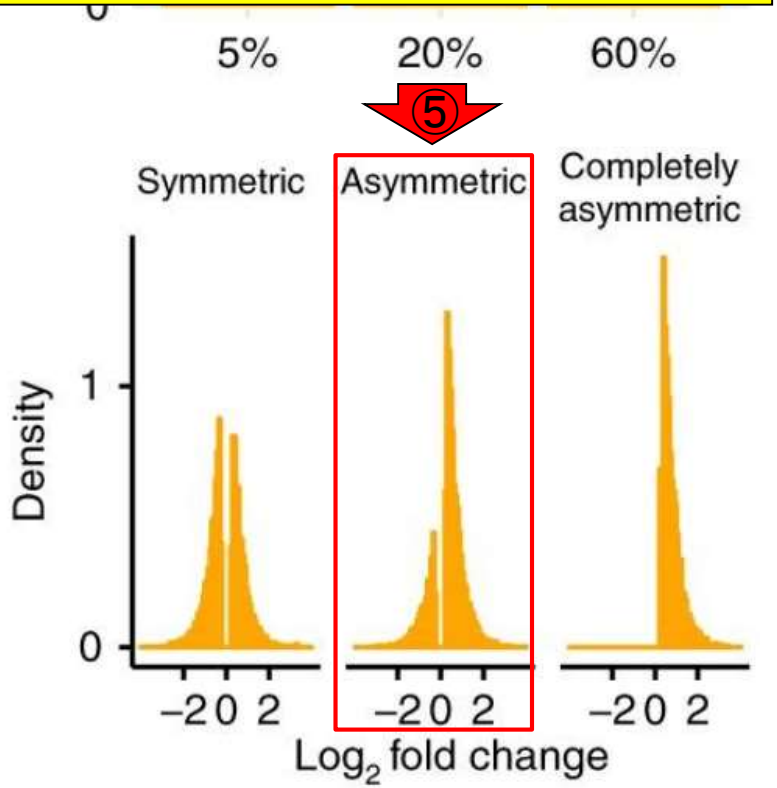
CONCLUSIEGs is essential for m
normalizatio ategy can widely be ap
types of tag



②

PMID: 22475125 PMCID: PMC3341196 DOI: 10.1186/1748-7188-7-5

（当時はbulkという表現はなかったが）①2012年の時点でbulk RNA-seq用のasymmetryに対応した正規化法に関する論文が出ている。②Fig. 1aを反時計回りに90度回転させたもの。③よりも④のほうにDEGが偏っていることがわかる。これがasymmetryと同じ意味。つまり、bulk RNA-seqの②は、scRNA-seqの⑤と本質的に同じ条件だということです。



DEGES論文5

①Effect sizeが小さい（DEGが少ない）という条件下では、②その内訳がどんな感じであろうと正規化に与える影響は軽微

Algorithms Mol Biol. 2012 Apr 5;7(1):5. doi: 10.1186/1748-7188-7-5.

A normalization strategy for comparing tag count data.

Kadota K[1], Nishiyama T, Shimizu K.

Author information

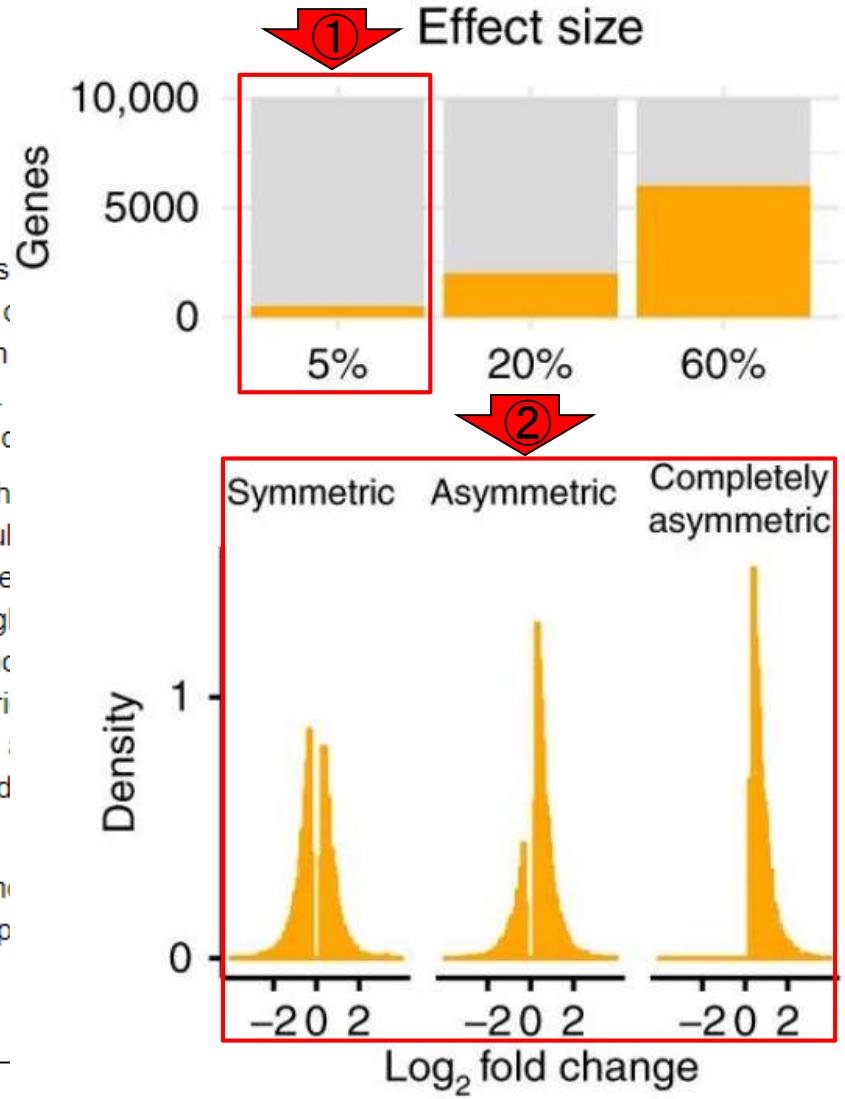
Abstract

BACKGROUND: High-throughput sequencing, such as ribonucleic acid sequencing (RNA-s
chromatin immunoprecipitation sequencing (ChIP-seq) analyses, enables various features c
be compared through tag counts. Recent studies have demonstrated that the normalization
seq data is critical for a more accurate subsequent analysis of differential gene expression.
of a more robust normalization method is desirable for identifying the true difference in tag c

RESULTS: We describe a strategy for normalizing tag count data, focusing on RNA-seq. Th
is to remove data assigned as potential differentially expressed genes (DEGs) before calcul
normalization factor. Several R packages for identifying DEGs are currently available, and e
uses its own normalization method and gene ranking algorithm. We compared a total of eig
combinations: four R packages (edgeR, DESeq, baySeq, and NBPSeq) with their default nc
settings and with our normalization strategy. Many synthetic datasets under various scenari
evaluated on the basis of the area under the curve (AUC) as a measure for both sensitivity
We found that packages using our strategy in the data normalization step overall performed
result was also observed for a real experimental dataset.

CONCLUSION: Our results showed that the elimination of potential DEGs is essential for m
normalization of RNA-seq data. The concept of this normalization strategy can widely be ap
types of tag count data and to microarray data.

PMID: 22475125 PMCID: PMC3341196 DOI: 10.1186/1748-7188-7-5



DEGES論文6

①Effect sizeが大きい（DEGが多い）という条件下であっても、②symmetricであれば、既存の正規化法でもうまくいく。

Algorithms Mol Biol. 2012 Apr 5;7(1):5. doi: 10.1186/1748-7188-7-5.

A normalization strategy for comparing tag count data.

Kadota K[1], Nishiyama T, Shimizu K.

Author information

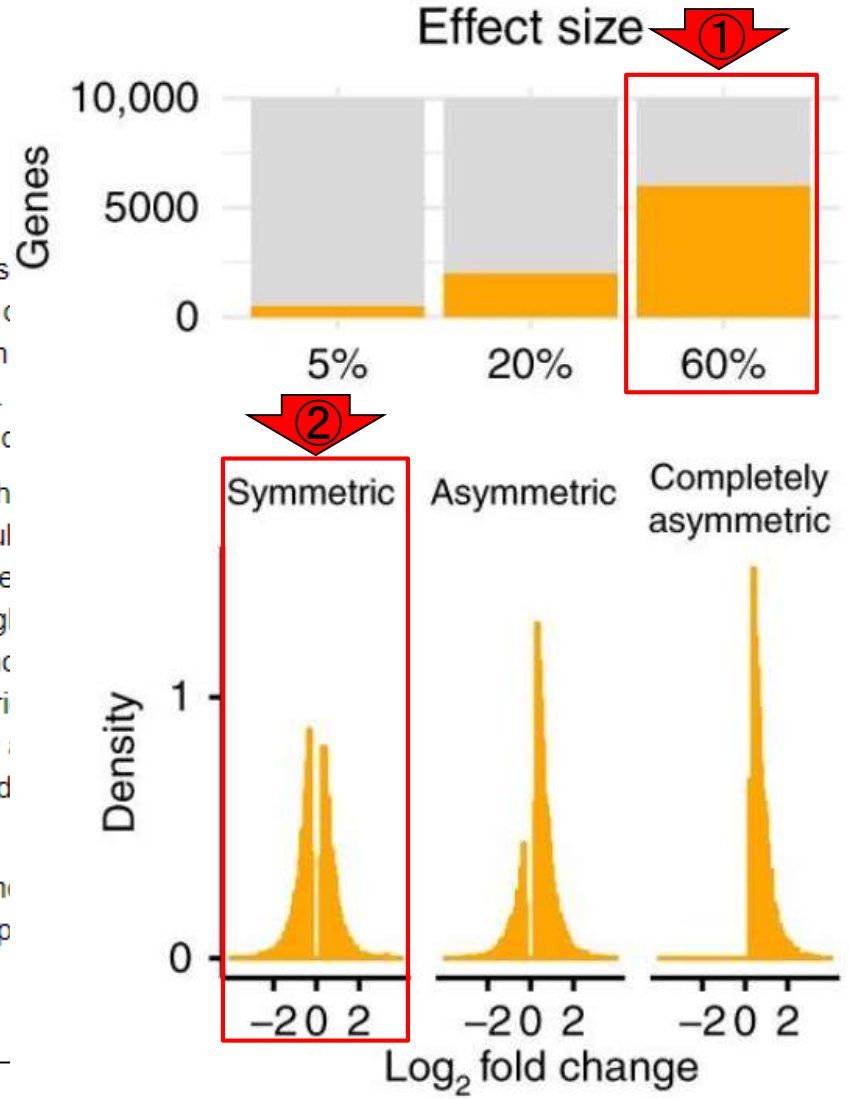
Abstract

BACKGROUND: High-throughput sequencing, such as ribonucleic acid sequencing (RNA-s chromatin immunoprecipitation sequencing (ChIP-seq) analyses, enables various features c be compared through tag counts. Recent studies have demonstrated that the normalization seq data is critical for a more accurate subsequent analysis of differential gene expression. of a more robust normalization method is desirable for identifying the true difference in tag c

RESULTS: We describe a strategy for normalizing tag count data, focusing on RNA-seq. Th is to remove data assigned as potential differentially expressed genes (DEGs) before calcul normalization factor. Several R packages for identifying DEGs are currently available, and e uses its own normalization method and gene ranking algorithm. We compared a total of eig combinations: four R packages (edgeR, DESeq, baySeq, and NBPSeq) with their default nc settings and with our normalization strategy. Many synthetic datasets under various scenari evaluated on the basis of the area under the curve (AUC) as a measure for both sensitivity We found that packages using our strategy in the data normalization step overall performed result was also observed for a real experimental dataset.

CONCLUSION: Our results showed that the elimination of potential DEGs is essential for m normalization of RNA-seq data. The concept of this normalization strategy can widely be ap types of tag count data and to microarray data.

PMID: 22475125 PMCID: PMC3341196 DOI: 10.1186/1748-7188-7-5



DEGES論文7

Algorithms Mol Biol. 2012 Apr 5;7(1):5. doi: 10.1186/1748-7188-7-5.

A normalization strategy for comparing tag count data.

Kadota K[1], Nishiyama T, Shimizu K.

Author information

Abstract

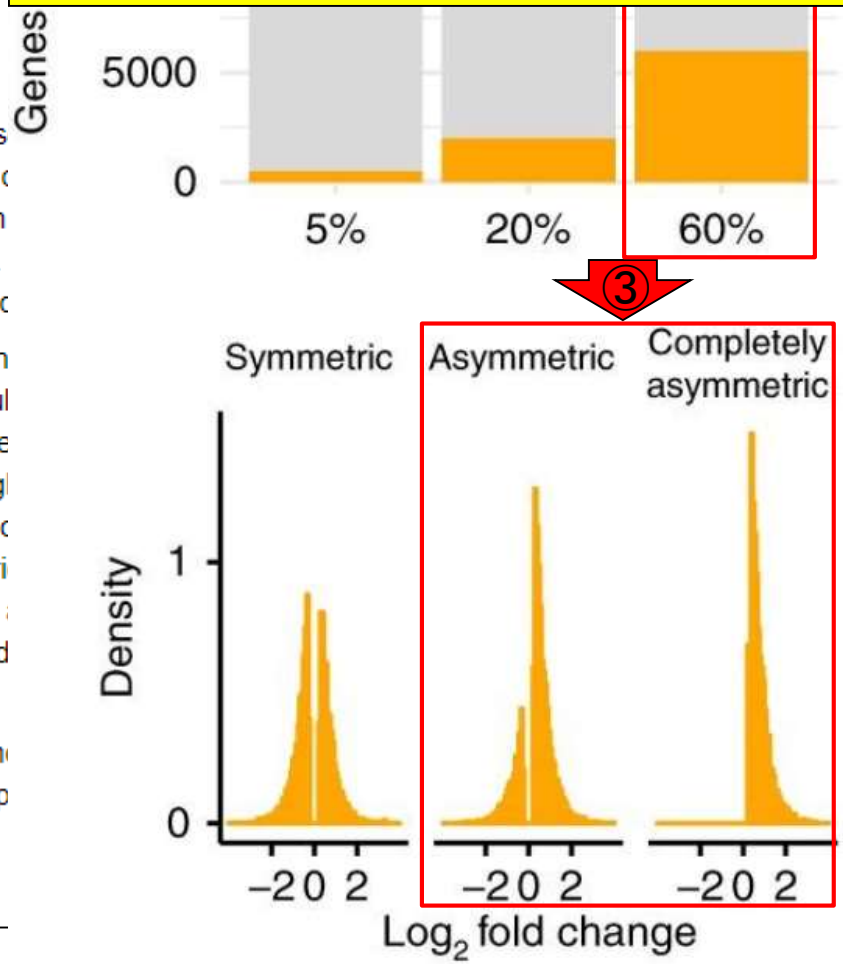
BACKGROUND: High-throughput sequencing, such as ribonucleic acid sequencing (RNA-s
chromatin immunoprecipitation sequencing (ChIP-seq) analyses, enables various features
be compared through tag counts. Recent studies have demonstrated that the normalization
seq data is critical for a more accurate subsequent analysis of differential gene expression.
of a more robust normalization method is desirable for identifying the true difference in tag

RESULTS: We describe a strategy for normalizing tag count data, focusing on RNA-seq. Th
is to remove data assigned as potential differentially expressed genes (DEGs) before calcul
normalization factor. Several R packages for identifying DEGs are currently available, and e
uses its own normalization method and gene ranking algorithm. We compared a total of eig
combinations: four R packages (edgeR, DESeq, baySeq, and NBPSeq) with their default n
settings and with our normalization strategy. Many synthetic datasets under various scenari
evaluated on the basis of the area under the curve (AUC) as a measure for both sensitivity
We found that packages using our strategy in the data normalization step overall performed
result was also observed for a real experimental dataset.

CONCLUSION: Our results showed that the elimination of potential DEGs is essential for m
normalization of RNA-seq data. The concept of this normalization strategy can widely be ap
types of tag count data and to microarray data.

PMID: 22475125 PMCID: PMC3341196 DOI: 10.1186/1748-7188-7-5

①Effect sizeが大きい（DEGが多い）という条件下であっても、②symmetricであれば、既存の正規化法でもうまくいく。しかし、③asymmetricな状況では「DEGの存在自体が正確な正規化を阻み、結果的にうまくDEGが検出できない」状況になる。



DEGES論文8

①Effect sizeが大きい（DEGが多い）という条件下であっても、②symmetricであれば、既存の正規化法でもうまくいく。しかし、③asymmetricな状況では「DEGの存在自体が正確な正規化を阻み、結果的にうまくDEGが検出できない」状況になる。
じゃあどうすればいいか？

Algorithms Mol Biol. 2012 Apr 5;7(1):5. doi: 10.1186/1748-7188-7-5.

A normalization strategy for comparing tag count data.

Kadota K[1], Nishiyama T, Shimizu K.

Author information

Abstract

BACKGROUND: High-throughput sequencing, such as ribonucleic acid sequencing (RNA-seq) and chromatin immunoprecipitation sequencing (ChIP-seq) analyses, enables various features of organisms to be compared through tag counts. Recent studies have demonstrated that the normalization step for RNA-seq data is critical for a more accurate subsequent analysis of differential gene expression. Development of a more robust normalization method is desirable for identifying the true difference in tag count data.

RESULTS: We describe a strategy for normalizing tag count data, focusing on RNA-seq. The key concept is to remove data assigned as potential differentially expressed genes (DEGs) before calculating the normalization factor. Several R packages for identifying DEGs are currently available, and each package uses its own normalization method and gene ranking algorithm. We compared a total of eight package combinations: four R packages (edgeR, DESeq, baySeq, and NBPSeq) with their default normalization settings and with our normalization strategy. Many synthetic datasets under various scenarios were evaluated on the basis of the area under the curve (AUC) as a measure for both sensitivity and specificity. We found that packages using our strategy in the data normalization step overall performed well. This result was also observed for a real experimental dataset.

CONCLUSION: Our results showed that the elimination of potential DEGs is essential for more accurate normalization of RNA-seq data. The concept of this normalization strategy can widely be applied to other types of tag count data and to microarray data.

PMID: 22475125 PMCID: PMC3341196 DOI: 10.1186/1748-7188-7-5

DEGES論文9

Algorithms Mol Biol. 2012 Apr 5;7(1):5. doi: 10.1186/1748-7188-7-5.

A normalization strategy for comparing tag count data.

Kadota K[1], Nishiyama T, Shimizu K.

Author information

Abstract

BACKGROUND: High-throughput sequencing, such as ribonucleic acid sequencing (RNA-s
chromatin immunoprecipitation sequencing (ChIP-seq) analyses, enables various features
be compared through tag counts. Recent studies have demonstrated that the normalization
seq data is critical for a more accurate subsequent analysis of differential gene expression.
of a more robust normalization method is desirable for identifying the true difference in tag count data.

RESULTS: We describe a strategy for normalizing tag count data, focusing on RNA-seq. The key concept is to remove data assigned as potential differentially expressed genes (DEGs) before calculating the normalization factor. Several R packages for identifying DEGs are currently available, and each package uses its own normalization method and gene ranking algorithm. We compared a total of eight package combinations: four R packages (edgeR, DESeq, baySeq, and NBPSeq) with their default normalization settings and with our normalization strategy. Many synthetic datasets under various scenarios were evaluated on the basis of the area under the curve (AUC) as a measure for both sensitivity and specificity. We found that packages using our strategy in the data normalization step overall performed well. This result was also observed for a real experimental dataset.

CONCLUSION: Our results showed that the elimination of potential DEGs is essential for more accurate normalization of RNA-seq data. The concept of this normalization strategy can widely be applied to other types of tag count data and to microarray data.

PMID: 22475125 PMCID: PMC3341196 DOI: 10.1186/1748-7188-7-5

①Effect sizeが大きい（DEGが多い）という条件下であっても、②symmetricであれば、既存の正規化法でもうまくいく。しかし、③asymmetricな状況では「DEGの存在自体が正確な正規化を阻み、結果的にうまくDEGが検出できない」状況になる。じゃあどうすればいいか？DEGっぽいやつを除いたデータのみで正規化すればいいんです。DEGESは「DEG Elimination Strategy」の略だが、後述するTCC論文で出てきます。

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

TCC論文1

2012年のDEGES法の考え方を一般化して、①RパッケージTCCとしてまとめた2013年の論文。

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. ①

Sun J[1], 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 NBPSeq) 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論文2

2012年のDEGES法の考え方を一般化して、①RパッケージTCCとしてまとめた2013年の論文。②DEGESはここで初めて出現します。

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 J[1], 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 NBPSeq) 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: 22827715 PMCID: PMC3716788 DOI: 10.1186/1471-2105-14-219

②

TCC論文3

2012年のDEGES法の考え方を一般化して、①RパッケージTCCとしてまとめた2013年の論文。②DEGESはここで初めて出現します。asymmetryな状況の場合に特にうまくいくということが明記されています。

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 n

Sun J[1], 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 NBPSeq) 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: 23827749 PMCID: PMC3716788 DOI: 10.1186/1471-2105-14-219

TCC論文4

2012年のDEGES法の考え方を一般化して、①RパッケージTCCとしてまとめた2013年の論文。②DEGESはここで初めて出現します。asymmetryな状況の場合に特にうまくいくということが明記されています。この論文中では、データ中のDEGの割合（Effect size）が5%, 15%, and 25%までしか性能評価していませんが…

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 n

Sun J[1], 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 NBPSeq) 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

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

bulk性能評価論文1

この性能評価論文中では、データ中のDEGの割合（Effect size）が5-95%まで、かつsymmetric and asymmetricを調べています。

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

bulk性能評価論文2

この性能評価論文中では、データ中のDEGの割合（Effect size）が5-95%まで、かつsymmetric and asymmetricを調べています。①2018年9月となっていますが、Published online自体は2017年2月です。

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

①

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

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

Author information

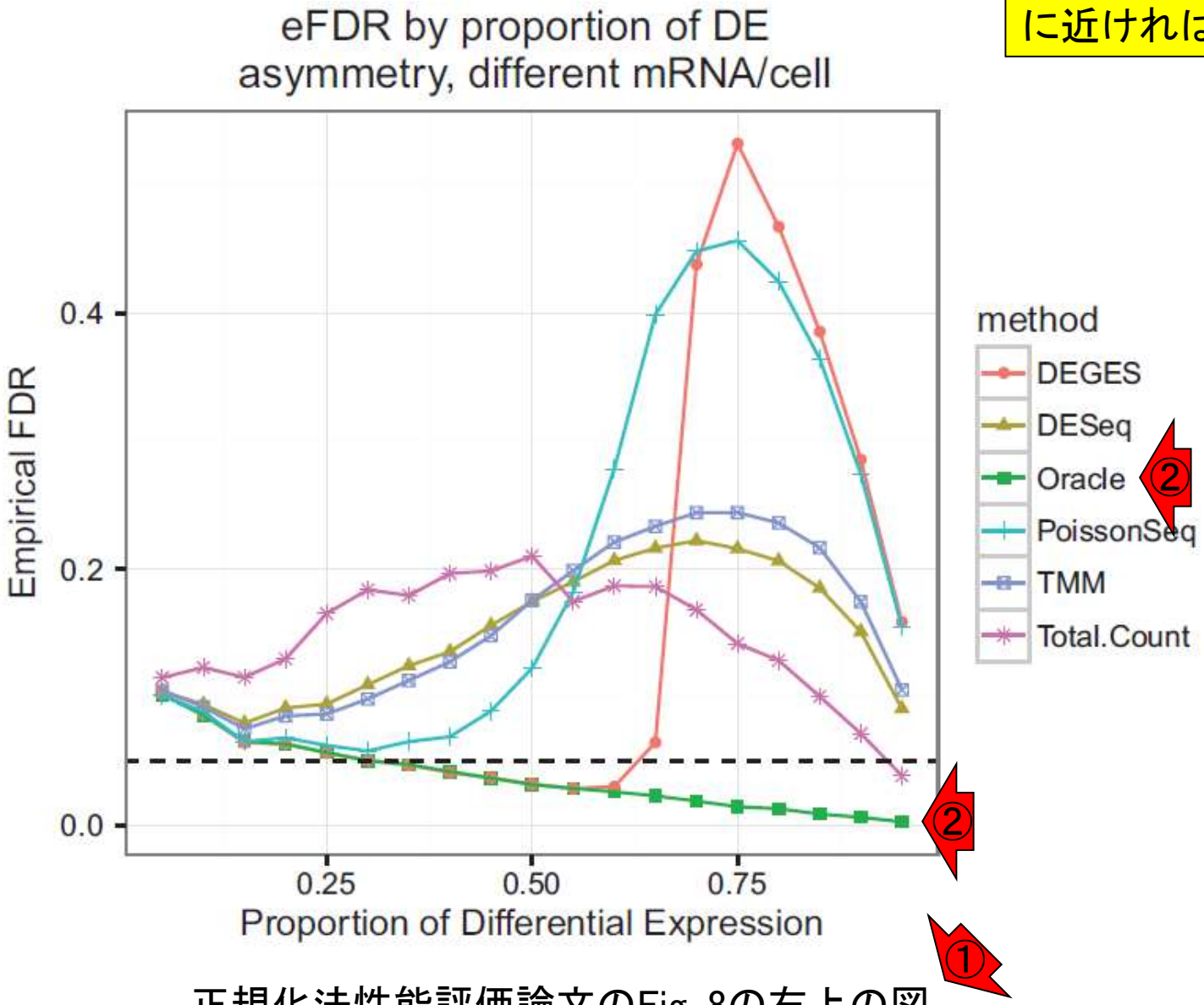
Abstract

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

bulk性能評価論文3

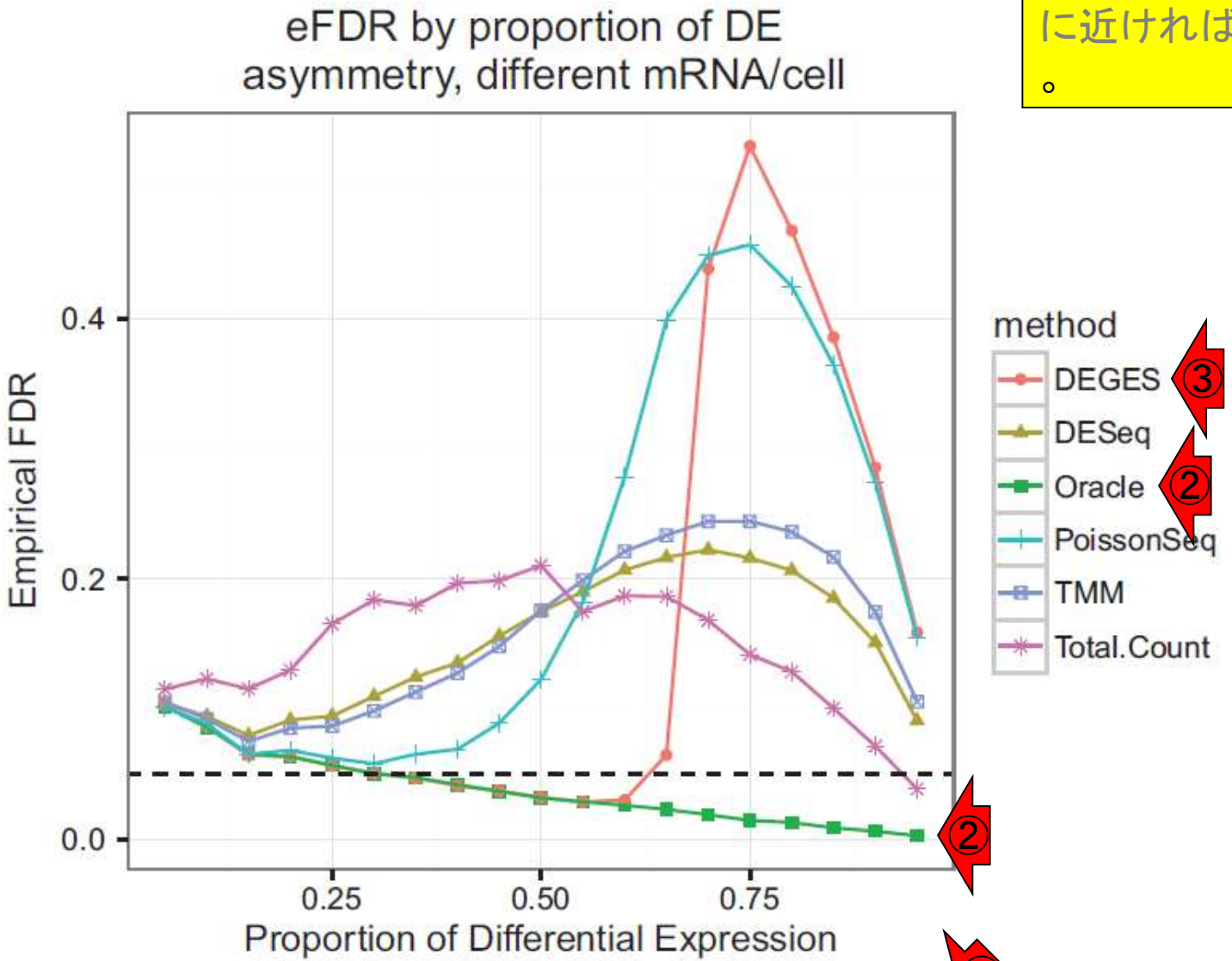
①他グループの性能評価論文のFig. 8。横軸がDEGの割合。縦軸の値が②Oracle（神託；理想値と解釈すればよい）に近ければよい。



正規化法性能評価論文のFig. 8の右上の図

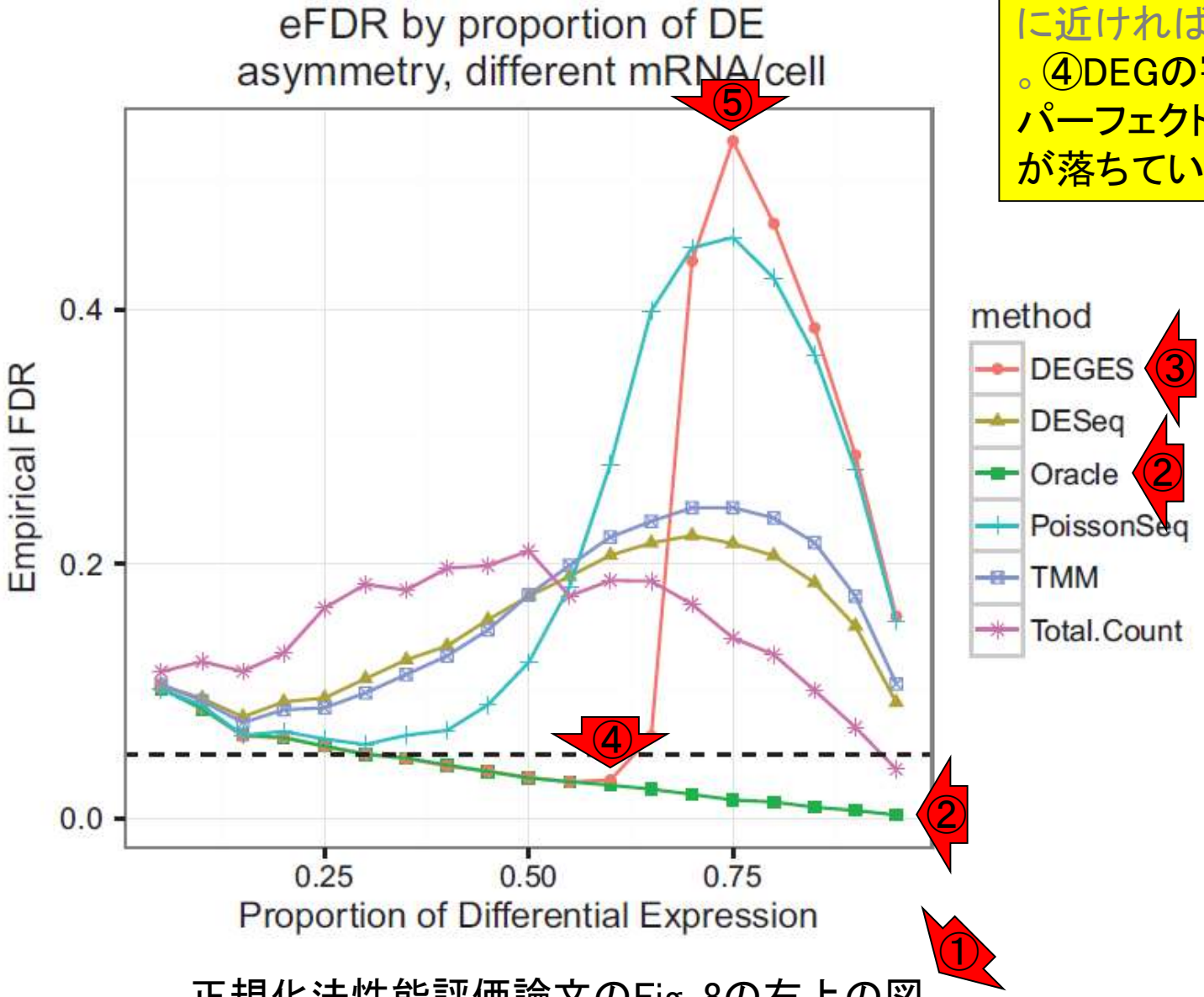
bulk性能評価論文4

①他グループの性能評価論文のFig. 8。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

bulk性能評価論文5



正規化法性能評価論文のFig. 8の右上の図

①他グループの性能評価論文のFig. 8。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。

Contents（2019年8月）

Bulk性能評価論文の話は、①2019年8月28日に「バイオインフォマティクス解析集中トレーニングコース」の一環?!で話した内容と同じです。

- 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!）①
- 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

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

何故次から次へと…1

①scRNA-seqのベストプラクティスに関する比較的最近の論文の場合は、過去の研究成果を正しく認識できていないことに起因（サーベイ不足）。これが1点目。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analysis pipelines.

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing (scRNA-seq) methods has created a large variety of experimental and computational pipelines for which best practices have not yet been established. Here, we use simulations based on five scRNA-seq library protocols in combination with nine realistic differential expression (DE) setups to systematically evaluate three mapping, four imputation, seven normalisation and four differential expression testing approaches resulting in ~3000 pipelines, allowing us to also assess interactions among pipeline steps. We find that choices of normalisation and library preparation protocols have the biggest impact on scRNA-seq analyses. Specifically, we find that library preparation determines the ability to detect symmetric expression differences, while normalisation dominates pipeline performance in asymmetric DE-setups. Finally, we illustrate the importance of informed choices by showing that a good scRNA-seq pipeline can have the same impact on detecting a biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-019-12266-7

何故次から次へと…1

①scRNA-seqのベストプラクティスに関する比較的最近の論文の場合は、過去の研究成果を正しく認識できていないことに起因（サーベイ不足）。例えば、②Introductionの赤下線部分など。Bulk RNA-seqでは取り組まれていないと書かれているが、実際にはasymmetry対応の正規化法も存在し、性能評価論文も存在する。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analy

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing
large variety of experimental and computational pipelin
yet been established. Here, we use simulations based
combination with nine realistic differential expression (
three mapping, four imputation, seven normalisation ar
approaches resulting in ~3000 pipelines, allowing us to
pipeline steps. We find that choices of normalisation ar
biggest impact on scRNA-seq analyses. Specifically, w
the ability to detect symmetric expression differences,
performance in asymmetric DE-setups. Finally, we illus
by showing that a good scRNA-seq pipeline can have
biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

何故次から次へと…2

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analy

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing
large variety of experimental and computational pipelin
yet been established. Here, we use simulations based
combination with nine realistic differential expression (
three mapping, four imputation, seven normalisation ar
approaches resulting in ~3000 pipelines, allowing us to
pipeline steps. We find that choices of normalisation ar
biggest impact on scRNA-seq analyses. Specifically, w
the ability to detect symmetric expression differences,
performance in asymmetric DE-setups. Finally, we illus
by showing that a good scRNA-seq pipeline can have
biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-

①

②

One main assumpti
expression are sym
genes is DE while t
constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

①scRNA-seqのベストプラクティスに関する比較的最近の論文の場合は、過去の研究成果を正しく認識できていないことに起因（サーベイ不足）。例えば、②Introductionの赤下線部分など。Bulk RNA-seqでは取り組まれていないと書かれているが、実際にはasymmetry対応の正規化法も存在し、性能評価論文も存在する。第一義的にはこの論文著者らの責任ではあるが、査読者がスルーしてしまっているのも残念なところ。これが2点目。

何故次から次へと…3

①scRNA-seqのベストプラクティスに関する比較的最近の論文の、②Fig. 1。3点目は、読み手側の能力に関する問題。大抵の読者は結論のみにフォーカスし、どのような条件で得られた結論なのかまでおそらくまともに見ていない。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analy

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

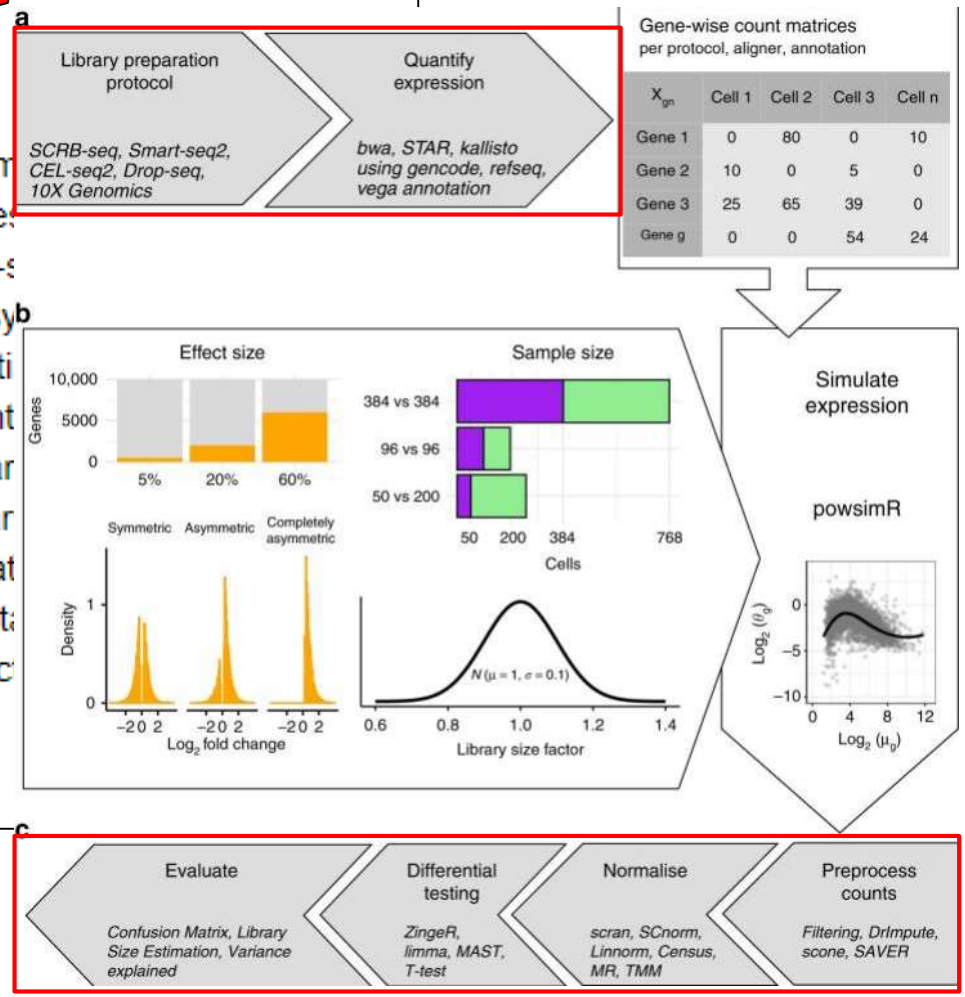
Author information

Abstract

The recent rapid spread of single cell RNA sequencing (scRNA-seq) m
large variety of experimental and computational pipelines for which be
yet been established. Here, we use simulations based on five scRNA-s
combination with nine realistic differential expression (DE) setups to sy
three mapping, four imputation, seven normalisation and four differenti
approaches resulting in ~3000 pipelines, allowing us to also assess int
pipeline steps. We find that choices of normalisation and library prepar
biggest impact on scRNA-seq analyses. Specifically, we find that librar
the ability to detect symmetric expression differences, while normalisat
performance in asymmetric DE-setups. Finally, we illustrate the import
by showing that a good scRNA-seq pipeline can have the same impac
biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-019-12266-7

②



何故次から次へと…3

①scRNA-seqのベストプラクティスに関する比較的最近の論文の、②Fig. 1。3点目は、読み手側の能力に関する問題。大抵の読者は結論のみにフォーカスし、どのような条件で得られた結論なのかまでおそらくまともに見ていない。例えば①の論文で得られたガイドラインは、赤枠内の比較対象に用いた方法や解析条件から外れるほど有効性は低下する傾向にあるはず…。しかしそういうことまで認識できていないヒトが多いのだろう。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analy

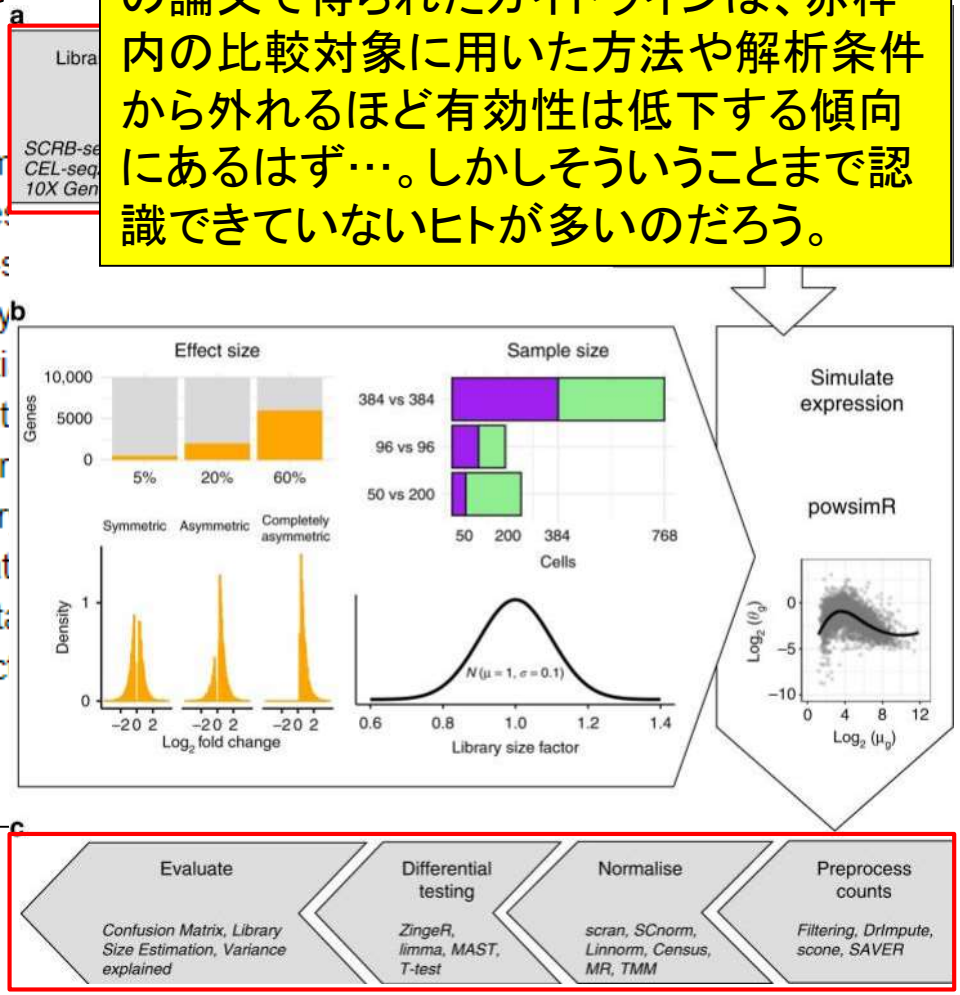
Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing (scRNA-seq) m
large variety of experimental and computational pipelines for which be
yet been established. Here, we use simulations based on five scRNA-s
combination with nine realistic differential expression (DE) setups to sy
three mapping, four imputation, seven normalisation and four differenti
approaches resulting in ~3000 pipelines, allowing us to also assess int
pipeline steps. We find that choices of normalisation and library prepar
biggest impact on scRNA-seq analyses. Specifically, we find that librar
the ability to detect symmetric expression differences, while normalisat
performance in asymmetric DE-setups. Finally, we illustrate the import
by showing that a good scRNA-seq pipeline can have the same impac
biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-019-12266-7



何故次から次へと…3.1

①scRNA-seq用発現変動解析プログラムDECENTの論文の、②シミュレーションデータの解析では、全遺伝子のうち10%までの発現変動遺伝子（DEGs）しか想定していない。

Bioinformatics. 2019 Dec 15;35(24):5155-5162. doi: 10.1093/bioinformatics/btz453.

DECENT: differential expression with capture efficiency adjustr
single-cell RNA-seq data.

Ye C[1,2,3], Speed TP[1,4], Salim A[1,5,6].

Author information

Abstract

MOTIVATION: Dropo
unaddressed it affect
expression (DE) ana
We develop DECEN
molecule capture pro

RESULTS: We show
not explicitly model d
datasets generated u
the capture process
performance without

AVAILABILITY AND
available from https:/

SUPPLEMENTARY I

PMID: 31197307 PMCI

①

3.1 Benchmarking using simulated data

We simulated 20 datasets, each consisting of 500 cells belonging to 2 cell types (224 cells from cell type 1 versus 276 cells from cell type 2) with 3000 endogenous genes and 50 spike-ins. The observed count were generated under the DECENT model using parameters estimated from Tung's dataset (see Supplementary Materials for details). In each dataset, we set ~10% of the genes to be DEGs. Figure 2 shows that DECENT estimates gene-specific pre-dropout proportion of zeroes and variance, as well as the actual pre-dropout counts unbiasedly. Figure 3 shows that DECENT's performance in detecting DEGs also appear to be competitive when compared with existing methods, namely SCDE (Kharchenko *et al.*, 2014), MAST (Finak *et al.*, 2015), Monocle2 (Qiu *et al.*, 2017; Trapnell *et al.*, 2014), ZINB-WaVE adjusted edgeR (Van den Berge *et al.*, 2018) and edgeR (McCarthy *et al.*, 2012). Over the 20 datasets, the mean(SD) of the partial area under the receiver operating characteristic (pAUROC) for DECENT is 0.708(0.001), followed by MAST with 0.687(0.001) (see

②

何故次から次へと…3.1

①scRNA-seq用発現変動解析プログラムDECENTの論文の、②シミュレーションデータの解析では、全遺伝子のうち10%までの発現変動遺伝子（DEGs）しか想定していない。この範囲から外れるほど、うまくいかないかもしれないと想像します。

Bioinformatics. 2019 Dec 15;35(24):5155-5162. doi: 10.1093/bioinformatics/btz453.

DECENT: differential expression with capture efficiency adjustr
single-cell RNA-seq data.

Ye C[1,2,3], Speed TP[1,4], Salim A[1,5,6].

Author information

Abstract

MOTIVATION: Dropo
unaddressed it affect
expression (DE) ana
We develop DECEN
molecule capture pro

RESULTS: We show
not explicitly model d
datasets generated u
the capture process
performance without

AVAILABILITY AND
available from https:/

SUPPLEMENTARY I

PMID: 31197307 PMCI

③

3.1 Benchmarking using simulated data

We simulated 20 datasets, each consisting of 500 cells belonging to 2 cell types (224 cells from cell type 1 versus 276 cells from cell type 2) with 3000 endogenous genes and 50 spike-ins. The observed count were generated under the DECENT model using parameters estimated from Tung's dataset (see Supplementary Materials for details). In each dataset, we set ~10% of the genes to be DEGs. Figure 2 shows that DECENT estimates gene-specific pre-dropout proportion of zeroes and variance, as well as the actual pre-dropout counts unbiasedly. Figure 3 shows that DECENT's performance in detecting DEGs also appear to be competitive when compared with existing methods, namely SCDE (Kharchenko *et al.*, 2014), MAST (Finak *et al.*, 2015), Monocle2 (Qiu *et al.*, 2017; Trapnell *et al.*, 2014), ZINB-WaVE adjusted edgeR (Van den Berge *et al.*, 2018) and edgeR (McCarthy *et al.*, 2012). Over the 20 datasets, the mean(SD) of the partial area under the receiver operating characteristic (pAUROC) for DECENT is 0.708(0.001), followed by MAST with 0.687(0.001) (see

④

何故次から次へと…4

4点目として、(主な)比較対象がedgeRやDESeq2などの有名な方法であり、最新の方法ではない場合が多い。それゆえ、「有名なedgeRやDESeq2より優れている」という結論からなる新規手法が出現し続けられる。比較的新しい方法も多数あるので、その中から自分が勝っている方法を選択的に抽出し、「新しい方法もちゃんと比較対象に含めました」というアリバイ工作をやることも原理的に可能。

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

Zero-inflated…1

①scRNA-seqは細胞あたりのRNA量が少ないので、ドロップアウト（dropouts）と呼ばれる「本当は発現しているのに検出できない遺伝子」が一定数存在する。

Nat Commun. 2018 Jan 18;9(1):284. doi: 10.1038/s41467-017-02554-5.

A general and flexible method for signal extraction from single-cell RNA-seq data.

Risso D[1], Perraudeau F[2], Gribkova S[3], Dudoit S[4,5], Vert JP[6,7,8,9].

Author information

Erratum in

Publisher Correction: A general and flexible method for signal extraction from single-cell RNA-seq data. [Nat Commun. 2019]

Abstract

Single-cell RNA-sequencing (scRNA-seq) is a powerful high-throughput technique that enables researchers to measure genome-wide transcription levels at the resolution of single cells. Because of the low amount of RNA present in a single cell, some genes may fail to be detected even though they are expressed; these genes are usually referred to as dropouts. Here, we present a general and flexible zero-inflated negative binomial model (ZINB-WaVE), which leads to low-dimensional representations of the data that account for zero inflation (dropouts), over-dispersion, and the count nature of the data. We demonstrate, with simulated and real data, that the model and its associated estimation procedure are able to give a more stable and accurate low-dimensional representation of the data than principal component analysis (PCA) and zero-inflated factor analysis (ZIFA), without the need for a preliminary normalization step.

PMID: 29348443 PMCID: PMC5773593 DOI: 10.1038/s41467-017-02554-5

[Indexed for MEDLINE] **Free PMC Article**

①

Zero-inflated…2

Nat Commun. 2018 Jan 18;9(1):284. doi: 10.1038/s41467-017-02554-5.

A general and flexible method for signal extraction from s cell RNA-seq data.

Risso D[1], Perraudeau F[2], Gribkova S[3], Dudoit S[4,5], Vert JP[6,7,8,9].

Author information

Erratum in

Publisher Correction: A general and flexible method for signal extraction from single-cell RNA-seq data. [Nat Commun. 2019]

Abstract

Single-cell RNA-sequencing (scRNA-seq) is a powerful high-throughput technique that enables researchers to measure genome-wide transcription levels at the resolution of single cells. Because of the low amount of RNA present in a single cell, some genes may fail to be detected even though they are expressed; these genes are usually referred to as dropouts. Here, we present a general and flexible zero-inflated negative binomial model (ZINB-WaVE), which leads to low-dimensional representations of the data that account for zero inflation (dropouts), over-dispersion, and the count nature of the data. We demonstrate, with simulated and real data, that the model and its associated estimation procedure are able to give a more stable and accurate low-dimensional representation of the data than principal component analysis (PCA) and zero-inflated factor analysis (ZIFA), without the need for a preliminary normalization step.

PMID: 29348443 PMCID: PMC5773593 DOI: 10.1038/s41467-017-02554-5

[Indexed for MEDLINE] **Free PMC Article**

②

③

①scRNA-seqは細胞あたりのRNA量が少ないので、ドロップアウト（dropouts）と呼ばれる「本当は発現しているのに検出できない遺伝子」が一定数存在する。②この論文が、③scRNA-seq分野でよく用いられる?!「ゼロ過剰負の二項分布（zero-inflated negative binomial distribution）」を本当に最初に提唱したかは不明だが…

Zero-inflated…3.1

Nat Commun. 2018 Jan 18;9(1):284. doi: 10.1038/s41467-017-02554-5.

A general and flexible method for signal extraction from s cell RNA-seq data.

Risso D[1], Perraudeau F[2], Gribkova S[3], Dudoit S[4,5], Vert JP[6,7,8,9].

Author information

Erratum in

Publisher Correction: A general and flexible method for signal extraction from single-ce data. [Nat Commun. 2019]

Abstract

Single-cell RNA-sequencing (scRNA-seq) is a powerful high-throughput technique that enables researchers to measure genome-wide transcription levels at the resolution of single cells. Because of the low amount of RNA present in a single cell, some genes may fail to be detected even though they are expressed; these genes are usually referred to as dropouts. Here, we present a general and flexible zero-inflated negative binomial model (ZINB-WaVE), which leads to low-dimensional representations of the data that account for zero inflation (dropouts), over-dispersion, and the count nature of the data. We demonstrate, with simulated and real data, that the model and its associated estimation procedure are able to give a more stable and accurate low-dimensional representation of the data than principal component analysis (PCA) and zero-inflated factor analysis (ZIFA), without the need for a preliminary normalization step.

PMID: 29348443 PMCID: PMC5773593 DOI: 10.1038/s41467-017-02554-5

[Indexed for MEDLINE] **Free PMC Article**

④

①scRNA-seqは細胞あたりのRNA量が少ないので、ドロップアウト（dropouts）と呼ばれる「本当は発現しているのに検出できない遺伝子」が一定数存在する。②この論文が、③scRNA-seq分野でよく用いられる?!「ゼロ過剰負の二項分布（zero-inflated negative binomial distribution）」を本当に最初に提唱したかは不明だが、④ドロップアウトを意味するzero inflationを考慮するのがscRNA-seqデータの主な特徴と理解しているヒトはおそらく多い。しかしながら…という話。

Zero-inflated…3.2

①

Genome Biol. 2015 Nov 2;16:241. doi: 10.1186/s13059-015-0805-z.

ZIFA: Dimensionality reduction for zero-inflated sing expression analysis.

Pierson E[1], Yau C[2,3].

Author information

Abstract

Single-cell RNA-seq data allows insight into normal cellular function and various disease states through molecular characterization of gene expression on the single cell level. Dimensionality reduction of such high-dimensional data sets is essential for visualization and analysis, but single-cell RNA-seq data are challenging for classical dimensionality-reduction methods because of the prevalence of dropout events, which lead to zero-inflated data. Here, we develop a dimensionality-reduction method, (Z)ero (I)nflated (F)actor (A)nalysis (ZIFA), which explicitly models the dropout characteristics, and show that it improves modeling accuracy on simulated and biological data sets.

PMID: 26527291 PMCID: PMC4630968 DOI: 10.1186/s13059-015-0805-z

その後、「ゼロ過剰（zero-inflated）」や「ドロップアウト（dropout）」と明記されているscRNA-seq論文を探してみました。①2015年11月のZIFAプログラムの論文がおそらく一番最初ですかねえ…。重要な点は、2013年のbulk用のPoisson-Tweedieモデルを実装したtweeDEseq論文のほうが、先にzero-inflatedについて言及しているということ（後述）。

Zero-inflated…4

①2013年に提唱された(bulk) RNA-seq用の新規カウントデータモデル。

BMC Bioinformatics. 2013 Aug 21;14:254. doi: 10.1186/1471-2105-14-254.

A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments.

Esnaola M[1], Puig P, Gonzalez D, Castelo R, Gonzalez JR.

Author information

Abstract

BACKGROUND: High-throughput RNA sequencing (RNA-seq) offers unprecedented power to capture the real dynamics of gene expression. Experimental designs with extensive biological replication present a unique opportunity to exploit this feature and distinguish expression profiles with higher resolution. RNA-seq data analysis methods so far have been mostly applied to data sets with few replicates and their default settings try to provide the best performance under this constraint. These methods are based on two well-known count data distributions: the Poisson and the negative binomial. The way to properly calibrate them with large RNA-seq data sets is not trivial for the non-expert bioinformatics user.

RESULTS: Here we show that expression profiles produced by extensively-replicated RNA-seq experiments lead to a rich diversity of count data distributions beyond the Poisson and the negative binomial, such as Poisson-Inverse Gaussian or Pólya-Aeppli, which can be captured by a more general family of count data distributions called the Poisson-Tweedie. The flexibility of the Poisson-Tweedie family enables a direct fitting of emerging features of large expression profiles, such as heavy-tails or zero-inflation, without the need to alter a single configuration parameter. We provide a software package for R called tweeDEseq implementing a new test for differential expression based on the Poisson-Tweedie family. Using simulations on synthetic

Zero-inflated…5

①2013年に提唱された(bulk) RNA-seq用の新規カウントデータモデル。②反復数（scRNA-seqでいうところの細胞数に相当）の多いRNA-seqカウントデータは、Poissonやnegative binomial (NB)分布では表現しきれない多様性がある。

BMC Bioinformatics. 2013 Aug 21;14:254. doi: 10.1186/1471-2105-14-254.

A flexible count data model to fit the wide diversity of exp
profiles arising from extensively replicated RNA-seq expe

Esnaola M[1], Puig P, Gonzalez D, Castelo R, Gonzalez JR.

Author information

Abstract

BACKGROUND: High-throughput RNA sequencing (RNA-seq) offers unprecedented power to capture the real dynamics of gene expression. Experimental designs with extensive biological replication present a unique opportunity to exploit this feature and distinguish expression profiles with higher resolution. RNA-seq data analysis methods so far have been mostly applied to data sets with few replicates and their default settings try to provide the best performance under this constraint. These methods are based on two well-known count data distributions: the Poisson and the negative binomial. The way to properly calibrate them with large RNA-seq data sets is not trivial for the non-expert bioinformatics user.

RESULTS: Here we show that expression profiles produced by extensively-replicated RNA-seq experiments lead to a rich diversity of count data distributions beyond the Poisson and the negative binomial, such as Poisson-Inverse Gaussian or Pólya-Aeppli, which can be captured by a more general family of count data distributions called the Poisson-Tweedie. The flexibility of the Poisson-Tweedie family enables a direct fitting of emerging features of large expression profiles, such as heavy-tails or zero-inflation, without the need to alter a single configuration parameter. We provide a software package for R called tweeDEseq implementing a new test for differential expression based on the Poisson-Tweedie family. Using simulations on synthetic

②

Zero-inflated…6

①2013年に提唱された(bulk) RNA-seq用の新規カウントデータモデル。②反復数（scRNA-seqでいうところの細胞数に相当）の多いRNA-seqカウントデータは、Poissonやnegative binomial (NB)分布では表現しきれない多様性がある。③Poisson-Tweedieを用いるとよい（うまくfitする → 帰無仮説の分布をうまく表現できる → 正確なp値を出せる）。

BMC Bioinformatics. 2013 Aug 21;14:254. doi: 10.1186/1471-2105-14-254.

A flexible count data model to fit the wide diversity of exp
profiles arising from extensively replicated RNA-seq expe

Esnaola M[1], Puig P, Gonzalez D, Castelo R, Gonzalez JR.

Author information

Abstract

BACKGROUND: High-throughput RNA sequencing (RNA-seq) offers unprecedented power to capture the real dynamics of gene expression. Experimental designs with extensive biological replication present a unique opportunity to exploit this feature and distinguish expression profiles with higher resolution. RNA-seq data analysis methods so far have been mostly applied to data sets with few replicates and their default settings try to provide the best performance under this constraint. These methods are based on two well-known count data distributions: the Poisson and the negative binomial. The way to properly calibrate them with large RNA-seq data sets is not trivial for the non-expert bioinformatics user.

RESULTS: Here we show that expression profiles produced by extensively-replicated RNA-seq experiments lead to a rich diversity of count data distributions beyond the Poisson and the negative binomial, such as Poisson-Inverse Gaussian or Pólya-Aeppli, which can be captured by a more general family of count data distributions called the Poisson-Tweedie. The flexibility of the Poisson-Tweedie family enables a direct fitting of emerging features of large expression profiles, such as heavy-tails or zero-inflation, without the need to alter a single configuration parameter. We provide a software package for R called tweeDEseq implementing a new test for differential expression based on the Poisson-Tweedie family. Using simulations on synthetic

③

Zero-inflated…7

BMC Bioinformatics. 2013 Aug 21;14:254. doi: 10.1186/1471-2105-14-254.

A flexible count data model to fit the wide diversity of exp
profiles arising from extensively replicated RNA-seq expe

Esnaola M[1], Puig P, Gonzalez D, Castelo R, Gonzalez JR.

Author information

Abstract

BACKGROUND: High-throughput RNA sequencing (RNA-seq) offers unprecedente
capture the real dynamics of gene expression. Experimental designs with extensive
replication present a unique opportunity to exploit this feature and distinguish expre
profiles with higher resolution. RNA-seq data analysis methods so far have been m
to data sets with few replicates and their default settings try to provide the best perf
under this constraint. These methods are based on two well-known count data distributions: the Poisson and the negative binomial. The way to properly calibrate them with large RNA-seq data sets is not trivial for the non-expert bioinformatics user.

RESULTS: Here we show that expression profiles produced by extensively-replicated RNA-seq experiments lead to a rich diversity of count data distributions beyond the Poisson and the negative binomial, such as Poisson-Inverse Gaussian or Pólya-Aeppli, which can be captured by a more general family of count data distributions called the Poisson-Tweedie. The flexibility of the Poisson-Tweedie family enables a direct fitting of emerging features of large expression profiles, such as heavy-tails or zero-inflation, without the need to alter a single configuration parameter. We provide a software package for R called tweeDEseq implementing a new test for differential expression based on the Poisson-Tweedie family. Using simulations on synthetic

①2013年に提唱された(bulk) RNA-seq用の新規カウントデータモデル。②反復数（scRNA-seqでいうところの細胞数に相当）の多いRNA-seqカウントデータは、Poissonやnegative binomial (NB)分布では表現しきれない多様性がある。③Poisson-Tweedieを用いるとよい（うまくfitする → 帰無仮説の分布をうまく表現できる → 正確なp値を出せる）。④heavy-tailsやzero-inflationのような現象をPoisson-Tweedieモデルはうまく取り込めるよ。つまり、zero-inflationは(bulk) RNA-seqデータの特徴でもある。

④

Zero-inflated…8

①Poisson-Tweedieモデル論文のBackground内の記述の一部。②zero-inflationは、確かにNB分布で想定されるゼロカウント数の期待値よりも多いという意味として使われていることがわかる。

Detection of DE genes using RNA-seq data was firstly based on using m
assuming a Poisson distribution [5] with one single parameter, the me
simultaneously determines the variance of the distribution. Given that the observed variation in read counts is much larger than the mean (overdispersion), researchers have proposed the use of negative binomial (NB) distributions [6-8] which are defined by two parameters: the mean and the dispersion. However, the larger power of RNA-seq to capture biological variability can poten ② ntroduce into count data not only overdispersion, but also oddities such as zero-inflation (i.e., in lowly expressed genes, the proportion of zero counts may be greater than expected under an NB distribution) and heavy tail behavior (i.e., a large dynamic range within the same expression profile), specially when many biological replicates are available. Under these circumstances even a two-parameter NB distribution may not provide an adequate fit to the data (see Figure 1). In turn, this may lead to incorrect statistical inferences resulting in lists of DE genes with a potentially increased number of false positive calls and poor reproducibility. To overcome this problem, methods based on the NB distribution [6-11] use sophisticated moderation techniques that borrow information across genes and exploit the mean-variance relationship in count data to improve the estimation of the NB dispersion parameter. This requires, however, that the parameter configuration is calibrated for the most appropriate moderation regime which may depend on features such as sample size, the magnitude of the fold-change, the variability of expression levels, the fraction of genes undergoing differential expression and the overall expression level.

①

Zero-inflated…9

①Poisson-Tweedieモデル論文のBackground内の記述の一部。②zero-inflationは、確かにNB分布で想定されるゼロカウント数の期待値よりも多いという意味として使われていることがわかる。
Poisson-Tweedie論文では、②zero-inflation現象（scRNA-seq分野でdropoutsと呼ばれていることと同義）を、③反復数の多いデータで見出していた。

Detection of DE genes using RNA-seq data was firstly based on using m
assuming a Poisson distribution [5] with one single parameter, the me
simultaneously determines the variance of the distribution. Given that
variation in read counts is much larger than the mean (overdispersion
have proposed the use of negative binomial (NB) distributions [6-8] wh
defined by two parameters: the mean and the dispersion. However, th
of RNA-seq to capture biological variability can poten ② ntroduce i
not only overdispersion, but also oddities such as zero-inflation (i.e., in lowly
expressed genes, the proportion of zero counts may be greater than expected under
an NB distribution) and heavy tail behavior (i.e., a large dynamic range within the
same expression profile), specially when many biological replicates are available. ③
Under these circumstances even a two-parameter NB distribution may not provide
an adequate fit to the data (see Figure 1). In turn, this may lead to incorrect statistical
inferences resulting in lists of DE genes with a potentially increased number of false
positive calls and poor reproducibility. To overcome this problem, methods based on
the NB distribution [6-11] use sophisticated moderation techniques that borrow
information across genes and exploit the mean-variance relationship in count data to
improve the estimation of the NB dispersion parameter. This requires, however, that
the parameter configuration is calibrated for the most appropriate moderation
regime which may depend on features such as sample size, the magnitude of the fold-
change, the variability of expression levels, the fraction of genes undergoing
differential expression and the overall expression level.

Zero-inflated…10

Detection of DE genes using RNA-seq data was firstly based on using m
assuming a Poisson distribution [5] with one single parameter, the me
simultaneously determines the variance of the distribution. Given tha
variation in read counts is much larger than the mean (overdispersion
have proposed the use of negative binomial (NB) distributions [6-8] wh
defined by two parameters: the mean and the dispersion. However, th
of RNA-seq to capture biological variability can potentially introduce i
not only overdispersion, but also oddities such as zero-inflation (i.e., in
expressed genes, the proportion of zero counts may be greater than ex
an NB distribution) and heavy tail behavior (i.e., a large dynamic rang
same expression profile), specially when many biological replicates ar
Under these circumstances even a two-parameter NB distribution may
an adequate fit to the data (see Figure 1). In turn, this may lead to incorrect statistical inferences resulting in lists of DE genes with a potentially increased number of false positive calls and poor reproducibility. To overcome this problem, methods based on the NB distribution [6-11] use sophisticated moderation techniques that borrow information across genes and exploit the mean-variance relationship in count data to improve the estimation of the NB dispersion parameter. This requires, however, that the parameter configuration is calibrated for the most appropriate moderation regime which may depend on features such as sample size, the magnitude of the fold-change, the variability of expression levels, the fraction of genes undergoing differential expression and the overall expression level.

①Poisson-Tweedieモデル論文のBackground内の記述の一部。②zero-inflationは、確かにNB分布で想定されるゼロカウント数の期待値よりも多いという意味として使われていることがわかる。Poisson-Tweedie論文では、②zero-inflation現象（scRNA-seq分野でdropoutsと呼ばれていることと同義）を、③反復数の多いデータで見出していた。しかし残念ながら、当時はまだRNA-seqが比較的高価な時代であり、反復数を稼ぐことができなかったため、それほど多くのヒトの目に留まることはなかったのだろう…。

Zero-inflated…11

①我々の2015年の(bulk) RNA-seq論文中でも…

BMC Bioinformatics. 2015 Nov 4;16:361. doi: 10.1186/s12859-015-0794-7.

Evaluation of methods for differential expression analysis on multi-group RNA-seq count data.

Tang M[1], Sun J[2], Shimizu K[3], Kadota K[4].

Author information

Abstract

BACKGROUND: RNA-seq is a powerful tool for measuring transcriptomes, especially for identifying differentially expressed genes or transcripts (DEGs) between sample groups. A number of methods have been developed for this task, and several evaluation studies have also been reported. However, those evaluations so far have been restricted to two-group comparisons. Accumulations of comparative studies for multi-group data are also desired.

METHODS: We compare 12 pipelines available in nine R packages for detecting differential expressions (DE) from multi-group RNA-seq count data, focusing on three-group data with or without replicates. We evaluate those pipelines on the basis of both simulation data and real count data.

RESULTS: As a result, the pipelines in the TCC package performed comparably to or better than other pipelines under various simulation scenarios. TCC implements a multi-step normalization strategy (called DEGES) that internally uses functions provided by other representative packages (edgeR, DESeq2, and so on). We found considerably different numbers of identified DEGs (18.5 ~ 45.7% of all genes) among the pipelines for the same real dataset but similar distributions of the classified expression patterns. We also found that DE results can roughly be estimated by the hierarchical dendrogram of sample clustering for the raw count data.

CONCLUSION: We confirmed the DEGES-based pipelines implemented in TCC performed well in a three-group comparison as well as a two-group comparison. We recommend using the DEGES-based pipeline that internally uses edgeR (here called the EEE-E pipeline) for count data with replicates (especially for small sample sizes). For data without replicates, the DEGES-based pipeline with DESeq2 (called SSS-S) can be recommended.

PMID: 26538400 PMCID: PMC4634584 DOI: 10.1186/s12859-015-0794-7

Zero-inflated…12

①我々の2015年の(bulk) RNA-seq論文中でも、②Background内でPoisson-Tweedieモデルに言及している。

BMC Bioinformatics. 2015 Nov 4;16:361. doi: 10.1186/s12859-015-0794-7.

Evaluation of methods for differential expression analysis on multi-group RNA-seq count data.

Tang M[1], Sun J[2], Shimizu K[3], Kadota K[4].

Author information

Abstract

BACKGROUND: RNA-seq is a powerful tool for meas
differentially expressed genes or transcripts (DEGs) I
been developed for this task, and several evaluation
evaluations so far have been restricted to two-group
multi-group data are also desired.

METHODS: We compare 12 pipelines available in nir
(DE) from multi-group RNA-seq count data, focusing
evaluate those pipelines on the basis of both simulati

RESULTS: As a result, the pipelines in the TCC pack
pipelines under various simulation scenarios. TCC im
DEGES) that internally uses functions provided by ot
on). We found considerably different numbers of iden
pipelines for the same real dataset but similar distribu
found that DE results can roughly be estimated by th
raw count data.

CONCLUSION: We confirmed the DEGES-based pip
group comparison as well as a two-group comparisor
that internally uses edgeR (here called the EEE-E pi
small sample sizes). For data without replicates, the
can be recommended.

PMID: 26538400 PMCID: PMC4634584 DOI: 10.1186/s1285

②

Read counts across technical replicates derived from a single source fit to a Poisson distribution [3, 21]. For data on biological replicates (BRs) derived from different individuals, the gene-level counts well fit to an over-dispersed Poisson distribution such as a negative-binomial (NB) model [10, 11, 22], beta-binomial (BB) model [5, 23], Poisson-Tweedie model [6], and so on. In particular, the Poisson-Tweedie model well captures the biological variation (especially for zero-inflation and heavy tail behavior, for details see [6]) when many BRs are available. As an increase in sample size (i.e., the number of replicate samples) precedes an increase in sequencing depth (i.e., the number of sequenced reads) [24–26], a more complex model such as Poisson-Tweedie may be the first choice for count data with many BRs. However, as many replicates are still difficult to take due to sequencing cost and the small amount of the target RNA sample, RNA-seq data with few BRs have mainly been stored. Two R packages based on the NB model (edgeR and DESeq) have been widely used as a common choice for DE analysis of RNA-seq data with few BRs [9–11, 27].

Zero-inflated…13

①我々の2015年の(bulk) RNA-seq論文中でも、②Background内でPoisson-Tweedieモデルに言及している。もちろん当時はscRNA-seqを意識してはいなかったが、反復数の多いカウントデータの場合はPoisson-Tweedieが第一選択となるだろうと述べている。

BMC Bioinformatics. 2015 Nov 4;16:361. doi: 10.1186/s12859-015-0794-7.

Evaluation of methods for differential expression analysis on multi-gro RNA-seq count data.

Tang M[1], Sun J[2], Shimizu K[3], Kadota K[4].

Author information

Abstract

BACKGROUND: RNA-seq is a powerful tool for meas differentially expressed genes or transcripts (DEGs) been developed for this task, and several evaluation evaluations so far have been restricted to two-group multi-group data are also desired.

METHODS: We compare 12 pipelines available in nir (DE) from multi-group RNA-seq count data, focusing evaluate those pipelines on the basis of both simulati

RESULTS: As a result, the pipelines in the TCC pack pipelines under various simulation scenarios. TCC im DEGES) that internally uses functions provided by ot on). We found considerably different numbers of iden pipelines for the same real dataset but similar distribu found that DE results can roughly be estimated by th raw count data.

CONCLUSION: We confirmed the DEGES-based pip group comparison as well as a two-group comparison that internally uses edgeR (here called the EEE-E pi small sample sizes). For data without replicates, the can be recommended.

PMID: 26538400 PMCID: PMC4634584 DOI: 10.1186/s1285

Read counts across technical replicates derived from a single source fit to a Poisson distribution [3, 21]. For data on biological replicates (BRs) derived from different individuals, the gene-level counts well fit to an over-dispersed Poisson distribution such as a negative-binomial (NB) model [10, 11, 22], beta-binomial (BB) model [5, 23], Poisson-Tweedie model [6], and so on. In particular, the Poisson-Tweedie model well captures the biological variation (especially for zero-inflation and heavy tail behavior, for details see [6]) when many BRs are available. As an increase in sample size (i.e., the number of replicate samples) precedes an increase in sequencing depth (i.e., the number of sequenced reads) [24–26], a more complex model such as Poisson-Tweedie may be the first choice for count data with many BRs. However, as many replicates are still difficult to take due to sequencing cost and the small amount of the target RNA sample, RNA-seq data with few BRs have mainly been stored. Two R packages based on the NB model (edgeR and DESeq) have been widely used as a common choice for DE analysis of RNA-seq data with few BRs [9–11, 27].

Zero-inflated…14

①我々の2015年の(bulk) RNA-seq論文中でも、②Background内でPoisson-Tweedieモデルに言及している。もちろん当時はscRNA-seqを意識してはいなかったが、反復数の多いカウントデータの場合はPoisson-Tweedieが第一選択となるだろうと述べている。当時は、反復数がそれほど多くなかったので、結果的にNBモデルに基づくRパッケージがよく利用されてきた。

BMC Bioinformatics. 2015 Nov 4;16:361. doi: 10.1186/s12859-015-0794-7.

Evaluation of methods for differential expression analysis on multi-gro RNA-seq count data.

Tang M[1], Sun J[2], Shimizu K[3], Kadota K[4].

Author information

Abstract

BACKGROUND: RNA-seq is a powerful tool for meas differentially expressed genes or transcripts (DEGs) been developed for this task, and several evaluation evaluations so far have been restricted to two-group multi-group data are also desired.

METHODS: We compare 12 pipelines available in nir (DE) from multi-group RNA-seq count data, focusing evaluate those pipelines on the basis of both simulati

RESULTS: As a result, the pipelines in the TCC pack pipelines under various simulation scenarios. TCC im DEGES) that internally uses functions provided by ot on). We found considerably different numbers of iden pipelines for the same real dataset but similar distribu found that DE results can roughly be estimated by th raw count data.

CONCLUSION: We confirmed the DEGES-based pip group comparison as well as a two-group comparisor that internally uses edgeR (here called the EEE-E pi small sample sizes). For data without replicates, the can be recommended.

PMID: 26538400 PMCID: PMC4634584 DOI: 10.1186/s1285

Read counts across technical r
Poisson distribution [3, 21]. Fo
from different individuals, the
dispersed Poisson distribution such as a negative-binomial (NB) model [10, 11, 22], beta-binomial (BB) model [5, 23], Poisson-Tweedie model [6], and so on. In particular, the Poisson-Tweedie model well captures the biological variation (especially for zero-inflation and heavy tail behavior, for details see [6]) when many BRs are available. As an increase in sample size (i.e., the number of replicate samples) precedes an increase in sequencing depth (i.e., the number of sequenced reads) [24–26], a more complex model such as Poisson-Tweedie may be the first choice for count data with many BRs. However, as many replicates are still difficult to take due to sequencing cost and the small amount of the target RNA sample, RNA-seq data with few BRs have mainly been stored. Two R packages based on the NB model (edgeR and DESeq) have been widely used as a common choice for DE analysis of RNA-seq data with few BRs [9–11, 27].

Zero-inflated…15

BMC Bioinformatics. 2015 Nov 4;16:361. doi: 10.1186/s12859-015-0794-7.

Evaluation of methods for differential expression analysis on multi-gro RNA-seq count data.

Tang M[1], Sun J[2], Shimizu K[3], Kadota K[4].

Author information

Abstract

BACKGROUND: RNA-seq is a powerful tool for meas differentially expressed genes or transcripts (DEGs) been developed for this task, and several evaluation evaluations so far have been restricted to two-group multi-group data are also desired.

METHODS: We compare 12 pipelines available in nir (DE) from multi-group RNA-seq count data, focusing evaluate those pipelines on the basis of both simulati

RESULTS: As a result, the pipelines in the TCC pack pipelines under various simulation scenarios. TCC in DEGES) that internally uses functions provided by ot on). We found considerably different numbers of iden pipelines for the same real dataset but similar distribu found that DE results can roughly be estimated by th raw count data.

CONCLUSION: We confirmed the DEGES-based pip group comparison as well as a two-group comparisor that internally uses edgeR (here called the EEE-E pi small sample sizes). For data without replicates, the can be recommended.

PMID: 26538400 PMCID: PMC4634584 DOI: 10.1186/s1285

Read counts across technical r
Poisson distribution [3, 21]. Fo
from different individuals, the
dispersed Poisson distribution
11, 22], beta-binomial (BB) mo
so on. In particular, the Poisso
biological variation (especially
for details see [6]) when many
size (i.e., the number of replica
sequencing depth (i.e., the nur
complex model such as Poisso
data with many BRs. Hoever
due to sequencing cost and the small amount of the target RNA sample, RNA-seq data with few BRs have mainly been stored. Two R packages based on the NB model (edgeR and DESeq) have been widely used as a common choice for DE analysis of RNA-seq data with few BRs [9–11, 27].

①我々の2015年の(bulk) RNA-seq論文中でも、②Background内でPoisson-Tweedieモデルに言及している。もちろん当時はscRNA-seqを意識してはいなかったが、反復数の多いカウントデータの場合はPoisson-Tweedieが第一選択となるだろうと述べている。当時は、反復数がそれほど多くなかったので、結果的にNBモデルに基づくRパッケージがよく利用されてきた。そして(bulk)RNA-seqデータで提唱されたPoisson-Tweedieが注目される前に、scRNA-seq分野で実質的にほぼ同じzero-inflated NBモデルが注目?!されることとなった。そして、比較対象はPoisson-Tweedieではなく有名なNBモデルに基づくRパッケージとなり、(bulk)RNA-seq用ではダメなのでscRNA-seq用のものを開発せねば…という流れになったのであろう。…残念。

Zero-inflated…16

(bulk) RNA-seqの発現変動解析用Rパッケージで有名な、①edgeRの筆頭著者と、②DESeq2の筆頭著者が含まれているのでそれを考慮する必要はあるが、③2018年2月の論文。

Genome Biol. 2018 Feb 26;19(1):24. doi: 10.1186/s13059-018-1406-4.

Observation weights unlock bulk RNA-seq tools for z and single-cell applications.

Van den Berge K[1,2], Perraudeau F[3], Soneson C[4,5], Love MI[6], Risso D[7], Vert JP[8,9,10,11], Robinson MD[4,5], Dudoit S[3,12], Clement L[1,2].

Author information

Abstract

Dropout events in single-cell RNA sequencing (scRNA-seq) cause many transcripts to go undetected and induce an excess of zero read counts, leading to power issues in differential expression (DE) analysis. This has triggered the development of bespoke scRNA-seq DE methods to cope with zero inflation. Recent evaluations, however, have shown that dedicated scRNA-seq tools provide no advantage compared to traditional bulk RNA-seq tools. We introduce a weighting strategy, based on a zero-inflated negative binomial model, that identifies excess zero counts and generates gene- and cell-specific weights to unlock bulk RNA-seq DE pipelines for zero-inflated data, boosting performance for scRNA-seq.

KEYWORDS: Differential expression; Single-cell RNA sequencing; Weights; Zero-inflated negative binomial

PMID: 29478411 PMCID: PMC6251479 DOI: 10.1186/s13059-018-1406-4

Zero-inflated…17

(bulk) RNA-seqの発現変動解析用Rパッケージで有名な、①edgeRの筆頭著者と、②DESeq2の筆頭著者が含まれているのでそれを考慮する必要はあるが、③2018年2月の論文。④本当は発現しているのに検出されないドロップアウト問題を解決すべく、scRNA-seqに特化した発現変動解析手法の開発が行われてきた。

Genome Biol. 2018 Feb 26;19(1):24. doi: 10.1186/s13059-018-1406-4.

Observation weights unlock bulk RNA-seq tools for z and single-cell applications.

Van den Berge K[1,2], Perraudeau F[3], Soneson C[4,5], Love MI[6], Risso D[7], Vert JP[8,9,10], Dudoit S[3,12], Clement L[1,2].

Author information

Abstract

Dropout events in single-cell RNA sequencing (scRNA-seq) cause many transcripts to go undetected and induce an excess of zero read counts, leading to power issues in differential expression (DE) analysis. This has triggered the development of bespoke scRNA-seq DE methods to cope with zero inflation. Recent evaluations, however, have shown that dedicated scRNA-seq tools provide no advantage compared to traditional bulk RNA-seq tools. We introduce a weighting strategy, based on a zero-inflated negative binomial model, that identifies excess zero counts and generates gene- and cell-specific weights to unlock bulk RNA-seq DE pipelines for zero-inflated data, boosting performance for scRNA-seq.

④

KEYWORDS: Differential expression; Single-cell RNA sequencing; Weights; Zero-inflated negative binomial

PMID: 29478411 PMCID: PMC6251479 DOI: 10.1186/s13059-018-1406-4

Zero-inflated…18

(bulk) RNA-seqの発現変動解析用Rパッケージで有名な、①edgeRの筆頭著者と、②DESeq2の筆頭著者が含まれているのでそれを考慮する必要はあるが、③2018年2月の論文。④本当は発現しているのに検出されないドロップアウト問題を解決すべく、scRNA-seqに特化した発現変動解析手法の開発が行われてきた。⑤しかしながら、旧来のbulk用と比べて優れているわけでもないことが示されている。

Genome Biol. 2018 Feb 26;19(1):24. doi: 10.1186/s13059-018-1406-4.

Observation weights unlock bulk RNA-seq tools for z and single-cell applications.

Van den Berge K[1,2], Perraudeau F[3], Soneson C[4,5], Love MI[6], Risso D[7], Vert JP[8,9,10], Dudoit S[3,12], Clement L[1,2].

Author information

Abstract

Dropout events in single-cell RNA sequencing (scRNA-seq) cause many transcripts to go undetected and induce an excess of zero read counts, leading to power issues in differential expression (DE) analysis. This has triggered the development of bespoke scRNA-seq DE methods to cope with zero inflation. Recent evaluations, however, have shown that dedicated scRNA-seq tools provide no advantage compared to traditional bulk RNA-seq tools. We introduce a weighting strategy, based on a zero-inflated negative binomial model, that identifies excess zero counts and generates gene- and cell-specific weights to unlock bulk RNA-seq DE pipelines for zero-inflated data, boosting performance for scRNA-seq.

KEYWORDS: Differential expression; Single-cell RNA sequencing; Weights; Zero-inflated negative binomial

PMID: 29478411 PMCID: PMC6251479 DOI: 10.1186/s13059-018-1406-4

Zero-inflated…19

Genome Biol. 2018 Feb 26;19(1):24. doi: 10.1186/s13059-018-1406-4.

Observation weights unlock bulk RNA-seq tools for z and single-cell applications.

Van den Berge K[1,2], Perraudeau F[3], Soneson C[4,5], Love MI[6], Risso D[7], Vert JP[8,9,10], Dudoit S[3,12], Clement L[1,2].

Author information

Abstract

Dropout events in single-cell RNA sequencing (scRNA-seq) cause many tra undetected and induce an excess of zero read counts, leading to power issu expression (DE) analysis. This has triggered the development of bespoke sc methods to cope with zero inflation. Recent evaluations, however, have shown that dedicated scRNA-seq tools provide no advantage compared to traditional bulk RNA-seq tools. We introduce a weighting strategy, based on a zero-inflated negative binomial model, that identifies excess zero counts and generates gene- and cell-specific weights to unlock bu NA-seq DE pipelines for zero-inflated data, boosting performance for scRNA-seq.

KEYWORDS: Differential expression; Single-cell RNA sequencing; Weights; Zero-inflated negative binomial

PMID: 29478411 PMCID: PMC6251479 DOI: 10.1186/s13059-018-1406-4

(bulk) RNA-seqの発現変動解析用Rパッケージで有名な、①edgeRの筆頭著者と、②DESeq2の筆頭著者が含まれているのでそれを考慮する必要はあるが、③2018年2月の論文。④本当は発現しているのに検出されないドロップアウト問題を解決すべく、scRNA-seqに特化した発現変動解析手法の開発が行われてきた。⑤しかしながら、旧来のbulk用と比べて優れているわけでもないことが示されている。⑥基本形の「負の二項分布（NB distribution）」に、ゼロ過剰の分を考慮したモデルをベースとしている。

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

①scRNA-seqのベストプラクティス論文。

反復数増やすとDEG増える

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analysis pipelines.

①

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing (scRNA-seq) methods has created a large variety of experimental and computational pipelines for which best practices have not yet been established. Here, we use simulations based on five scRNA-seq library protocols in combination with nine realistic differential expression (DE) setups to systematically evaluate three mapping, four imputation, seven normalisation and four differential expression testing approaches resulting in ~3000 pipelines, allowing us to also assess interactions among pipeline steps. We find that choices of normalisation and library preparation protocols have the biggest impact on scRNA-seq analyses. Specifically, we find that library preparation determines the ability to detect symmetric expression differences, while normalisation dominates pipeline performance in asymmetric DE-setups. Finally, we illustrate the importance of informed choices by showing that a good scRNA-seq pipeline can have the same impact on detecting a biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-019-12266-7

反復数増やすとDEG増え

①scRNA-seqのベストプラクティス論文。②Introductionの一部抜粋。③scRNA-seqでのcell types間比較では、DEGが60%程度に達する場合もあるし、トータルのmRNA量も細胞間で異なる。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analy

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing
large variety of experimental and computational pipelin
yet been established. Here, we use simulations based
combination with nine realistic differential expression (
three mapping, four imputation, seven normalisation a
approaches resulting in ~3000 pipelines, allowing us t
pipeline steps. We find that choices of normalisation
biggest impact on scRNA-seq analyses. Specifically
the ability to detect symmetric expression differences,
performance in asymmetric DE-setups. Finally, we illus
by showing that a good scRNA-seq pipeline can have
biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-

②

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

③

参考

反復数増やすとDEG増え

①scRNA-seqのベストプラクティス論文。②Introductionの一部抜粋。③scRNA-seqでのcell types間比較では、DEGが60%程度に達する場合もあるし、トータルのmRNA量も細胞間で異なる。④ここでのasymmetryという用語は、「トータルのmRNA量が細胞間で異なる」という意味で用いられており、この論文中のFig. bとは若干意味合いが異なっているという指摘があるかもしれない。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analy

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing
large variety of experimental and computational pipelir
yet been established. Here, we use simulations based
combination with nine realistic differential expression (
three mapping, four imputation, seven normalisation ar
approaches resulting in ~3000 pipelines, allowing us to
pipeline steps. We find that choices of normalisation ar
biggest impact on scRNA-seq analyses. Specifically, w
the ability to detect symmetric expression differences
performance in asymmetric DE-setups. Finally, we i
by showing that a good scRNA-seq pipeline can hav
biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-

④

One main assumpti
expression are sym
genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

反復数増やすとDEG増え

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

Selecting between-sample RNA-Seq normalization meth
perspective of their assumptions.

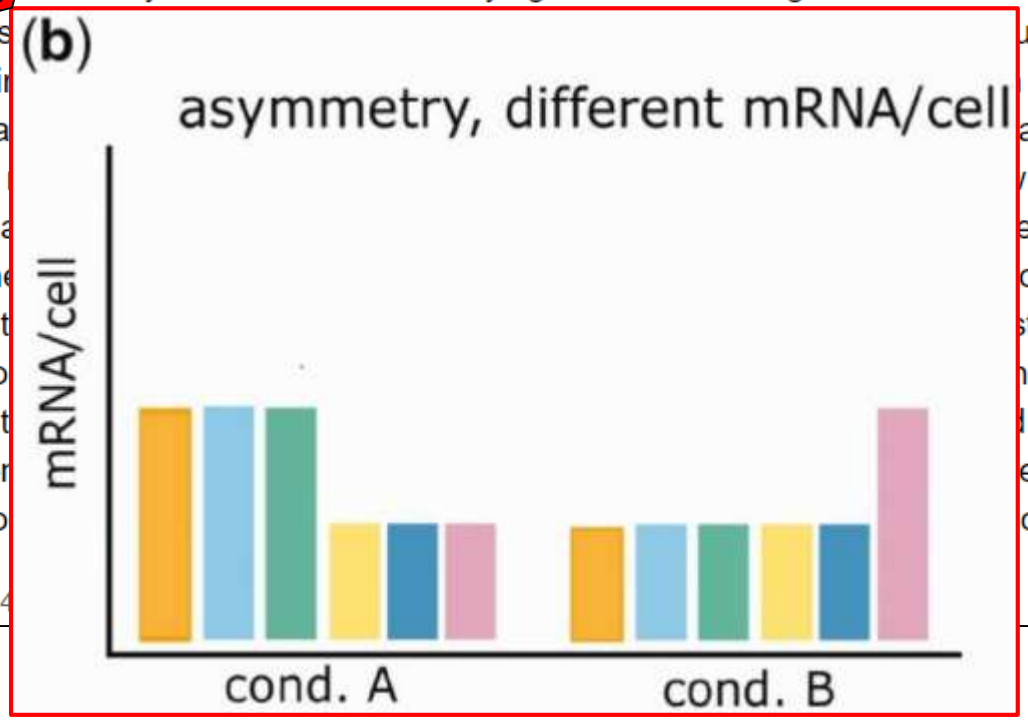
Evans C[1], Hardin J[2], Stoebel DM[3].

⑤ **or information**

Abstract

RNA-S⑥s a widely used method for studying the behavior of genes under diffe
essential s us
prevent di h
downstrea ar
feature of the validity of these assumptions
can have a e, we explain how assumptions
provide the of gene expression. We examine
normalizat standing of methodological
assumptio nd. Furthermore, we discuss why
normalizat how this causes problems in
subsequer ect a normalization method with
assumptio on for the given experiment.

PMID: 28334



①scRNA-seqのベストプラクティス論文。②Introductionの一部抜粋。③scRNA-seqでのcell types間比較では、DEGが60%程度に達する場合もあるし、トータルのmRNA量も細胞間で異なる。④ここでのasymmetryという用語は、「トータルのmRNA量が細胞間で異なる」という意味で用いられており、この論文中のFig. bとは若干意味合いが異なっているという指摘があるかもしれない。しかし、それでも⑤Evansらの2018年のbulk性能評価論文の、⑥Fig. 3bで示されている条件が上記の意味でのasymmetryに合致する。

反復数増やすとDEG増える

①scRNA-seq発現変動解析用Rパッケージ NBIDの論文。

Genome Biol. 2018 May 31;19(1):70. doi: 10.1186/s13059-018-1438-9.

UMI-count modeling and differential expression analysis for single-cell RNA sequencing.

Chen W[1], Li Y[2], Easton J[1], Finkelstein D[1], Wu G[1], Chen X[3].

Author information

Abstract

Read counting and unique molecular identifier (UMI) counting are the principal gene expression quantification schemes used in single-cell RNA-sequencing (scRNA-seq) analysis. By using multiple scRNA-seq datasets, we reveal distinct distribution differences between these schemes and conclude that the negative binomial model is a good approximation for UMI counts, even in heterogeneous populations. We further propose a novel differential expression analysis algorithm based on a negative binomial model with independent dispersions in each group (NBID). Our results show that this properly controls the FDR and achieves better power for UMI counts when compared to other recently developed packages for scRNA-seq analysis.

KEYWORDS: Differential expression analysis; Negative binomial; Unique molecular identifier

PMID: 29855333 PMCID: PMC5984373 DOI: 10.1186/s13059-018-1438-9

反復数増やすとDEG増え

①scRNA-seq発現変動解析用Rパッケージ NBIDの論文。②scRNA-seqは、まず探索的な解析（クラスタリング）により似た発現パターンを示す細胞をグループ化する。この段階ですでに「同一グループ内の細胞間の全体的な類似度が高いことが保証されている」ので、適当なp値やFDR閾値を満たす発現変動遺伝子（DEG）が多数得られてしまうので注意してね、ということが書かれている。が、おそらくこれ以外にも多数のDEGが得られる要因がある。

Genome Biol. 2018 May 31;19(1):70. doi: 10.1186/s13059-018-1438-9.

UMI-count modeling and differential expression anal
single-cell RNA sequencing.

Chen W[1], Li Y[2], Easton J[1], Finkelstein D[1]

Author information

Abstract

Read counting and unique molecular
expression quantification schemes us
By using multiple scRNA-seq dataset
these schemes and conclude that the
UMI counts, even in heterogeneous p
expression analysis algorithm based
dispersions in each group (NBID). Ou
achieves better power for UMI counts
for scRNA-seq analysis.

KEYWORDS: Differential expression ana

PMID: 29855333 PMCID: PMC5984373

single-cell analysis. Even though
the current study, the general for
tested simultaneously, as in the g

Differential expression analysis
samples run on separate lanes or
batch effects are expected to overlap with biological differences in this setting. It is better to account for batch effects by proper experimental design, such as by including multiple biological replicates for each group. Another typical application of differential expression analysis is to identify potential biomarkers for inferred cell subpopulations. One potential caveat in this setting is that cells are usually clustered from the same data. Therefore, *p* values or FDR values derived from the differential expression analysis might be overly optimistic. However, the result is still useful for prioritizing potential biomarkers for further validation.

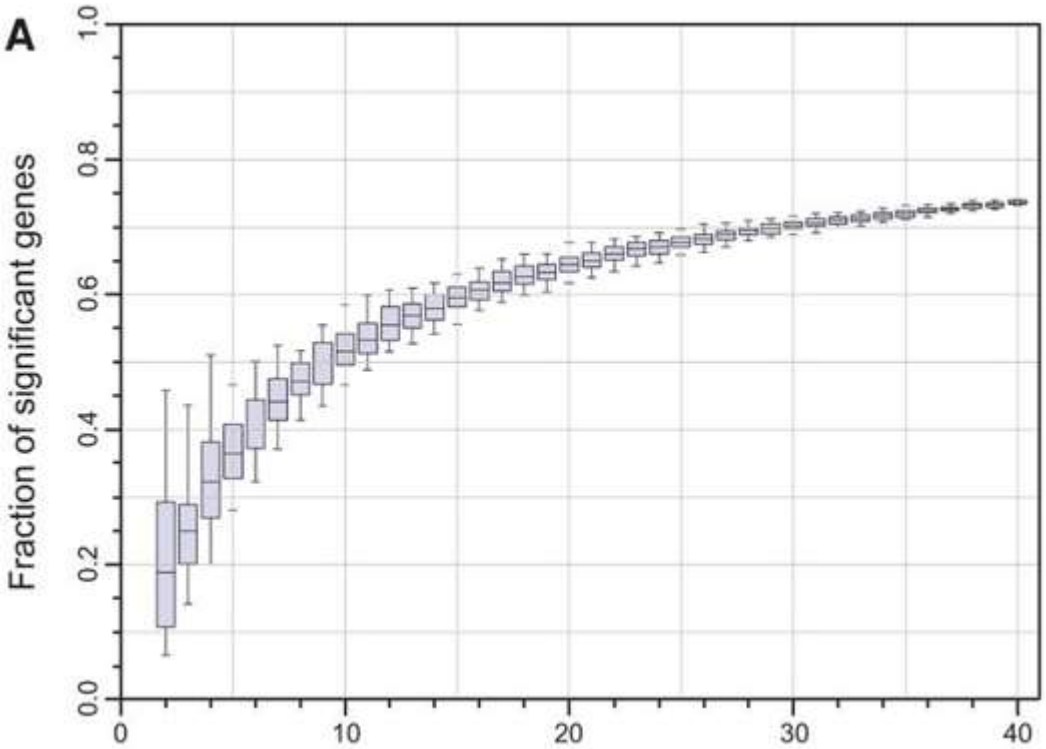
②

反復数増やすとDEG増え

①scRNA-seq発現変動解析用Rパッケージ NBIDの論文。②scRNA-seqは、まず探索的な解析（クラスタリング）により似た発現パターンを示す細胞をグループ化する。この段階ですでに「同一グループ内の細胞間の全体的な類似度が高いことが保証されている」ので、適当なp値やFDR閾値を満たす発現変動遺伝子（DEG）が多数得られてしまうので注意してね、ということが書かれている。が、おそらくこれ以外にも多数のDEGが得られる要因がある。それが③これ。

反復数増やすとDEG増え

①の(bulk)RNA-seq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。



③

②

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

Contents（2019年8月）

2019年8月28日に「バイオインフォマティクス解析集中トレーニングコース」の一環?!としてお話しさせていただいた内容の中の、①の部分です。

- 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）①
- 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

細胞数増やすとDEG増え

(bulk) RNA-seqデータは比較的少数の反復数からなる群間比較であり、得られるDEG数（正確には全遺伝子に占めるDEGの割合；Effect size）はそれほど多くなかった。反復数を増やすと得られるDEG数が増える傾向にあり、それと同じことがscRNA-seqデータについてもおそらく言えるはず。scRNA-seqの場合は多数の細胞数からなるcell types間比較。比較的多数のDEG数が得られるのはおそらく反復数に相当する細胞数が多いから。

Contents（2019年8月）

(bulk) RNA-seqデータを例に話した①の内容を、仮想scRNA-seqだと読み替えて（一部省略しながら）再度説明します。

- 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）①
- 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

反復数増→DEG数増1

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 H[1], 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 PIASY (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

反復数増→DEG数増2

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。これが実データ。

Pro群　Ras群

60,234 genes

	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	0	0	0	1	0	0
ENSG00000000419	282	354	208	165	301	209
ENSG00000000457	201	254	183	166	296	148
ENSG00000000460	114	112	101	55	81	59
ENSG00000000938	0	0	0	2	2	1
ENSG00000000971	747	914	605	252	414	147
…						

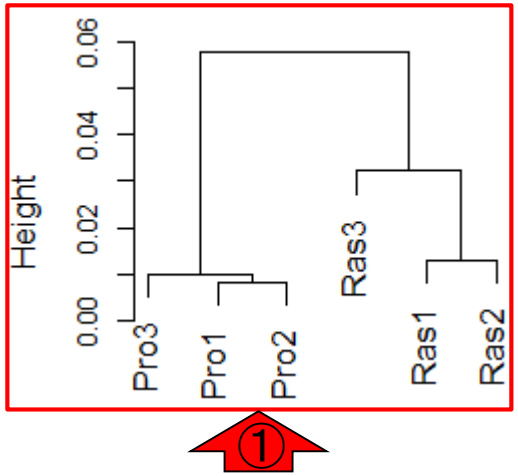
反復数増→DEG数増3

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。これが実データ。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、DEGが沢山得られることが期待される。

Pro群　　Ras群

60,234 genes

	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	0	0	0	1	0	0
ENSG00000000419	282	354	208	165	301	209
ENSG00000000457	201	254	183	166	296	148
ENSG00000000460	114	112	101	55	81	59
ENSG00000000938	0	0	0	2	2	1
ENSG00000000971	747	914	605	252	414	147
…						



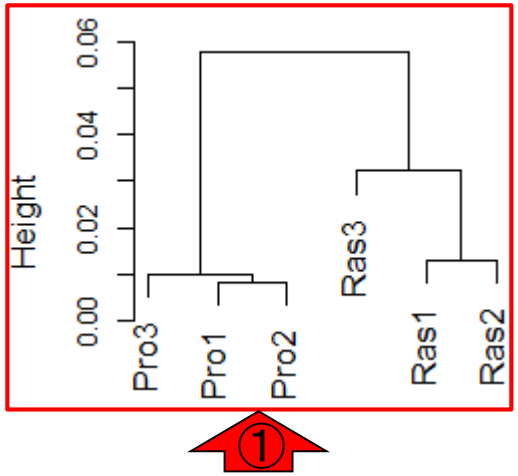
反復数増→DEG数増4

このデータは「各群につき3反復」ですが、scRNA-seqに読み替えると「各細胞型につき3細胞」となります。

Pro群　Ras群

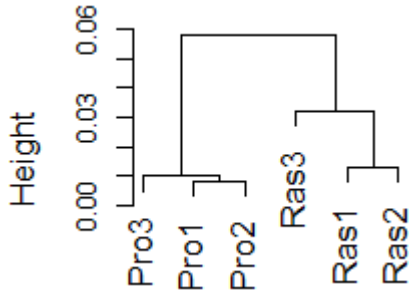
60,234 genes

	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	0	0	0	1	0	0
ENSG00000000419	282	354	208	165	301	209
ENSG00000000457	201	254	183	166	296	148
ENSG00000000460	114	112	101	55	81	59
ENSG00000000938	0	0	0	2	2	1
ENSG00000000971	747	914	605	252	414	147
…						



反復数増→DEG数増5

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。



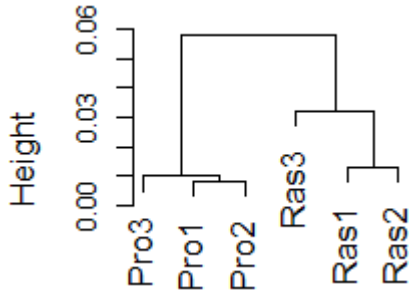
Pro群　Ras群　①

60,234 genes

	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	q.value	rank
ENSG00000000003	480	513	366	124	271	366	0.67708	12666
ENSG00000000005	0	0	0	1	0	0	1.00000	43869.5
ENSG00000000419	282	354	208	165	301	209	1.00000	24087
ENSG00000000457	201	254	183	166	296	148	1.00000	14902
ENSG00000000460	114	112	101	55	81	59	0.66930	12598
ENSG00000000938	0	0	0	2	2	1	0.24774	8736
ENSG00000000971	747	914	605	252	414	147	0.00312	3557
ENSG00000001036	2976	2899	2360	1621	2697	2367	1.00000	23058
ENSG00000001084	811	938	600	433	759	421	1.00000	18548

反復数増→DEG数増6

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

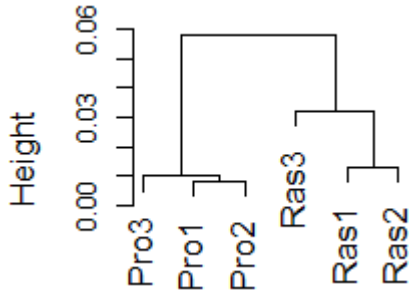
60,234 genes

	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

①　②

反復数増→DEG数増7

有意水準に相当する偽陽性率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。



Pro群　　Ras群

①

60,234 genes

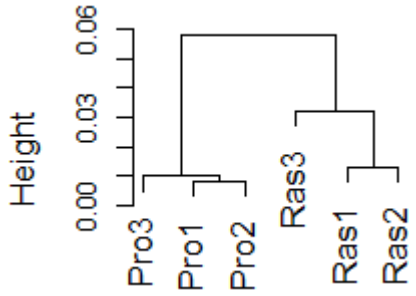
	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	q.value	rank
ENSG00000006715	978	1035	705	378	620	401	0.09980	6754
ENSG00000205060	1207	1369	1052	528	877	412	0.09980	6756
ENSG00000271075	2	6	1	0	0	0	0.09980	6755
ENSG00000150753	4292	4142	3305	3297	5334	4732	0.09986	6757
ENSG00000150907	25	21	20	6	9	9	0.09987	6758
ENSG00000233247	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

②

参考

反復数増→DEG数増8

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。



Pro群　Ras群

①

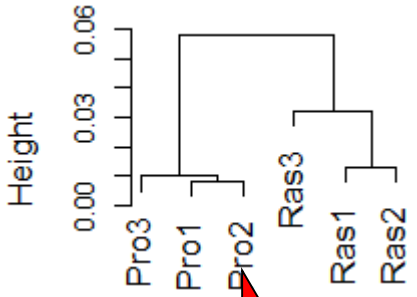
60,234 genes

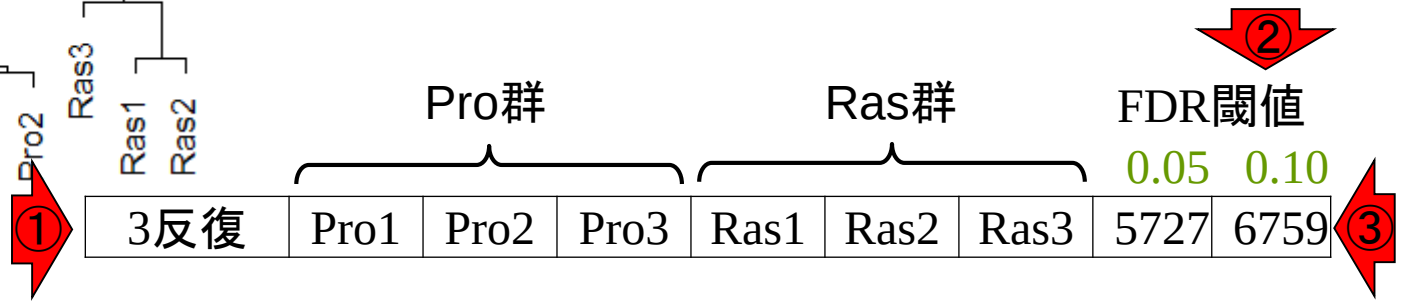
	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	q.value	rank
ENSG00000006715	978	1035	705	378	620	401	0.09980	6754
ENSG00000205060	1207	1369	1052	528	877	412	0.09980	6756
ENSG00000271075	2	6	1	0	0	0	0.09980	6755
ENSG00000150753	4292	4142	3305	3297	5334	4732	0.09986	6757
ENSG00000150907	25	21	20	6	9	9	0.09987	6758
ENSG00000233247	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

②

反復数増→DEG数増9

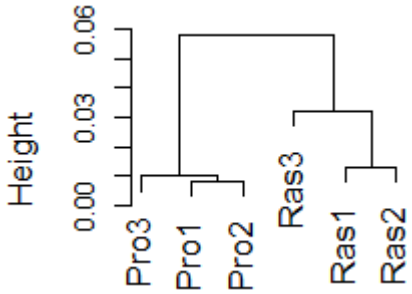
①3反復の2群間比較の結果として、②10% FDRを満たす遺伝子数は、③6,759個であった。





反復数増→DEG数増10

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。

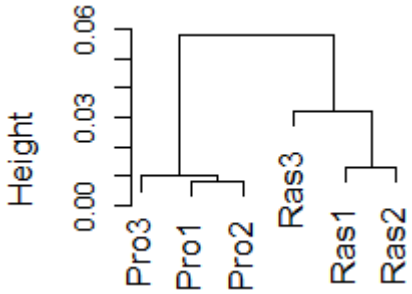


	Pro群			Ras群			FDR閾値 0.05	FDR閾値 0.10
3反復	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	5727	6759
2反復	Pro1	Pro2		Ras1	Ras2		8086	9026
2反復	Pro1	Pro2		Ras1		Ras3	3550	4371
2反復	Pro1	Pro2			Ras2	Ras3	3282	4059
2反復	Pro1		Pro3	Ras1	Ras2		7739	8578
2反復	Pro1		Pro3	Ras1		Ras3	3330	3986
2反復	Pro1		Pro3		Ras2	Ras3	3186	3889
2反復		Pro2	Pro3	Ras1	Ras2		6545	7444
2反復		Pro2	Pro3	Ras1		Ras3	3210	3883
2反復		Pro2	Pro3		Ras2	Ras3	3120	3821

④

反復数増→DEG数増11

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。



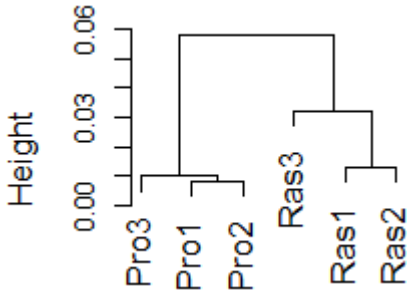
	Pro群			Ras群			DEG数 FDR 0.05	DEG数 FDR 0.10
3反復	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	5727	6759
2反復	Pro1	Pro2		Ras1	Ras2		8086	9026
2反復	Pro1	Pro2		Ras1		Ras3	3550	4371
2反復	Pro1	Pro2			Ras2	Ras3	3282	4059
2反復	Pro1		Pro3	Ras1	Ras2		7739	8578
2反復	Pro1		Pro3	Ras1		Ras3	3330	3986
2反復	Pro1		Pro3		Ras2	Ras3	3186	3889
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

反復数増→DEG数増12

①反復数が減ると(3→2)、FDR閾値を満たす遺伝子数も減る。

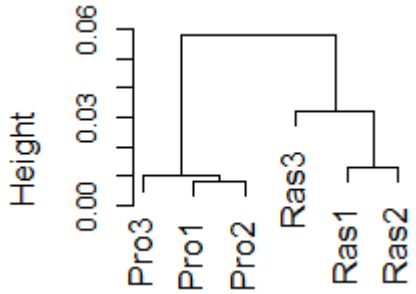


	Pro群			Ras群			FDR閾値	
							0.05	0.10
3反復	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	5727	6759
2反復							4672	5451

①

反復数増→DEG数増13

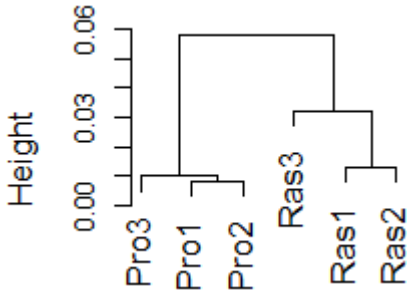
①反復数が減ると(3→2)、FDR閾値を満たす遺伝子数も減る。②反復なしにすると大幅に減る。



	Pro群			Ras群			FDR閾値 0.05	FDR閾値 0.10
3反復	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	5727	6759
2反復							4672	5451 ①
	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 ②

反復数増→DEG数増14

「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。

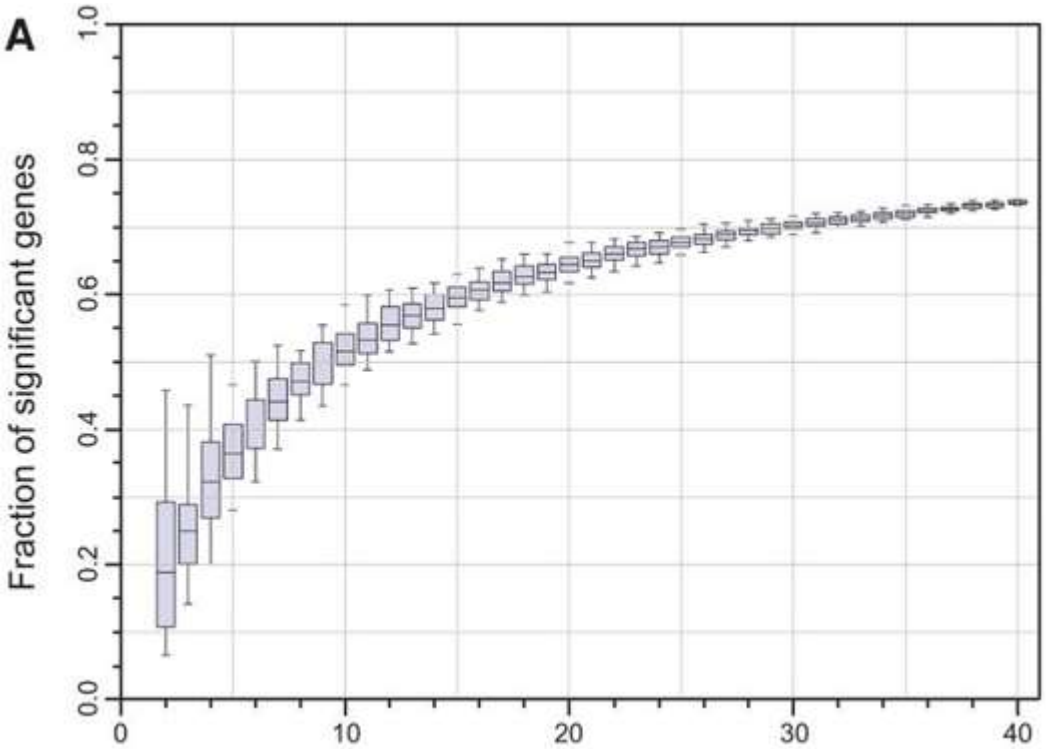


	Pro群			Ras群			FDR閾値 0.05	FDR閾値 0.10
3反復	Pro1	Pro2	Pro3	Ras1	Ras2	Ras3	5727	6759
2反復							4672	5451
反復なし							85	109

①

反復数増→DEG数増15

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。

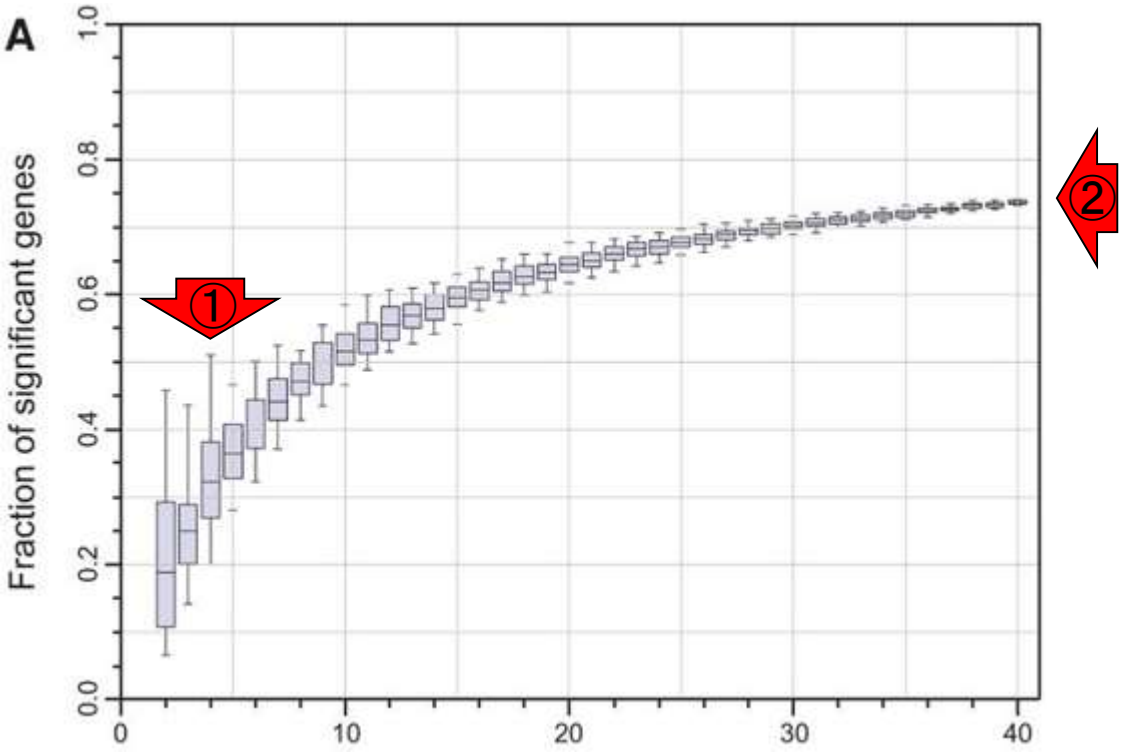


③

②

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

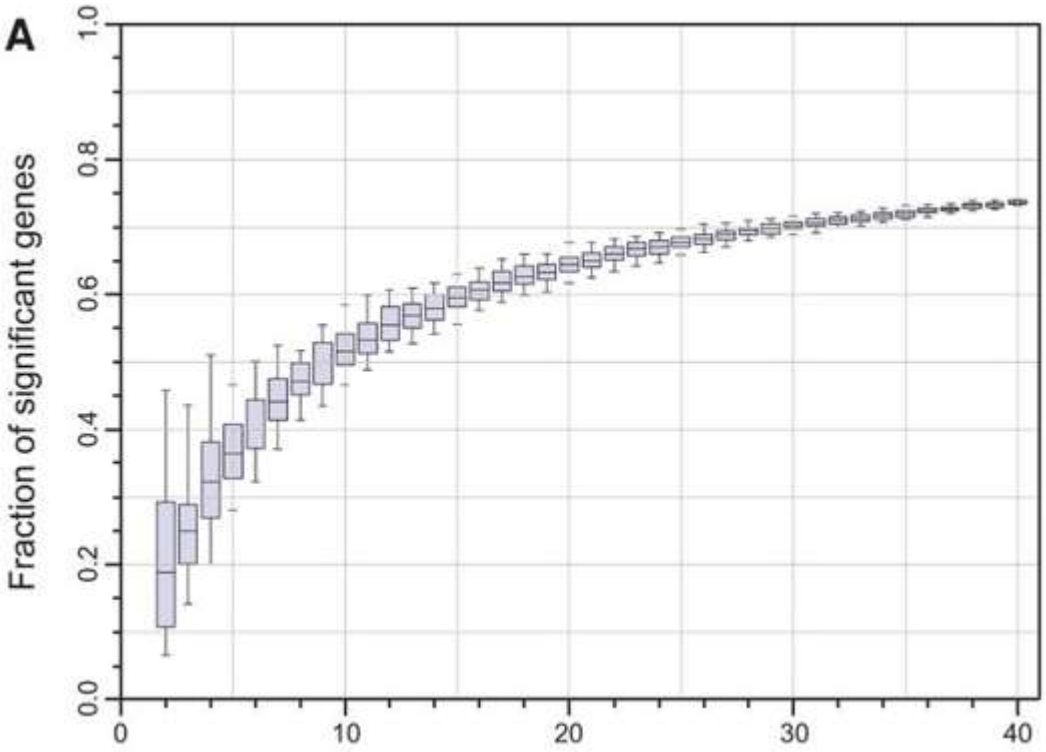
反復数増→DEG数増16



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。

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

反復数増→DEG数増17



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

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bulk RNA-seqデータの話。

反復数増→DEG数増17

①scRNA-seqのベストプラクティス論文。②Introductionの一部抜粋。③scRNA-seqでのcell types間比較では、DEGが60%程度に達する場合もあるし、トータルのmRNA量も細胞間で異なる。

Nat Commun. 2019 Oct 11;10(1):4667. doi: 10.1038/s41467-019-12266-7.

A systematic evaluation of single cell RNA-seq analy

Vieth B[1], Parekh S[2], Ziegenhain C[3], Enard W[1], Hellmann I[4].

Author information

Abstract

The recent rapid spread of single cell RNA sequencing
large variety of experimental and computational pipelir
yet been established. Here, we use simulations based
combination with nine realistic differential expression (
three mapping, four imputation, seven normalisation ar
approaches resulting in ~3000 pipelines, allowing us to
pipeline steps. We find that choices of normalisatio
biggest impact on scRNA-seq analyses. Specifically, w
the ability to detect symmetric expression differences,
performance in asymmetric DE-setups. Finally, we illus
by showing that a good scRNA-seq pipeline can have
biological signal as quadrupling the sample size.

PMID: 31604912 PMCID: PMC6789098 DOI: 10.1038/s41467-

③

One main assumption in traditional DE-analysis is that differences in expression are symmetric. This implies that either a small fraction of genes is DE while the expression of the majority of genes remains constant or similar numbers of genes are up- and down-regulated so that the mean total mRNA content does differ between groups[11]. This assumption is no longer true when diverse cell types are considered. For example, Zeisel et al.[12] find up to 60% DE genes and differing amounts of total mRNA levels between cell types. This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far. Therefore, we simulate varying numbers of DE-genes in conjunction with small to large differences in mRNA content including the entire spectrum of possible DE-settings.

反復数増→DEG数増18

①scRNA-seq発現変動解析用Rパッケージ NBIDの論文。②scRNA-seqは、まず探索的な解析（クラスタリング）により似た発現パターンを示す細胞をグループ化する。この段階ですでに「同一グループ内の細胞間の全体的な類似度が高いことが保証されている」ので、適当なp値やFDR閾値を満たす発現変動遺伝子（DEG）が多数得られてしまうので注意してね、ということが書かれている。

Genome Biol. 2018 May 31;19(1):70. doi: 10.1186/s13059-018-1438-9.

UMI-count modeling and differential expression anal
single-cell RNA sequencing.

Chen W[1], Li Y[2], Easton J[1], Finkelstein D[1]

Author information

Abstract

Read counting and unique molecular
expression quantification schemes us
By using multiple scRNA-seq dataset
these schemes and conclude that the
UMI counts, even in heterogeneous p
expression analysis algorithm based
dispersions in each group (NBID). Ou
achieves better power for UMI counts
for scRNA-seq analysis.

KEYWORDS: Differential expression ana

PMID: 29855333 PMCID: PMC5984373

single-cell analysis. Even though
the current study, the general for
tested simultaneously, as in the g

Differential expression analysis can be used to reveal differences among samples run on separate lanes or plates. However, it should be pointed out that batch effects are expected to overlap with biological differences in this setting. It is better to account for batch effects by proper experimental design, such as by including multiple biological replicates for each group. Another typical application of differential expression analysis is to identify potential biomarkers for inferred cell subpopulations. One potential caveat in this setting is that cells are usually clustered from the same data. Therefore, *p* values or FDR values derived from the differential expression analysis might be overly optimistic. However, the result is still useful for prioritizing potential biomarkers for further validation.

②

反復数増→DEG数増19

①scRNA-seq発現変動解析用Rパッケージ NBIDの論文。②scRNA-seqは、まず探索的な解析（クラスタリング）により似た発現パターンを示す細胞をグループ化する。この段階ですでに「同一グループ内の細胞間の全体的な類似度が高いことが保証されている」ので、適当なp値やFDR閾値を満たす発現変動遺伝子（DEG）が多数得られてしまうので注意してね、ということが書かれている。② 以外の多数のDEGが得られる要因が…

Genome Biol. 2018 May 31;19(1):70. doi: 10.1186/s13059-018-1438-9.

UMI-count modeling and differential expression anal
single-cell RNA sequencing.

Chen W[1], Li Y[2], Easton J[1], Finkelstein D[1]

Author information

Abstract

Read counting and unique molecular
expression quantification schemes us
By using multiple scRNA-seq dataset
these schemes and conclude that the
UMI counts, even in heterogeneous p
expression analysis algorithm based
dispersions in each group (NBID). Ou
achieves better power for UMI counts
for scRNA-seq analysis.

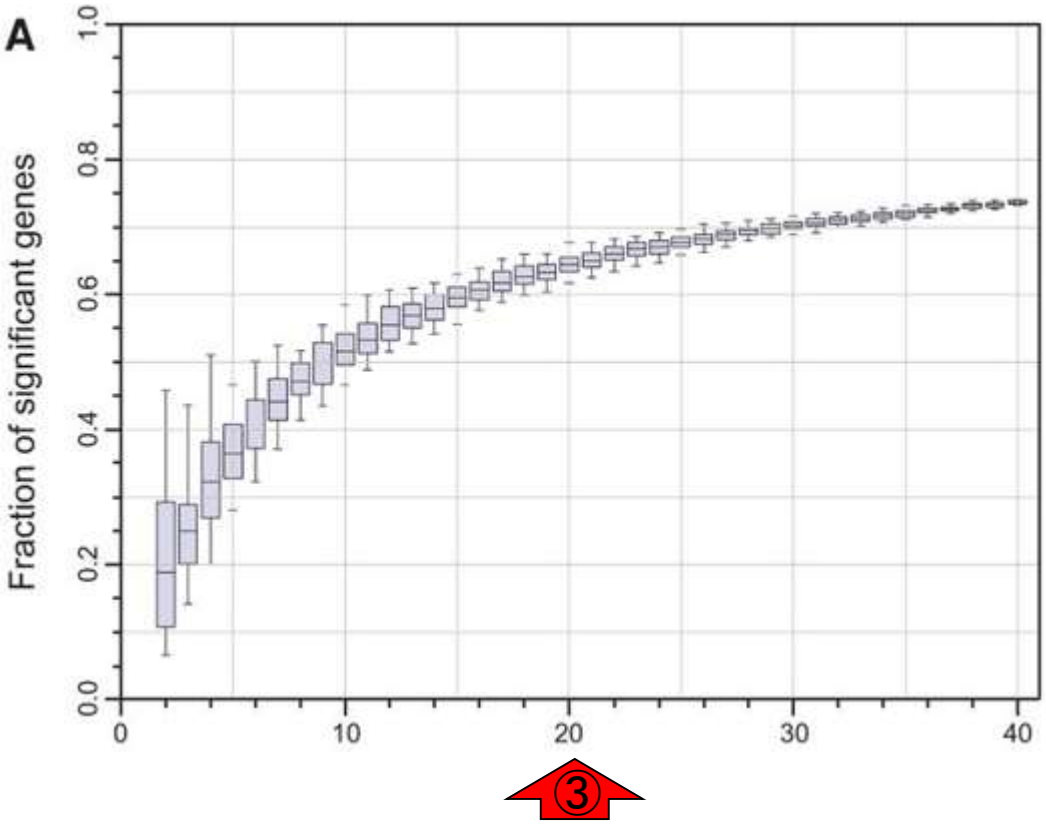
KEYWORDS: Differential expression ana

PMID: 29855333 PMCID: PMC5984373

single-cell analysis. Even though
the current study, the general for
tested simultaneously, as in the g

Differential expression analysis
samples run on separate lanes or plates. However, it should be pointed out that batch effects are expected to overlap with biological differences in this setting. It is better to account for batch effects by proper experimental design, such as by including multiple biological replicates for each group. Another typical application of differential expression analysis is to identify potential biomarkers for inferred cell subpopulations. One potential caveat in this setting is that cells are usually clustered from the same data. Therefore, *p* values or FDR values derived from the differential expression analysis might be overly optimistic. However, the result is still useful for prioritizing potential biomarkers for further validation.

反復数増→DEG数増20

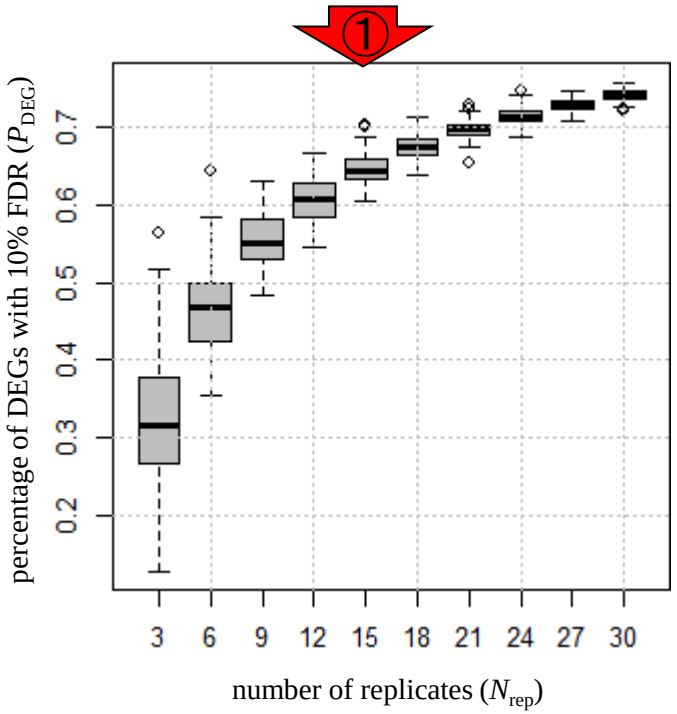


Schurch et al., *RNA*, **22**: 839–851, 2016

①scRNA-seq発現変動解析用Rパッケージ NBIDの論文。②scRNA-seqは、まず探索的な解析（クラスタリング）により似た発現パターンを示す細胞をグループ化する。この段階ですでに「同一グループ内の細胞間の全体的な類似度が高いことが保証されている」ので、適当なp値やFDR閾値を満たす発現変動遺伝子（DEG）が多数得られてしまうので注意してね、ということが書かれている。②以外の多数のDEGが得られる要因が、左図でいうところの③横軸に相当する数。これはbulk RNA-seqなので反復数に相当するが、scRNA-seqの場合は細胞数に置き換える。そして「DEGが存在するデータの場合、cell type間で発現変動解析を行うと、cell type内の細胞数が多いほど多数のDEGが得られる傾向にある（はず）」ということを正しく認識しておかねばなりません。

反復数増→DEG数増21

①さきほどと同じデータに対して、別のパッケージであるTCCを用いて発現変動解析を行った結果。②原著論文。TCCでも同じような結果が得られることを言いたいだけ。

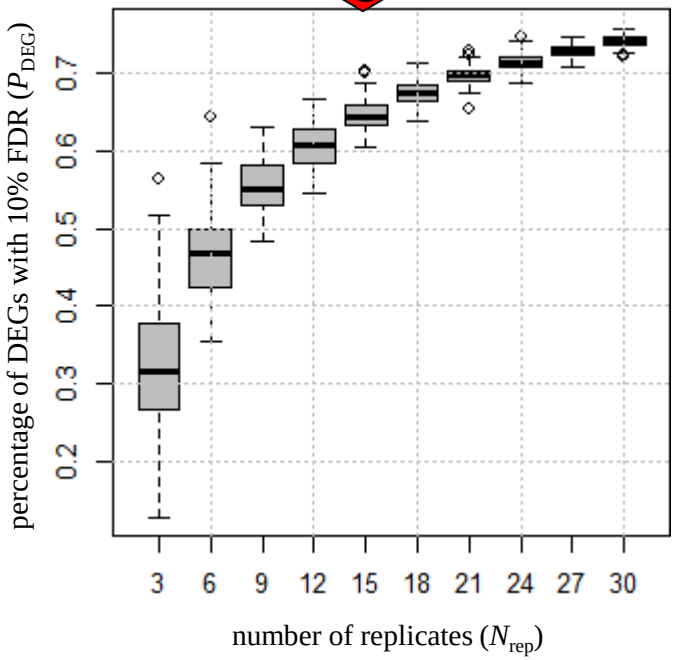


② Zhao et al., *Biol. Proc. Online*, **20**: 5, 2018
のAdditional file 3a

反復数増→DEG数増22

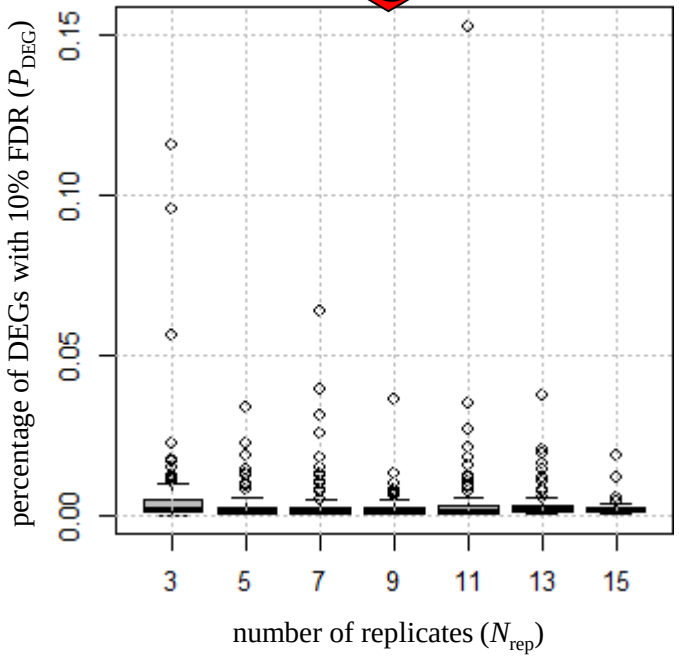
①さきほどと同じデータに対して、別のパッケージであるTCCを用いて発現変動解析を行った結果。②原著論文。TCCでも同じような結果が得られることを言いたいだけ。③が本題。DEGがほぼ存在しない場合は、反復数は無関係です。

①



② Zhao et al., *Biol. Proc. Online*, **20**: 5, 2018 のAdditional file 3a

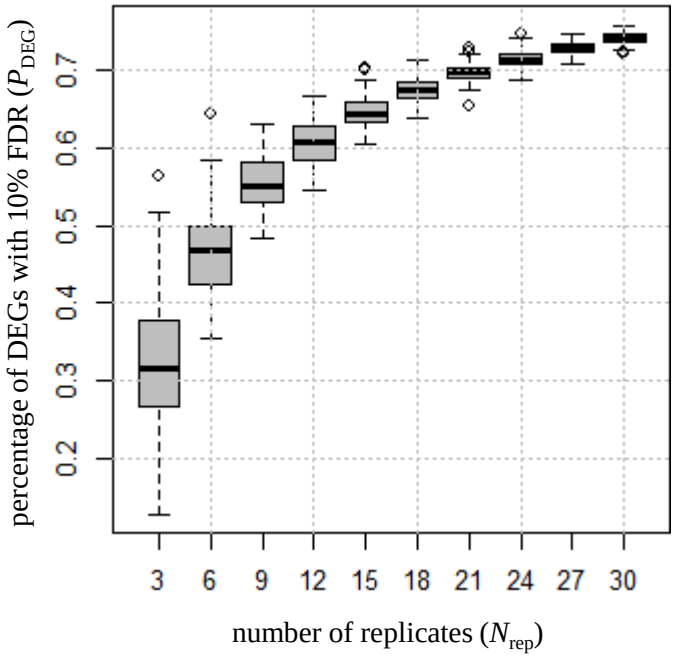
③



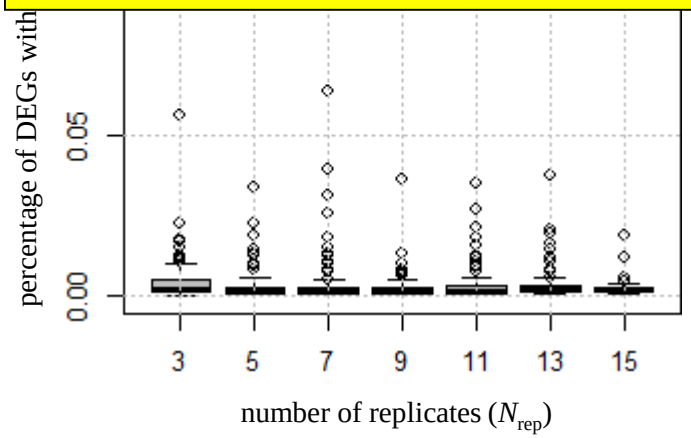
② Zhao et al., *Biol. Proc. Online*, **20**: 5, 2018 のAdditional file 5a

反復数増→DEG数増23

①さきほどと同じデータに対して、別のパッケージであるTCCを用いて発現変動解析を行った結果。②原著論文。TCCでも同じような結果が得られることを言いたいだけ。③が本題。DEGがほぼ存在しない場合は、反復数は無関係です。
scRNA-seqで最初にクラスタリングでグループ化したものを用いて発現変動解析を行う場合は、グループ内の高類似度の効果によってDEGが一定数得られる場合がほとんどだろうとは思います。



Zhao et al., *Biol. Proc. Online*, **20**: 5, 2018
のAdditional file 3a



Zhao et al., *Biol. Proc. Online*, **20**: 5, 2018
のAdditional file 5a

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

ここまでのまとめ

- 非対称性（asymmetry）の問題への対応は、データ正規化も含めてbulkでやられている。
- Zero inflationの特徴発見はおそらくbulkが先であり、数少ない「反復の多いbulkデータ」で見出されている。scRNA-seqは細胞数が多いのでよりはっきりと見出されたのであろう。
- scRNA-seqデータのうち、UMIカウントデータのほうは、bulkと同じNBモデルに従う可能性が高い。
  細胞数が少ないデータでの発見なだけかも?!
- scRNA-seqの発現変動解析系の評価は未だ2群間比較が中心。プログラム自体はGLMでmulti-group対応のものもちらほら。

①科学者は常に正直、誠実に判断し、行動し、…科学研究によって生み出される知の正確さや正当性を科学的に示す最善の努力を払うべきでしょう。何を比較対象とし、どのような条件で得られた結論なのかを冷徹に見極めることが重要だと、私は思います。私の知見や見解をまとめると左記のようになります。

①

https://www.jsps.go.jp/j-kousei/data/rinri.pdf

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

GLMベースの3群間比較

①scRNA-seq用の発現変動解析プログラムNBIDの論文は、②一般化線形モデル（generalized linear model; GLM）の枠組みで検定を行うのでmulti-group対応。

Genome Biol. 2018 May 31;19(1):70. doi: 10.1186/s13059-018-1438-9.

UMI-count modeling and differential expression analysis for single-cell RNA sequencing.

Chen W[1], Li Y[2], Easton J[1], Finkelstein D[1], Wu G[1], Chen X[3].

Author information

Abstract

Read counting and unique molecular identifi
expression quantification schemes used in s
By using multiple scRNA-seq datasets, we re
these schemes and conclude that the negati
UMI counts, even in heterogeneous populati
expression analysis algorithm based on a ne
dispersions in each group (NBID). Our result
achieves better power for UMI counts when
for scRNA-seq analysis.

KEYWORDS: Differential expression analysis; Ne

PMID: 29855333 PMCID: PMC5984373 DOI: 10.1

①

single-cell analysis. Even though only pairwise analyses were considered in the current study, the general form of NBID allows multiple groups to be tested simultaneously, as in the generalized linear model framework.

②

Differential expression analysis can be used to reveal differences among samples run on separate lanes or plates. However, it should be pointed out that batch effects are expected to overlap with biological differences in this setting. It is better to account for batch effects by proper experimental design, such as by including multiple biological replicates for each group. Another typical application of differential expression analysis is to identify potential biomarkers for inferred cell subpopulations. One potential caveat in this setting is that cells are usually clustered from the same data. Therefore, *p* values or FDR values derived from the differential expression analysis might be overly optimistic. However, the result is still useful for prioritizing potential biomarkers for further validation.

GLMベースの3群間比較

①scRNA-seq用の発現変動解析プログラムNBIDの論文は、②一般化線形モデル（generalized linear model; GLM）の枠組みで検定を行うのでmulti-group対応。
私の認識している一般的なGLMの弱点をbulk RNA-seq用のTCCパッケージ実行結果で示し、その解決策について述べます。

Genome Biol. 2018 May 31;19(1):70. doi: 10.1186/s13059-018-1438-9.

UMI-count modeling and differential expression analysis for single-cell RNA sequencing.

Chen W[1], Li Y[2], Easton J[1], Finkelstein D[1], Wu G[1], Chen X[3].

Author information

Abstract

Read counting and unique molecular identifi
expression quantification schemes used in s
By using multiple scRNA-seq datasets, we re
these schemes and conclude that the negati
UMI counts, even in heterogeneous populati
expression analysis algorithm based on a ne
dispersions in each group (NBID). Our result
achieves better power for UMI counts when
for scRNA-seq analysis.

KEYWORDS: Differential expression analysis; Ne

PMID: 29855333 PMCID: PMC5984373 DOI: 10.1

single-cell analysis. Even thoug
the current study, the general form of NBID allows multiple groups to be tested simultaneously, as in the generalized linear model framework.

Differential expression analysis can be used to reveal differences among samples run on separate lanes or plates. However, it should be pointed out that batch effects are expected to overlap with biological differences in this setting. It is better to account for batch effects by proper experimental design, such as by including multiple biological replicates for each group. Another typical application of differential expression analysis is to identify potential biomarkers for inferred cell subpopulations. One potential caveat in this setting is that cells are usually clustered from the same data. Therefore, *p* values or FDR values derived from the differential expression analysis might be overly optimistic. However, the result is still useful for prioritizing potential biomarkers for further validation.

GLMベースの3群間比較

①scRNA-seq用の発現変動解析プログラムNBIDの論文は、②一般化線形モデル（generalized linear model; GLM）の枠組みで検定を行うのでmulti-group対応。私の認識している一般的なGLMの弱点をbulk RNA-seq用のTCCパッケージ実行結果で示し、その解決策について述べます。①NBIDと③DECENTともにmulti-group対応であり、もしかしたら以降のスライドで課題として挙げている弱点が補強されているかもしれませんので、予めご了承ください。

Genome Biol. 2018 May 31;19(1):70. doi: 10.1186/s13059-018-1438-9.

UMI-count modeling and differential expression analysis for single-cell RNA sequencing.

①

Chen W[1], Li Y[2], Easton J[1], Finkelstein D[1], Wu G[1], Chen X[3].

Author information

Abstract

Read counting and unique molecular identifier (UMI) counting are the principal gene expression quantification schemes used in s
By using multiple scRNA-seq datasets, we r
these schemes and conclude that the negati
UMI counts, even in heterogeneous populati
expression analysis algorithm based on a ne
dispersions in each group (NBID). Our result
achieves better power for UMI counts when
for scRNA-seq analysis.

KEYWORDS: Differential expression analysis; Ne

PMID: 29855333 PMCID: PMC5984373 DOI: 10.1

Bioinformatics. 2019 Dec 15;35(24):5155-5162. doi: 10.1

DECENT: differential expression single-cell RNA-seq data.

③

Ye C[1,2,3], Speed TP[1,4], Salim A[1,5,6].

Author information

Abstract

MOTIVATION: Dropout is a common phenomenon in single-cell RNA-seq (scRNA-seq) data, and when left unaddressed it affects the validity of the statistical analyses. Despite this, few current methods for differential expression (DE) analysis of scRNA-seq data explicitly model the process that gives rise to the dropout events. We develop DECENT, a method for DE analysis of scRNA-seq data that explicitly and accurately models the molecule capture process in scRNA-seq experiments.

RESULTS: We show that DECENT demonstrates improved DE performance over existing DE methods that do not explicitly model dropout. This improvement is consistently observed across several public scRNA-seq datasets generated using different technological platforms. The gain in improvement is especially large when the capture process is overdispersed. DECENT maintains type I error well while achieving better sensitivity. Its performance without spike-ins is almost as good as when spike-ins are used to calibrate the capture model.

GLMベースの3群間比較

A群 vs. B群 vs. C群のようなデータの場合、GLMベースの方法では…

	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

…

GLMベースの3群間比較

A群 vs. B群 vs. C群のようなデータの場合、GLMベースの方法では、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			
	A1	A2	A3	B1	B2	B3	C1	C2	C2	q.value
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

…

GLMベースの3群間比較

A群 vs. B群 vs. C群のようなデータの場合、GLMベースの方法では、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			
	A1	A2	A3	B1	B2	B3	C1	C2	C2	q.value
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

①

GLMベースの3群間比較

A群 vs. B群 vs. C群のようなデータの場合、GLMベースの方法では、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			
	A1	A2	A3	B1	B2	B3	C1	C2	C2	q.value
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

①

②

GLMベースの3群間比較

A群 vs. B群 vs. C群のようなデータの場合、GLMベースの方法では、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			
	A1	A2	A3	B1	B2	B3	C1	C2	C2	q.value
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

①
②
③

GLMベースの3群間比較

GLMベースのbulk RNA-seq用プログラムである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			
	A1	A2	A3	B1	B2	B3	C1	C2	C2	q.value
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

①
②
③

GLMベースの3群間比較

GLMベースのbulk RNA-seq用プログラムである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				④
	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

GLMベースの3群間比較

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				④
	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

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

Osabe法で…1

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。基本はbulk RNA-seq用です。筆頭著者の名前を冠して、Osabe法と勝手に命名

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

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

Osabe T[1], 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法で…2

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です。

①

②

	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法で…3

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です。③は従来法の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	③ q.value	④ 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法で…4

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です。③は従来法の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

	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法で…5

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…

①

	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

Osabe法で…6

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…

①

	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

Osabe法で…7

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。

①

②

	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

Osabe法で…8

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

Osabe法で…9

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ます。

②

	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

①

Osabe法で…10

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ます。例えば、③「A>other」というパターンは、「A群のみで高発現」というパターンです。

	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	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

③

Osabe法で…11

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ます。例えば、③「A>other」というパターンは、「A群のみで高発現」というパターンです。④確かにそうなってますね。

	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	④.000	0.613	0.002	0.001	0.384	DEG_A	A>other

Osabe法で…12

実用上は、②のorderingsという列の情報に対して、「どのパターンのものがいくつあったか?」という情報が欲しいです。

②

	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

Osabe法で…13

実用上は、②のorderingsという列の情報に対して、「どのパターンのものがいくつあったか?」という情報が欲しいです。Rでは、例えば②orderings列の情報からなるベクトルを、③入力として、④tableという関数を実行することで…

R Console

```
> table(out$MAP)

 DEG_A  DEG_B  DEG_C DEGall nonDEG
   ④      ③     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
>
```

②

										G_B	DEG_C	DEGall	pattern	orderings
										.000	0.000	0.001	DEG_A	A>other
										.000	0.000	0.004	DEG_A	A>other
										.000	0.000	0.003	DEG_A	A>other
										.993	0.000	0.007	DEG_B	B>other
										.000	0.000	0.155	DEG_A	A>other
										.000	0.000	0.083	DEG_A	A>other
gene_2333	118	112	133	1430	1440	1393	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

Osabe法で…14

実用上は、②のorderingsという列の情報に対して、「どのパターンのものがいくつあったか?」という情報が欲しいです。Rでは、例えば②orderings列の情報からなるベクトルを、③入力として、④tableという関数を実行することで、⑤赤枠のような結果が得られます。

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
>
```

⑤

											G_B	DEG_C	DEGall	pattern	orderings
											.000	0.000	0.001	DEG_A	A>other
											.000	0.000	0.004	DEG_A	A>other
											.000	0.000	0.003	DEG_A	A>other
											.993	0.000	0.007	DEG_B	B>other
											.000	0.000	0.155	DEG_A	A>other
											.000	0.000	0.083	DEG_A	A>other
gene_2333	118	112	133	1430	1440	1393	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

Osabe法で…15

実用上は、②のorderingsという列の情報に対して、「どのパターンのものがいくつあったか?」という情報が欲しいです。Rでは、例えば②orderings列の情報からなるベクトルを、③入力として、④tableという関数を実行することで、⑤赤枠のような結果が得られます。例えば⑥は、③で見た「A>other」というパターンが1245個、そして「other>A」というパターンが48個あったことを示しています。

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
>
```

⑥

gene_2333	118	112	133	1430	1440	1393	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

③

Osabe法で…16

実用上は、②のorderingsという列の情報に対して、「どのパターンのものがいくつあったか?」という情報が欲しいです。Rでは、例えば②orderings列の情報からなるベクトルを、③入力として、④tableという関数を実行することで、⑤赤枠のような結果が得られます。例えば⑥は、③で見た「A>other」というパターンが1245個、そして「other>A」というパターンが48個あったことを示しています。これらは①「DEG_A」パターンに属しますが…

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>oth
   7493       3       4     1245       2     6
  C>A>B   C>B>A C>other  other>A other>B other>C
      1       4     458       48      36      17
>
```

⑥

gene_2333	118	112	133	1430	1440	1393	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

①

Osabe法で…17

⑦ここの列に相当する情報である、⑧out$MAPというベクトルを入力として同様の作業を行うことで…

⑧

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
>
```

⑦

										G_B	DEG_C	DEGall	pattern	orderings
										.000	0.000	0.001	DEG_A	A>other
										.000	0.000	0.004	DEG_A	A>other
										.000	0.000	0.003	DEG_A	A>other
										.993	0.000	0.007	DEG_B	B>other
										.000	0.000	0.155	DEG_A	A>other
										.000	0.000	0.083	DEG_A	A>other
gene_2333	118	112	133	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

Osabe法で…18

⑦ここの列に相当する情報である、⑧out$MAPというベクトルを入力として同様の作業を行うことで、⑨「DEG_A」パターンが1293個という結果を得ることもできます。

R Console

```
> table(out$MAP)
⑨
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
>
```

											G_B	DEG_C	DEGall	pattern	orderings
											.000	0.000	0.001	DEG_A	A>other
											.000	0.000	0.004	DEG_A	A>other
											.000	0.000	0.003	DEG_A	A>other
											.993	0.000	0.007	DEG_B	B>other
											.000	0.000	0.155	DEG_A	A>other
											.000	0.000	0.083	DEG_A	A>other
gene_2333	118	112	133	1430	1440	1393	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

Osabe法で…19

⑦ここの列に相当する情報である、⑧out$MAPというベクトルを入力として同様の作業を行うことで、⑨「DEG_A」パターンが1293個という結果を得ることもできます。⑥のパターン数を足した結果と一致してますね。

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
>
```

⑥

													G_B	DEG_C	DEGall	pattern	orderings
													.000	0.000	0.001	DEG_A	A>other
													.000	0.000	0.004	DEG_A	A>other
													.000	0.000	0.003	DEG_A	A>other
													.993	0.000	0.007	DEG_B	B>other
													.000	0.000	0.155	DEG_A	A>other
													.000	0.000	0.083	DEG_A	A>other
gene_2333	118	112	133	1430	1440	1393	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	

Osabe法で…20

ちなみに、①DEGallは、全ての群間で発現変動しているパターン。

①

	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

Osabe法で…21

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、③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	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

③ ②

Osabe法で…22

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、③の発現パターン（A >> C > B）を見れば納得できる。
実際には④「DEG_A」パターンのほうが居心地がよさそうだとbaySeqが判断したので、⑤のように判定されたのだと納得すればよい。

	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	④	0.000	0.995	0.005	DEG_C	C>⑤
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

Contents（今回）

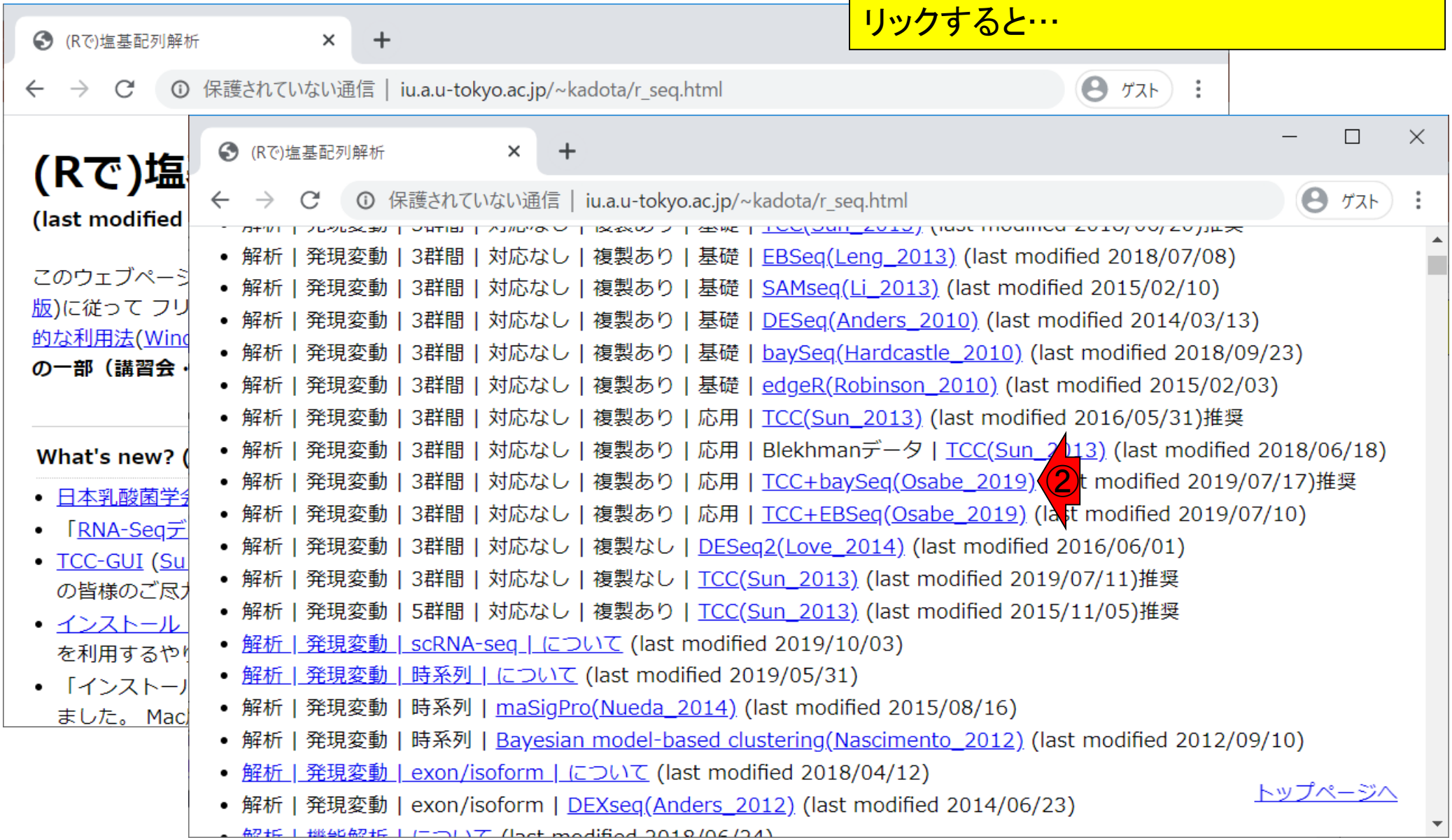
- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

Osabe法は、①で提供しています。

Osabe法の実践1

Osabe法の実践2

Osabe法は、①で提供しています。沢山項目がありますが、②が(bulk) RNA-seqで評価した場合の推奨手法です。②をクリックすると…



Osabe法の実践3

Osabe法は、①で提供しています。沢山項目がありますが、②が(bulk) RNA-seqで評価した場合の推奨手法です。②をクリックすると、こんな感じになります。

解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 応用 |

TCC+baySeq(Osabe_2019)

TCCとbaySeq を組み合わせたやり方を示します。**例題8以降が推奨パイプライン(Osabe et al., Bioinform. Biol. Insight, 2019)です(2019年7月10日追記)。** 例題1-7までは、「TCC中のiDEGES/edgeR正規化で得られた正規化係数をbaySeqに与えた解析パイプライン」です。 TCC原著論文中のiDEGES/edgeR-baySeqという解析パイプラインに相当し、 多群間比較用の推奨ガイドライン提唱論文(Tang et al., 2015)の表記法に従うとEEE-bに相当します。

その後、このパイプライン(EEE-b)は、TCCのデフォルトの解析パイプライン(EEE-E)よりも全体的な発現変動のランキングの点で劣っていることが判明しました(Osabe et al., 2019)。 しかし、EEE-Eはどこかの群間で発現変動しているというANOVA的な結果までしか返さないのに対して、EEE-bは発現変動パターンの割当て(や分類)の点で優位です。 理由は、TCCやedgeRやDESeq2を用いて3群間比較で発現変動パターンの割当まで行う場合、「G1 vs. G2、G1 vs. G3、G2 vs. G3の3通りの2群間比較を独立に行ってから、 その結果に基づいて発現変動パターンを構築する」とか、あるいは 「(G1+G2) vs. G3、(G1+G3) vs. G2、(G2+G3) vs. G1を行ってから構築する」など**やろうと思えばやれないことはないが現実には結構大変**だからです。 **例題8は、全体的な発現変動の度合い(ANOVA的などこかの群間で発現変動している度合いでのランキング)をEEE-Eで行い、 発現変動パターンの同定をEEE-b(長部らの原著論文中ではEEE-baySEqと表記)で行った結果を出力**しています。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

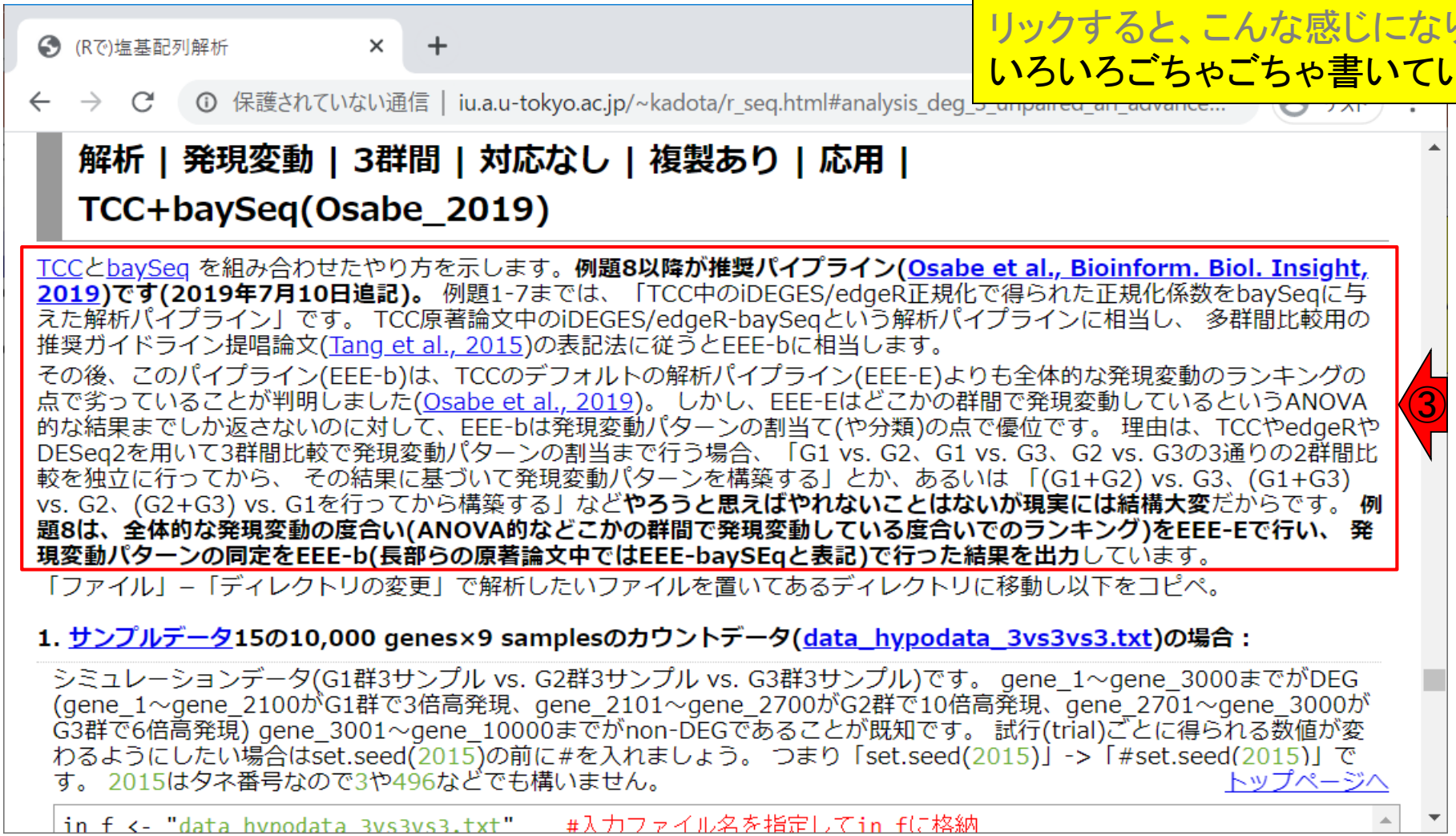
1. サンプルデータ15の10,000 genes×9 samplesのカウントデータ(data_hypodata_3vs3vs3.txt)の場合：

シミュレーションデータ(G1群3サンプル vs. G2群3サンプル vs. G3群3サンプル)です。 gene_1～gene_3000までがDEG (gene_1～gene_2100がG1群で3倍高発現、gene_2101～gene_2700がG2群で10倍高発現、gene_2701～gene_3000がG3群で6倍高発現) gene_3001～gene_10000までがnon-DEGであることが既知です。 試行(trial)ごとに得られる数値が変わるようにしたい場合はset.seed(2015)の前に#を入れましょう。 つまり「set.seed(2015)」->「#set.seed(2015)」です。 2015はタネ番号なので3や496などでも構いません。 トップページへ

```
in_f <- "data_hypodata_3vs3vs3.txt"    #入力ファイル名を指定してin_fに格納
```

Osabe法の実践4

Osabe法は、①で提供しています。沢山項目がありますが、②が(bulk) RNA-seqで評価した場合の推奨手法です。②をクリックすると、こんな感じになります。③いろいろごちゃごちゃ書いていますが…



(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#analysis_deg_5_unpaired_an_advance...

解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 応用 |

TCC+baySeq(Osabe_2019)

TCCとbaySeq を組み合わせたやり方を示します。**例題8以降が推奨パイプライン(Osabe et al., Bioinform. Biol. Insight, 2019)です(2019年7月10日追記)。** 例題1-7までは、「TCC中のiDEGES/edgeR正規化で得られた正規化係数をbaySeqに与えた解析パイプライン」です。 TCC原著論文中のiDEGES/edgeR-baySeqという解析パイプラインに相当し、 多群間比較用の推奨ガイドライン提唱論文(Tang et al., 2015)の表記法に従うとEEE-bに相当します。

その後、このパイプライン(EEE-b)は、TCCのデフォルトの解析パイプライン(EEE-E)よりも全体的な発現変動のランキングの点で劣っていることが判明しました(Osabe et al., 2019)。 しかし、EEE-Eはどこかの群間で発現変動しているというANOVA的な結果までしか返さないのに対して、EEE-bは発現変動パターンの割当て(や分類)の点で優位です。 理由は、TCCやedgeRやDESeq2を用いて3群間比較で発現変動パターンの割当まで行う場合、「G1 vs. G2、G1 vs. G3、G2 vs. G3の3通りの2群間比較を独立に行ってから、 その結果に基づいて発現変動パターンを構築する」とか、あるいは 「(G1+G2) vs. G3、(G1+G3) vs. G2、(G2+G3) vs. G1を行ってから構築する」など**やろうと思えばやれないことはないが現実には結構大変**だからです。 **例題8は、全体的な発現変動の度合い(ANOVA的などこかの群間で発現変動している度合いでのランキング)をEEE-Eで行い、 発現変動パターンの同定をEEE-b(長部らの原著論文中ではEEE-baySEqと表記)で行った結果を出力**しています。

③

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

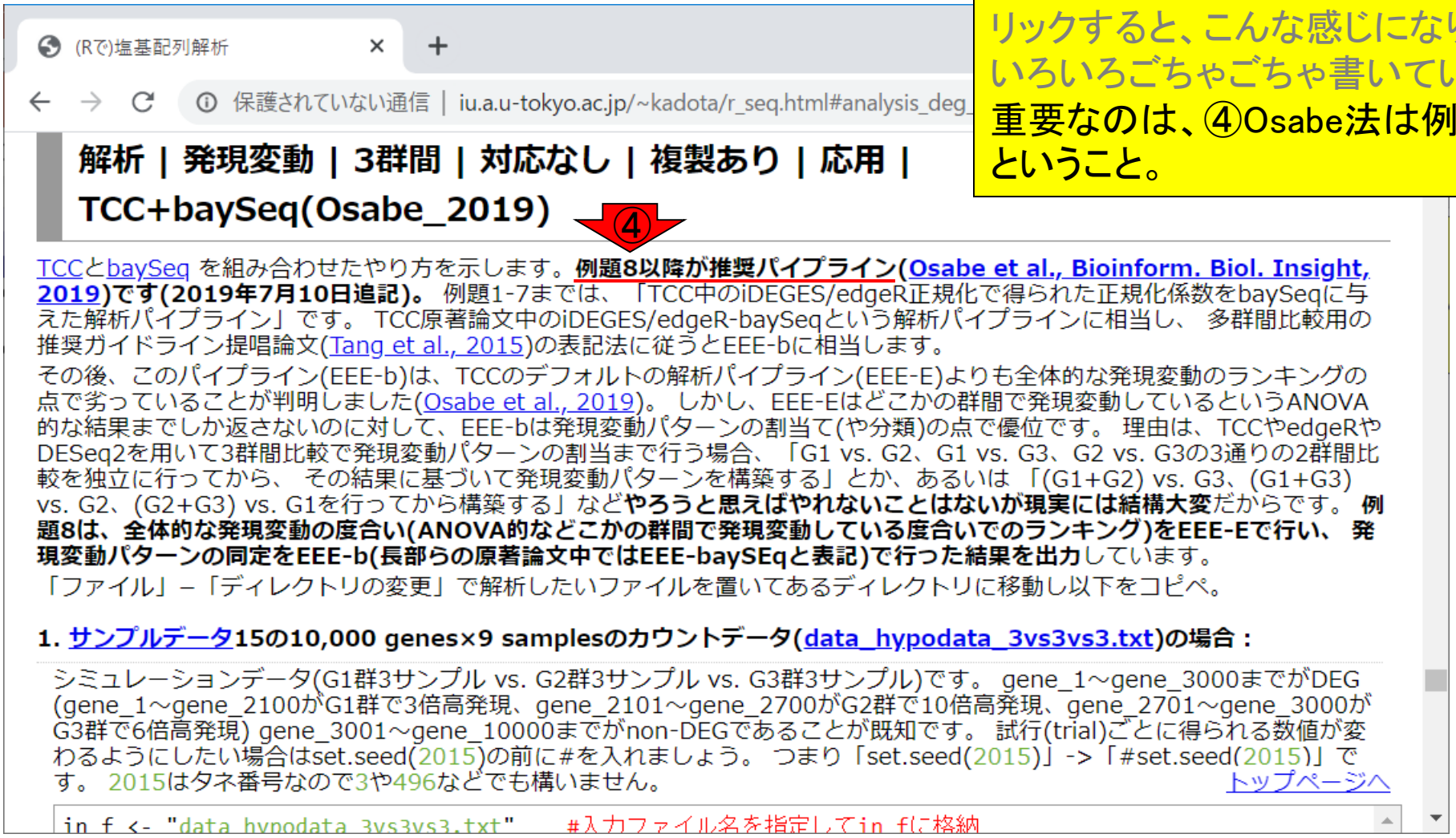
1. サンプルデータ15の10,000 genes×9 samplesのカウントデータ(data_hypodata_3vs3vs3.txt)の場合：

シミュレーションデータ(G1群3サンプル vs. G2群3サンプル vs. G3群3サンプル)です。 gene_1～gene_3000までがDEG (gene_1～gene_2100がG1群で3倍高発現、gene_2101～gene_2700がG2群で10倍高発現、gene_2701～gene_3000がG3群で6倍高発現) gene_3001～gene_10000までがnon-DEGであることが既知です。 試行(trial)ごとに得られる数値が変わるようにしたい場合はset.seed(2015)の前に#を入れましょう。 つまり「set.seed(2015)」->「#set.seed(2015)」です。 2015はタネ番号なので3や496などでも構いません。 トップページへ

```
in_f <- "data_hypodata_3vs3vs3.txt"    #入力ファイル名を指定してin_fに格納
```

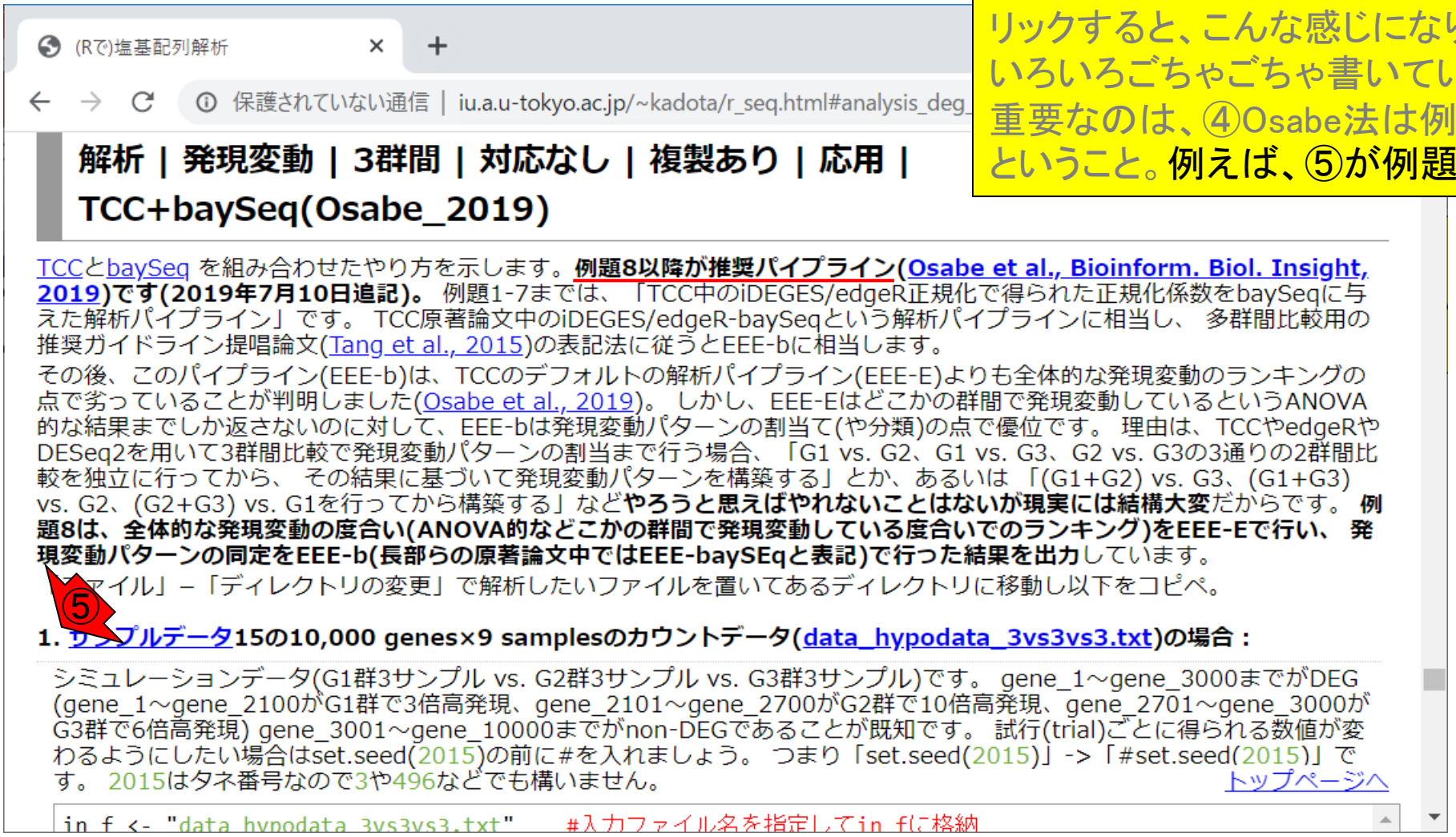
Osabe法の実践5

Osabe法は、①で提供しています。沢山項目がありますが、②が(bulk) RNA-seqで評価した場合の推奨手法です。②をクリックすると、こんな感じになります。③いろいろごちゃごちゃ書いていますが、重要なのは、④Osabe法は例題8以降だということ。



Osabe法の実践6

Osabe法は、①で提供しています。沢山項目がありますが、②が(bulk) RNA-seqで評価した場合の推奨手法です。②をクリックすると、こんな感じになります。③いろいろごちゃごちゃ書いていますが、重要なのは、④Osabe法は例題8以降だということ。例えば、⑤が例題1です。



(R で)塩基配列解析

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

解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 応用 | TCC+baySeq(Osabe_2019)

TCCとbaySeq を組み合わせたやり方を示します。**例題8以降が推奨パイプライン(Osabe et al., Bioinform. Biol. Insight, 2019)です(2019年7月10日追記)。** 例題1-7までは、「TCC中のiDEGES/edgeR正規化で得られた正規化係数をbaySeqに与えた解析パイプライン」です。 TCC原著論文中のiDEGES/edgeR-baySeqという解析パイプラインに相当し、 多群間比較用の推奨ガイドライン提唱論文(Tang et al., 2015)の表記法に従うとEEE-bに相当します。

その後、このパイプライン(EEE-b)は、TCCのデフォルトの解析パイプライン(EEE-E)よりも全体的な発現変動のランキングの点で劣っていることが判明しました(Osabe et al., 2019)。 しかし、EEE-Eはどこかの群間で発現変動しているというANOVA的な結果までしか返さないのに対して、EEE-bは発現変動パターンの割当て(や分類)の点で優位です。 理由は、TCCやedgeRやDESeq2を用いて3群間比較で発現変動パターンの割当まで行う場合、「G1 vs. G2、G1 vs. G3、G2 vs. G3の3通りの2群間比較を独立に行ってから、 その結果に基づいて発現変動パターンを構築する」とか、あるいは 「(G1+G2) vs. G3、(G1+G3) vs. G2、(G2+G3) vs. G1を行ってから構築する」など**やろうと思えばやれないことはないが現実には結構大変**だからです。 **例題8は、全体的な発現変動の度合い(ANOVA的などこかの群間で発現変動している度合いでのランキング)をEEE-Eで行い、 発現変動パターンの同定をEEE-b(長部らの原著論文中ではEEE-baySEqと表記)で行った結果を出力**しています。

⑤

ァイル」−「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. サンプルデータ15の10,000 genes×9 samplesのカウントデータ(data_hypodata_3vs3vs3.txt)の場合：

シミュレーションデータ(G1群3サンプル vs. G2群3サンプル vs. G3群3サンプル)です。 gene_1～gene_3000までがDEG (gene_1～gene_2100がG1群で3倍高発現、gene_2101～gene_2700がG2群で10倍高発現、gene_2701～gene_3000がG3群で6倍高発現) gene_3001～gene_10000までがnon-DEGであることが既知です。 試行(trial)ごとに得られる数値が変わるようにしたい場合はset.seed(2015)の前に#を入れましょう。 つまり「set.seed(2015)」->「#set.seed(2015)」です。 2015はタネ番号なので3や496などでも構いません。 トップページへ

```
in_f <- "data_hypodata_3vs3vs3.txt"    #入力ファイル名を指定してin_fに格納
```

Osabe法の実践7

Osabe法は、①で提供しています。沢山項目がありますが、②が(bulk) RNA-seqで評価した場合の推奨手法です。②をクリックすると、こんな感じになります。③いろいろごちゃごちゃ書いていますが、重要なのは、④Osabe法は例題8以降だということ。例えば、⑤が例題1です。このページ自体が非常にデカいので、例えば⑥の項目内の例題8を見つける際は、私はいつも⑦「下矢印キー」を押していって探します。

(Rで)塩基配列解析

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

解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 応用 |

TCC+baySeq(Osabe_2019)

⑥

TCCとbaySeq を組み合わせたやり方を示します。**例題8以降が推奨パイプライン(Osabe 2019)です(2019年7月10日追記)。** 例題1-7までは、「TCC中のiDEGES/edgeR正規化
えた解析パイプライン」です。 TCC原著論文中のiDEGES/edgeR-baySeqという解析パ
推奨ガイドライン提唱論文(Tang et al., 2015)の表記法に従うとEEE-bに相当します。
その後、このパイプライン(EEE-b)は、TCCのデフォルトの解析パイプライン(EEE-E)よ
点で劣っていることが判明しました(Osabe et al., 2019)。 しかし、EEE-Eはどこかの群間で発現変動しているという
的な結果までしか返さないのに対して、EEE-bは発現変動パターンの割当て(や分類)の点で優位です。 理由は、TCCや
DESeq2を用いて3群間比較で発現変動パターンの割当まで行う場合、「G1 vs. G2、G1 vs. G3、G2 vs. G3の3通り
較を独立に行ってから、 その結果に基づいて発現変動パターンを構築する」とか、あるいは 「(G1+G2) vs. G3、(G
vs. G2、(G2+G3) vs. G1を行ってから構築する」など**やろうと思えばやれないことはないが現実には結構大変**だから
**題8は、全体的な発現変動の度合い(ANOVA的などこかの群間で発現変動している度合いでのランキング)をEEE-Eで
現変動パターンの同定をEEE-b(長部らの原著論文中ではEEE-baySEqと表記)で行った結果を出力**しています。
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. サンプルデータ15の10,000 genes×9 samplesのカウントデータ(data_hypodata_3vs3vs3.txt)の場合：

シミュレーションデータ(G1群3サンプル vs. G2群3サンプル vs. G3群3サンプル)です。 gene_1～gene_3000まで
(gene_1～gene_2100がG1群で3倍高発現、gene_2101～gene_2700がG2群で10倍高発現、gene_2701～gene
G3群で6倍高発現) gene_3001～gene_10000までがnon-DEGであることが既知です。 試行(trial)ごとに得られる
わるようにしたい場合はset.seed(2015)の前に#を入れましょう。 つまり「set.seed(2015)」->「#set.seed(20
す。 2015はタネ番号なので3や496などでも構いません。 トッ

```
in_f <- "data_hypodata_3vs3vs3.txt"    #入力ファイル名を指定してin_fに格納
```

①例題8を発見。

Osabe法の実践8

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#analysis_deg_3_unpaired_ari_advance...

ゲスト

①

8. サンプルデータ15の10,000 genes×9 samplesのカウントデータ(data_hypodata_3vs3vs3.txt)の場合：

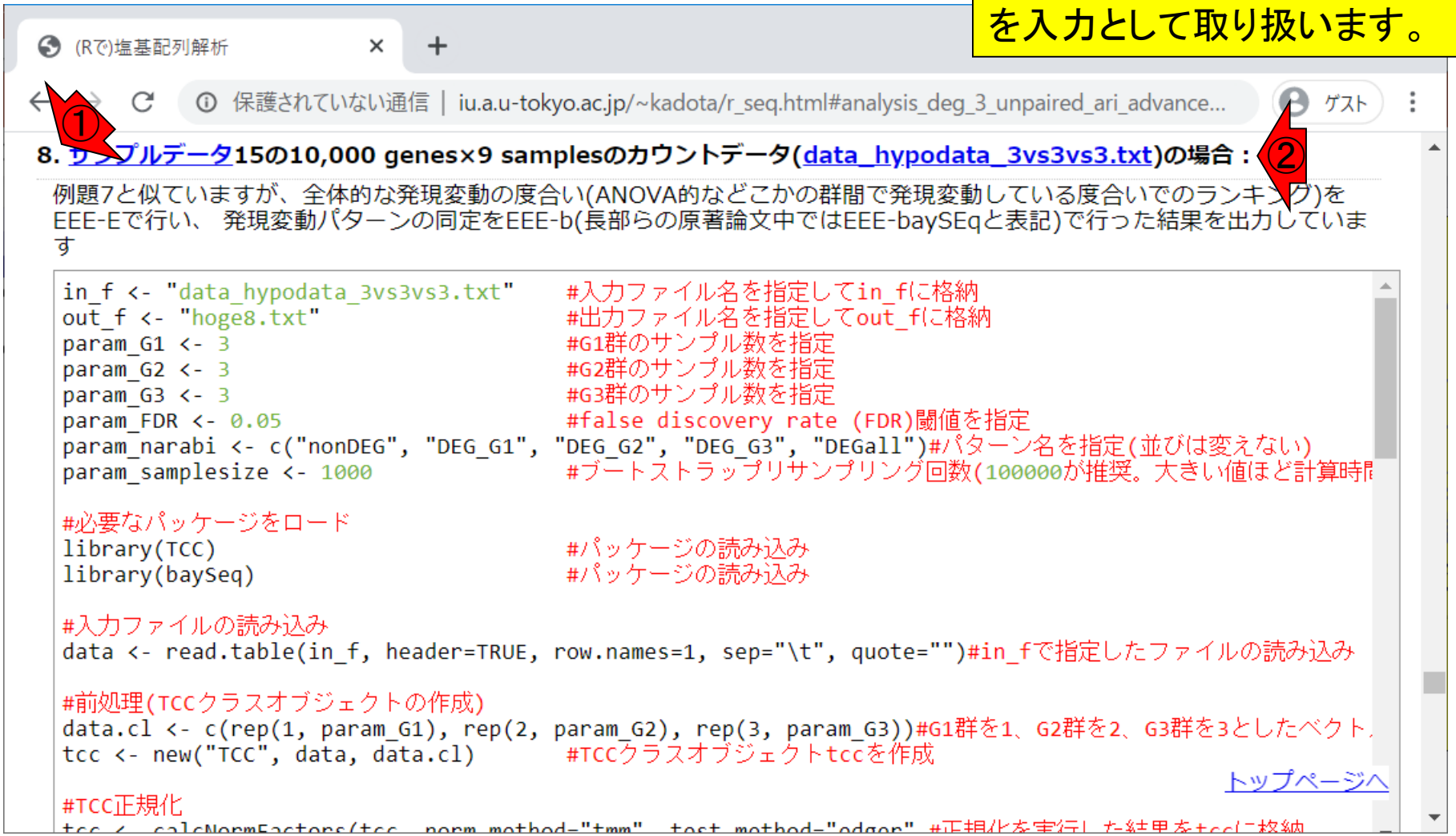
例題7と似ていますが、全体的な発現変動の度合い(ANOVA的などこかの群間で発現変動している度合いでのランキング)をEEE-Eで行い、 発現変動パターンの同定をEEE-b(長部らの原著論文中ではEEE-baySEqと表記)で行った結果を出力しています

```
in_f <- "data_hypodata_3vs3vs3.txt"     #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.txt"                    #出力ファイル名を指定してout_fに格納
param_G1 <- 3                           #G1群のサンプル数を指定
param_G2 <- 3                           #G2群のサンプル数を指定
param_G3 <- 3                           #G3群のサンプル数を指定
param_FDR <- 0.05                       #false discovery rate (FDR)閾値を指定
param_narabi <- c("nonDEG", "DEG_G1", "DEG_G2", "DEG_G3", "DEGall")#パターン名を指定(並びは変えな
param_samplesize <- 1000                #ブートストラップリサンプリング回数(100000が推奨。大きい値ほ

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み
library(baySeq)                         #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの

#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2), rep(3, param_G3))#G1群を1、G2群を2、G3群を3とし
tcc <- new("TCC", data, data.cl)        #TCCクラスオブジェクトtccを作成

#TCC正規化
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger" #正規化を実行した結果をtccに
```

トッ

Osabe法の実践9

①例題8を発見。この例題では、②で示されているように、10,000遺伝子×9サンプル（各群3サンプル）からなるファイルを入力として取り扱います。



Osabe法の実践10

①例題8を発見。この例題では、②で示されているように、10,000遺伝子×9サンプル（各群3サンプル）からなるファイルを入力として取り扱います。赤枠内で右クリックして、③保存。

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#analysis_deg_5_unpaired_an_advance...

8. サンプルデータ15の10,000 genes×9 samplesのカウントデータ(data_hypodata_3vs3vs3.txt)の場合：

例題7と似ていますが、全体的な発現変動の度合い(ANOVA的などこかの群間で発現変動してい
EEE-Eで行い、 発現変動パターンの同定をEEE-b(長部らの原著論文中ではEEE-baySEqと表記)
す

```
in_f <- "data_hypodata_3vs3vs3.txt"     #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.txt"                    #出力ファイル名を指定してout_fに格納
param_G1 <- 3                           #G1群のサンプル数を指定
param_G2 <- 3                           #G2群のサンプル数を指定
param_G3 <- 3                           #G3群のサンプル数を指定
param_FDR <- 0.05                       #false discovery rate (FDR)閾値を指定
param_narabi <- c("nonDEG", "DEG_G1", "DEG_G2", "DEG_G3", "DEGall")#パターン名
param_samplesize <- 1000                #ブートストラップリサンプリング回数(100000が推奨。大きい値ほど計算時
#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み
library(baySeq)                         #パッケージの読み込み
#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの読み込み
#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2), rep(3, param_G3))#G1群を1、G2群を2、G3群を3としたベクト
tcc <- new("TCC", data, data.cl)        #TCCクラスオブジェクトtccを作成
#TCC正規化
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger" #正規化を実行した結果をtccに格納
```

トップページへ

新しいタブで開く(T)
新しいウィンドウで開く(W)
シークレット ウィンドウで開く(G)
名前を付けてリンク先を保存(K)...
リンクのアドレスをコピー(E)
検証(I) Ctrl+Shift+I

③

Osabe法の実践11

①例題8を発見。この例題では、②で示されているように、10,000遺伝子×9サンプル（各群3サンプル）からなるファイルを入力として取り扱います。赤枠内で右クリックして、③保存。ここでは、④デスクトップに、⑤という名前（デフォルト）で、⑥保存します。

(Rで)塩基配列解析

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

8. サンプルデータ15の10,000 genes×9 samplesのカウントデータ(data_hypodat

例題7と似ていますが、全体的な発現変動の度合い(ANOVA的などこかの群間で発現変動している度合いでのランキング)をEEE-Eで行い、 発現変動パターンの同定をEEE-b(長部らの原著論文中ではEEE-baySEqと表記)で行った結果を出力しています

```
in_f <- "data_hypodata_3vs3vs3.txt"     #入力ファイル名を指定してin_fに格納
out_f
param
param
param
param
param
param

#必要
libra
libra

#入力
data

#前処
data.
tcc <- new("TCC", data, data.cl)        #TCCクラスオブジェクトtccを作成

#TCC正規化
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger" #正規化を実行した結果をtccに格納
```

トップページへ

名前を付けて保存

PC › デスクトップ ④

デスクトップの検索

ファイル名(N): data_hypodata_3vs3vs3.txt ⑤

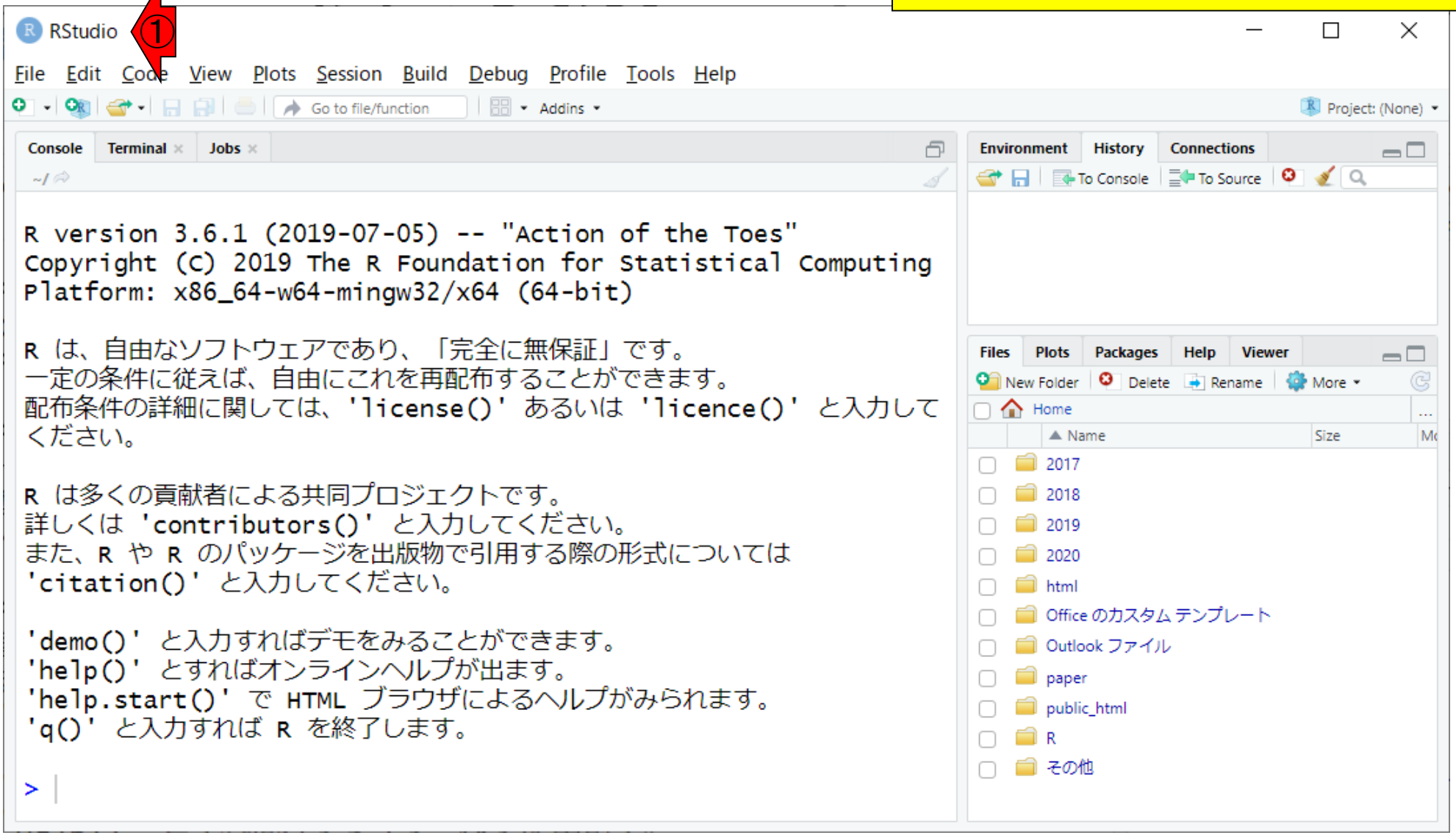
ファイルの種類(T): Text Document (*.txt)

フォルダーの参照(B)

⑥ 保存(S) キャンセル

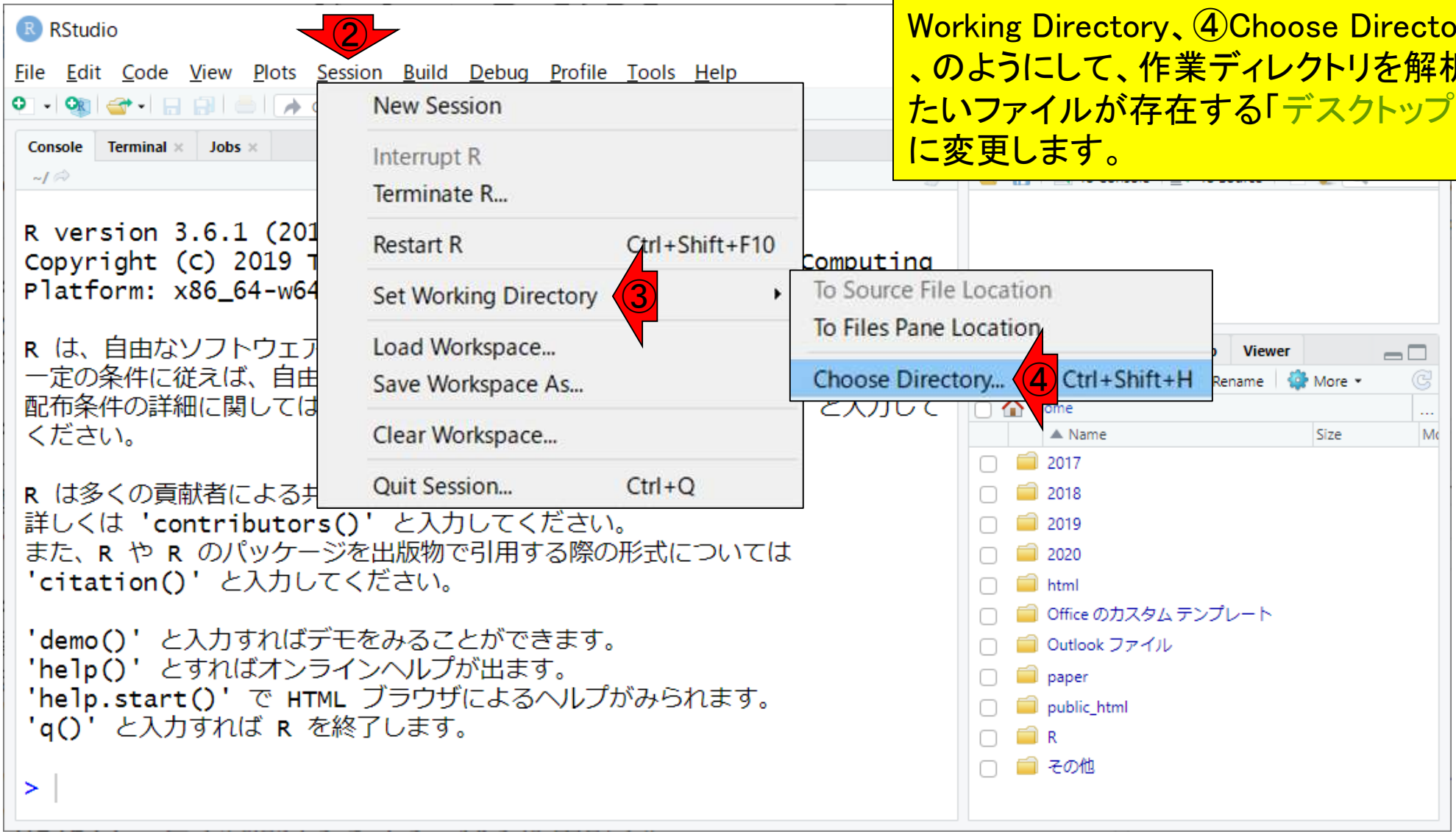
Osabe法の実践12

①RStudioの起動。昨年度第3回（2019年3月15日分）の講義資料の最後のほうにも解説あり。



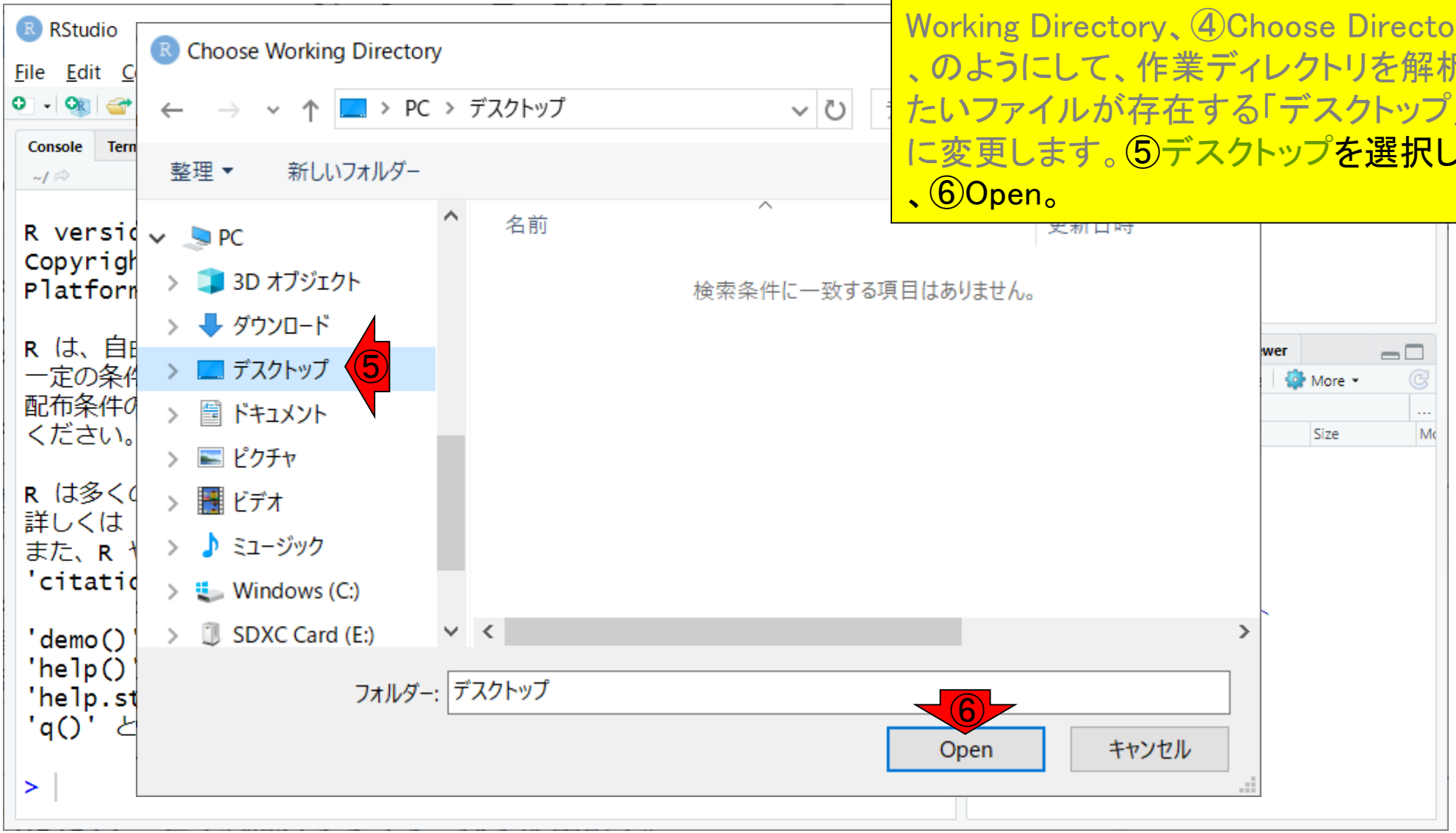
Osabe法の実践13

①RStudioの起動。昨年度第3回（2019年3月15日分）の講義資料の最後のほうにも解説あり。②Session、③Set Working Directory、④Choose Directory、のようにして、作業ディレクトリを解析したいファイルが存在する「デスクトップ」に変更します。



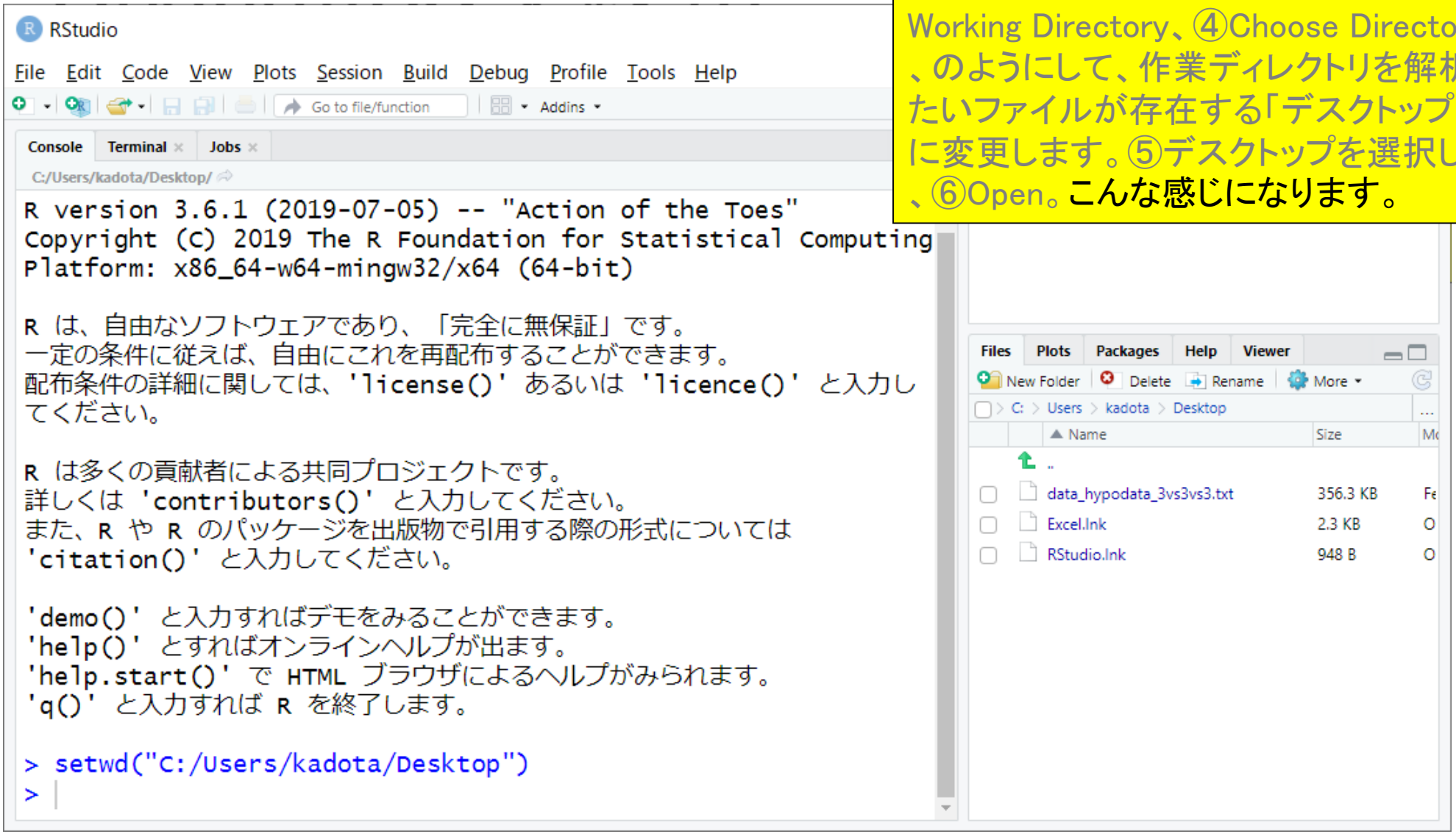
Osabe法の実践14

①RStudioの起動。昨年度第3回（2019年3月15日分）の講義資料の最後のほうにも解説あり。②Session、③Set Working Directory、④Choose Directory、のようにして、作業ディレクトリを解析したいファイルが存在する「デスクトップ」に変更します。⑤デスクトップを選択して、⑥Open。

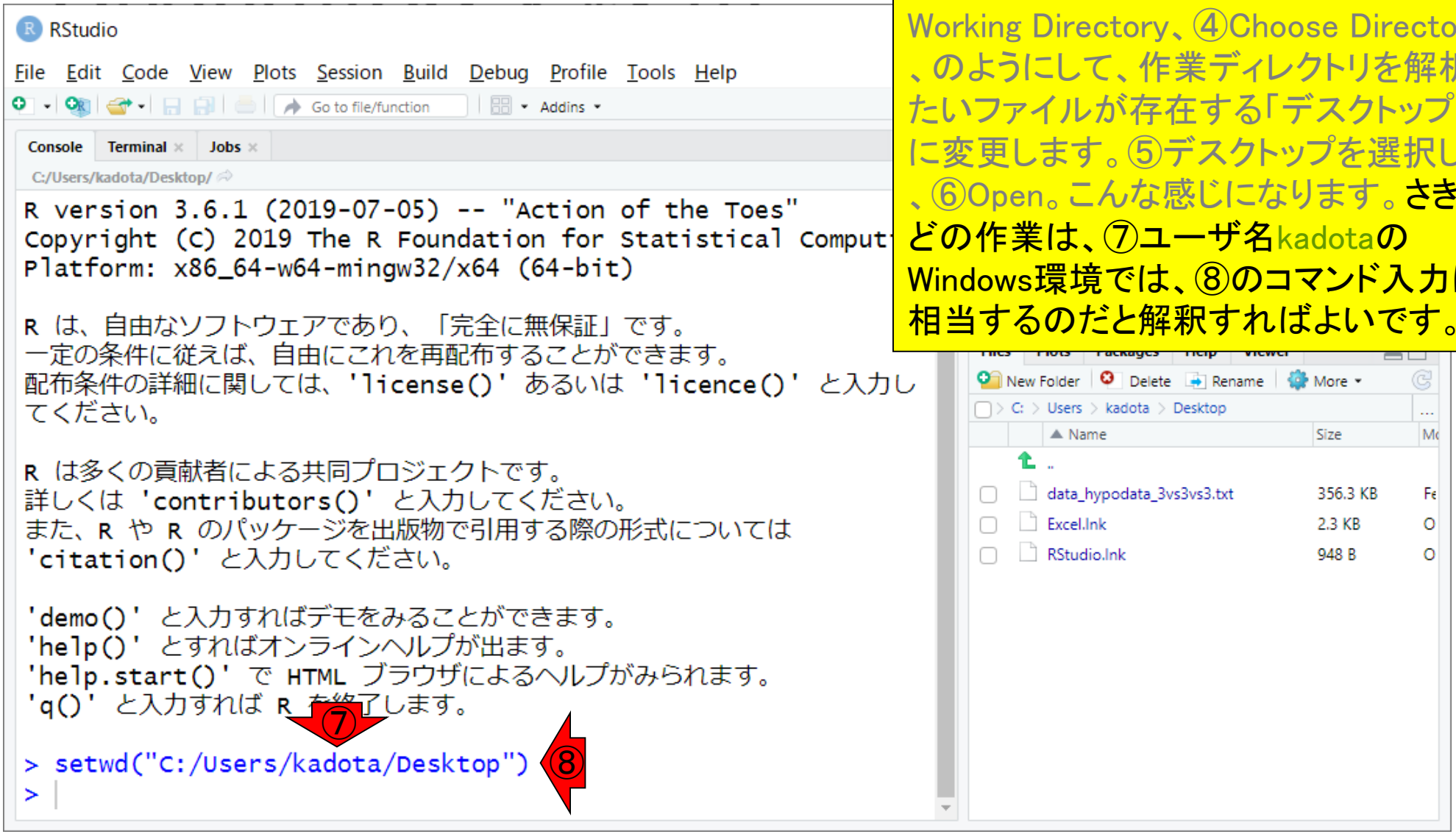


Osabe法の実践15

①RStudioの起動。昨年度第3回（2019年3月15日分）の講義資料の最後のほうにも解説あり。②Session、③Set Working Directory、④Choose Directory、のようにして、作業ディレクトリを解析したいファイルが存在する「デスクトップ」に変更します。⑤デスクトップを選択して、⑥Open。こんな感じになります。



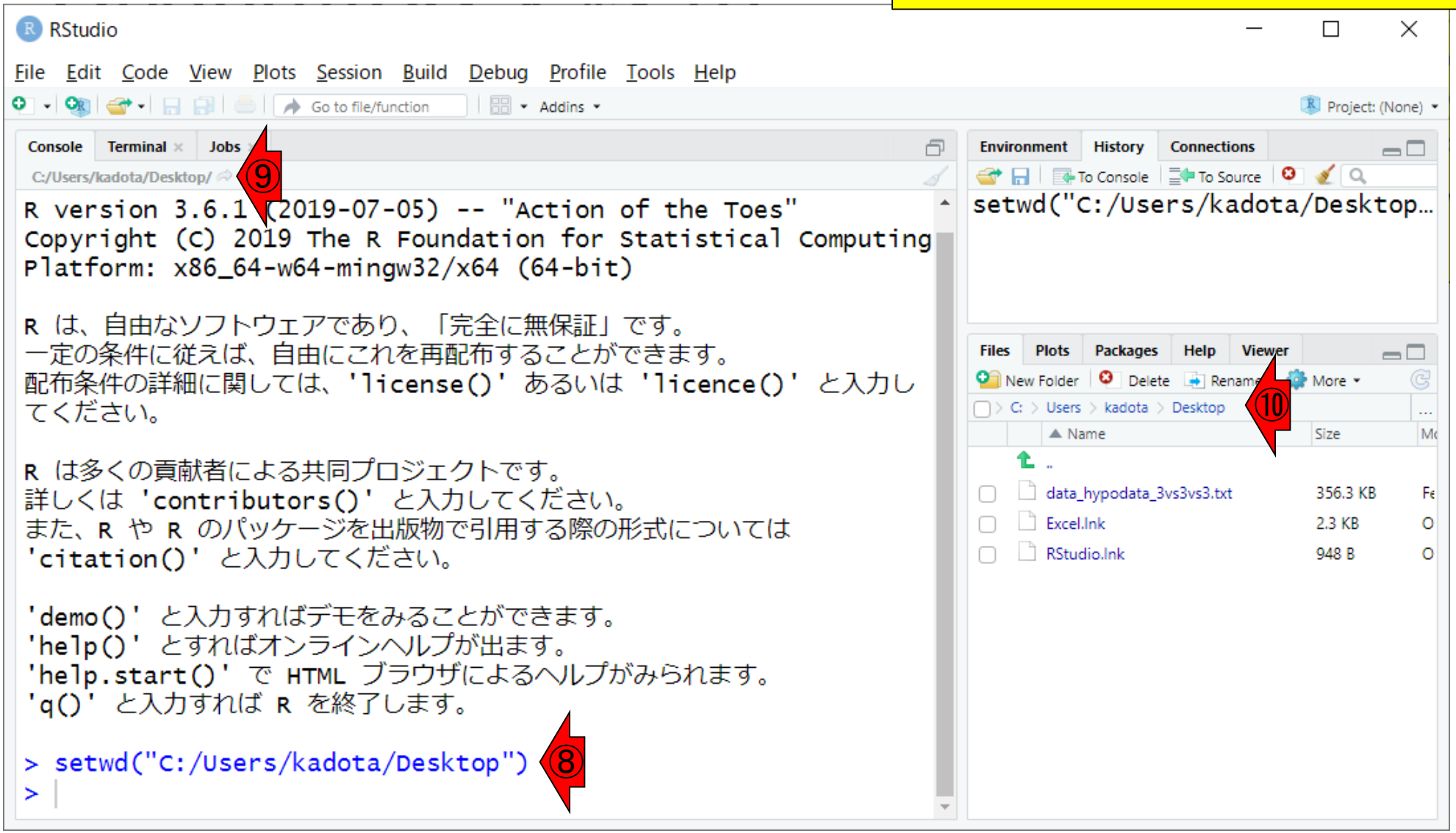
Osabe法の実践16



①RStudioの起動。昨年度第3回（2019年3月15日分）の講義資料の最後のほうにも解説あり。②Session、③Set Working Directory、④Choose Directory、のようにして、作業ディレクトリを解析したいファイルが存在する「デスクトップ」に変更します。⑤デスクトップを選択して、⑥Open。こんな感じになります。さきほどの作業は、⑦ユーザ名kadotaのWindows環境では、⑧のコマンド入力に相当するのだと解釈すればよいです。

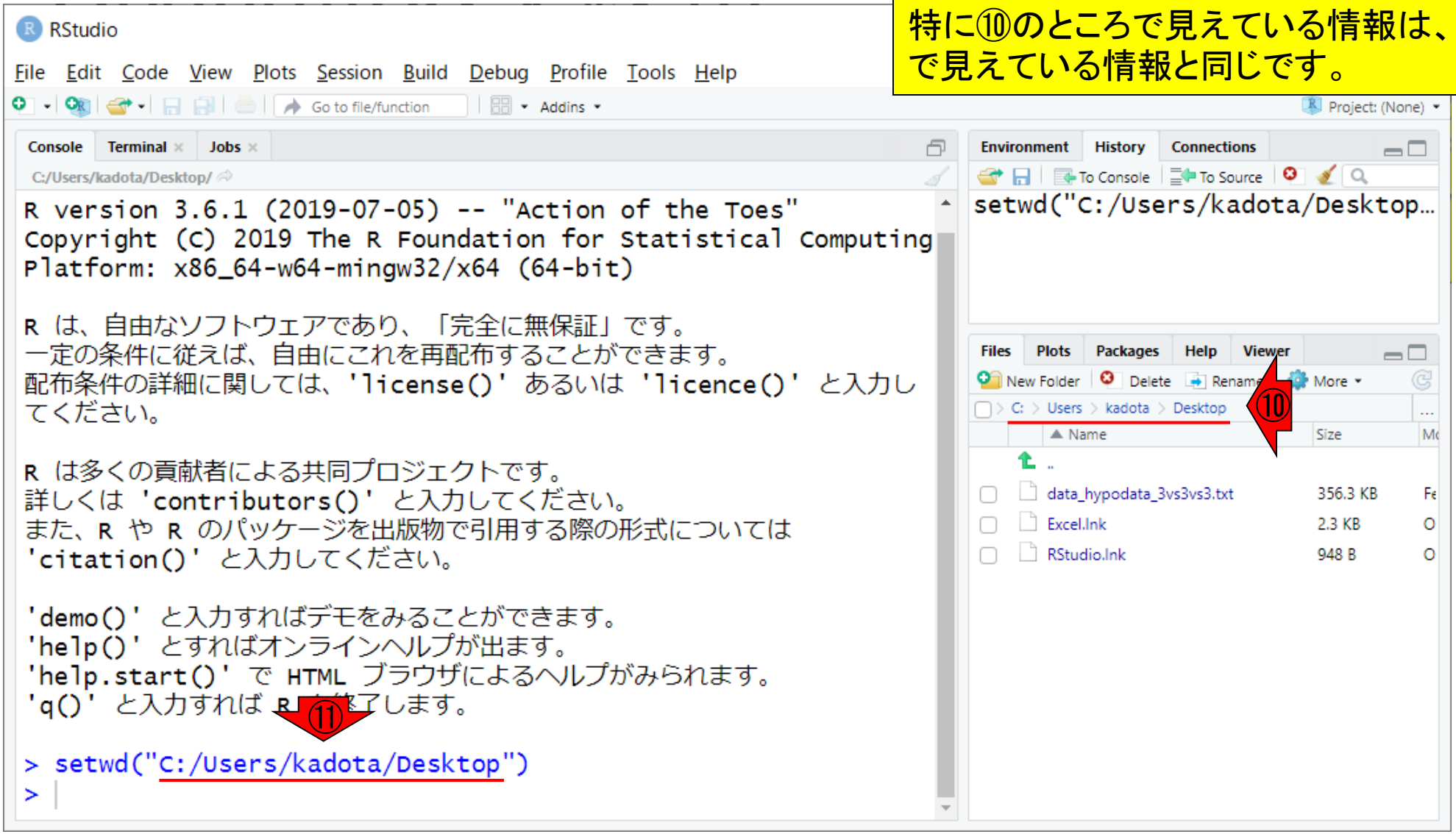
Osabe法の実践17

⑧のコマンド実行に相当する「デスクトップへの作業ディレクトリの変更」の結果として、⑨と⑩の部分も変わっています。



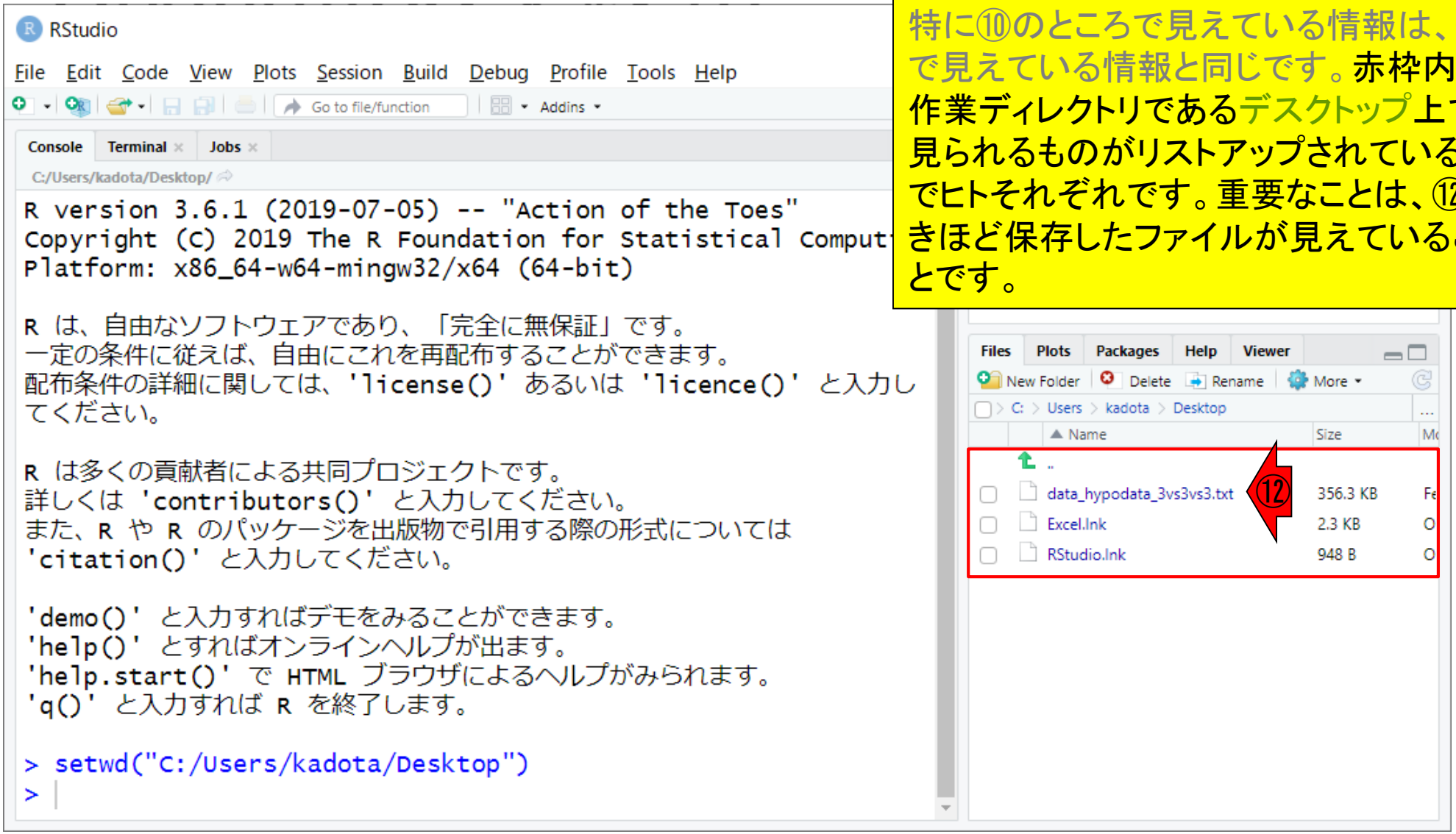
Osabe法の実践18

⑧のコマンド実行に相当する「デスクトップへの作業ディレクトリの変更」の結果として、⑨と⑩の部分も変わっています。
特に⑩のところで見えている情報は、⑪で見えている情報と同じです。



Osabe法の実践19

⑧のコマンド実行に相当する「デスクトップへの作業ディレクトリの変更」の結果として、⑨と⑩の部分も変わっています。特に⑩のところで見えている情報は、⑪で見えている情報と同じです。赤枠内は作業ディレクトリであるデスクトップ上で見られるものがリストアップされているのでヒトそれぞれです。重要なことは、⑫さきほど保存したファイルが見えていることです。



Osabe法の実践20

①例題8で用意されているスクリプトの簡単な解説。②がさきほど見えていたものと同じファイル名の文字列であり、これが入力ファイルです。③出力ファイル名はhoge8.txtとしていると読み解きます。つまり、無事解析が終われば、hoge8.txtという名前のファイルが作成されるということです。

(Rで)塩基配列解析

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

①

8. サンプルデータ15の10,000 genes×9 samplesのカウントデータ(data_hypodat

例題7と似ていますが、全体的な発現変動の度合い(ANOVA的などこかの群間で発現変動
EEE-Eで行い、 発現変動パターンの同定をEEE-b(長部らの原著論文中ではEEE-baySEq
す

②

③

```
in_f <- "data_hypodata_3vs3vs3.txt"  #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.txt"                 #出力ファイル名を指定してout_fに格納
param_G1 <- 3                        #G1群のサンプル数を指定
param_G2 <- 3                        #G2群のサンプル数を指定
param_G3 <- 3                        #G3群のサンプル数を指定
param_FDR <- 0.05                    #false discovery rate (FDR)閾値を指定
param_narabi <- c("nonDEG", "DEG_G1", "DEG_G2", "DEG_G3", "DEGall")#パターン名を指定(並びは変えない)
param_samplesize <- 1000             #ブートストラップリサンプリング回数(100000が推奨。大きい値ほど計算時間

#必要なパッケージをロード
library(TCC)                         #パッケージの読み込み
library(baySeq)                      #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの読み込み

#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2), rep(3, param_G3))#G1群を1、G2群を2、G3群を3としたベクト
tcc <- new("TCC", data, data.cl)     #TCCクラスオブジェクトtccを作成

#TCC正規化
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger" #正規化を実行した結果をtccに格納
```

トップページへ

Osabe法の実践21

①例題8で用意されているスクリプトの簡単な解説。②がさきほど見えていたものと同じファイル名の文字列であり、これが入力ファイルです。③出力ファイル名はhoge8.txtとしていると読み解きます。つまり、無事解析が終われば、hoge8.txtという名前のファイルが作成されるということです。赤枠内のスクリプト全体をコピーします。左上の④からスタートして…

(Rで)塩基配列解析

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

8. サンプルデータ15の10,000 genes×9 samplesのカウントデータ(data_hypodat

例題7と似ていますが、全体的な発現変動の度合い(ANOVA的などこかの群間で発現変動
EEE-Eで行い、 発現変動パターンの同定をEEE-b(長部らの原著論文中ではEEE-baySE

④

```
in_f <- "data_hypodata_3vs3vs3.txt"     #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.txt"                    #出力ファイル名を指定してout_fに格納
param_G1 <- 3                           #G1群のサンプル数を指定
param_G2 <- 3                           #G2群のサンプル数を指定
param_G3 <- 3                           #G3群のサンプル数を指定
param_FDR <- 0.05                       #false discovery rate (FDR)閾値を指定
param_narabi <- c("nonDEG", "DEG_G1", "DEG_G2", "DEG_G3", "DEGall")#パターン名を指定(並びは変えない)
param_samplesize <- 1000                #ブートストラップリサンプリング回数(100000が推奨。大きい値ほど計算時間

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み
library(baySeq)                         #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの読み込み

#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2), rep(3, param_G3))#G1群を1、G2群を2、G3群を3としたベクト
tcc <- new("TCC", data, data.cl)        #TCCクラスオブジェクトtccを作成

#TCC正規化
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger" #正規化を実行した結果をtccに格納
```

トップページへ

Osabe法の実践22

①例題8で用意されているスクリプトの簡単な解説。②がさきほど見えていたものと同じファイル名の文字列であり、これが入力ファイルです。③出力ファイル名はhoge8.txtとしていると読み解きます。つまり、無事解析が終われば、hoge8.txtという名前のファイルが作成されるということです。赤枠内のスクリプト全体をコピーします。左上の④からスタートして、こんな感じで、⑤例題9が見えるところの、⑥ここまで反転させて…

(Rで)塩基配列解析

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

```
    orderings <- append(orderings, as.character(ba@orderings[i, max.co
}

#ファイルに保存(テキストファイル)
tmp <- cbind(rownames(data), data, out$PP, out$MAP, orderings, ranking
write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)#tmpの

#結果のまとめ(全遺伝子)
head(out$MAP)                          #最初の6遺伝子分のパターンを表示
table(out$MAP)                         #パターンごとの出現頻度を表示
table(out$MAP)/length(out$MAP)         #パターンごとの出現確率を表示

#結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを判定した結果をobjに格納
sum(obj)                               #条件を満たす数を表示
table(out$MAP[obj])                    #パターンごとの出現頻度を表示(条件を満たすもののみ)
table(out$MAP[obj])/length(out$MAP[obj])#パターンごとの出現確率を表示(条件を満たすもののみ)
```

⑥

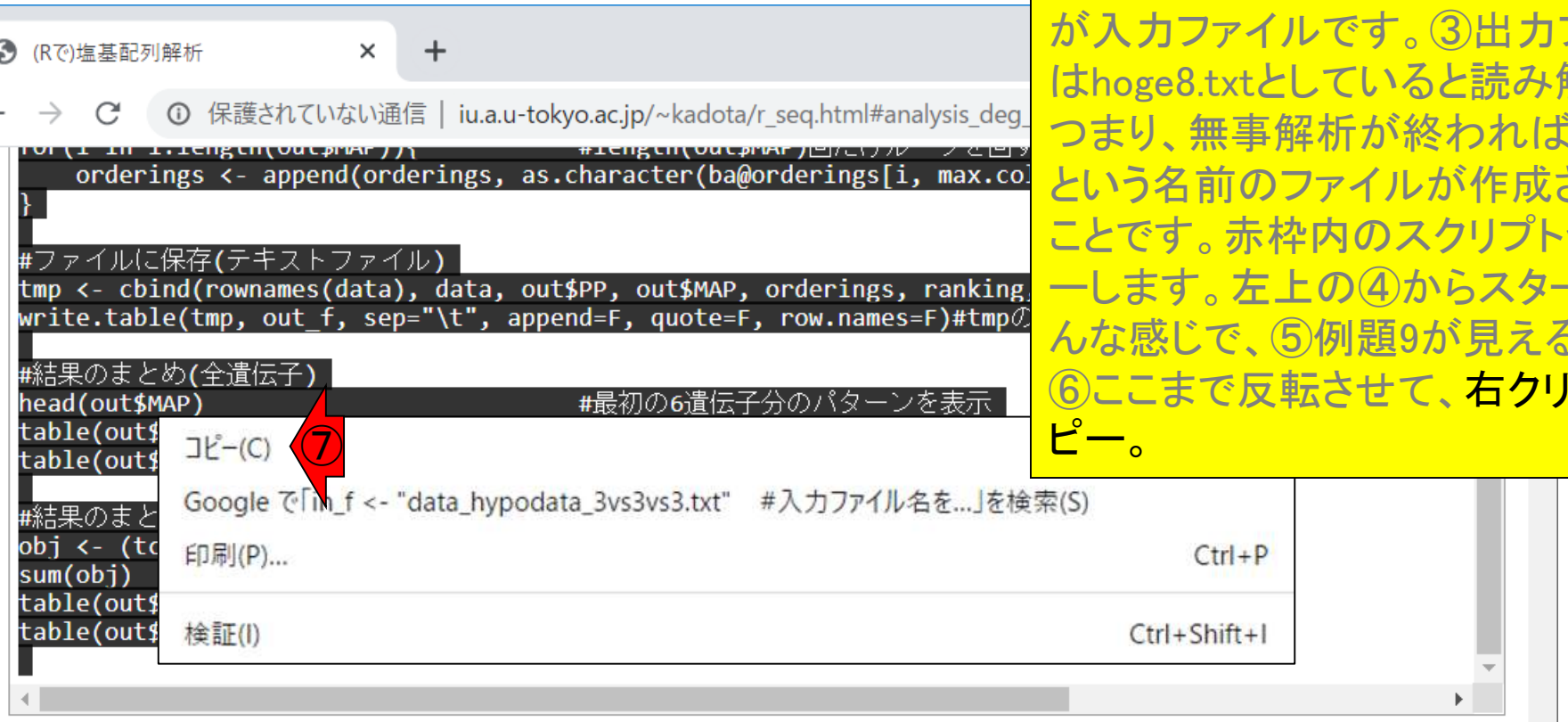
⑤

9. 10,000 genes×9 samplesのカウントデータ(Simulation_3group_sheep.txt)の場合：

例題8とは基本的に入力ファイルが異なるだけです。入力ファイルは、1～1,000行目がG1群で4倍高発現パターン(DEG_G1)、 1,001～2,000行目がG2群で4倍高発現パターン(DEG_G2)、2,001～3,000行目がG3群で4倍高発現パターン(DEG_G3)、残りがnon-DEGです。 他は、TCC-GUI (Su et al., 2019)の デフォルトのFDR閾値(= 0.10)に合 トップページへ す。

Osabe法の実践23

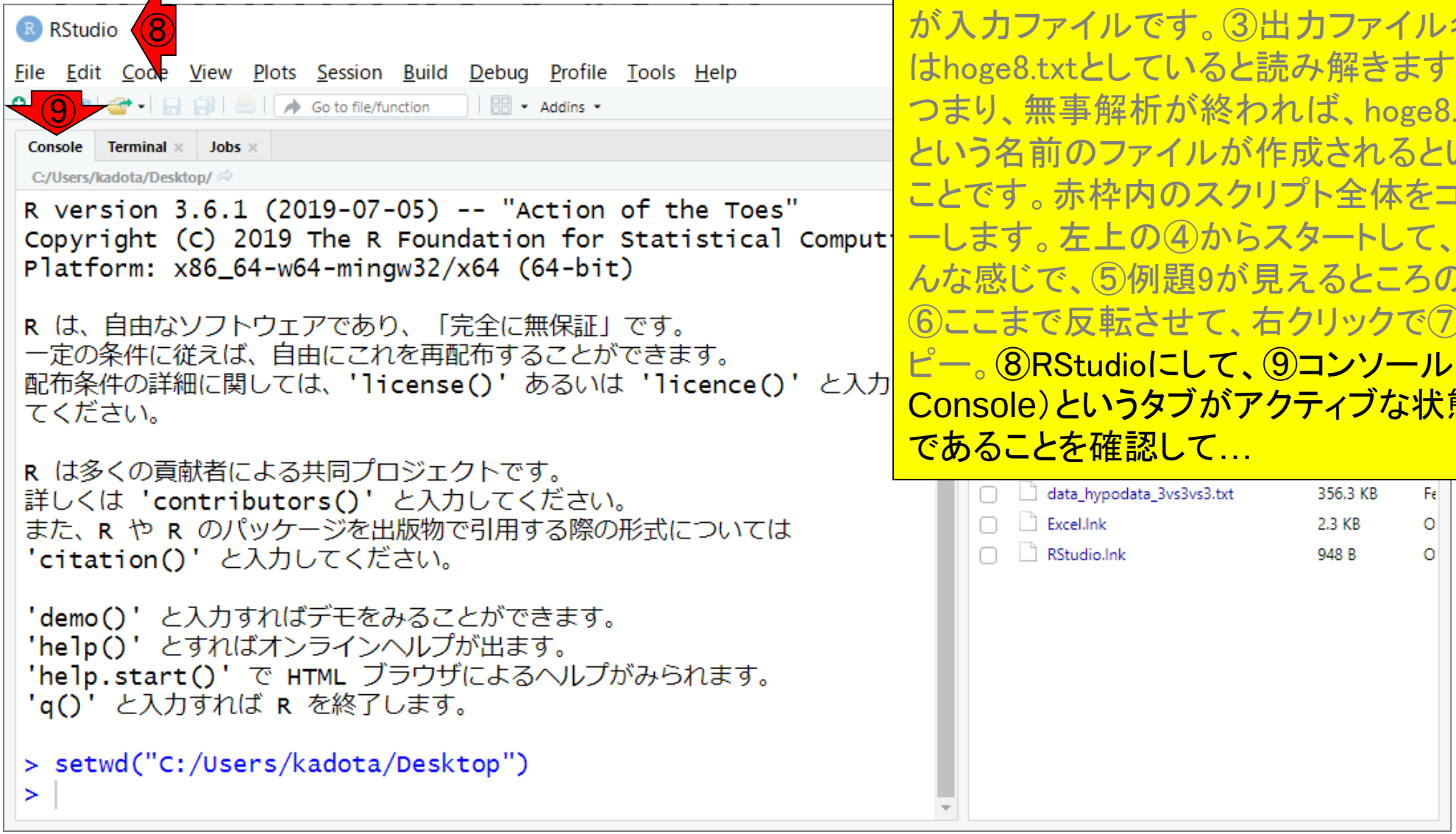
①例題8で用意されているスクリプトの簡単な解説。②がさきほど見えていたものと同じファイル名の文字列であり、これが入力ファイルです。③出力ファイル名はhoge8.txtとしていると読み解きます。つまり、無事解析が終われば、hoge8.txtという名前のファイルが作成されるということです。赤枠内のスクリプト全体をコピーします。左上の④からスタートして、こんな感じで、⑤例題9が見えるところの、⑥ここまで反転させて、右クリックで⑦コピー。



9. 10,000 genes×9 samplesのカウントデータ(Simulation_3group_sheep.txt)の場合：

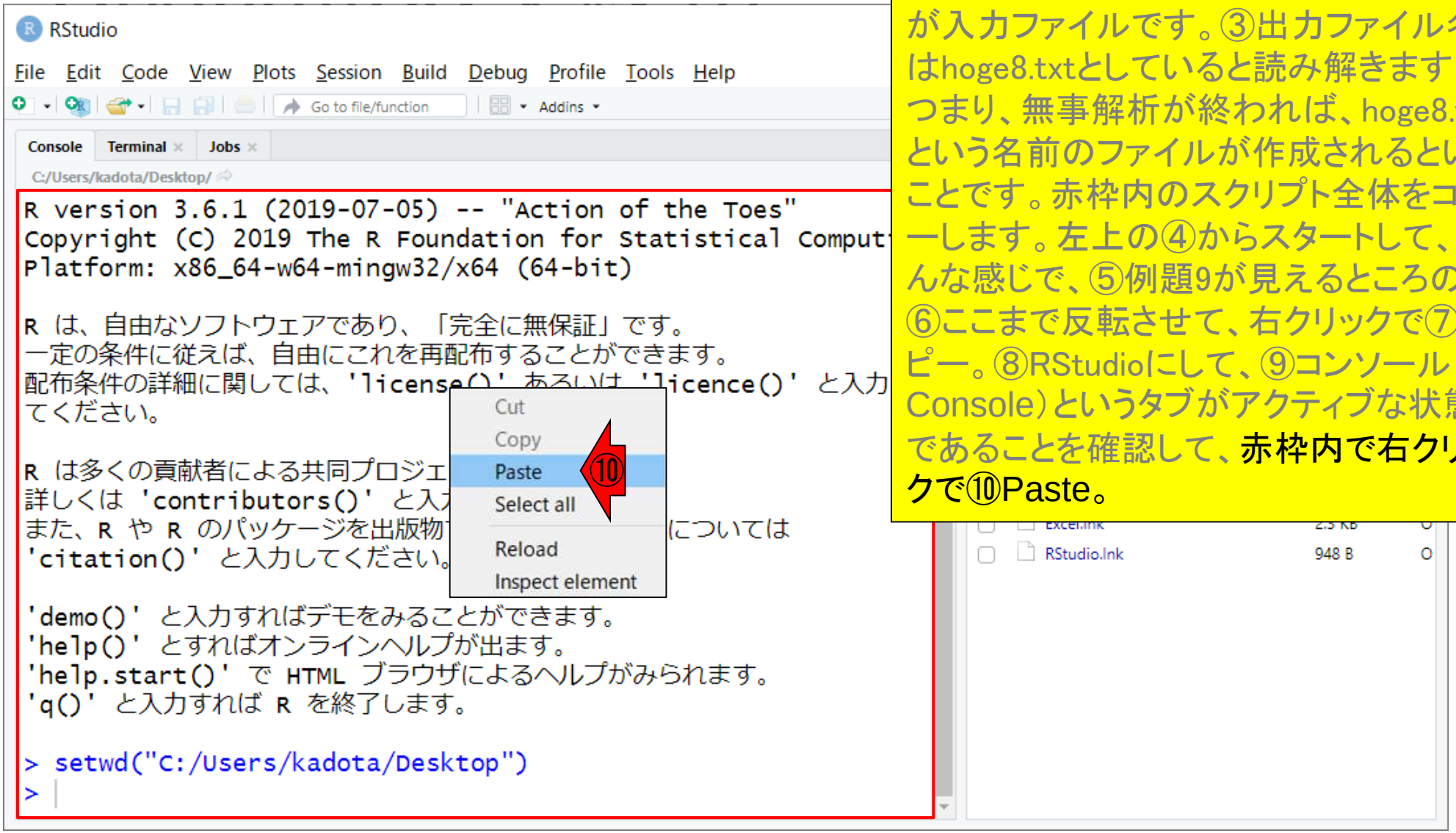
例題8とは基本的に入力ファイルが異なるだけです。入力ファイルは、1～1,000行目がG1群で4倍高発現パターン(DEG_G1)、 1,001～2,000行目がG2群で4倍高発現パターン(DEG_G2)、2,001～3,000行目がG3群で4倍高発現パターン(DEG_G3)、残りがnon-DEGです。 他は、TCC-GUI (Su et al., 2019)の デフォルトのFDR閾値(= 0.10)に合トップページへす。

Osabe法の実践24



①例題8で用意されているスクリプトの簡単な解説。②がさきほど見えていたものと同じファイル名の文字列であり、これが入力ファイルです。③出力ファイル名はhoge8.txtとしていると読み解きます。つまり、無事解析が終われば、hoge8.txtという名前のファイルが作成されるということです。赤枠内のスクリプト全体をコピーします。左上の④からスタートして、こんな感じで、⑤例題9が見えるところの、⑥ここまで反転させて、右クリックで⑦コピー。⑧RStudioにして、⑨コンソール（Console）というタブがアクティブな状態であることを確認して…

Osabe法の実践25



①例題8で用意されているスクリプトの簡単な解説。②がさきほど見えていたものと同じファイル名の文字列であり、これが入力ファイルです。③出力ファイル名はhoge8.txtとしていると読み解きます。つまり、無事解析が終われば、hoge8.txtという名前のファイルが作成されるということです。赤枠内のスクリプト全体をコピーします。左上の④からスタートして、こんな感じで、⑤例題9が見えるところの、⑥ここまで反転させて、右クリックで⑦コピー。⑧RStudioにして、⑨コンソール（Console）というタブがアクティブな状態であることを確認して、赤枠内で右クリックで⑩Paste。

こんな感じになる。①リターン。

Osabe法の実践26

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Console Terminal Jobs

C:/Users/kadota/Desktop/

```
write.table(tmp, out_f, sep="\t", append=F, quote=F, row
.names=F)#tmpの中身を指定したファイル名で保存

#結果のまとめ(全遺伝子)
head(out$MAP)                          #最初の6遺伝子分の
パターンを表示
table(out$MAP)                         #パターンごとの出現
頻度を表示
table(out$MAP)/length(out$MAP)         #パターンごとの出現
確率を表示

#結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうか
を判定した結果をobjに格納
sum(obj)                               #条件を満たす数を表
示
table(out$MAP[obj])                    #パターンごとの出現
頻度を表示(条件を満たすもののみ)
table(out$MAP[obj])/length(out$MAP[obj])#パターンごとの出
現確率を表示(条件を満たすもののみ)
```

Environment History Connections

To Console | To Source

```
setwd("C:/Users/kadota/Desktop...
```

Files Plots Packages

New Folder | Delete

C: > Users > kadota

Name

..

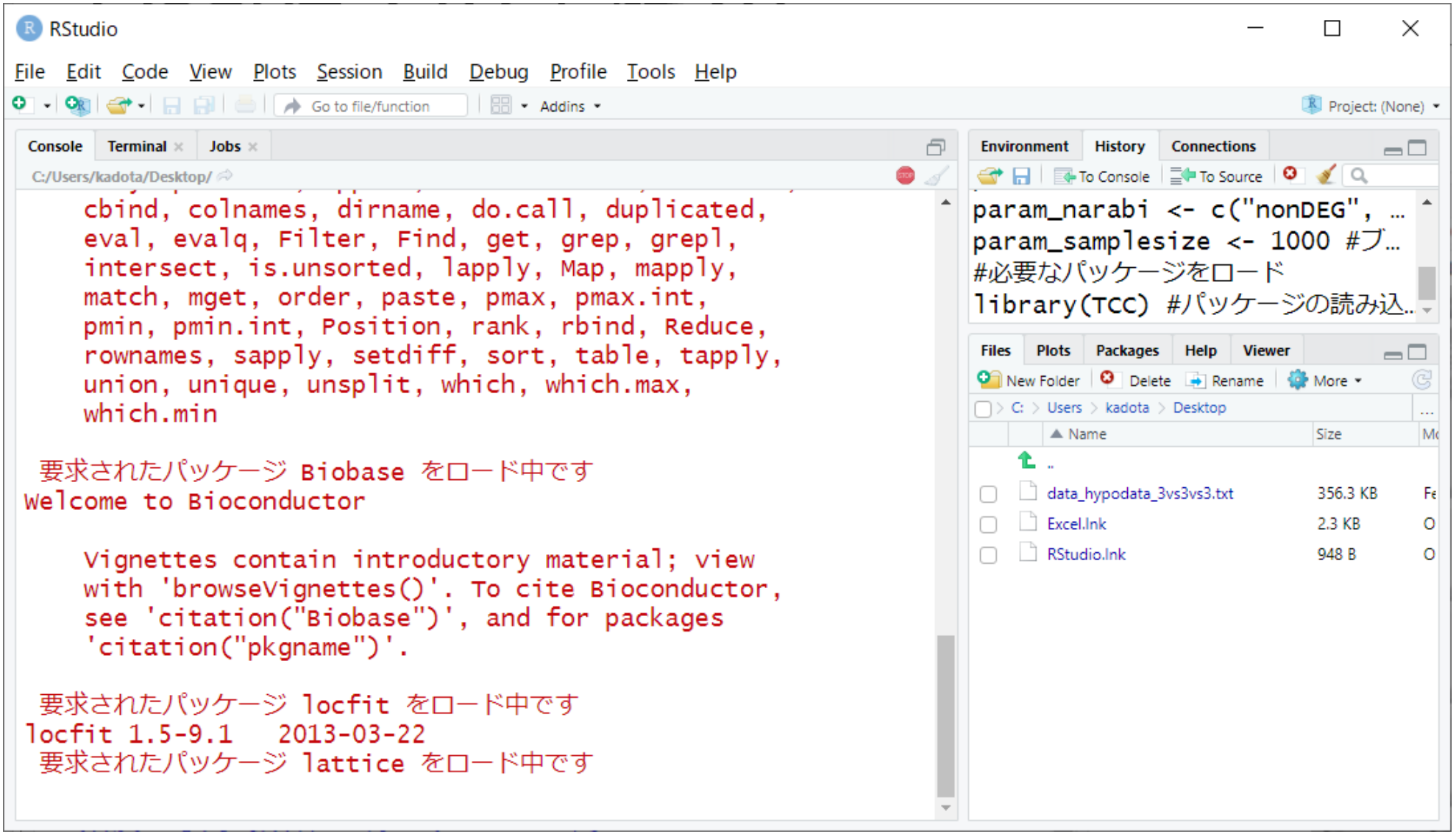
data_hypodata_3

Excel.lnk

RStudio.lnk

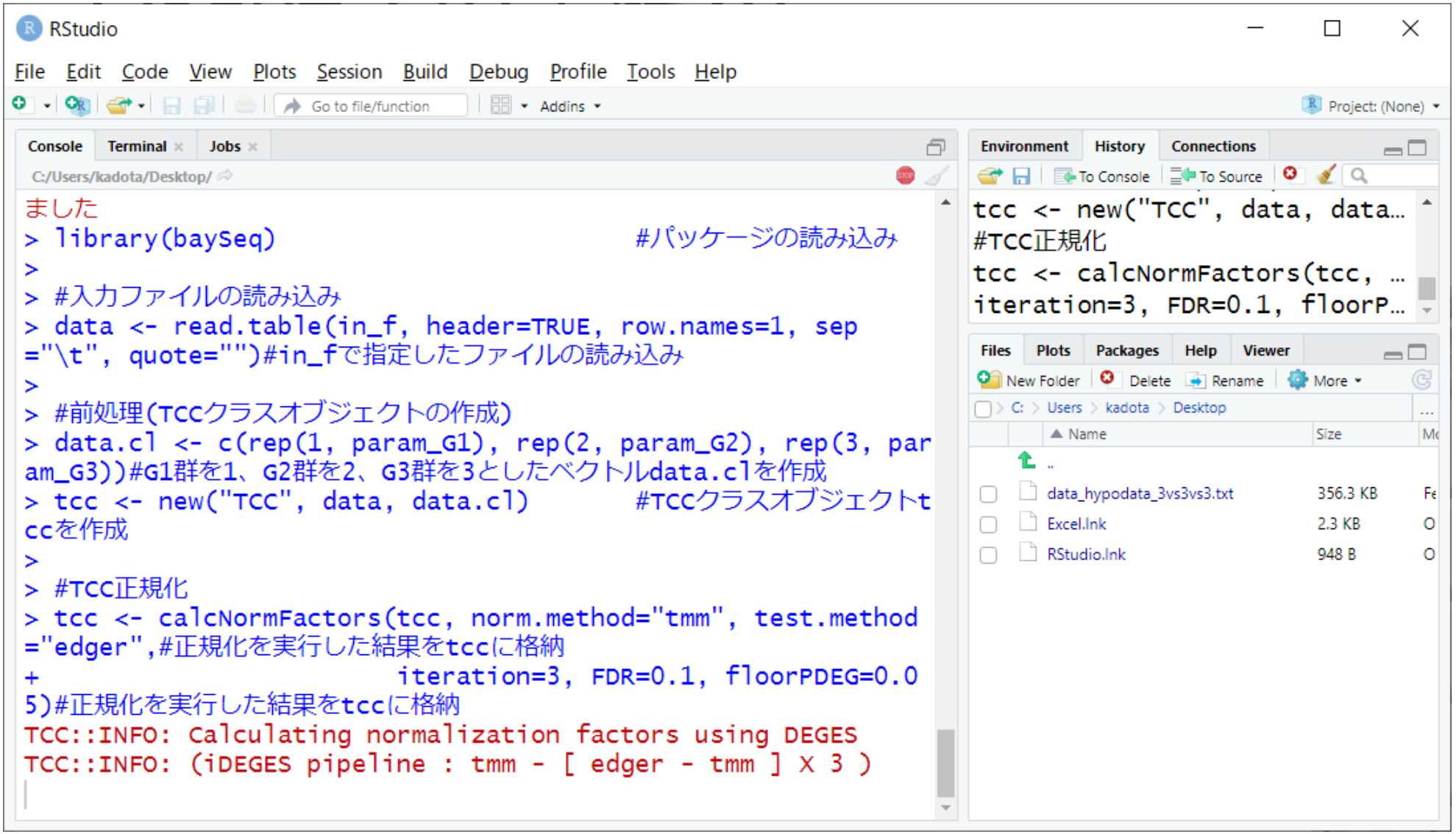
Osabe法の実践27

こんな感じになる。①リターン。実行中
1…



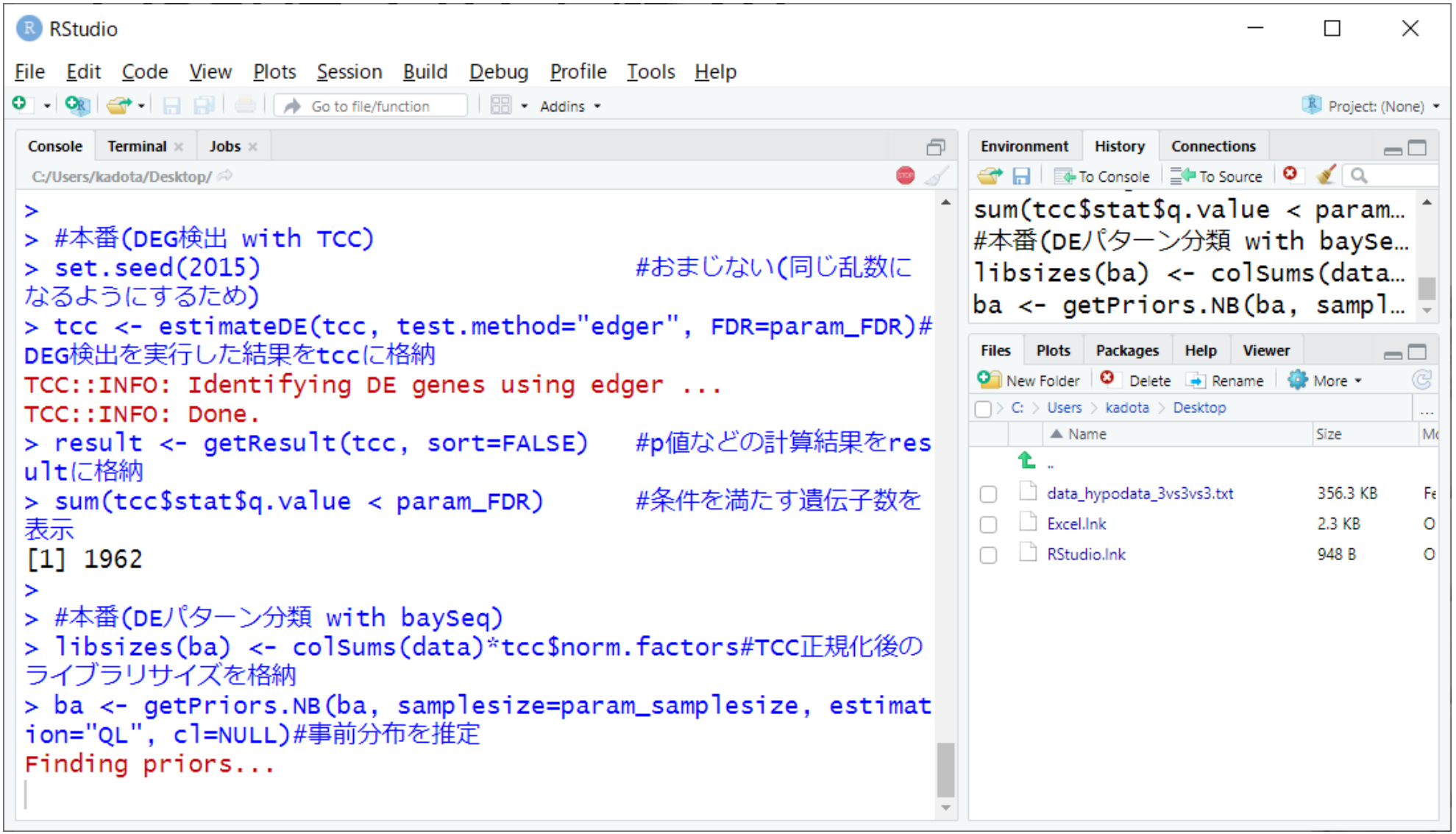
Osabe法の実践28

こんな感じになる。①リターン。実行中1…実行中2…



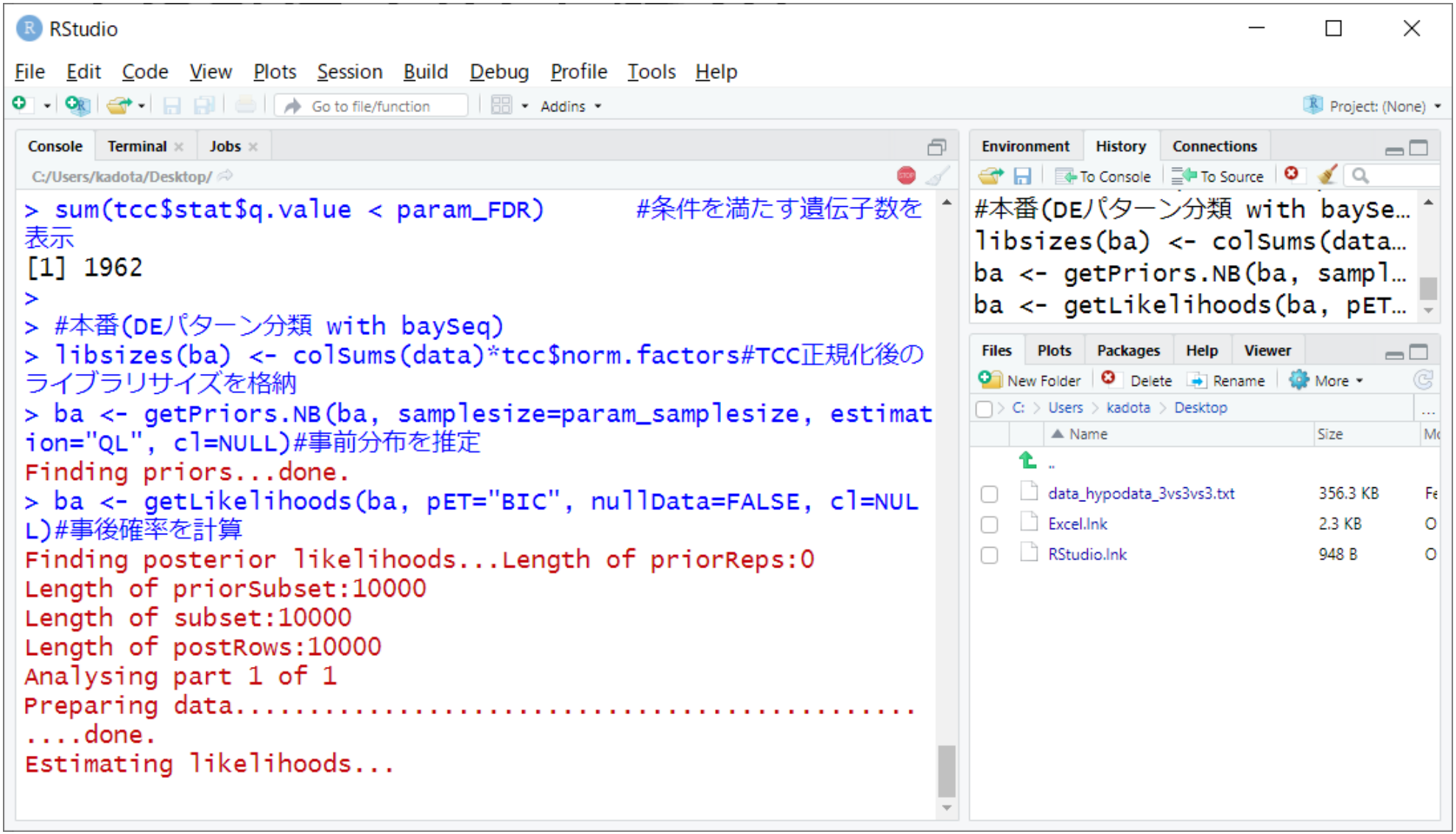
Osabe法の実践29

こんな感じになる。①リターン。実行中1…実行中2…実行中3…



Osabe法の実践30

こんな感じになる。①リターン。実行中1…実行中2…実行中3…実行中4…



Osabe法の実践31

こんな感じになる。①リターン。実行中1…実行中2…実行中3…実行中4…実行完了後の状態。ノートPCで約5分。

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Console Terminal Jobs

C:/Users/kadota/Desktop/

```
0.1157 0.0513 0.0220 0.0002 0.8108
>
> #結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
> obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを判定した結果をobjに格納
> sum(obj)                                    #条件を満たす数を表示
[1] 1962
> table(out$MAP[obj])                         #パターンごとの出現頻度を表示(条件を満たすもののみ)

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
   955    491    200       2    314
> table(out$MAP[obj])/length(out$MAP[obj])#パターンごとの出現確率を表示(条件を満たすもののみ)

     DEG_G1      DEG_G2      DEG_G3      DEGall
0.486748216 0.250254842 0.101936799 0.001019368
     nonDEG
0.160040775
>
>
```

Environment History Connections

To Console | To Source

```
obj <- (tcc$stat$q.value < p…
sum(obj) #条件を満たす数を表示
table(out$MAP[obj]) #パターン…
table(out$MAP[obj])/length(o…
```

Files Plots Packages Help Viewer

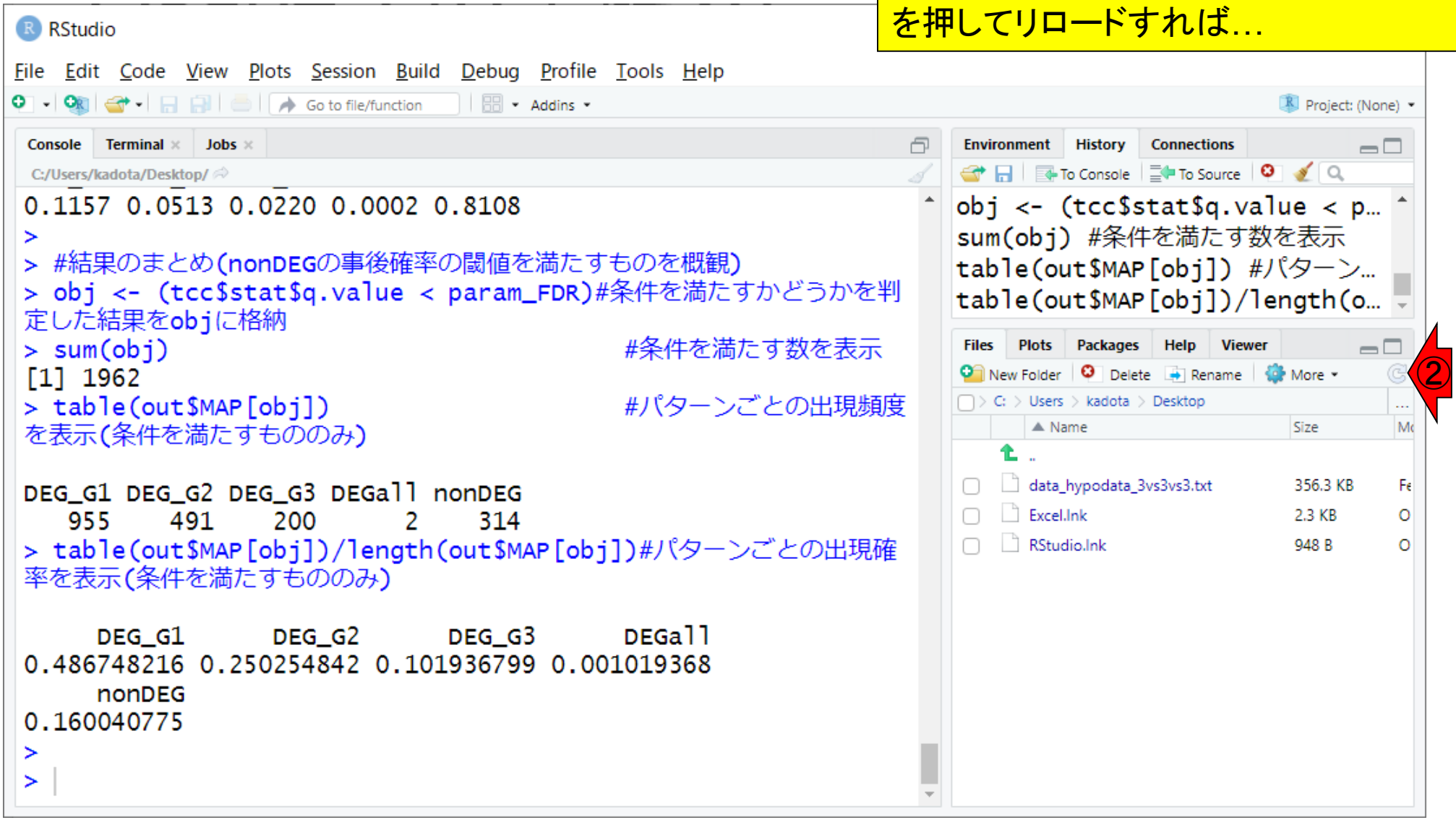
New Folder | Delete | Rename | More

C: > Users > kadota > Desktop

Name	Size
..	
data_hypodata_3vs3vs3.txt	356.3 KB
Excel.lnk	2.3 KB
RStudio.lnk	948 B

Osabe法の実践32

こんな感じになる。①リターン。実行中1…実行中2…実行中3…実行中4…実行完了後の状態。ノートPCで約5分。②を押してリロードすれば…



Osabe法の実践33

こんな感じになる。①リターン。実行中1…実行中2…実行中3…実行中4…実行完了後の状態。ノートPCで約5分。②を押してリロードすれば、③出力ファイルであるhoge8.txtが見られます。

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Console Terminal Jobs

C:/Users/kadota/Desktop/

```
0.1157 0.0513 0.0220 0.0002 0.8108
>
> #結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
> obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを判定した結果をobjに格納
> sum(obj)                                    #条件を満たす数を表示
[1] 1962
> table(out$MAP[obj])                         #パターンごとの出現頻度を表示(条件を満たすもののみ)

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
   955    491    200       2    314
> table(out$MAP[obj])/length(out$MAP[obj])#パターンごとの出現確率を表示(条件を満たすもののみ)

     DEG_G1      DEG_G2      DEG_G3      DEGall
0.486748216 0.250254842 0.101936799 0.001019368
     nonDEG
0.160040775
>
>
```

Environment History Connections

To Console | To Source

```
obj <- (tcc$stat$q.value < p…
sum(obj) #条件を満たす数を表示
table(out$MAP[obj]) #パターン…
table(out$MAP[obj])/length(o…
```

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

C: > Users > kadota > Desktop

Name	Size
..	
data_hypodata_3vs3vs3.txt	356.3 KB
Excel.lnk	2.3 KB
hoge8.txt	1.5 MB
RStudio.lnk	948 B

③

Osabe法の実践34

こんな感じになる。①リターン。実行中1…実行中2…実行中3…実行中4…実行完了後の状態。ノートPCで約5分。②を押してリロードすれば、③出力ファイルであるhoge8.txtが見られます。このファイルを直接Excelなどで眺めてもよいですが、④Console画面上に表示されているのが要約情報なので結果の概要を把握しやすい。

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

④

Go to file/function | Addins

Console | Terminal | Jobs

C:/Users/kadota/Desktop/

```
0.1157 0.0513 0.0220 0.0002 0.8108
>
> #結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
> obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを判定した結果をobjに格納
> sum(obj)                                    #条件を満たす数を表示
[1] 1962
> table(out$MAP[obj])                         #パターンごとの出現頻度を表示(条件を満たすもののみ)

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
   955    491    200      2    314
> table(out$MAP[obj])/length(out$MAP[obj])#パターンごとの出現確率を表示(条件を満たすもののみ)

     DEG_G1      DEG_G2      DEG_G3      DEGall
0.486748216 0.250254842 0.101936799 0.001019368
     nonDEG
0.160040775
>
>
```

table(out$MAP[obj]) #パターン…

table(out$MAP[obj])/length(o…

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

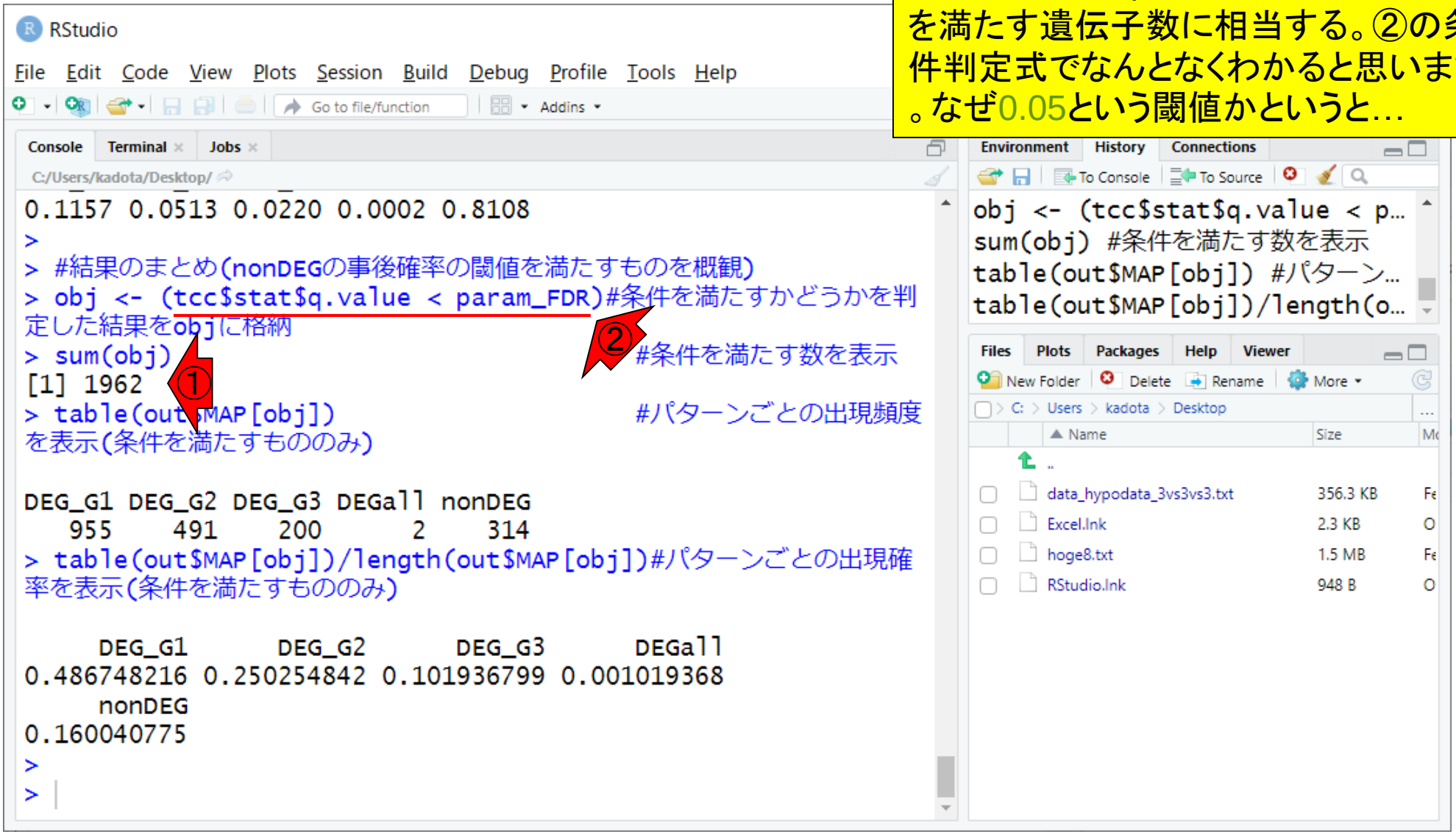
C: > Users > kadota > Desktop

Name	Size
..	
data_hypodata_3vs3vs3.txt	356.3 KB
Excel.lnk	2.3 KB
hoge8.txt	1.5 MB
RStudio.lnk	948 B

③

Osabe法の実践35

例えば①この1962という情報は、GLMベースの発現変動解析結果から得られたものであり、q値が0.05未満という閾値を満たす遺伝子数に相当する。②の条件判定式でなんとなくわかると思います。なぜ0.05という閾値かというと…



Osabe法の実践36

例えば①この1962という情報は、GLMベースの発現変動解析結果から得られたものであり、q値が0.05未満という閾値を満たす遺伝子数に相当する。②の条件判定式でなんとなくわかると思います。なぜ0.05という閾値かというと、③param_FDRの中身が0.05だからです。

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Console | Terminal | Jobs

C:/Users/kadota/Desktop/

```
0.1157 0.0513 0.0220 0.0002 0.8108
>
> #結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
> obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを判定した結果をobjに格納
> sum(obj)                                    #条件を満たす数を表示
[1] 1962
> table(out$MAP[obj])                         #パターンごとの出現頻度を表示(条件を満たすもののみ)

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
   955    491    200       2    314
> table(out$MAP[obj])/length(out$MAP[obj])#パターンごとの出現確率を表示(条件を満たすもののみ)

     DEG_G1      DEG_G2      DEG_G3      DEGall
0.486748216 0.250254842 0.101936799 0.001019368
     nonDEG
0.160040775
>
>
```

③

```
obj <- (tcc$stat$q.value < p...
sum(obj) #条件を満たす数を表示
table(out$MAP[obj]) #パターン...
table(out$MAP[obj])/length(o...
```

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

C: > Users > kadota > Desktop

Name	Size
..	
data_hypodata_3vs3vs3.txt	356.3 KB
Excel.lnk	2.3 KB
hoge8.txt	1.5 MB
RStudio.lnk	948 B

Osabe法の実践37

例えば①この1962という情報は、GLMベースの発現変動解析結果から得られたものであり、q値が0.05未満という閾値を満たす遺伝子数に相当する。②の条件判定式でなんとなくわかると思います。なぜ0.05という閾値かというと、③param_FDRの中身が0.05だからです。④この部分で0.05という数値をparam_FDRに代入していることがわかります。

(Rで)塩基配列解析

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

8. サンプルデータ15の10,000 genes×9 samplesのカウントデータ(data_hypodat

例題7と似ていますが、全体的な発現変動の度合い(ANOVA的などこかの群間で発現変動
EEE-Eで行い、 発現変動パターンの同定をEEE-b(長部らの原著論文中ではEEE-baySE
す

```
in_f <- "data_hypodata_3vs3vs3.txt"     #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.txt"                    #出力ファイル名を指定してout_fに格納
param_G1 <- 3                           #G1群のサンプル数を指定
param_G2 <- 3                           #G2群のサンプル数を指定
param_G3 <- 3                           #G3群のサンプル数を指定
param_FDR <- 0.05                       #false discovery rate (FDR)閾値を指定
param_narabi <- c("nonDEG", "DEG_G1", "DEG_G2", "DEG_G3", "DEGall")#パターン名を指定(並びは変えない)
param_samplesize <- 1000                #ブートストラップリサンプリング回数(100000が推奨。大きい値ほど計算時間

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み
library(baySeq)                         #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの読み込み

#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2), rep(3, param_G3))#G1群を1、G2群を2、G3群を3としたベクト
tcc <- new("TCC", data, data.cl)        #TCCクラスオブジェクトtccを作成

#TCC正規化
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger" #正規化を実行した結果をtccに格納
```

トップページへ

Osabe法の実践38

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Console | Terminal | Jobs

C:/Users/kadota/Desktop/

```
0.1157 0.0513 0.0220 0.0002 0.8108
>
> #結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
> obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを
定した結果をobjに格納
> sum(obj)                                    #条件を満たす数を表示
[1] 1962
> table(out$MAP[obj])                         #パターンごとの出現頻
を表示(条件を満たすもののみ)

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
   955    491    200       2    314
> table(out$MAP[obj])/length(out$MAP[obj])#パターンごとの出現確
率を表示(条件を満たすもののみ)

     DEG_G1      DEG_G2      DEG_G3      DEGall
0.486748216 0.250254842 0.101936799 0.001019368
     nonDEG
0.160040775
>
>
```

⑤

Excel.lnk 2.3 KB

hoge8.txt 1.5 MB

RStudio.lnk 948 B

例えば①この1962という情報は、GLMベースの発現変動解析結果から得られたものであり、q値が0.05未満という閾値を満たす遺伝子数に相当する。②の条件判定式でなんとなくわかると思います。なぜ0.05という閾値かというと、③param_FDRの中身が0.05だからです。④この部分で0.05という数値をparam_FDRに代入していることがわかります。FDRは、False Discovery Rate（誤発見率）の略なので、⑤例えば5% FDRという閾値で得られた1962個という結果は、1962×(1 - 0.05) = 1863.9個が本物のDEGで、残りの1962×0.05 = 98.1個が偽物だということ。

Osabe法の実践39

①は、5% FDR閾値を満たした1962個に対して、どの発現パターンのものがいくつあったかを示したもの。

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Console Terminal Jobs

C:/Users/kadota/Desktop/

```
0.1157 0.0513 0.0220 0.0002 0.8108
>
> #結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
> obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを判定した結果をobjに格納
> sum(obj)                                    #条件を満たす数を表示
[1] 1962
> table(out$MAP[obj])                         #パターンごとの出現頻度を表示(条件を満たすもののみ)

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
   955    491    200      2    314
> table(out$MAP[obj])/length(out$MAP[obj])#パターンごとの出現確率を表示(条件を満たすもののみ)

     DEG_G1      DEG_G2      DEG_G3      DEGall
0.486748216 0.250254842 0.101936799 0.001019368
     nonDEG
0.160040775
>
>
```

①

Environment History Connections

To Console | To Source

```
obj <- (tcc$stat$q.value < p...
sum(obj) #条件を満たす数を表示
table(out$MAP[obj]) #パターン...
table(out$MAP[obj])/length(o...
```

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

C: > Users > kadota > Desktop

Name	Size
..	
data_hypodata_3vs3vs3.txt	356.3 KB
Excel.lnk	2.3 KB
hoge8.txt	1.5 MB
RStudio.lnk	948 B

Osabe法の実践40

①は、5% FDR閾値を満たした1962個に対して、どの発現パターンのものがいくつあったかを示したもの。①発現パターンの割当て自体は「TCC正規化+baySeq」というパイプラインで独立に実行した結果であるため…

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Console | Terminal | Jobs

C:/Users/kadota/Desktop/

```
0.1157 0.0513 0.0220 0.0002 0.8108
>
> #結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
> obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを判定した結果をobjに格納
> sum(obj)                                    #条件を満たす数を表示
[1] 1962
> table(out$MAP[obj])                         #パターンごとの出現頻度を表示(条件を満たすもののみ)

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
   955    491    200      2    314
> table(out$MAP[obj])/length(out$MAP[obj])#パターンごとの出現確率を表示(条件を満たすもののみ)

     DEG_G1      DEG_G2      DEG_G3      DEGall
0.486748216 0.250254842 0.101936799 0.001019368
     nonDEG
0.160040775
>
>
```

①

Environment | History | Connections

To Console | To Source

```
obj <- (tcc$stat$q.value < p...
sum(obj) #条件を満たす数を表示
table(out$MAP[obj]) #パターン...
table(out$MAP[obj])/length(o...
```

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

C: > Users > kadota > Desktop

Name	Size
..	
data_hypodata_3vs3vs3.txt	356.3 KB
Excel.lnk	2.3 KB
hoge8.txt	1.5 MB
RStudio.lnk	948 B

Osabe法の実践41

①は、5% FDR閾値を満たした1962個に対して、どの発現パターンのものがいくつあったかを示したもの。①発現パターンの割当て自体は「TCC正規化+baySeq」というパイプラインで独立に実行した結果であるため、②この枠組みでnonDEGパターンに割り当てられた314個という数値が…

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Console | Terminal | Jobs

C:/Users/kadota/Desktop/

```
0.1157 0.0513 0.0220 0.0002 0.8108
>
> #結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
> obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを判定した結果をobjに格納
> sum(obj)                                    #条件を満たす数を表示
[1] 1962
> table(out$MAP[obj])                         #パターンごとの出現頻度を表示(条件を満たすもののみ)

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
   955    491    200      2    314
> table(out$MAP[obj])/length(out$MAP[obj])#パターンごとの出現確率を表示(条件を満たすもののみ)

     DEG_G1      DEG_G2      DEG_G3      DEGall
0.486748216 0.250254842 0.101936799 0.001019368
     nonDEG
0.160040775
>
>
```

②

```
sum(obj) #条件を満たす数を表示
table(out$MAP[obj]) #パターン…
table(out$MAP[obj])/length(o…
```

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

C: > Users > kadota > Desktop

Name	Size
..	
data_hypodata_3vs3vs3.txt	356.3 KB
Excel.lnk	2.3 KB
hoge8.txt	1.5 MB
RStudio.lnk	948 B

Osabe法の実践42

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Console | Terminal | Jobs

C:/Users/kadota/Desktop/

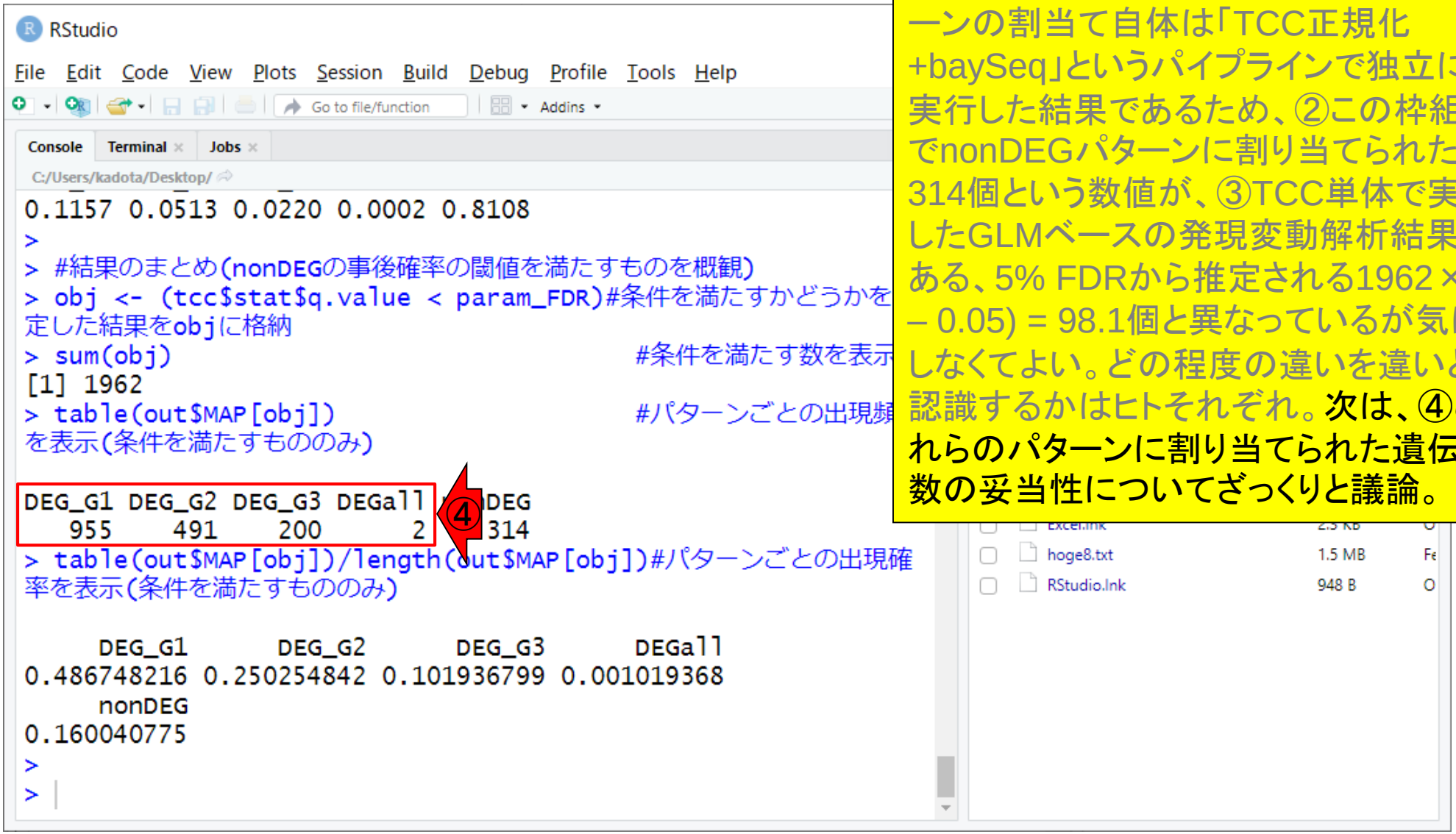
```
0.1157 0.0513 0.0220 0.0002 0.8108
>
> #結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
> obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを
定した結果をobjに格納
> sum(obj)                                    #条件を満たす数を表示
[1] 1962
> table(out$MAP[obj])                         #パターンごとの出現頻
を表示(条件を満たすもののみ)

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
   955    491    200       2    314
> table(out$MAP[obj])/length(out$MAP[obj])#パターンごとの出現確
率を表示(条件を満たすもののみ)

     DEG_G1      DEG_G2      DEG_G3      DEGall
0.486748216 0.250254842 0.101936799 0.001019368
     nonDEG
0.160040775
>
>
```

③

..		
data_hypodata_3vs3vs3.txt	356.3 KB	Fe
Excel.lnk	2.3 KB	O
hoge8.txt	1.5 MB	Fe
RStudio.lnk	948 B	O

①は、5% FDR閾値を満たした1962個に対して、どの発現パターンのものがいくつあったかを示したもの。①発現パターンの割当て自体は「TCC正規化+baySeq」というパイプラインで独立に実行した結果であるため、②この枠組みでnonDEGパターンに割り当てられた314個という数値が、③TCC単体で実行したGLMベースの発現変動解析結果である、5% FDRから推定される1962×(1 – 0.05) = 98.1個と異なっているが気にしなくてよい。どの程度の違いを違いと認識するかはヒトそれぞれ。

Osabe法の実践43



①は、5% FDR閾値を満たした1962個に対して、どの発現パターンのものがいくつあったかを示したもの。①発現パターンの割当て自体は「TCC正規化+baySeq」というパイプラインで独立に実行した結果であるため、②この枠組みでnonDEGパターンに割り当てられた314個という数値が、③TCC単体で実行したGLMベースの発現変動解析結果である、5% FDRから推定される1962 × (1 – 0.05) = 98.1個と異なっているが気にしなくてよい。どの程度の違いを違いと認識するかはヒトそれぞれ。**次は、④これらのパターンに割り当てられた遺伝子数の妥当性についてざっくりと議論。**

Osabe法の実践44

①入力ファイルは、②「サンプルデータ」という項目の、③例題15で作成したシミュレーションデータなので、答えが分かっています。

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#analysis_deg_3_unpaired_ari_advance...

ゲスト

②

8. サンプルデータ15の10,000 genes×9 samplesのカウントデータ(data_hypodata_3vs3vs3.txt)の場合：

例題7と似ているが、全体的な発現変動の度合い(ANOVA的などこかの群間で発現変動している度合いでのランキング)をEEE-Eで行い、発現変動パターンの同定をEEE-b(長部らの原著論文中ではEEE-baySEqと表記)で行った結果を出力しています

③

①

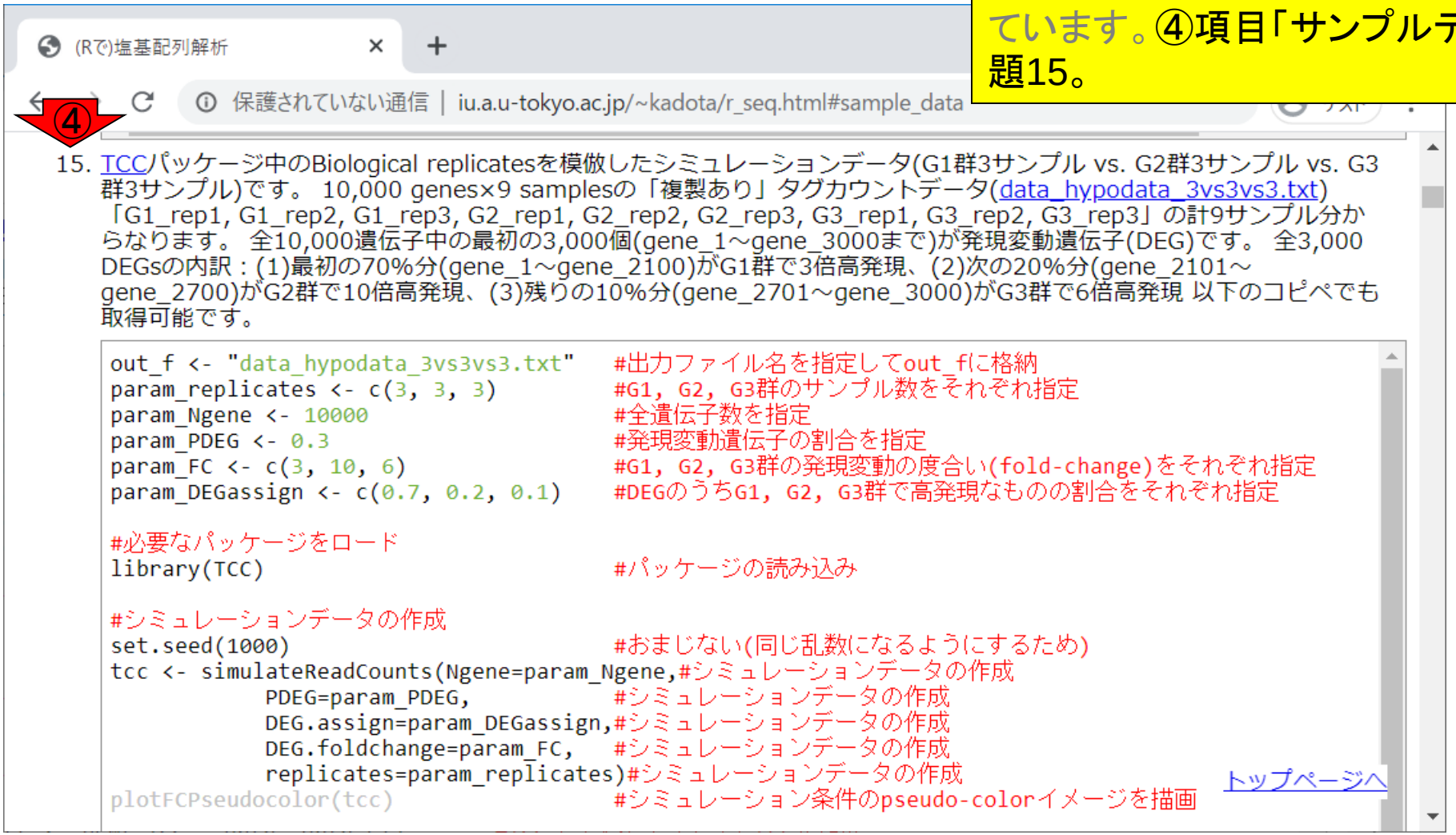
```
in_f <- "data_hypodata_3vs3vs3.txt"          #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.txt"                         #出力ファイル名を指定してout_fに格納
param_G1 <- 3                                #G1群のサンプル数を指定
param_G2 <- 3                                #G2群のサンプル数を指定
param_G3 <- 3                                #G3群のサンプル数を指定
param_FDR <- 0.05                            #false discovery rate (FDR)閾値を指定
param_narabi <- c("nonDEG", "DEG_G1", "DEG_G2", "DEG_G3", "DEGall")#パターン名を指定(並びは変えない)
param_samplesize <- 1000                     #ブートストラップリサンプリング回数(100000が推奨。大きい値ほど計算時間

#必要なパッケージをロード
library(TCC)                                 #パッケージの読み込み
library(baySeq)                              #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの読み込み

#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2), rep(3, param_G3))#G1群を1、G2群を2、G3群を3としたベクト
tcc <- new("TCC", data, data.cl)             #TCCクラスオブジェクトtccを作成
```

トップページへ

```
#TCC正規化
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger" #正規化を実行した結果をtccに格納
```

Osabe法の実践45

①入力ファイルは、②「サンプルデータ」という項目の、③例題15で作成したシミュレーションデータなので、答えが分かっています。④項目「サンプルデータ」の例題15。



15. TCCパッケージ中のBiological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル vs. G3群3サンプル)です。 10,000 genes×9 samplesの「複製あり」タグカウントデータ(data_hypodata_3vs3vs3.txt)「G1_rep1, G1_rep2, G1_rep3, G2_rep1, G2_rep2, G2_rep3, G3_rep1, G3_rep2, G3_rep3」の計9サンプル分からなります。 全10,000遺伝子中の最初の3,000個(gene_1～gene_3000まで)が発現変動遺伝子(DEG)です。 全3,000 DEGsの内訳：(1)最初の70%分(gene_1～gene_2100)がG1群で3倍高発現、(2)次の20%分(gene_2101～gene_2700)がG2群で10倍高発現、(3)残りの10%分(gene_2701～gene_3000)がG3群で6倍高発現 以下のコピペでも取得可能です。

```
out_f <- "data_hypodata_3vs3vs3.txt"    #出力ファイル名を指定してout_fに格納
param_replicates <- c(3, 3, 3)          #G1, G2, G3群のサンプル数をそれぞれ指定
param_Ngene <- 10000                    #全遺伝子数を指定
param_PDEG <- 0.3                       #発現変動遺伝子の割合を指定
param_FC <- c(3, 10, 6)                 #G1, G2, G3群の発現変動の度合い(fold-change)をそれぞれ指定
param_DEGassign <- c(0.7, 0.2, 0.1)     #DEGのうちG1, G2, G3群で高発現なものの割合をそれぞれ指定

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み

#シミュレーションデータの作成
set.seed(1000)                          #おまじない(同じ乱数になるようにするため)
tcc <- simulateReadCounts(Ngene=param_Ngene,#シミュレーションデータの作成
            PDEG=param_PDEG,            #シミュレーションデータの作成
            DEG.assign=param_DEGassign,#シミュレーションデータの作成
            DEG.foldchange=param_FC,    #シミュレーションデータの作成
            replicates=param_replicates)#シミュレーションデータの作成
plotFCPseudocolor(tcc)                  #シミュレーション条件のpseudo-colorイメージを描画
```

トップページへ

Osabe法の実践46

①入力ファイルは、②「サンプルデータ」という項目の、③例題15で作成したシミュレーションデータなので、答えが分かっています。④項目「サンプルデータ」の例題15。⑤出力ファイル名。⑥各群につき3反復からなる3群間比較。

(Rで)塩基配列解析

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

15. TCCパッケージ中のBiological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル vs. G3群3サンプル)です。 10,000 genes×9 samplesの「複製あり」タグカウントデータ(data_hypodata_3vs3vs3.txt)「G1_rep1, G1_rep2, G1_rep3, G2_rep1, G2_rep2, G2_rep3, G3_rep1, G3_rep2, G3_rep3」の計9サンプル分からなります。 全10,000遺伝子中の最初の3,000個(gene_1～gene_3000まで)が発現変動遺伝子(DEG)です。 全3,000 DEGsの内訳：(1)最初の70%分(gene_1～gene_2100)がG1群で3倍高発現、(2)次の20%分(gene_2101～gene_2700)がG2群で10倍高発現、(3)残りの10%分(gene_2701～gene_3000)がG3群で6倍高発現 以下のコピペでも取得可能です。

⑤

⑥

```
out_f <- "data_hypodata_3vs3vs3.txt"  #出力ファイル名を指定してout_fに格納
param_replicates <- c(3, 3, 3)        #G1, G2, G3群のサンプル数をそれぞれ指定
param_Ngene <- 10000                  #全遺伝子数を指定
param_PDEG <- 0.3                     #発現変動遺伝子の割合を指定
param_FC <- c(3, 10, 6)               #G1, G2, G3群の発現変動の度合い(fold-change)をそれぞれ指定
param_DEGassign <- c(0.7, 0.2, 0.1)   #DEGのうちG1, G2, G3群で高発現なものの割合をそれぞれ指定

#必要なパッケージをロード
library(TCC)                          #パッケージの読み込み

#シミュレーションデータの作成
set.seed(1000)                        #おまじない(同じ乱数になるようにするため)
tcc <- simulateReadCounts(Ngene=param_Ngene,#シミュレーションデータの作成
            PDEG=param_PDEG,          #シミュレーションデータの作成
            DEG.assign=param_DEGassign,#シミュレーションデータの作成
            DEG.foldchange=param_FC,  #シミュレーションデータの作成
            replicates=param_replicates)#シミュレーションデータの作成
plotFCPseudocolor(tcc)                #シミュレーション条件のpseudo-colorイメージを描画
```

トップページへ

Osabe法の実践47

①入力ファイルは、②「サンプルデータ」という項目の、③例題15で作成したシミュレーションデータなので、答えが分かっています。④項目「サンプルデータ」の例題15。⑤出力ファイル名。⑥各群につき3反復からなる3群間比較。⑦遺伝子数は10000個で、そのうち30%が発現変動遺伝子（DEG）。

(Rで)塩基配列解析

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

15. TCCパッケージ中のBiological replicatesを模倣したシミュレーションデータ(G1群
群3サンプル)です。 10,000 genes×9 samplesの「複製あり」タグカウントデー
「G1_rep1, G1_rep2, G1_rep3, G2_rep1, G2_rep2, G2_rep3, G3_rep1, G3
らなります。 全10,000遺伝子中の最初の3,000個(gene_1～gene_3000まで)が
DEGsの内訳：(1)最初の70%分(gene_1～gene_2100)がG1群で3倍高発現、(2)次の20%分(gene_2101～
gene_2700)がG2群で10倍高発現、(3)残りの10%分(gene_2701～gene_3000)がG3群で6倍高発現 以下のコピペでも
取得可能です。

```
out_f <- "data_hypodata_3vs3vs3.txt"    #出力ファイル名を指定してout_fに格納
param_replicates <- c(3, 3, 3)          #G1, G2, G3群のサンプル数をそれぞれ指定
param_Ngene <- 10000                    #全遺伝子数を指定
param_PDEG <- 0.3                       #発現変動遺伝子の割合を指定
param_FC <- c(3, 10, 6)                 #G1, G2, G3群の発現変動の度合い(fold-change)をそれぞれ指定
param_DEGassign <- c(0.7, 0.2, 0.1)     #DEGのうちG1, G2, G3群で高発現なものの割合をそれぞれ指定

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み

#シミュレーションデータの作成
set.seed(1000)                          #おまじない(同じ乱数になるようにするため)
tcc <- simulateReadCounts(Ngene=param_Ngene,#シミュレーションデータの作成
            PDEG=param_PDEG,            #シミュレーションデータの作成
            DEG.assign=param_DEGassign,#シミュレーションデータの作成
            DEG.foldchange=param_FC,    #シミュレーションデータの作成
            replicates=param_replicates)#シミュレーションデータの作成
plotFCPseudocolor(tcc)                  #シミュレーション条件のpseudo-colorイメージを描画
```

⑦

トップページへ

Osabe法の実践48

①入力ファイルは、②「サンプルデータ」という項目の、③例題15で作成したシミュレーションデータなので、答えが分かっています。④項目「サンプルデータ」の例題15。⑤出力ファイル名。⑥各群につき3反復からなる3群間比較。⑦遺伝子数は10000個で、そのうち30%が発現変動遺伝子（DEG）。⑧G1群で3倍、G2群で10倍、G3群で6倍高発現となるように設定。

(Rで)塩基配列解析

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

15. TCCパッケージ中のBiological replicatesを模倣したシミュレーションデータ(G1群
群3サンプル)です。 10,000 genes×9 samplesの「複製あり」タグカウントデー
「G1_rep1, G1_rep2, G1_rep3, G2_rep1, G2_rep2, G2_rep3, G3_rep1, G3_
らなります。 全10,000遺伝子中の最初の3,000個(gene_1～gene_3000まで)が
DEGsの内訳：(1)最初の70%分(gene_1～gene_2100)がG1群で3倍高発現、(2)
gene_2700)がG2群で10倍高発現、(3)残りの10%分(gene_2701～gene_3000)
取得可能です。

```
out_f <- "data_hypodata_3vs3vs3.txt"    #出力ファイル名を指定してout_fに格納
param_replicates <- c(3, 3, 3)          #G1, G2, G3群のサンプル数をそれぞれ指定
param_Ngene <- 10000                    #全遺伝子数を指定
param_PDEG <- 0.3                       #発現変動遺伝子の割合を指定
param_FC <- c(3, 10, 6)                 #G1, G2, G3群の発現変動の度合い(fold-change)をそれぞれ指定
param_DEGassign <- c(0.7, 0.2, 0.1)     #DEGのうちG1, G2, G3群で高発現なものの割合をそれぞれ指定

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み

#シミュレーションデータの作成
set.seed(1000)                          #おまじない(同じ乱数になるようにするため)
tcc <- simulateReadCounts(Ngene=param_Ngene,#シミュレーションデータの作成
            PDEG=param_PDEG,            #シミュレーションデータの作成
            DEG.assign=param_DEGassign,#シミュレーションデータの作成
            DEG.foldchange=param_FC,    #シミュレーションデータの作成
            replicates=param_replicates)#シミュレーションデータの作成
plotFCPseudocolor(tcc)                  #シミュレーション条件のpseudo-colorイメージを描画
```

⑧

トップページへ

Osabe法の実践49

(Rで)塩基配列解析

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

15. TCCパッケージ中のBiological replicatesを模倣したシミュレーションデータ(G1群
群3サンプル)です。 10,000 genes×9 samplesの「複製あり」タグカウントデー
「G1_rep1, G1_rep2, G1_rep3, G2_rep1, G2_rep2, G2_rep3, G3_rep1, G3_
らなります。 全10,000遺伝子中の最初の3,000個(gene_1～gene_3000まで)が
DEGsの内訳：(1)最初の70%分(gene_1～gene_2100)がG1群で3倍高発現、(2)
gene_2700)がG2群で10倍高発現、(3)残りの10%分(gene_2701～gene_3000)
取得可能です。

```
out_f <- "data_hypodata_3vs3vs3.txt"    #出力ファイル名を指定してout_f
param_replicates <- c(3, 3, 3)          #G1, G2, G3群のサンプル数をそ
param_Ngene <- 10000                    #全遺伝子数を指定
param_PDEG <- 0.3                       #発現変動遺伝子の割合を指定
param_FC <- c(3, 10, 6)                 #G1, G2, G3群の発現変動の度合
param_DEGassign <- c(0.7, 0.2, 0.1)     #DEGのうちG1, G2, G3群で高発現

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み

#シミュレーションデータの作成
set.seed(1000)                          #おまじない(同じ乱数になるようにするため)
tcc <- simulateReadCounts(Ngene=param_Ngene,#シミュレーションデータの作成
            PDEG=param_PDEG,            #シミュレーションデータの作成
            DEG.assign=param_DEGassign,#シミュレーションデータの作成
            DEG.foldchange=param_FC,    #シミュレーションデータの作成
            replicates=param_replicates)#シミュレーションデータの作成
plotFCPseudocolor(tcc)                  #シミュレーション条件のpseudo-colorイメージを描画
```

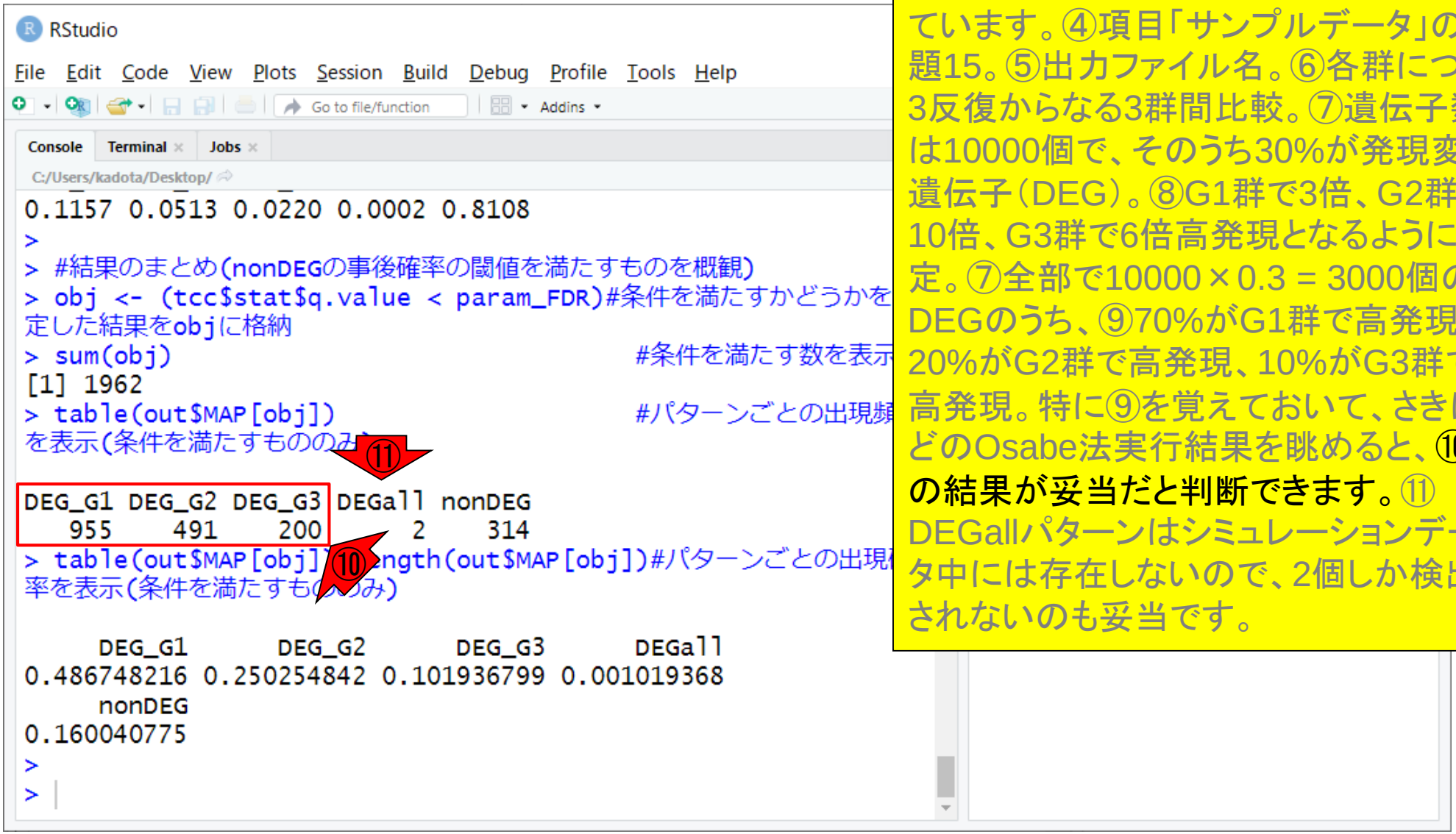
⑦

⑨

トップページへ

①入力ファイルは、②「サンプルデータ」という項目の、③例題15で作成したシミュレーションデータなので、答えが分かっています。④項目「サンプルデータ」の例題15。⑤出力ファイル名。⑥各群につき3反復からなる3群間比較。⑦遺伝子数は10000個で、そのうち30%が発現変動遺伝子（DEG）。⑧G1群で3倍、G2群で10倍、G3群で6倍高発現となるように設定。⑦全部で10000×0.3 = 3000個のDEGのうち、⑨70%がG1群で高発現、20%がG2群で高発現、10%がG3群で高発現。特に⑨を覚えておいて、さきほどのOsabe法実行結果を眺めると、…

Osabe法の実践50



①入力ファイルは、②「サンプルデータ」という項目の、③例題15で作成したシミュレーションデータなので、答えが分かっています。④項目「サンプルデータ」の例題15。⑤出力ファイル名。⑥各群につき3反復からなる3群間比較。⑦遺伝子数は10000個で、そのうち30%が発現変動遺伝子（DEG）。⑧G1群で3倍、G2群で10倍、G3群で6倍高発現となるように設定。⑦全部で10000×0.3 = 3000個のDEGのうち、⑨70%がG1群で高発現、20%がG2群で高発現、10%がG3群で高発現。特に⑨を覚えておいて、さきほどのOsabe法実行結果を眺めると、⑩の結果が妥当だと判断できます。⑪DEGallパターンはシミュレーションデータ中には存在しないので、2個しか検出されないのも妥当です。

Osabe法の実践51

①が画面の下のほうになるようにページ上部に移動。

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Console Terminal Jobs

C:/Users/kadota/Desktop/

```
0.1157 0.0513 0.0220 0.0002 0.8108
>
> #結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
> obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを判定した結果をobjに格納
> sum(obj)                                    #条件を満たす数を表示
[1] 1962
> table(out$MAP[obj])                         #パターンごとの出現頻度を表示(条件を満たすもののみ)

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
   955    491    200       2    314
> table(out$MAP[obj])/length(out$MAP[obj])#パターンごとの出現確率を表示(条件を満たすもののみ)

     DEG_G1      DEG_G2      DEG_G3      DEGall
0.486748216 0.250254842 0.101936799 0.001019368
     nonDEG
0.160040775
>
>
```

① (赤い矢印)

Environment History Connections

To Console | To Source

```
obj <- (tcc$stat$q.value < p...
sum(obj) #条件を満たす数を表示
table(out$MAP[obj]) #パターン...
table(out$MAP[obj])/length(o...
```

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

C: > Users > kadota > Desktop

Name	Size
..	
data_hypodata_3vs3vs3.txt	356.3 KB
Excel.lnk	2.3 KB
hoge8.txt	1.5 MB
RStudio.lnk	948 B

Osabe法の実践52

①が画面の下のほうになるようにページ上部に移動。こんな感じ。

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Console Terminal Jobs

C:/Users/kadota/Desktop/

```
> #結果のまとめ(全遺伝子)
> head(out$MAP)                                    #最初の6遺伝子分のパターンを表示
[1] "DEG_G1" "nonDEG" "DEG_G1" "DEG_G1" "nonDEG" "nonDEG"
> table(out$MAP)                                   #パターンごとの出現頻度を表示

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
  1157    513    220      2   8108
> table(out$MAP)/length(out$MAP)                   #パターンごとの出現確率を表示

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
0.1157 0.0513 0.0220 0.0002 0.8108
>
> #結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
> obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを判定した結果をobjに格納
> sum(obj)                                         #条件を満たす数を表示
[1] 1962
> table(out$MAP[obj])                              #パターンごとの出現頻度
```

①

Environment History Connections

To Console | To Source

```
obj <- (tcc$stat$q.value < p...
sum(obj) #条件を満たす数を表示
table(out$MAP[obj]) #パターン...
table(out$MAP[obj])/length(o...
```

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

C: > Users > kadota > Desktop

Name	Size
..	
data_hypodata_3vs3vs3.txt	356.3 KB
Excel.lnk	2.3 KB
RStudio.lnk	948 B

Osabe法の実践53

①が画面の下のほうになるようにページ上部に移動。こんな感じ。②(1157 + 513 + 220 + 2) = 1892個がDEGと判定されていることが分かる。実際のDEGは3000個なので少ないと感じるところだが、Osabe法では(このスクリプトの場合は)5% FDR閾値を満たさなかった遺伝子は全てnonDEGパターンに強制的に移行させるので、妥当と言えば妥当。

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Console | Terminal | Jobs

C:/Users/kadota/Desktop/

```
> #結果のまとめ(全遺伝子)
> head(out$MAP)                                  #最初の6遺伝子分のパターンを表示
[1] "DEG_G1" "nonDEG" "DEG_G1" "DEG_G1" "nonDEG" "nonDEG"
> table(out$MAP)                                 #パターンごとの出現頻度を表示

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
  1157    513    220      2   8108
> table(out$MAP)/length(out$MAP)                 #パターンごとの出現確率を表示

DEG_G1 DEG_G2 DEG_G3 DEGall nonDEG
0.1157 0.0513 0.0220 0.0002 0.8108
>
> #結果のまとめ(nonDEGの事後確率の閾値を満たすものを概観)
> obj <- (tcc$stat$q.value < param_FDR)#条件を満たすかどうかを判定した結果をobjに格納
> sum(obj)                                       #条件を満たす数を表示
[1] 1962
> table(out$MAP[obj])                            #パターンごとの出現頻度
```

②

①

table(out$MAP[obj]) #パターン...

table(out$MAP[obj])/length(o...

Files | Plots | Packages | Help | Viewer

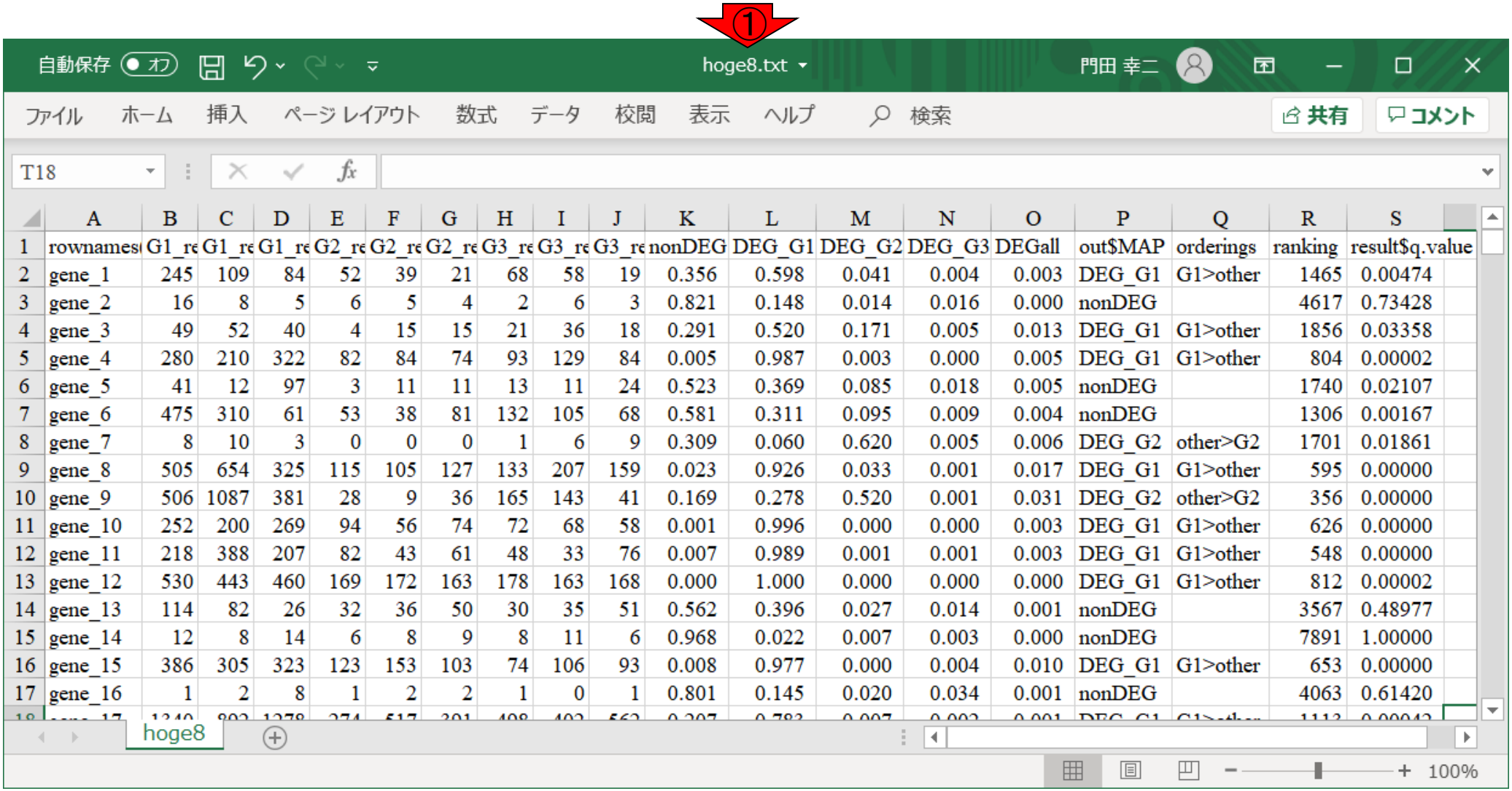
New Folder | Delete | Rename | More

C: > Users > kadota > Desktop

Name	Size
..	
data_hypodata_3vs3vs3.txt	356.3 KB
Excel.lnk	2.3 KB
RStudio.lnk	948 B

Osabe法の実践54

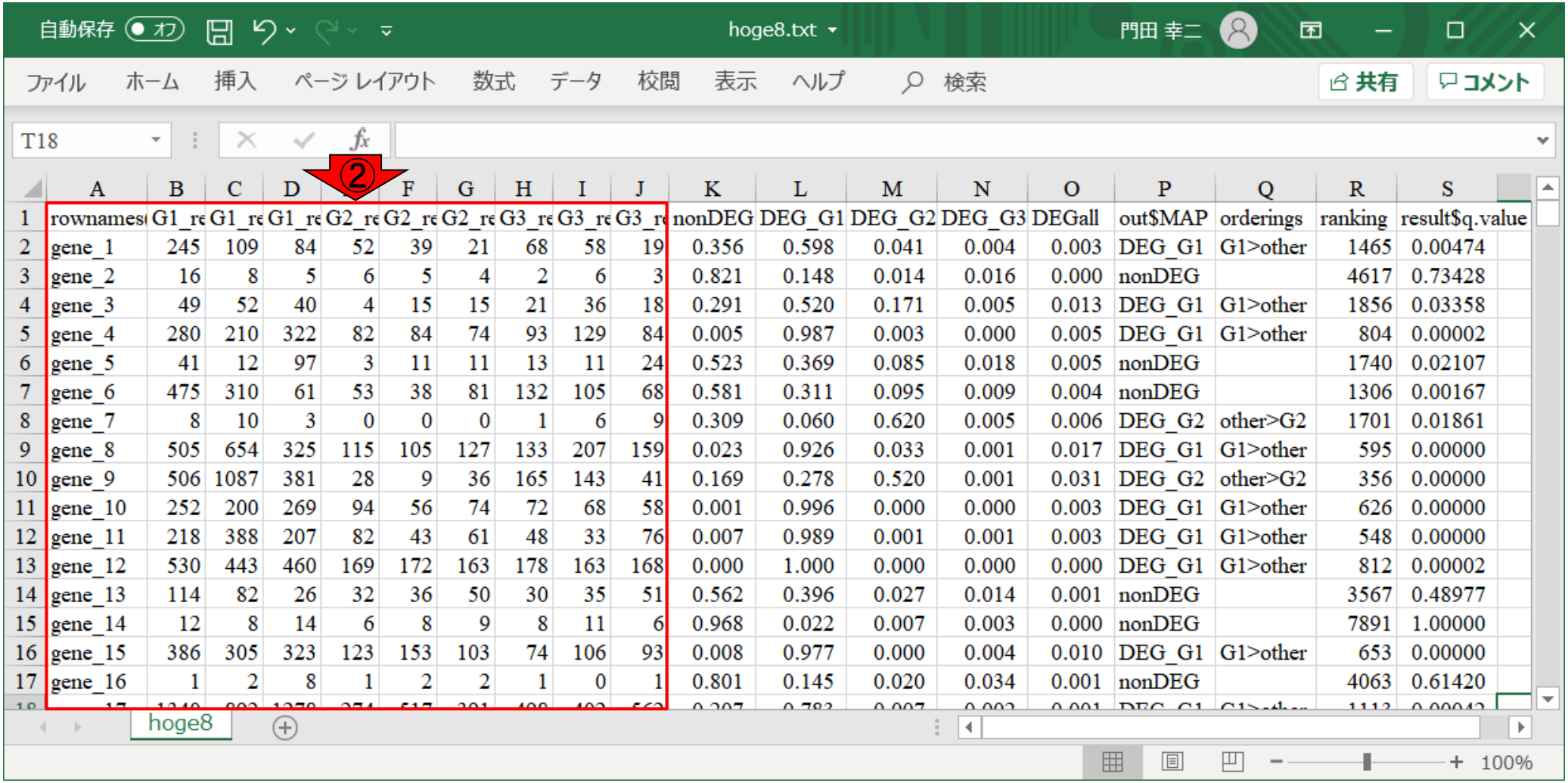
①出力ファイルであるhoge8.txtをExcelで開いたところ。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	rownames(	G1_re	G1_re	G1_re	G2_re	G2_re	G2_re	G3_re	G3_re	G3_re	nonDEG	DEG_G1	DEG_G2	DEG_G3	DEGall	out$MAP	orderings	ranking	result$q.value
2	gene_1	245	109	84	52	39	21	68	58	19	0.356	0.598	0.041	0.004	0.003	DEG_G1	G1>other	1465	0.00474
3	gene_2	16	8	5	6	5	4	2	6	3	0.821	0.148	0.014	0.016	0.000	nonDEG		4617	0.73428
4	gene_3	49	52	40	4	15	15	21	36	18	0.291	0.520	0.171	0.005	0.013	DEG_G1	G1>other	1856	0.03358
5	gene_4	280	210	322	82	84	74	93	129	84	0.005	0.987	0.003	0.000	0.005	DEG_G1	G1>other	804	0.00002
6	gene_5	41	12	97	3	11	11	13	11	24	0.523	0.369	0.085	0.018	0.005	nonDEG		1740	0.02107
7	gene_6	475	310	61	53	38	81	132	105	68	0.581	0.311	0.095	0.009	0.004	nonDEG		1306	0.00167
8	gene_7	8	10	3	0	0	0	1	6	9	0.309	0.060	0.620	0.005	0.006	DEG_G2	other>G2	1701	0.01861
9	gene_8	505	654	325	115	105	127	133	207	159	0.023	0.926	0.033	0.001	0.017	DEG_G1	G1>other	595	0.00000
10	gene_9	506	1087	381	28	9	36	165	143	41	0.169	0.278	0.520	0.001	0.031	DEG_G2	other>G2	356	0.00000
11	gene_10	252	200	269	94	56	74	72	68	58	0.001	0.996	0.000	0.000	0.003	DEG_G1	G1>other	626	0.00000
12	gene_11	218	388	207	82	43	61	48	33	76	0.007	0.989	0.001	0.001	0.003	DEG_G1	G1>other	548	0.00000
13	gene_12	530	443	460	169	172	163	178	163	168	0.000	1.000	0.000	0.000	0.000	DEG_G1	G1>other	812	0.00002
14	gene_13	114	82	26	32	36	50	30	35	51	0.562	0.396	0.027	0.014	0.001	nonDEG		3567	0.48977
15	gene_14	12	8	14	6	8	9	8	11	6	0.968	0.022	0.007	0.003	0.000	nonDEG		7891	1.00000
16	gene_15	386	305	323	123	153	103	74	106	93	0.008	0.977	0.000	0.004	0.010	DEG_G1	G1>other	653	0.00000
17	gene_16	1	2	8	1	2	2	1	0	1	0.801	0.145	0.020	0.034	0.001	nonDEG		4063	0.61420

Osabe法の実践55

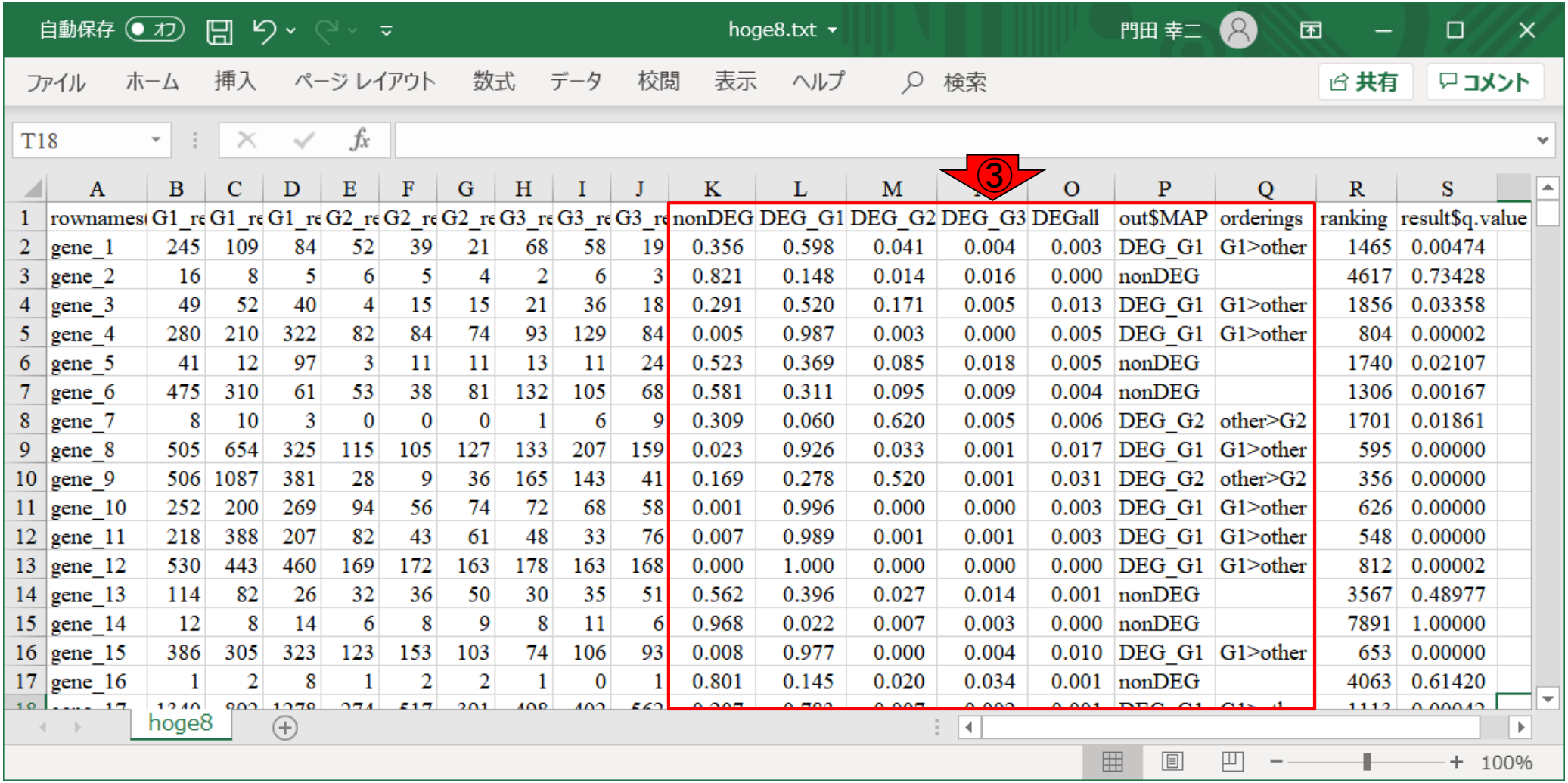
①出力ファイルであるhoge8.txtをExcelで開いたところ。②入力データ。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	rownames(	G1_re	G1_re	G1_re	G2_re	G2_re	G2_re	G3_re	G3_re	G3_re	nonDEG	DEG_G1	DEG_G2	DEG_G3	DEGall	out$MAP	orderings	ranking	result$q.value
2	gene_1	245	109	84	52	39	21	68	58	19	0.356	0.598	0.041	0.004	0.003	DEG_G1	G1>other	1465	0.00474
3	gene_2	16	8	5	6	5	4	2	6	3	0.821	0.148	0.014	0.016	0.000	nonDEG		4617	0.73428
4	gene_3	49	52	40	4	15	15	21	36	18	0.291	0.520	0.171	0.005	0.013	DEG_G1	G1>other	1856	0.03358
5	gene_4	280	210	322	82	84	74	93	129	84	0.005	0.987	0.003	0.000	0.005	DEG_G1	G1>other	804	0.00002
6	gene_5	41	12	97	3	11	11	13	11	24	0.523	0.369	0.085	0.018	0.005	nonDEG		1740	0.02107
7	gene_6	475	310	61	53	38	81	132	105	68	0.581	0.311	0.095	0.009	0.004	nonDEG		1306	0.00167
8	gene_7	8	10	3	0	0	0	1	6	9	0.309	0.060	0.620	0.005	0.006	DEG_G2	other>G2	1701	0.01861
9	gene_8	505	654	325	115	105	127	133	207	159	0.023	0.926	0.033	0.001	0.017	DEG_G1	G1>other	595	0.00000
10	gene_9	506	1087	381	28	9	36	165	143	41	0.169	0.278	0.520	0.001	0.031	DEG_G2	other>G2	356	0.00000
11	gene_10	252	200	269	94	56	74	72	68	58	0.001	0.996	0.000	0.000	0.003	DEG_G1	G1>other	626	0.00000
12	gene_11	218	388	207	82	43	61	48	33	76	0.007	0.989	0.001	0.001	0.003	DEG_G1	G1>other	548	0.00000
13	gene_12	530	443	460	169	172	163	178	163	168	0.000	1.000	0.000	0.000	0.000	DEG_G1	G1>other	812	0.00002
14	gene_13	114	82	26	32	36	50	30	35	51	0.562	0.396	0.027	0.014	0.001	nonDEG		3567	0.48977
15	gene_14	12	8	14	6	8	9	8	11	6	0.968	0.022	0.007	0.003	0.000	nonDEG		7891	1.00000
16	gene_15	386	305	323	123	153	103	74	106	93	0.008	0.977	0.000	0.004	0.010	DEG_G1	G1>other	653	0.00000
17	gene_16	1	2	8	1	2	2	1	0	1	0.801	0.145	0.020	0.034	0.001	nonDEG		4063	0.61420

Osabe法の実践56

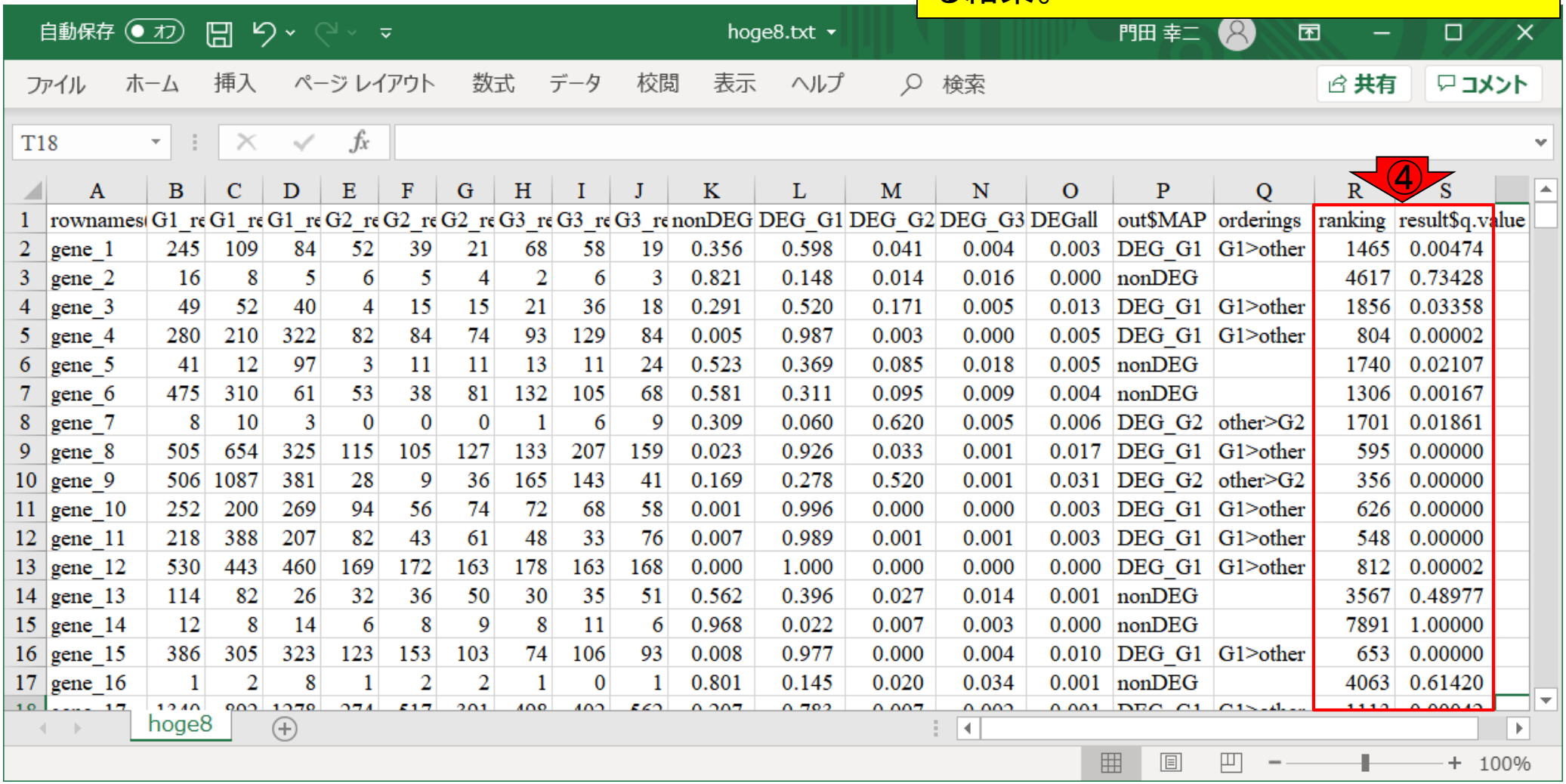
①出力ファイルであるhoge8.txtをExcelで開いたところ。②入力データ。③「TCC正規化+baySeq」で得られた発現パターン分類結果。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	rownames(	G1_re	G1_re	G1_re	G2_re	G2_re	G2_re	G3_re	G3_re	G3_re	nonDEG	DEG_G1	DEG_G2	DEG_G3	DEGall	out$MAP	orderings	ranking	result$q.value
2	gene_1	245	109	84	52	39	21	68	58	19	0.356	0.598	0.041	0.004	0.003	DEG_G1	G1>other	1465	0.00474
3	gene_2	16	8	5	6	5	4	2	6	3	0.821	0.148	0.014	0.016	0.000	nonDEG		4617	0.73428
4	gene_3	49	52	40	4	15	15	21	36	18	0.291	0.520	0.171	0.005	0.013	DEG_G1	G1>other	1856	0.03358
5	gene_4	280	210	322	82	84	74	93	129	84	0.005	0.987	0.003	0.000	0.005	DEG_G1	G1>other	804	0.00002
6	gene_5	41	12	97	3	11	11	13	11	24	0.523	0.369	0.085	0.018	0.005	nonDEG		1740	0.02107
7	gene_6	475	310	61	53	38	81	132	105	68	0.581	0.311	0.095	0.009	0.004	nonDEG		1306	0.00167
8	gene_7	8	10	3	0	0	0	1	6	9	0.309	0.060	0.620	0.005	0.006	DEG_G2	other>G2	1701	0.01861
9	gene_8	505	654	325	115	105	127	133	207	159	0.023	0.926	0.033	0.001	0.017	DEG_G1	G1>other	595	0.00000
10	gene_9	506	1087	381	28	9	36	165	143	41	0.169	0.278	0.520	0.001	0.031	DEG_G2	other>G2	356	0.00000
11	gene_10	252	200	269	94	56	74	72	68	58	0.001	0.996	0.000	0.000	0.003	DEG_G1	G1>other	626	0.00000
12	gene_11	218	388	207	82	43	61	48	33	76	0.007	0.989	0.001	0.001	0.003	DEG_G1	G1>other	548	0.00000
13	gene_12	530	443	460	169	172	163	178	163	168	0.000	1.000	0.000	0.000	0.000	DEG_G1	G1>other	812	0.00002
14	gene_13	114	82	26	32	36	50	30	35	51	0.562	0.396	0.027	0.014	0.001	nonDEG		3567	0.48977
15	gene_14	12	8	14	6	8	9	8	11	6	0.968	0.022	0.007	0.003	0.000	nonDEG		7891	1.00000
16	gene_15	386	305	323	123	153	103	74	106	93	0.008	0.977	0.000	0.004	0.010	DEG_G1	G1>other	653	0.00000
17	gene_16	1	2	8	1	2	2	1	0	1	0.801	0.145	0.020	0.034	0.001	nonDEG		4063	0.61420

Osabe法の実践57

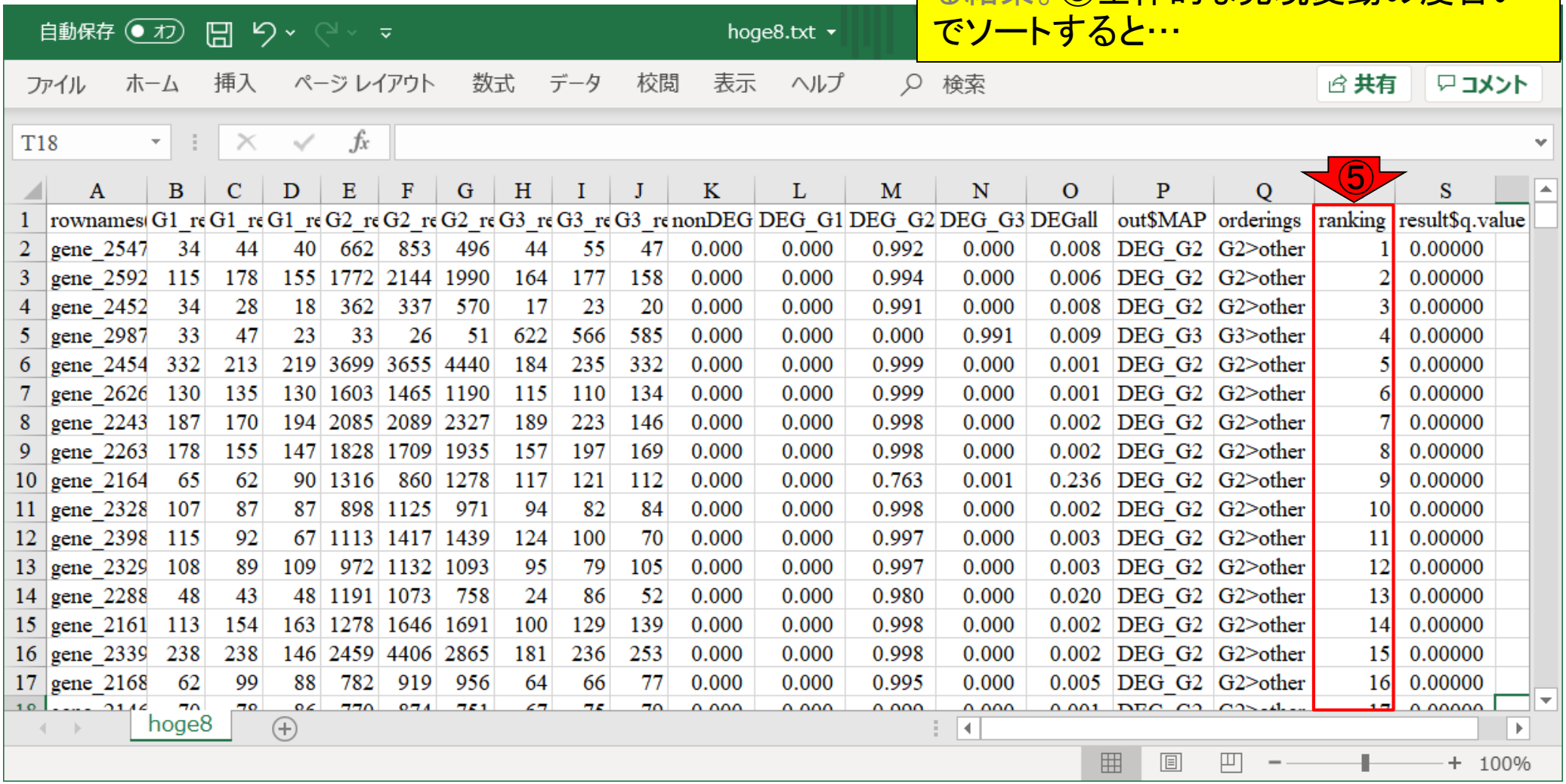
①出力ファイルであるhoge8.txtをExcelで開いたところ。②入力データ。③「TCC正規化+baySeq」で得られた発現パターン分類結果。④従来法のTCCで得られる結果。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	rownames(	G1_re	G1_re	G1_re	G2_re	G2_re	G2_re	G3_re	G3_re	G3_re	nonDEG	DEG_G1	DEG_G2	DEG_G3	DEGall	out$MAP	orderings	ranking	result$q.value
2	gene_1	245	109	84	52	39	21	68	58	19	0.356	0.598	0.041	0.004	0.003	DEG_G1	G1>other	1465	0.00474
3	gene_2	16	8	5	6	5	4	2	6	3	0.821	0.148	0.014	0.016	0.000	nonDEG		4617	0.73428
4	gene_3	49	52	40	4	15	15	21	36	18	0.291	0.520	0.171	0.005	0.013	DEG_G1	G1>other	1856	0.03358
5	gene_4	280	210	322	82	84	74	93	129	84	0.005	0.987	0.003	0.000	0.005	DEG_G1	G1>other	804	0.00002
6	gene_5	41	12	97	3	11	11	13	11	24	0.523	0.369	0.085	0.018	0.005	nonDEG		1740	0.02107
7	gene_6	475	310	61	53	38	81	132	105	68	0.581	0.311	0.095	0.009	0.004	nonDEG		1306	0.00167
8	gene_7	8	10	3	0	0	0	1	6	9	0.309	0.060	0.620	0.005	0.006	DEG_G2	other>G2	1701	0.01861
9	gene_8	505	654	325	115	105	127	133	207	159	0.023	0.926	0.033	0.001	0.017	DEG_G1	G1>other	595	0.00000
10	gene_9	506	1087	381	28	9	36	165	143	41	0.169	0.278	0.520	0.001	0.031	DEG_G2	other>G2	356	0.00000
11	gene_10	252	200	269	94	56	74	72	68	58	0.001	0.996	0.000	0.000	0.003	DEG_G1	G1>other	626	0.00000
12	gene_11	218	388	207	82	43	61	48	33	76	0.007	0.989	0.001	0.001	0.003	DEG_G1	G1>other	548	0.00000
13	gene_12	530	443	460	169	172	163	178	163	168	0.000	1.000	0.000	0.000	0.000	DEG_G1	G1>other	812	0.00002
14	gene_13	114	82	26	32	36	50	30	35	51	0.562	0.396	0.027	0.014	0.001	nonDEG		3567	0.48977
15	gene_14	12	8	14	6	8	9	8	11	6	0.968	0.022	0.007	0.003	0.000	nonDEG		7891	1.00000
16	gene_15	386	305	323	123	153	103	74	106	93	0.008	0.977	0.000	0.004	0.010	DEG_G1	G1>other	653	0.00000
17	gene_16	1	2	8	1	2	2	1	0	1	0.801	0.145	0.020	0.034	0.001	nonDEG		4063	0.61420

Osabe法の実践58

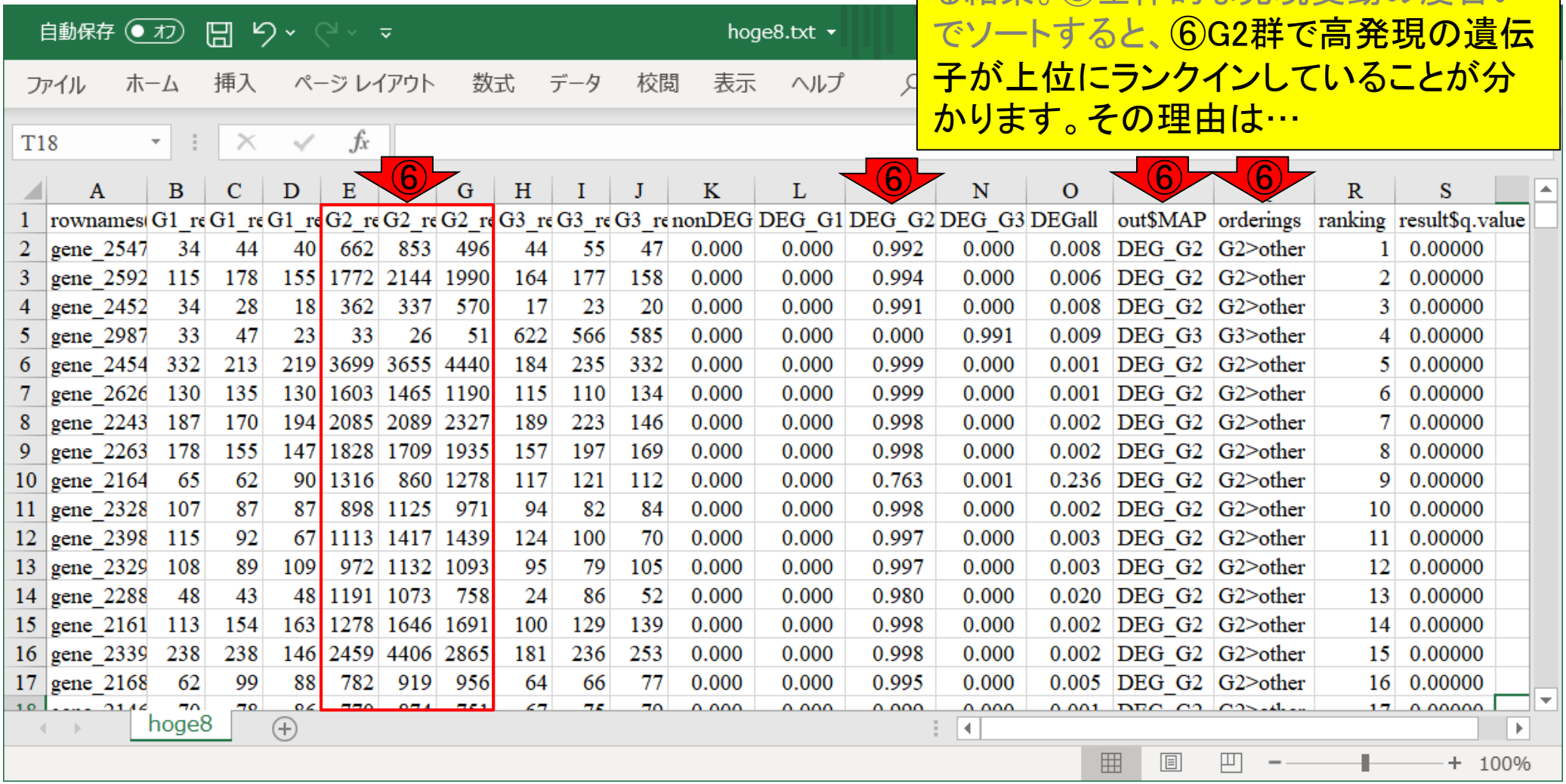
①出力ファイルであるhoge8.txtをExcelで開いたところ。②入力データ。③「TCC正規化+baySeq」で得られた発現パターン分類結果。④従来法のTCCで得られる結果。⑤全体的な発現変動の度合いでソートすると…



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	rownames(	G1_re	G1_re	G1_re	G2_re	G2_re	G2_re	G3_re	G3_re	G3_re	nonDEG	DEG_G1	DEG_G2	DEG_G3	DEGall	out$MAP	orderings	ranking	result$q.value
2	gene_2547	34	44	40	662	853	496	44	55	47	0.000	0.000	0.992	0.000	0.008	DEG_G2	G2>other	1	0.00000
3	gene_2592	115	178	155	1772	2144	1990	164	177	158	0.000	0.000	0.994	0.000	0.006	DEG_G2	G2>other	2	0.00000
4	gene_2452	34	28	18	362	337	570	17	23	20	0.000	0.000	0.991	0.000	0.008	DEG_G2	G2>other	3	0.00000
5	gene_2987	33	47	23	33	26	51	622	566	585	0.000	0.000	0.000	0.991	0.009	DEG_G3	G3>other	4	0.00000
6	gene_2454	332	213	219	3699	3655	4440	184	235	332	0.000	0.000	0.999	0.000	0.001	DEG_G2	G2>other	5	0.00000
7	gene_2626	130	135	130	1603	1465	1190	115	110	134	0.000	0.000	0.999	0.000	0.001	DEG_G2	G2>other	6	0.00000
8	gene_2243	187	170	194	2085	2089	2327	189	223	146	0.000	0.000	0.998	0.000	0.002	DEG_G2	G2>other	7	0.00000
9	gene_2263	178	155	147	1828	1709	1935	157	197	169	0.000	0.000	0.998	0.000	0.002	DEG_G2	G2>other	8	0.00000
10	gene_2164	65	62	90	1316	860	1278	117	121	112	0.000	0.000	0.763	0.001	0.236	DEG_G2	G2>other	9	0.00000
11	gene_2328	107	87	87	898	1125	971	94	82	84	0.000	0.000	0.998	0.000	0.002	DEG_G2	G2>other	10	0.00000
12	gene_2398	115	92	67	1113	1417	1439	124	100	70	0.000	0.000	0.997	0.000	0.003	DEG_G2	G2>other	11	0.00000
13	gene_2329	108	89	109	972	1132	1093	95	79	105	0.000	0.000	0.997	0.000	0.003	DEG_G2	G2>other	12	0.00000
14	gene_2288	48	43	48	1191	1073	758	24	86	52	0.000	0.000	0.980	0.000	0.020	DEG_G2	G2>other	13	0.00000
15	gene_2161	113	154	163	1278	1646	1691	100	129	139	0.000	0.000	0.998	0.000	0.002	DEG_G2	G2>other	14	0.00000
16	gene_2339	238	238	146	2459	4406	2865	181	236	253	0.000	0.000	0.998	0.000	0.002	DEG_G2	G2>other	15	0.00000
17	gene_2168	62	99	88	782	919	956	64	66	77	0.000	0.000	0.995	0.000	0.005	DEG_G2	G2>other	16	0.00000

Osabe法の実践59

①出力ファイルであるhoge8.txtをExcelで開いたところ。②入力データ。③「TCC正規化+baySeq」で得られた発現パターン分類結果。④従来法のTCCで得られる結果。⑤全体的な発現変動の度合いでソートすると、⑥G2群で高発現の遺伝子が上位にランクインしていることが分かります。その理由は…



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	rownames(	G1_re	G1_re	G1_re	G2_re	G2_re	G2_re	G3_re	G3_re	G3_re	nonDEG	DEG_G1	DEG_G2	DEG_G3	DEGall	out$MAP	orderings	ranking	result$q.value
2	gene_2547	34	44	40	662	853	496	44	55	47	0.000	0.000	0.992	0.000	0.008	DEG_G2	G2>other	1	0.00000
3	gene_2592	115	178	155	1772	2144	1990	164	177	158	0.000	0.000	0.994	0.000	0.006	DEG_G2	G2>other	2	0.00000
4	gene_2452	34	28	18	362	337	570	17	23	20	0.000	0.000	0.991	0.000	0.008	DEG_G2	G2>other	3	0.00000
5	gene_2987	33	47	23	33	26	51	622	566	585	0.000	0.000	0.000	0.991	0.009	DEG_G3	G3>other	4	0.00000
6	gene_2454	332	213	219	3699	3655	4440	184	235	332	0.000	0.000	0.999	0.000	0.001	DEG_G2	G2>other	5	0.00000
7	gene_2626	130	135	130	1603	1465	1190	115	110	134	0.000	0.000	0.999	0.000	0.001	DEG_G2	G2>other	6	0.00000
8	gene_2243	187	170	194	2085	2089	2327	189	223	146	0.000	0.000	0.998	0.000	0.002	DEG_G2	G2>other	7	0.00000
9	gene_2263	178	155	147	1828	1709	1935	157	197	169	0.000	0.000	0.998	0.000	0.002	DEG_G2	G2>other	8	0.00000
10	gene_2164	65	62	90	1316	860	1278	117	121	112	0.000	0.000	0.763	0.001	0.236	DEG_G2	G2>other	9	0.00000
11	gene_2328	107	87	87	898	1125	971	94	82	84	0.000	0.000	0.998	0.000	0.002	DEG_G2	G2>other	10	0.00000
12	gene_2398	115	92	67	1113	1417	1439	124	100	70	0.000	0.000	0.997	0.000	0.003	DEG_G2	G2>other	11	0.00000
13	gene_2329	108	89	109	972	1132	1093	95	79	105	0.000	0.000	0.997	0.000	0.003	DEG_G2	G2>other	12	0.00000
14	gene_2288	48	43	48	1191	1073	758	24	86	52	0.000	0.000	0.980	0.000	0.020	DEG_G2	G2>other	13	0.00000
15	gene_2161	113	154	163	1278	1646	1691	100	129	139	0.000	0.000	0.998	0.000	0.002	DEG_G2	G2>other	14	0.00000
16	gene_2339	238	238	146	2459	4406	2865	181	236	253	0.000	0.000	0.998	0.000	0.002	DEG_G2	G2>other	15	0.00000
17	gene_2168	62	99	88	782	919	956	64	66	77	0.000	0.000	0.995	0.000	0.005	DEG_G2	G2>other	16	0.00000

Osabe法の実践60

(Rで)塩基配列解析

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

15. TCCパッケージ中のBiological replicatesを模倣したシミュレーションデータ(G1群
群3サンプル)です。 10,000 genes×9 samplesの「複製あり」タグカウントデー
「G1_rep1, G1_rep2, G1_rep3, G2_rep1, G2_rep2, G2_rep3, G3_rep1, G3_
らなります。 全10,000遺伝子中の最初の3,000個(gene_1～gene_3000まで)が
DEGsの内訳：(1)最初の70%分(gene_1～gene_2100)がG1群で3倍高発現、(2)
gene_2700)がG2群で10倍高発現、(3)残りの10%分(gene_2701～gene_3000)
取得可能です。

```
out_f <- "data_hypodata_3vs3vs3.txt"    #出力ファイル名を指定してout_f
param_replicates <- c(3, 3, 3)          #G1, G2, G3群のサンプル数をそ
param_Ngene <- 1                        #全遺伝子数を指定
param_PDEG <- 0.3                       #発現変動遺伝子の割合を指定
param_FC <- c(3, 10, 6)                 #G1, G2, G3群の発現変動の度合
param_DEGassign <- c(0.7, 0.2, 0.1)     #DEGのうちG1, G2, G3群で高発現なものの割合をそれぞれ指定

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み

#シミュレーションデータの作成
set.seed(1000)                          #おまじない(同じ乱数になるようにするため)
tcc <- simulateReadCounts(Ngene=param_Ngene,#シミュレーションデータの作成
            PDEG=param_PDEG,            #シミュレーションデータの作成
            DEG.assign=param_DEGassign,#シミュレーションデータの作成
            DEG.foldchange=param_FC,    #シミュレーションデータの作成
            replicates=param_replicates)#シミュレーションデータの作成
plotFCPseudocolor(tcc)                  #シミュレーション条件のpseudo-colorイメージを描画
```

⑧

⑦

トップページへ

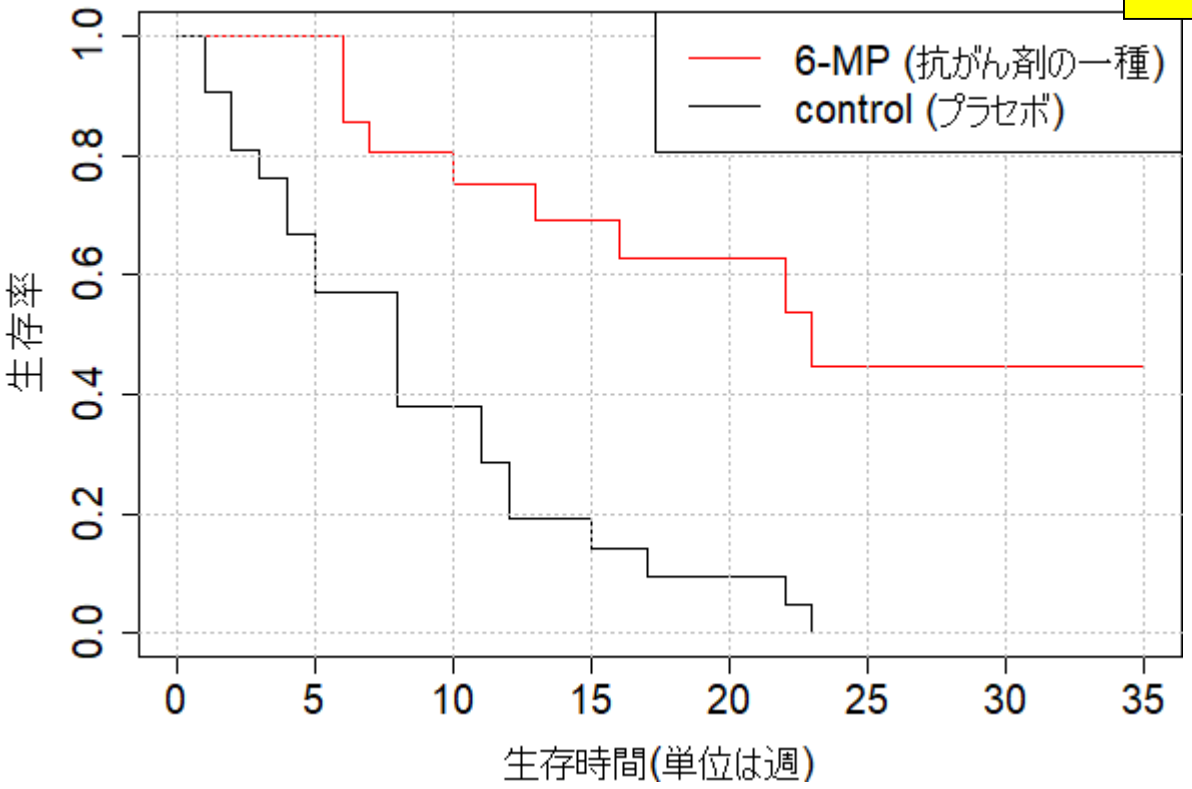
①出力ファイルであるhoge8.txtをExcelで開いたところ。②入力データ。③「TCC正規化+baySeq」で得られた発現パターン分類結果。④従来法のTCCで得られる結果。⑤全体的な発現変動の度合いでソートすると、⑥G2群で高発現の遺伝子が上位にランクインしていることが分かります。その理由は、⑦シミュレーションデータ作成時に、⑧G2群の発現変動の度合いを最も高くしていたからです。任意の条件でシミュレーションデータを生成できると、簡単な性能評価もできてよいですよ。

Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

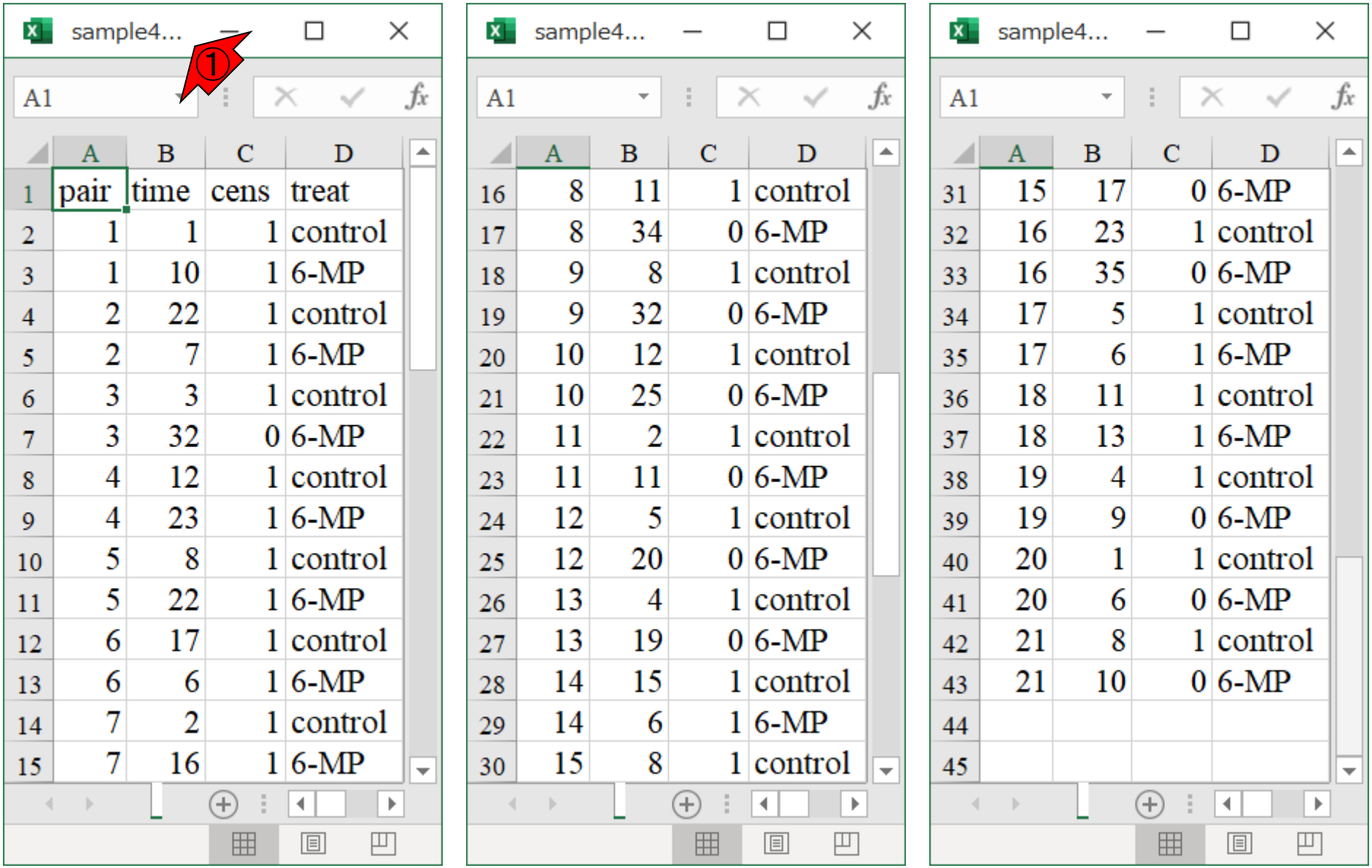
生存曲線の描画1

生存曲線の描画を行います。これは、白血病患者への6-MP（6-メルカプトプリン）投与群21例とプラセボ投与群21例の生存率に違いがあることを示した有名?!なgehanデータ（ファイル名：sample48.txt）と呼ばれるものです。



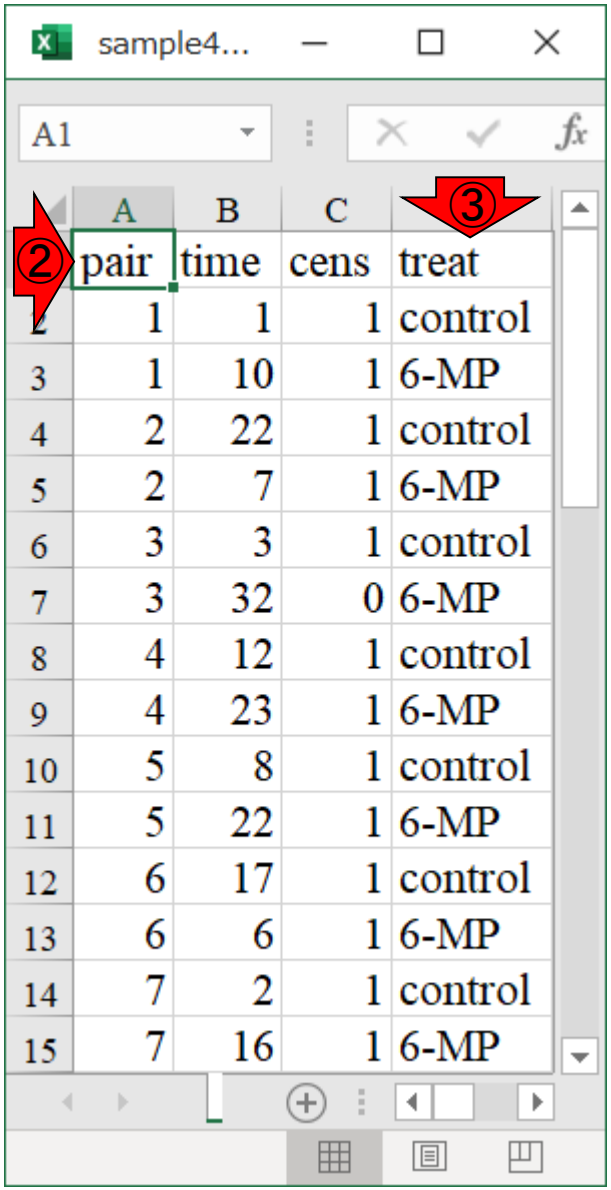
①sample48.txtをExcelで眺めた結果。

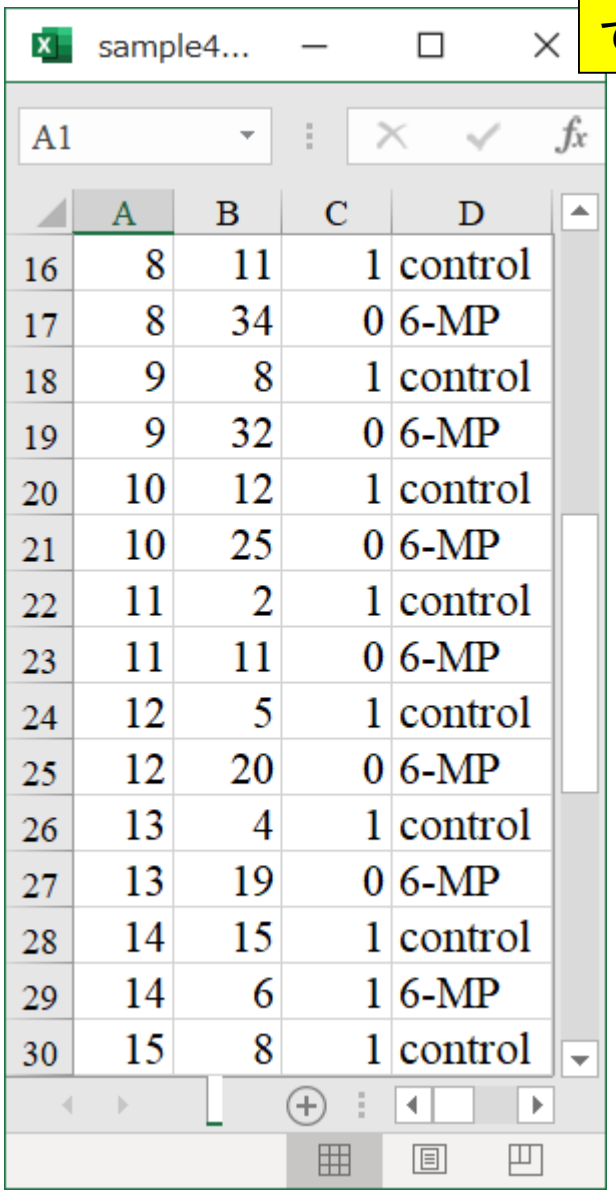
生存曲線の描画2

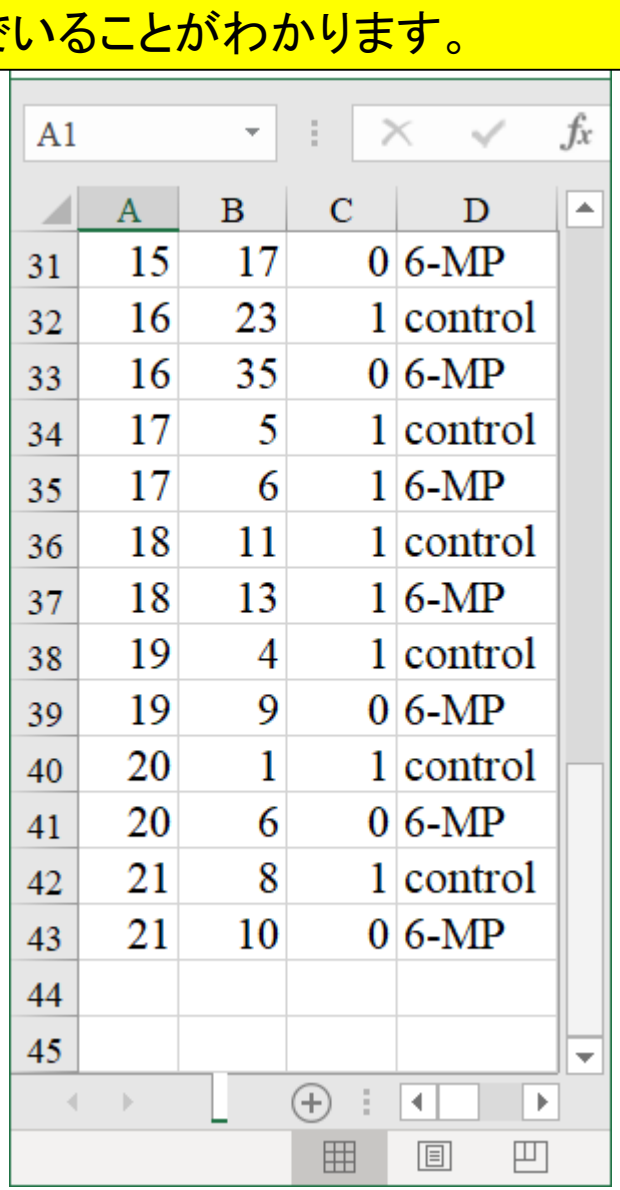


生存曲線の描画3

①sample48.txtをExcelで眺めた結果。②ヘッダー行を含んでおり、③treat列が抗がん剤投与の有無に関する情報を含んでいることがわかります。







生存曲線の描画4

①sample48.txtをExcelで眺めた結果。②ヘッダー行を含んでおり、③treat列が抗がん剤投与の有無に関する情報を含んでいることがわかります。④cens列が打ち切り(censoring)に関する情報であることが列名から感覚的にわかる。

sample4...

A1

	A	B	④ cens	③ treat
1	pair	time	cens	treat
2	1	1	1	control
3	1	10	1	6-MP
4	2	22	1	control
5	2	7	1	6-MP
6	3	3	1	control
7	3	32	0	6-MP
8	4	12	1	control
9	4	23	1	6-MP
10	5	8	1	control
11	5	22	1	6-MP
12	6	17	1	control
13	6	6	1	6-MP
14	7	2	1	control
15	7	16	1	6-MP

sample4...

A1

	A	B	C	D
16	8	11	1	control
17	8	34	0	6-MP
18	9	8	1	control
19	9	32	0	6-MP
20	10	12	1	control
21	10	25	0	6-MP
22	11	2	1	control
23	11	11	0	6-MP
24	12	5	1	control
25	12	20	0	6-MP
26	13	4	1	control
27	13	19	0	6-MP
28	14	15	1	control
29	14	6	1	6-MP
30	15	8	1	control

	A	B	C	D
31	15	17	0	6-MP
32	16	23	1	control
33	16	35	0	6-MP
34	17	5	1	control
35	17	6	1	6-MP
36	18	11	1	control
37	18	13	1	6-MP
38	19	4	1	control
39	19	9	0	6-MP
40	20	1	1	control
41	20	6	0	6-MP
42	21	8	1	control
43	21	10	0	6-MP
44				
45				

生存曲線の描画5

④cens列の数値は0 or 1。どちらが打ち切りあり or なしなのか一見よくわからないが、0が打ち切りあり、1が打ち切りなしです。

sample4...

A1

	A	B	④ C	D
1	pair	time	cens	treat
2	1	1	1	control
3	1	10	1	6-MP
4	2	22	1	control
5	2	7	1	6-MP
6	3	3	1	control
7	3	32	0	6-MP
8	4	12	1	control
9	4	23	1	6-MP
10	5	8	1	control
11	5	22	1	6-MP
12	6	17	1	control
13	6	6	1	6-MP
14	7	2	1	control
15	7	16	1	6-MP

sample4...

A1

	A	B	C	D
16	8	11	1	control
17	8	34	0	6-MP
18	9	8	1	control
19	9	32	0	6-MP
20	10	12	1	control
21	10	25	0	6-MP
22	11	2	1	control
23	11	11	0	6-MP
24	12	5	1	control
25	12	20	0	6-MP
26	13	4	1	control
27	13	19	0	6-MP
28	14	15	1	control
29	14	6	1	6-MP
30	15	8	1	control

A1

	A	B	C	D
31	15	17	0	6-MP
32	16	23	1	control
33	16	35	0	6-MP
34	17	5	1	control
35	17	6	1	6-MP
36	18	11	1	control
37	18	13	1	6-MP
38	19	4	1	control
39	19	9	0	6-MP
40	20	1	1	control
41	20	6	0	6-MP
42	21	8	1	control
43	21	10	0	6-MP
44				
45				

生存曲線の描画6

④cens列の数値は0 or 1。どちらが打ち切りあり or なしなのか一見よくわからないが、0が打ち切りあり、1が打ち切りなしです。公共データを取り扱う場合に、説明が不十分なこともある。1つの判断基準は、⑤観測期間最長（35週）のデータの打ち切り情報。生存しているヒトは打ち切りアリなので、0がアリだと判断。

sample4...

A1

	A	B	④ cens	D
1	pair	time	cens	treat
2	1	1	1	control
3	1	10	1	6-MP
4	2	22	1	control
5	2	7	1	6-MP
6	3	3	1	control
7	3	32	0	6-MP
8	4	12	1	control
9	4	23	1	6-MP
10	5	8	1	control
11	5	22	1	6-MP
12	6	17	1	control
13	6	6	1	6-MP
14	7	2	1	control
15	7	16	1	6-MP

sample4...

A1

	A	B	C	D
16	8	11	1	control
17	8	34	0	6-MP
18	9	8	1	control
19	9	32	0	6-MP
20	10	12	1	control
21	10	25	0	6-MP
22	11	2	1	control
23	11	11	0	6-MP
24	12	5	1	control
25	12	20	0	6-MP
26	13	4	1	control
27	13	19	0	6-MP
28	14	15	1	control
29	14	6	1	6-MP
30	15	8	1	control

32	16	23	1	control
33	16	35	0	6-MP
34	17	5	1	control
35	17	6	1	6-MP
36	18	11	1	control
37	18	13	1	6-MP
38	19	4	1	control
39	19	9	0	6-MP
40	20	1	1	control
41	20	6	0	6-MP
42	21	8	1	control
43	21	10	0	6-MP
44				
45				

⑤

生存曲線の描画7

各列の説明の続き。①treat列が抗がん剤投与の有無に関する情報。②cens列が打ち切り情報（0がアリ、1がナシ）。③が生存時間情報。④はペアに関する情報だが、おそらく本質的ではない。観測を実施する病院や病棟によっても傾向が違うかもしれないので、そのあたりを評価しようと思えばできる状況にしているだけだろう、程度。

sample4...

A1

④ ③ ② ①

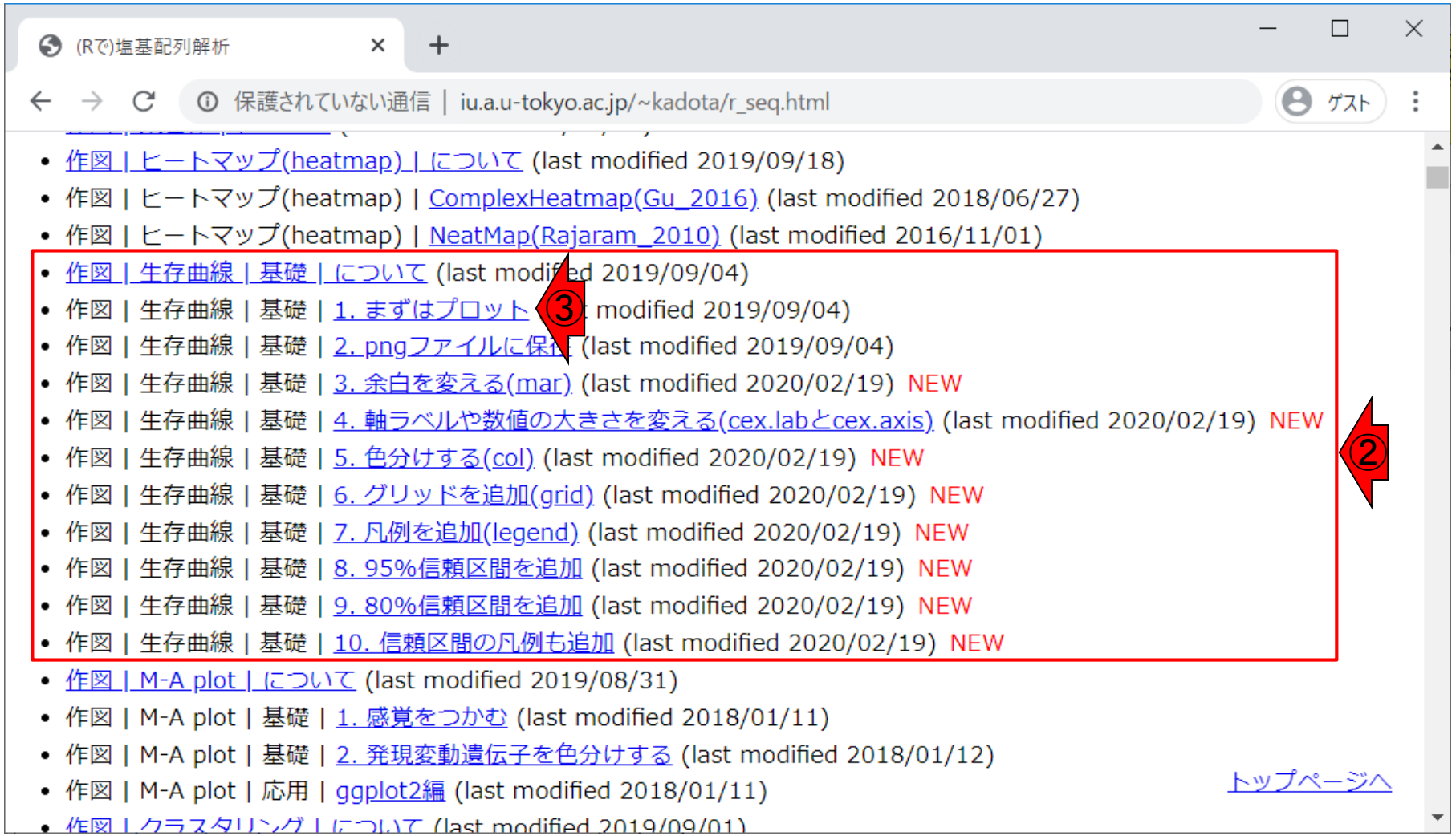
	pair	time	cens	treat
1	1	1	1	control
2	1	10	1	6-MP
3	2	22	1	control
4	2	7	1	6-MP
5	3	3	1	control
6	3	32	0	6-MP
7	4	12	1	control
8	4	23	1	6-MP
9	5	8	1	control
10	5	22	1	6-MP
11	6	17	1	control
12	6	6	1	6-MP
13	7	2	1	control
14	7	16	1	6-MP
15	8	11	1	control
16	8	34	0	6-MP
17	9	8	1	control
18	9	32	0	6-MP
19	10	12	1	control
20	10	25	0	6-MP
21	11	2	1	control
22	11	11	0	6-MP
23	12	5	1	control
24	12	20	0	6-MP
25	13	4	1	control
26	13	19	0	6-MP
27	14	15	1	control
28	14	6	1	6-MP
29	15	8	1	control
30	16	35	0	6-MP
31	17	5	1	control
32	17	6	1	6-MP
33	18	11	1	control
34	18	13	1	6-MP
35	19	4	1	control
36	19	9	0	6-MP
37	20	1	1	control
38	20	6	0	6-MP
39	21	8	1	control
40	21	10	0	6-MP
41				
42				
43				
44				
45				

生存曲線の描画8


(Rで)塩基配列解析
保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html
ゲスト
(Rで)塩基配列解析
①
(last modified 2020/02/19, 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 (Windows2019.10.09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)
日本乳酸菌学会誌のNGS関連連載の第14回分原稿PDFを公開しました。ウェブ資料も公開しました。(2019/12/23)
「RNA-Seqデータ解析　WETラボのための鉄板レシピ（編：坊農秀雅）」が出版されています。(2019/12/23)
TCC-GUI (Su et al., BMC Res. Notes, 2019) の解説動画が統合TVで公開されました。DBCLSの小野さんはじめ関係者の皆様のご尽力に深謝m(_ _)m(2019/11/08)
インストール | についての推奨手順をとりあえずWindows版(R_install_win.pdf) のみですがアップデートし、RStudioを利用するやり方に変更しました。(2019/10/09)
「インストール | R本体 | 最新版 | Win用」の項目名を 「インストール | R本体とRStudio | 最新版 | Win用」トップページへ
た。 Mac用についても同様です。(2019/10/08)

生存曲線の描画9

①(Rで)塩基配列解析の、②ここの例題を順にやっていきます。③をクリック。



生存曲線の描画10

①(Rで)塩基配列解析の、②ここの例題を順にやっていきます。③をクリック。こんな感じになります。

作図 | 生存曲線 | 基礎 | 1. まずはプロット

survivalパッケージを用いた生存曲線（カプランマイヤー曲線） の描画を行うやり方を示します。
「ファイル」−「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. サンプルデータ48のsample48.txtの場合：

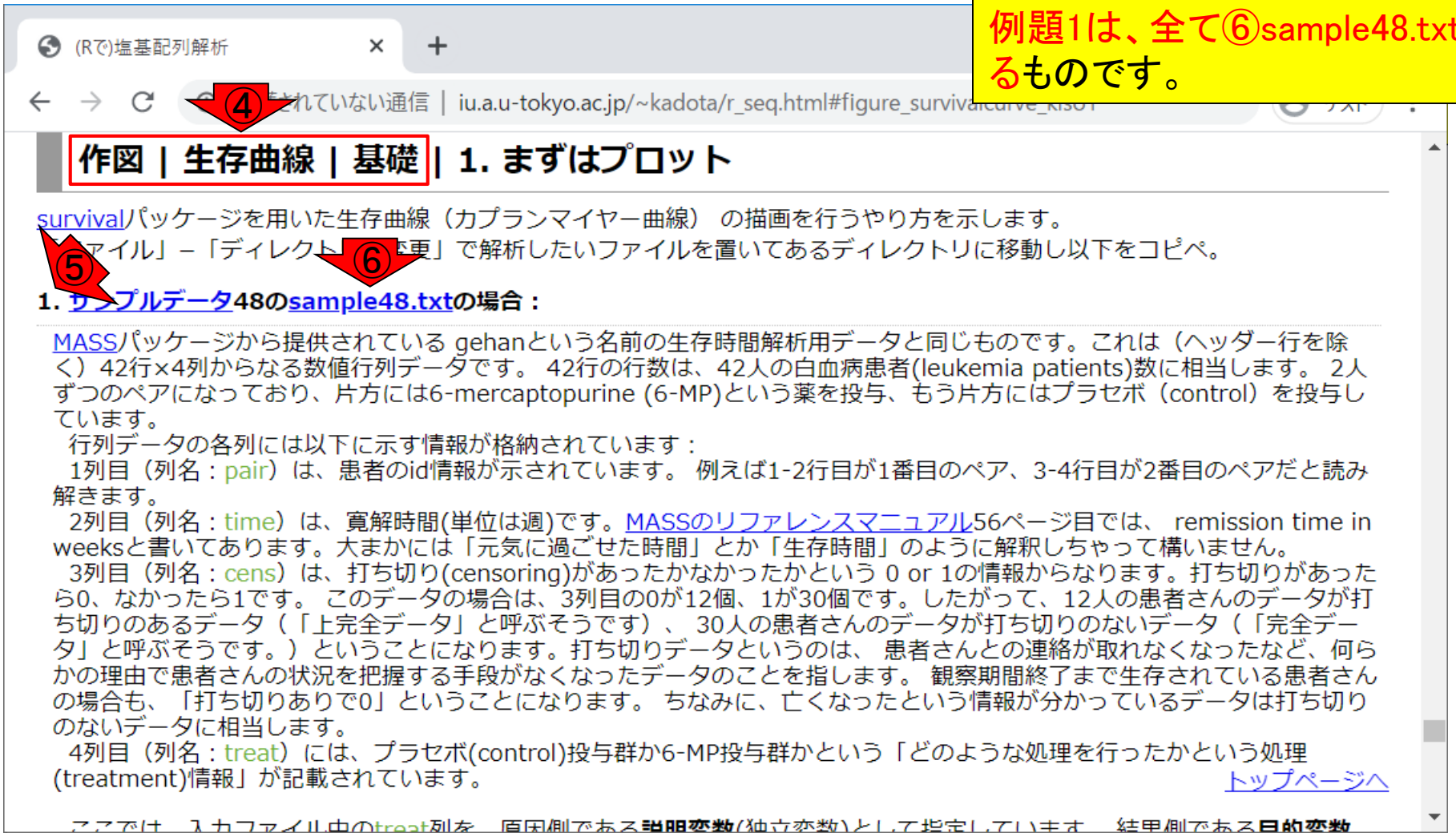
MASSパッケージから提供されている gehanという名前の生存時間解析用データと同じものです。これは（ヘッダー行を除く）42行×4列からなる数値行列データです。 42行の行数は、42人の白血病患者(leukemia patients)数に相当します。 2人ずつのペアになっており、片方には6-mercaptopurine (6-MP)という薬を投与、もう片方にはプラセボ（control）を投与しています。
行列データの各列には以下に示す情報が格納されています：
1列目（列名：pair）は、患者のid情報が示されています。 例えば1-2行目が1番目のペア、3-4行目が2番目のペアだと読み解きます。
2列目（列名：time）は、寛解時間(単位は週)です。MASSのリファレンスマニュアル56ページ目では、 remission time in weeksと書いてあります。大まかには「元気に過ごせた時間」とか「生存時間」のように解釈しちゃって構いません。
3列目（列名：cens）は、打ち切り(censoring)があったかなかったかという 0 or 1の情報からなります。打ち切りがあったら0、なかったら1です。 このデータの場合は、3列目の0が12個、1が30個です。したがって、12人の患者さんのデータが打ち切りのあるデータ（「上完全データ」と呼ぶそうです）、 30人の患者さんのデータが打ち切りのないデータ（「完全データ」と呼ぶそうです。）ということになります。打ち切りデータというのは、 患者さんとの連絡が取れなくなったなど、何らかの理由で患者さんの状況を把握する手段がなくなったデータのことを指します。 観察期間終了まで生存されている患者さんの場合も、「打ち切りありで0」ということになります。 ちなみに、亡くなったという情報が分かっているデータは打ち切りのないデータに相当します。
4列目（列名：treat）には、プラセボ(control)投与群か6-MP投与群かという「どのような処理を行ったかという処理(treatment)情報」が記載されています。

トップページへ

ここでは、入力ファイル中のtreat列を、原因側である**説明変数**(独立変数)として指定しています。結果側である**目的変数**

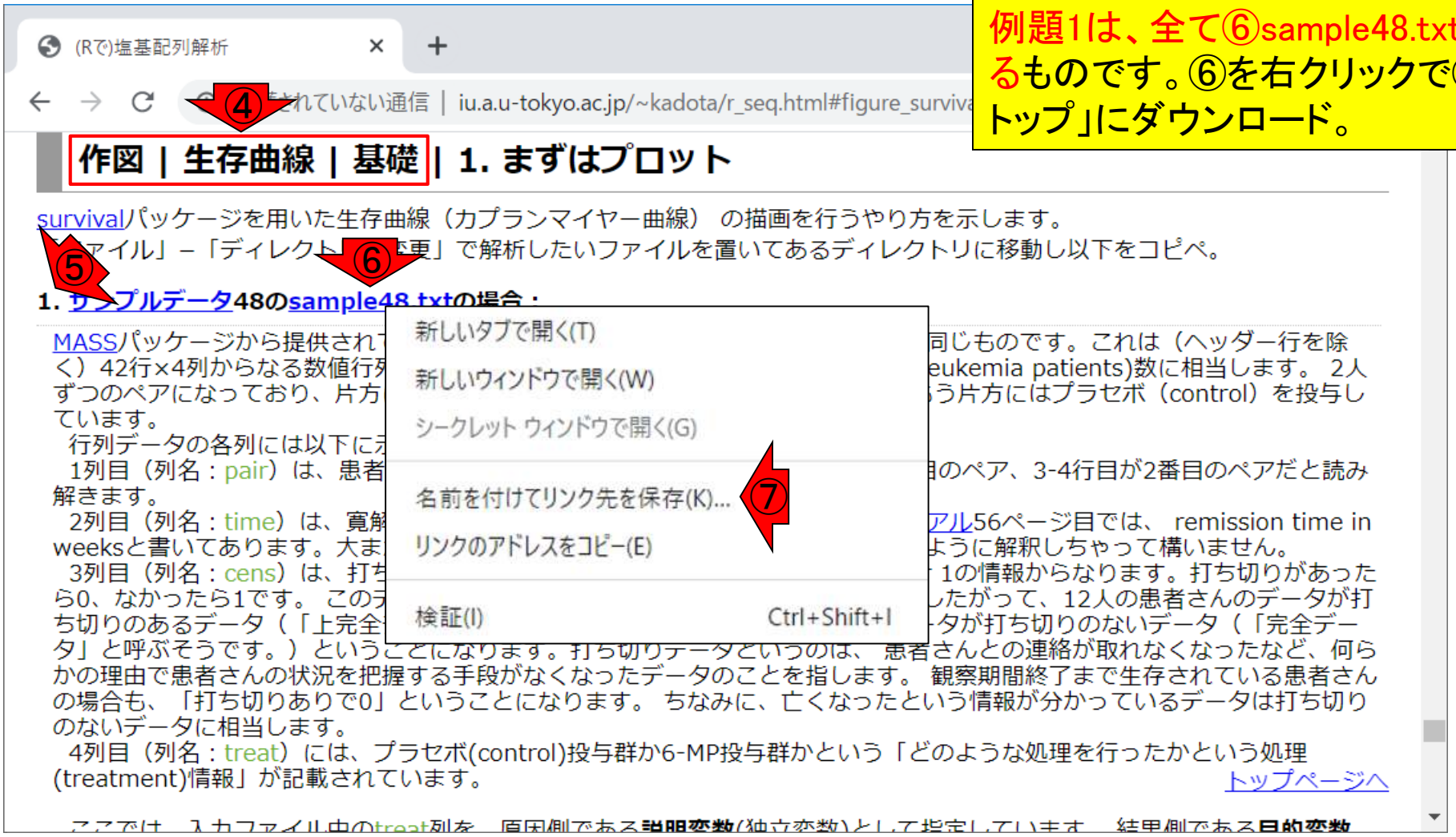
生存曲線の描画11

①(Rで)塩基配列解析の、②ここの例題を順にやっていきます。③をクリック。こんな感じになります。④の項目群の、⑤例題1は、全て⑥sample48.txtを入力とするものです。



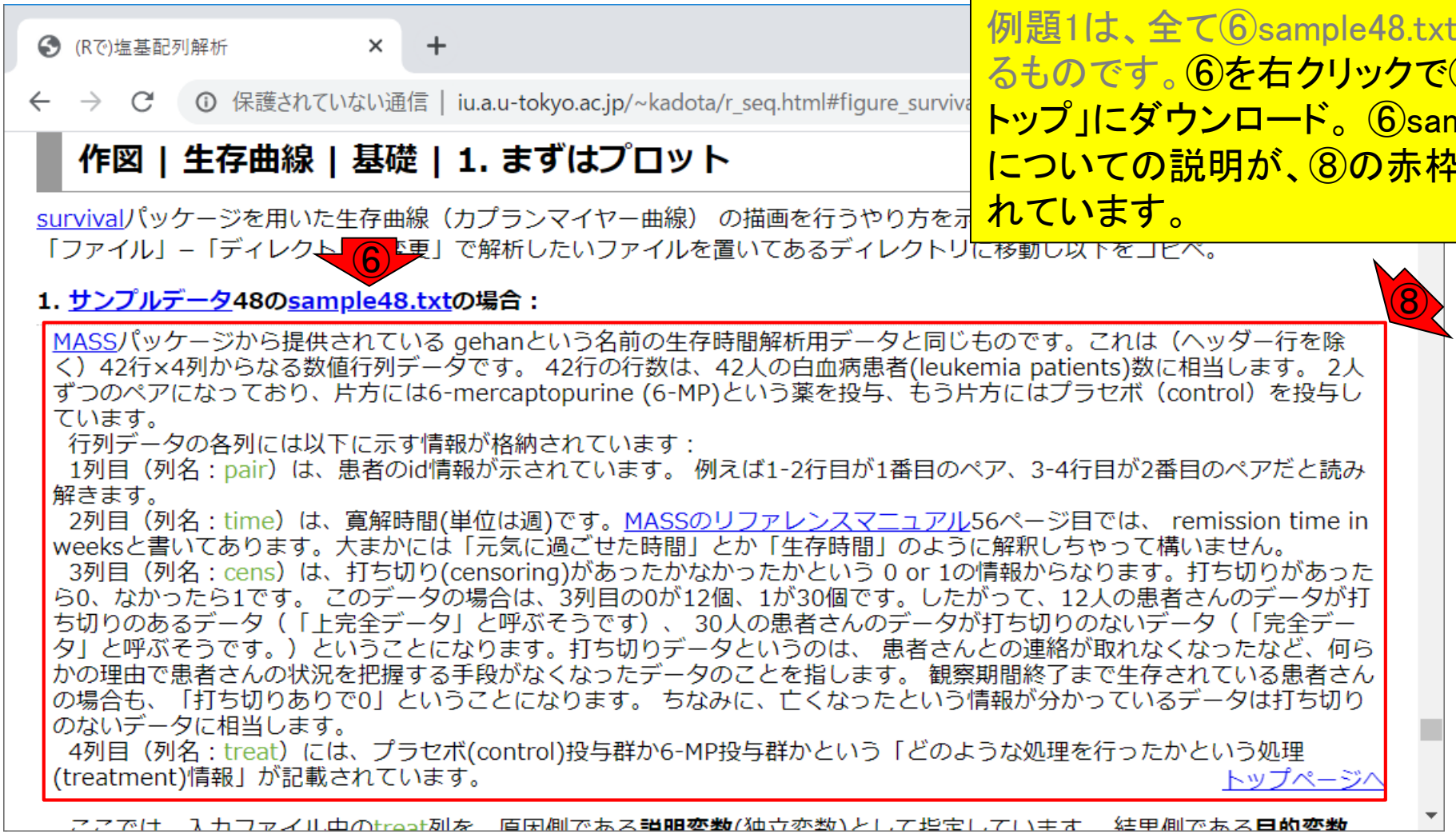
生存曲線の描画12

①（Rで）塩基配列解析の、②ここの例題を順にやっていきます。③をクリック。こんな感じになります。④の項目群の、⑤例題1は、全て⑥sample48.txtを入力とするものです。⑥を右クリックで⑦「デスクトップ」にダウンロード。



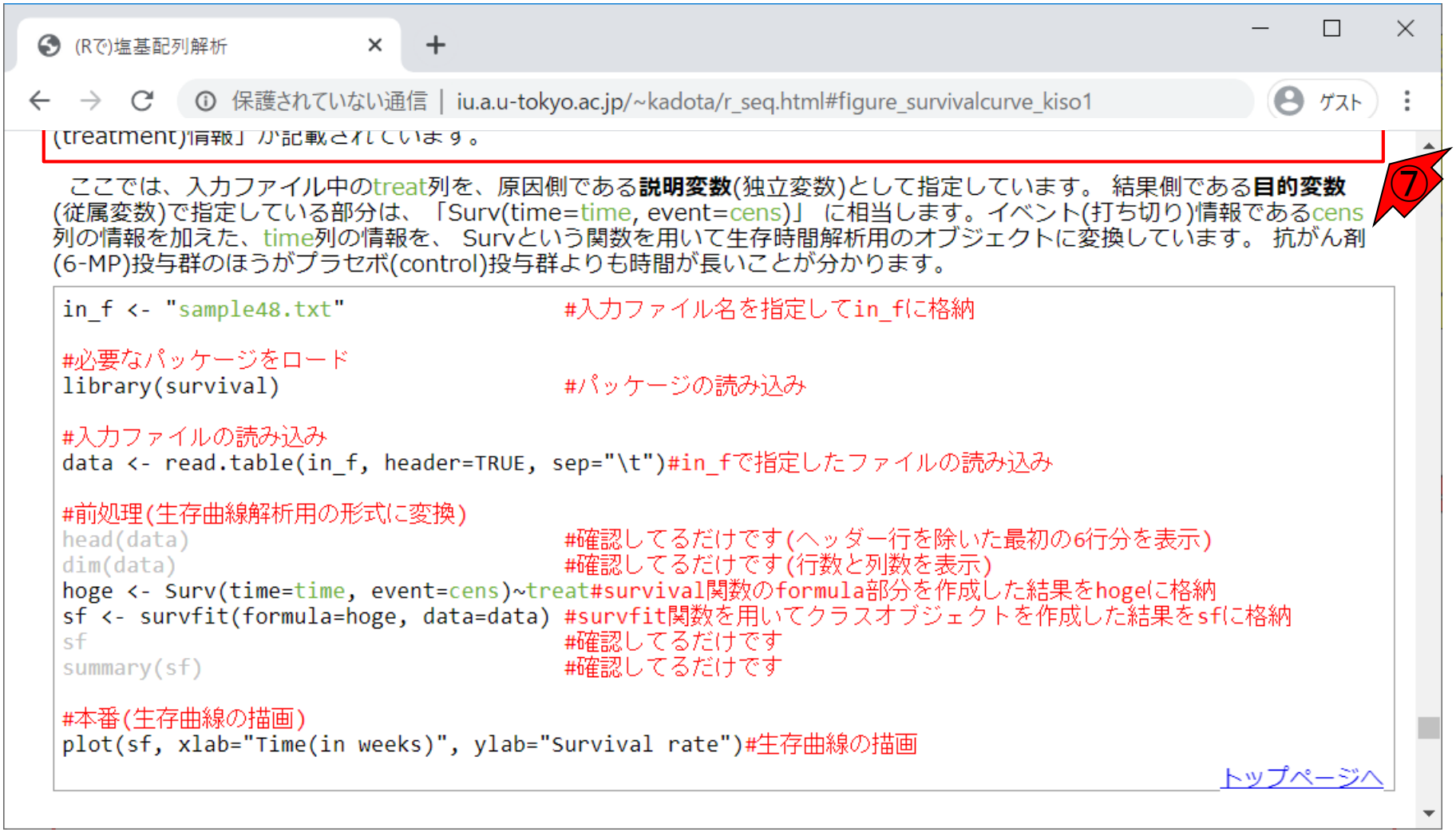
生存曲線の描画13

①（Rで）塩基配列解析の、②ここの例題を順にやっていきます。③をクリック。こんな感じになります。④の項目群の、⑤例題1は、全て⑥sample48.txtを入力とするものです。⑥を右クリックで⑦「デスクトップ」にダウンロード。⑥sample48.txtについての説明が、⑧の赤枠内でなされています。



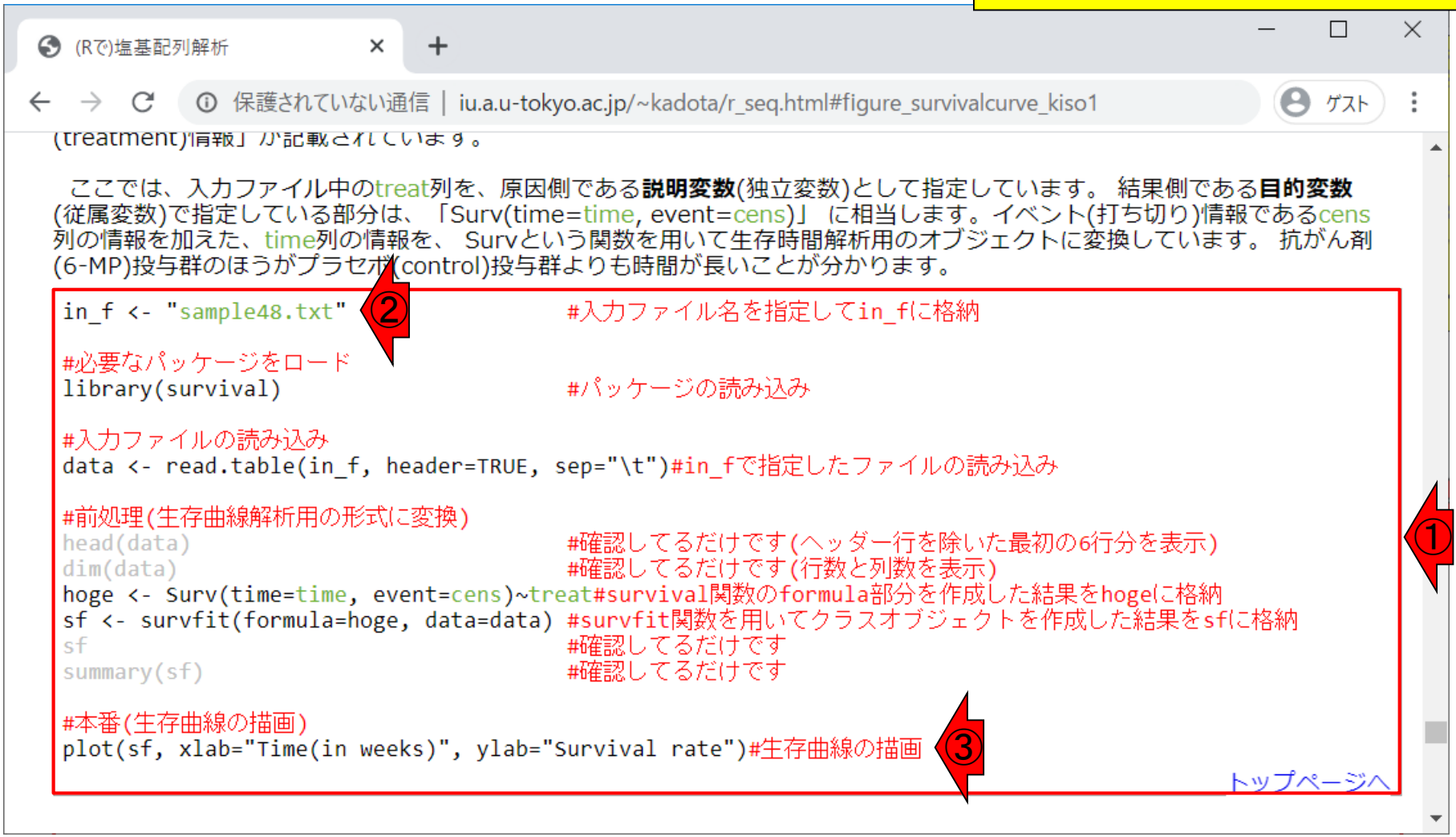
生存曲線の描画14

1ページ分ほど下部に移動（⑦の赤枠下部がギリギリ上のほうに見える程度）。



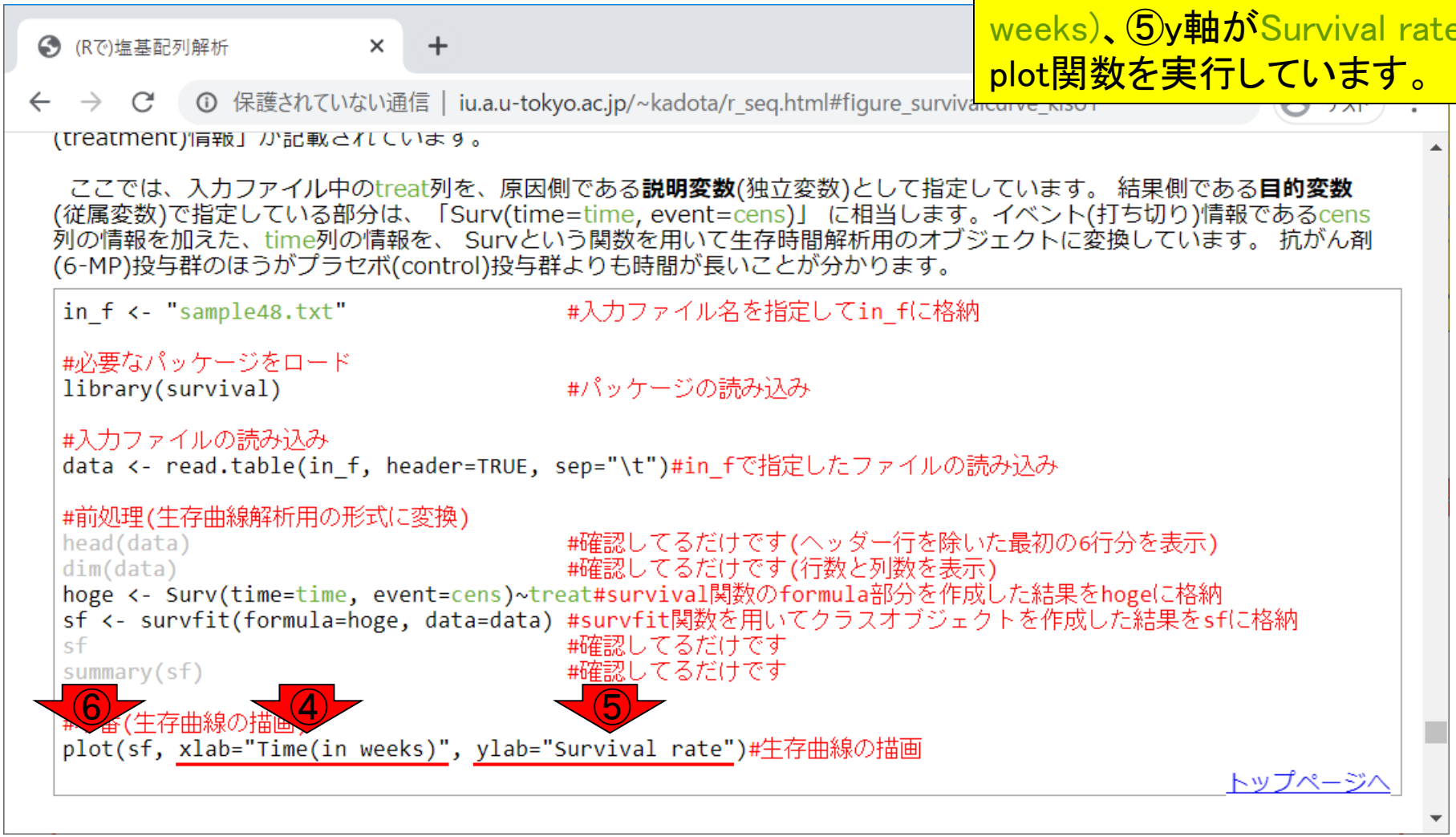
生存曲線の描画15

①の赤枠内が、②sample48.txtを入力として、③最終的に生存曲線を描画する一連のスクリプト。



生存曲線の描画16

①の赤枠内が、②sample48.txtを入力として、③最終的に生存曲線を描画する一連のスクリプト。④x軸がTime(in weeks)、⑤y軸がSurvival rateとして、⑥plot関数を実行しています。



生存曲線の描画17

①の赤枠内が、②sample48.txtを入力として、③最終的に生存曲線を描画する一連のスクリプト。④x軸がTime(in weeks)、⑤y軸がSurvival rateとして、⑥plot関数を実行しています。②入力ファイルの中身はこんな感じでした。

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.

(treatment)情報」が記載されています。

ここでは、入力ファイル中のtreat列を、原因側で[redacted]定しています。 結果側である**目的変数**(従属変数)で指定している部分は、「Surv(time=[redacted]す。イベント(打ち切り)情報であるcens列の情報を加えた、time列の情報を、 Survという[redacted]ブジェクトに変換しています。 抗がん剤(6-MP)投与群のほうがプラセボ(control)投与群よ[redacted]

②

```
in_f <- "sample48.txt"                    #[redacted]納

#必要なパッケージをロード
library(survival)                         #[redacted]

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, se[redacted]の読み込み

#前処理(生存曲線解析用の形式に変換)
head(data)                                #[redacted]除いた最初の6行分を表示)
dim(data)                                 #確認してるだけです(行数と列数を表示)
hoge <- Surv(time=time, event=cens)~treat#survival関数のformula部分を作成した結果をhogeに格納
sf <- survfit(formula=hoge, data=data) #survfit関数を用いてクラスオブジェクトを作成した結果をsfに格納
sf                                        #確認してるだけです
summary(sf)                               #確認してるだけです

#本番(生存曲線の描画)
plot(sf, xlab="Time(in weeks)", ylab="Survival rate")#生存曲線の描画
```

トップページへ

pair	time	cens	treat
1	1	1	control
1	10	1	6-MP
2	22	1	control
2	7	1	6-MP
3	3	1	control
3	32	0	6-MP
4	12	1	control
4	23	1	6-MP
5	8	1	control
5	22	1	6-MP
6	17	1	control

生存曲線の描画18

①の赤枠内が、②sample48.txtを入力として、③最終的に生存曲線を描画する一連のスクリプト。④x軸がTime(in weeks)、⑤y軸がSurvival rateとして、⑥plot関数を実行しています。②入力ファイルの中身はこんな感じでした。抗がん剤を使ったかどうかという原因側の情報が、⑦treat列。それによってどれだけ長生きできたかという結果側の情報が、主に⑧time列。この際、⑨cens列の打ち切りの有無(0 or 1)という結果の情報も入力として与えている。

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.

(treatment)情報」が記載されています。

ここでは、入力ファイル中のtreat列を、原因側で
(従属変数)で指定している部分は、「Surv(time=
列の情報を加えた、time列の情報を、 Survという
(6-MP)投与群のほうがプラセボ(control)投与群よ

⑧ ⑨ ⑦

pair	time	cens	treat
1	1	1	control
1	10	1	6-MP
2	22	1	control
2	7	1	6-MP
3	3	1	control
3	32	0	6-MP
4	12	1	control

```
in_f <- "sample48.txt"                          #入力ファイルを納

#必要なパッケージをロード
library(survival)                               #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#in_fで指定したファイルの読み込み

#前処理(生存曲線解析用の形式に変換)
head(data)                                      #確認してるだけです(ヘッダー行を除いた最初の6行分を表示)
dim(data)                                       #確認してるだけです(行数と列数を表示)
hoge <- Surv(time=time, event=cens)~treat#survival関数のformula部分を作成した結果をhogeに格納
sf <- survfit(formula=hoge, data=data) #survfit関数を用いてクラスオブジェクトを作成した結果をsfに格納
sf                                              #確認してるだけです
summary(sf)                                     #確認してるだけです

#本番(生存曲線の描画)
plot(sf, xlab="Time(in weeks)", ylab="Survival rate")#生存曲線の描画
```

⑧ ⑨ ⑦

トップページへ

生存曲線の描画19

①の赤枠内が、②sample48.txtを入力として、③最終的に生存曲線を描画する一連のスクリプト。④x軸がTime(in weeks)、⑤y軸がSurvival rateとして、⑥plot関数を実行しています。②入力ファイルの中身はこんな感じでした。抗がん剤を使ったかどうかという原因側の情報が、①treat列。それによってどれだけ長生きできたかという結果側の情報が、主に⑧time列。この際、⑨cens列の打ち切りの有無(0 or 1)という結果の情報も入力として与えている。**実用上は、どこにどの列名情報を与えるかが分かっていればOK。**

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.

(treatment)情報」が記載されています。

ここでは、入力ファイル中のtreat列を、原因側で
(従属変数)で指定している部分は、「Surv(time=
列の情報を加えた、time列の情報を、 Survという
(6-MP)投与群のほうがプラセボ(control)投与群よ

⑧ ⑨ ⑦

pair	time	cens	treat
1	1	1	control
1	10	1	6-MP
2	22	1	control
2	7	1	6-MP
3	3	1	control
3	32	0	6-MP
4	12	1	control

```
in_f <- "sample48.txt"                                  #入

#必要なパッケージをロード
library(survival)                                       #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#in_fで指定したファイルの読

#前処理(生存曲線解析用の形式に変換)
head(data)                                              #確認してるだけです(ヘッダー行を除いた最初の6行分を表示)
dim(data)                                               #確認してるだけです(行数と列数を表示)
hoge <- Surv(time=time, event=cens)~treat               #survival関数のformula部分を作成した結果をhogeに格納
sf <- survfit(formula=hoge, data=data) #survival関数を用いてクラスオブジェクトを作成した結果をsfに格納
sf                                                      #確認してるだけです
summary(sf)

#本番(生存曲線の描画)
plot(sf, xlab="Time(in weeks)", ylab="Survival rate")#生存曲線の描画
```

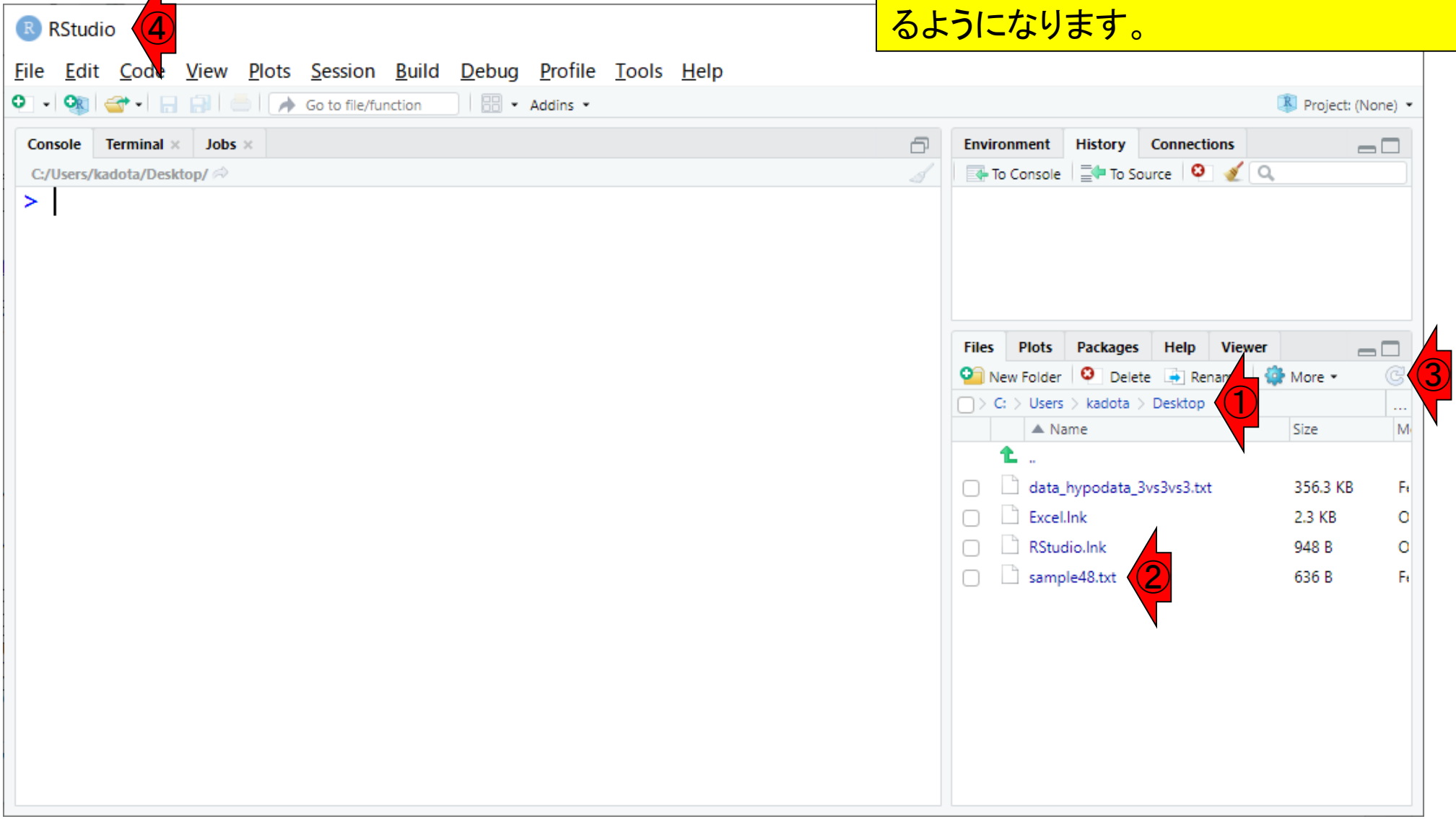
⑧ ⑨ ⑦

```
Surv(time=time, event=cens)~treat
```

トップページへ

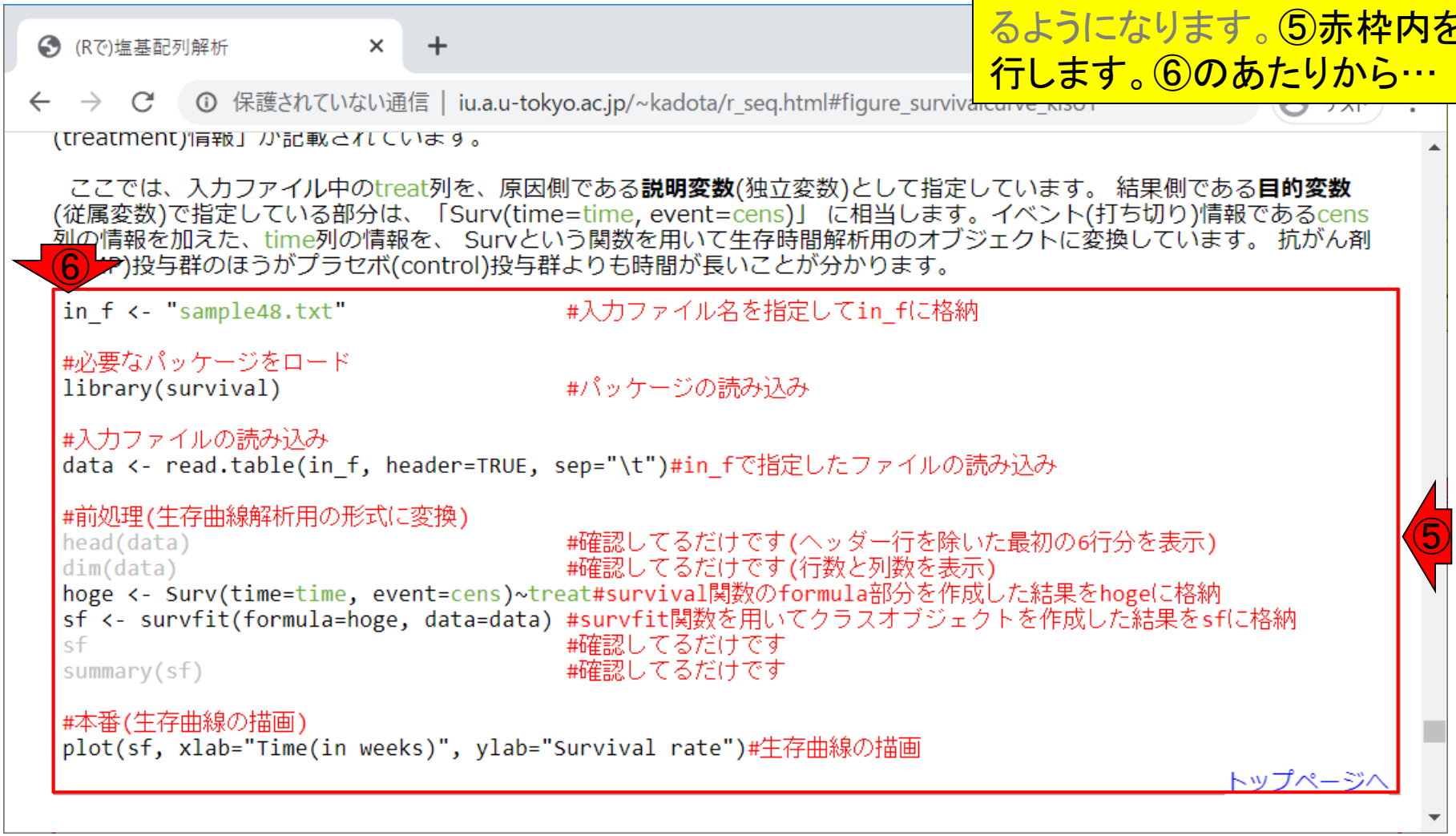
生存曲線の描画20

さきほど①デスクトップに、②sample48.txtをダウンロードしたので、③リロードすれば、④RStudio上でも見られるようになります。



生存曲線の描画21

さきほど①デスクトップに、②sample48.txtをダウンロードしたので、③リロードすれば、④RStudio上でも見られるようになります。⑤赤枠内をコピペ実行します。⑥のあたりから…



生存曲線の描画22

さきほど①デスクトップに、②sample48.txtをダウンロードしたので、③リロードすれば、④RStudio上でも見られるようになります。⑤赤枠内をコピペ実行します。⑥のあたりから、⑦のあたりまで反転させて…

(R で)塩基配列解析

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

(treatment)情報」が記載されています。

ここでは、入力ファイル中のtreat列を、原因側である**説明変数**(独立変数)として指定しています。 結果側である**目的変数**(従属変数)で指定している部分は、「Surv(time=time, event=cens)」 に相当します。イベント(打ち切り)情報であるcens列の情報を加えた、time列の情報を、 Survという関数を用いて生存時間解析用のオブジェクトに変換しています。 抗がん剤(DP)投与群のほうがプラセボ(control)投与群よりも時間が長いことが分かります。

⑥

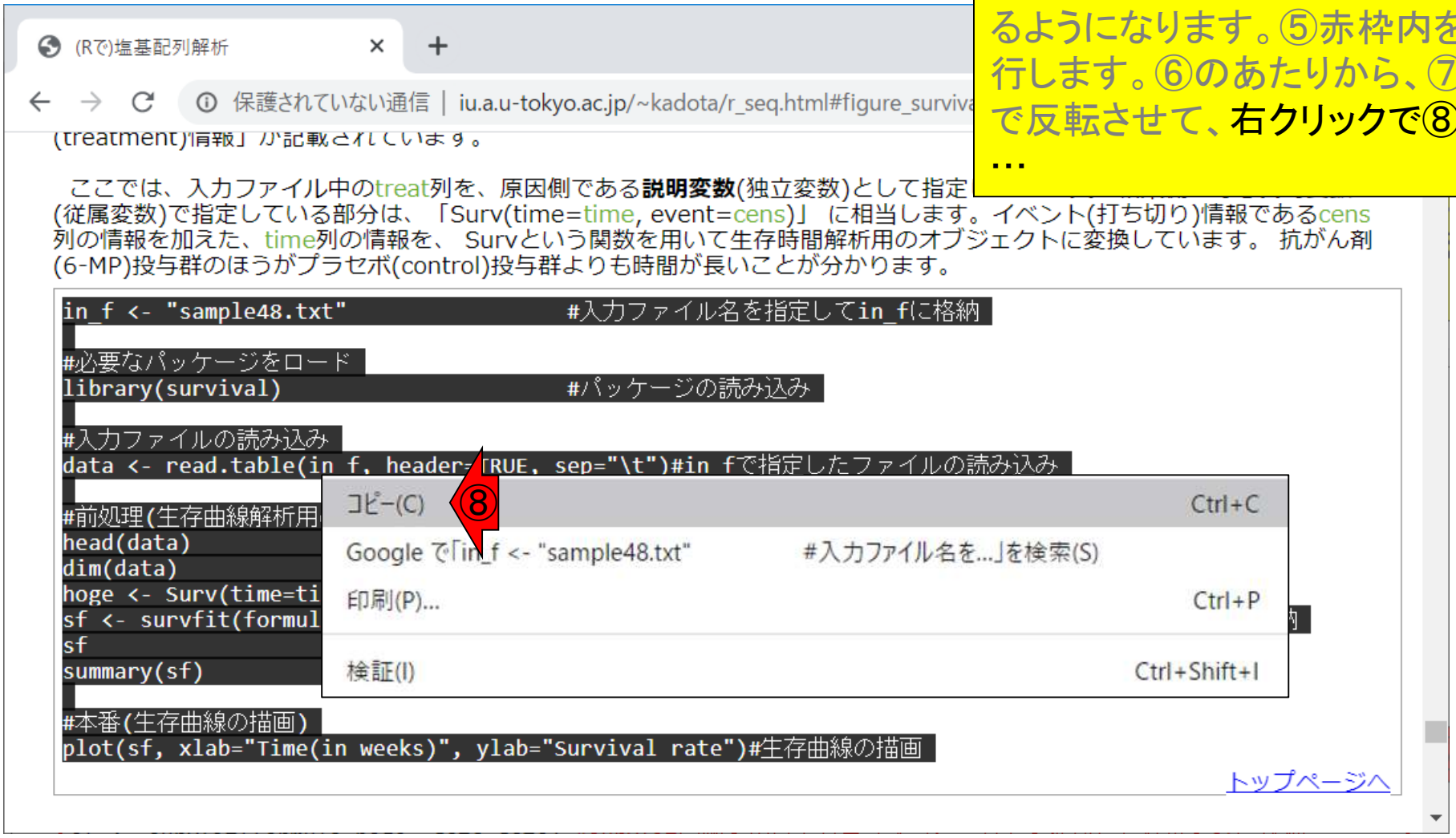
```
in_f <- "sample48.txt"                    #入力ファイル名を指定してin_fに格納

#必要なパッケージをロード
library(survival)                         #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#in_fで指定したファイルの読み込み

#前処理(生存曲線解析用の形式に変換)
head(data)                                #確認してるだけです(ヘッダー行を除いた最初の6行分を表示)
dim(data)                                 #確認してるだけです(行数と列数を表示)
hoge <- Surv(time=time, event=cens)~treat#survival関数のformula部分を作成した結果をhogeに格納
sf <- survfit(formula=hoge, data=data) #survfit関数を用いてクラスオブジェクトを作成した結果をsfに格納
sf                                        #確認してるだけです
summary(sf)                               #確認してるだけです

#本番(生存曲線の描画)
plot(sf, xlab="Time(in weeks)", ylab="Survival rate")#生存曲線の描画
```

⑦

トップページへ

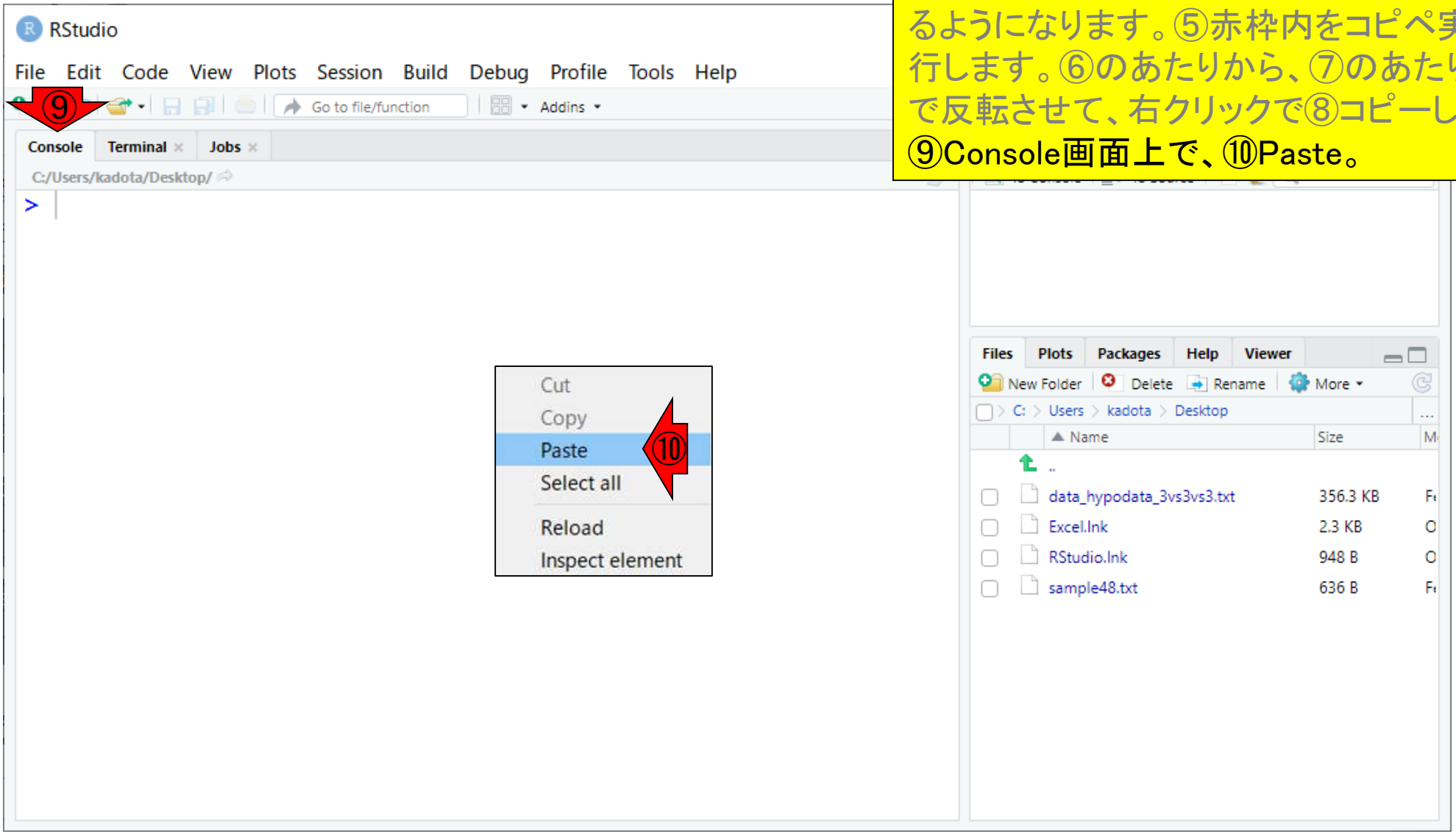
生存曲線の描画23

さきほど①デスクトップに、②sample48.txtをダウンロードしたので、③リロードすれば、④RStudio上でも見られるようになります。⑤赤枠内をコピペ実行します。⑥のあたりから、⑦のあたりまで反転させて、右クリックで⑧コピーして…



生存曲線の描画24

さきほど①デスクトップに、②sample48.txtをダウンロードしたので、③リロードすれば、④RStudio上でも見られるようになります。⑤赤枠内をコピペ実行します。⑥のあたりから、⑦のあたりまで反転させて、右クリックで⑧コピーして、⑨Console画面上で、⑩Paste。



生存曲線の描画25

さきほど①デスクトップに、②sample48.txtをダウンロードしたので、③リロードすれば、④RStudio上でも見られるようになります。⑤赤枠内をコピペ実行します。⑥のあたりから、⑦のあたりまで反転させて、右クリックで⑧コピーして、⑨Console画面上で、⑩Paste。こんな感じになります。⑪Enter。

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Console | Terminal | Jobs

C:/Users/kadota/Desktop/

```
data <- read.table(in_f, header=TRUE, sep="\t")#in_fで指定したファイルの読み込み

#前処理(生存曲線解析用の形式に変換)
head(data)                                    #確認してるだけです(ヘッダー行を除いた最初の6行分を表示)
dim(data)                                     #確認してるだけです(行数と列数を表示)
hoge <- Surv(time=time, event=cens)~treat#survival関数のformula部分を作成した結果をhogeに格納
sf <- survfit(formula=hoge, data=data) #survfit関数を用いてクラスオブジェクトを作成した結果をsfに格納
sf                                            #確認してるだけです
summary(sf)                                   #確認してるだけです

#本番(生存曲線の描画)
plot(sf, xlab="Time(in weeks)", ylab="Survival rate")#生存曲線の描画
```

Files | Plots | Packages

New Folder | Delete

C: > Users > kadota

Name

..

data_hypodata_3

Excel.lnk

RStudio.lnk

sample48.txt

生存曲線の描画26

さきほど①デスクトップに、②sample48.txtをダウンロードしたので、③リロードすれば、④RStudio上でも見られるようになります。⑤赤枠内をコピペ実行します。⑥のあたりから、⑦のあたりまで反転させて、右クリックで⑧コピーして、⑨Console画面上で、⑩Paste。こんな感じになります。⑪Enter。**実行結果。**

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Console | Terminal | Jobs

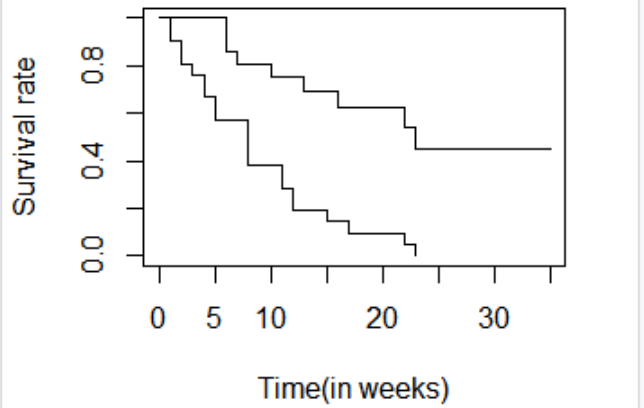
C:/Users/kadota/Desktop/

```
   23      1        1    0.0000      NaN              NA
 upper 95% CI
        1.000
        0.996
        0.968
        0.902
        0.828
        0.657
        0.562
        0.460
        0.407
        0.356
        0.322
           NA

>
> #本番(生存曲線の描画)
> plot(sf, xlab="Time(in weeks)", ylab="Survival rate")#生存曲線の描画
>
>
```

```
summary(sf) #確認してるだけです
#本番(生存曲線の描画)
plot(sf, xlab="Time(in weeks…
```

Files | Plots | Packages | Help | Viewer

Zoom | Export



生存曲線の描画27

さきほどまでは①Filesタブがアクティブな状態でしたが、②Plotsタブに自動で切り替わり、③生存曲線が描画されていることがわかります。

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Console | Terminal | Jobs

C:/Users/kadota/Desktop/

```
   23      1        1    0.0000      NaN              NA
 upper 95% CI
        1.000
        0.996
        0.968
        0.902
        0.828
        0.657
        0.562
        0.460
        0.407
        0.356
        0.322
           NA

>
> #本番(生存曲線の描画)
> plot(sf, xlab="Time(in weeks)", ylab="Survival rate")#生存曲線の描画
>
>
```

Environment | History | Connections

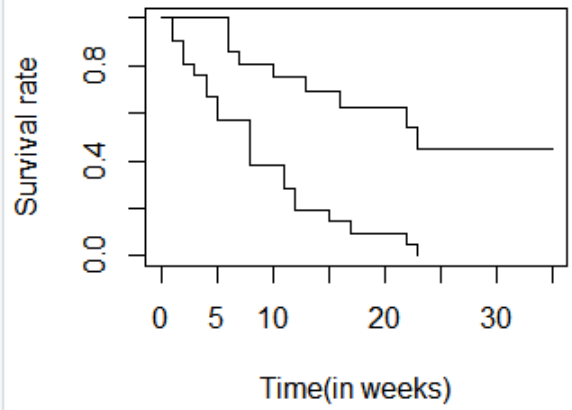
To Console | To Source

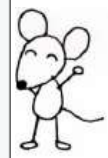
```
sf #確認してるだけです
summary(sf) #確認してるだけです
#本番(生存曲線の描画)
plot(sf, xlab="Time(in weeks...
```

② Files | Plots | Packages | Help | Viewer

① Zoom | Export

③





生存曲線の描画28

さきほどまでは①Filesタブがアクティブな状態でしたが、②Plotsタブに自動で切り替わり、③生存曲線が描画されていることがわかります。ただ、②Plotsタブ上で③のように表示されているだけではうれしくないので、pngファイルで保存したい。

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Console | Terminal | Jobs

C:/Users/kadota/Desktop/

```
   23      1        1    0.0000      NaN              NA
 upper 95% CI
        1.000
        0.996
        0.968
        0.902
        0.828
        0.657
        0.562
        0.460
        0.407
        0.356
        0.322
           NA

>
> #本番(生存曲線の描画)
> plot(sf, xlab="Time(in weeks)", ylab="Survival rate")#生存曲線の描画
>
>
```

Environment | History | Connections

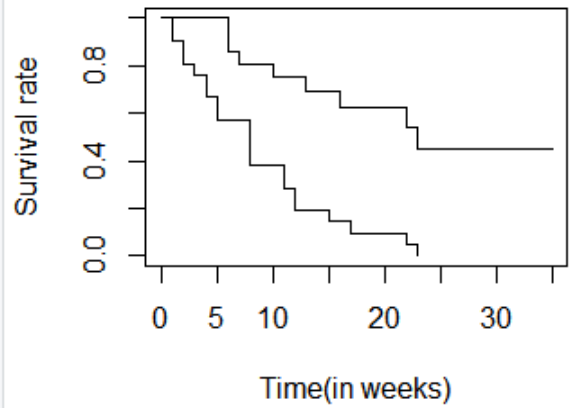
To Console | To Source

```
sf #確認してるだけです
summary(sf) #確認してるだけです
#本番(生存曲線の描画)
plot(sf, xlab="Time(in weeks...
```

② Files | Plots | Packages | Help | Viewer

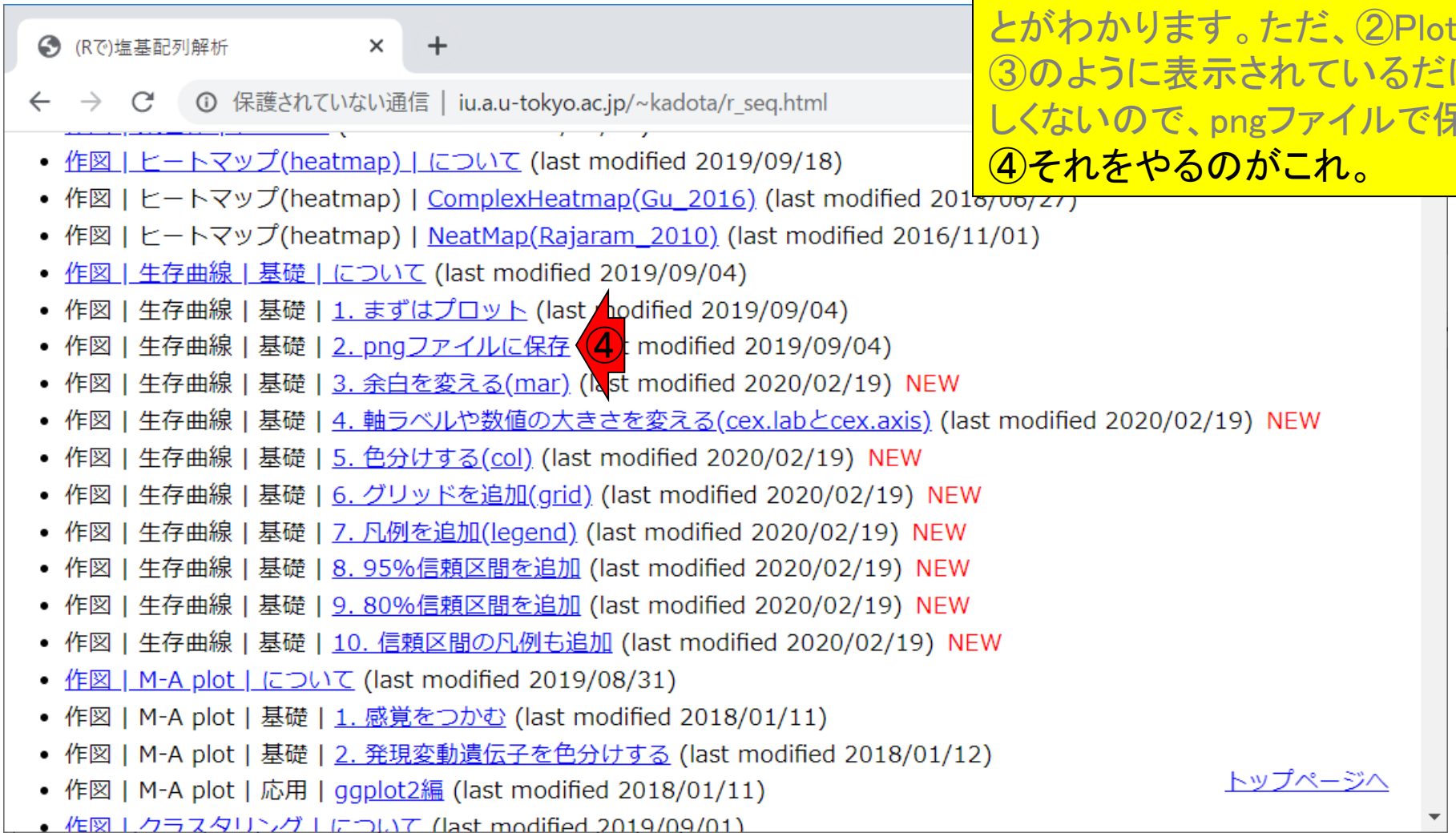
Zoom | Export

③



生存曲線の描画29

さきほどまでは①Filesタブがアクティブな状態でしたが、②Plotsタブに自動で切り替わり、③生存曲線が描画されていることがわかります。ただ、②Plotsタブ上で③のように表示されているだけではうれしくないので、pngファイルで保存したい。
④それをやるのがこれ。



生存曲線の描画30

さきほどまでは①Filesタブがアクティブな状態でしたが、②Plotsタブに自動で切り替わり、③生存曲線が描画されていることがわかります。ただ、②Plotsタブ上で③のように表示されているだけではうれしくないので、pngファイルで保存したい。④それをやるのがこれ。⑤リンク先の、⑥例題1。

(Rで)塩基配列解析

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

作図 | 生存曲線 | 基礎 | 2. pngファイルに保存 ⑤

ここでは、「1. まずはプロット」の続きとして、縦横のサイズを指定してPNG形式ファイ
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. サンプルデータ48のsample48.txtの場合： ⑥

```
in_f <- "sample48.txt"                  #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.png"                    #出力ファイル名を指定してout_fに格納
param_fig <- c(500, 400)                #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(survival)                       #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#in_fで指定したファイルの読み込み

#前処理(生存曲線解析用の形式に変換)
head(data)                              #確認してるだけです(ヘッダー行を除いた最初の6行分を表示)
dim(data)                               #確認してるだけです(行数と列数を表示)
hoge <- Surv(time=time, event=cens)~treat#survival関数のformula部分を作成した結果をhogeに格納
sf <- survfit(formula=hoge, data=data) #survfit関数を用いてクラスオブジェクトを作成した結果をsfに格納
sf                                      #確認してるだけです
summary(sf)                             #確認してるだけです

#ファイルに保存
```

トップページへ

生存曲線の描画31

さきほどまでは①Filesタブがアクティブな状態でしたが、②Plotsタブに自動で切り替わり、③生存曲線が描画されていることがわかります。ただ、②Plotsタブ上で③のように表示されているだけではうれしくないので、pngファイルで保存したい。④それをやるのがこれ。⑤リンク先の、⑥例題1。このスクリプトは、描画結果を⑦hoge1.pngというファイルに保存するものであり、そのサイズは⑧500×400ピクセルの大きさ。

生存曲線の描画32

さきほどまでは①Filesタブがアクティブな状態でしたが、②Plotsタブに自動で切り替わり、③生存曲線が描画されていることがわかります。ただ、②Plotsタブ上で③のように表示されているだけではうれしくないので、pngファイルで保存したい。④それをやるのがこれ。⑤リンク先の、⑥例題1。このスクリプトは、描画結果を⑦hoge1.pngというファイルに保存するものであり、そのサイズは⑧500×400ピクセルの大きさ。少し下に移動しただけ。

(Rで)塩基配列解析

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

1. サンプルデータ48のsample48.txtの場合： ⑥

```
in_f <- "sample48.tx                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.png"                    #出力ファイル名を指定してout_fに格
param_fig <- c(500, 400)                #ファイル出力時の横幅と縦幅を指定(単

#必要なパッケージをロード
library(survival)                       #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#in_fで指定したファイルの読み込み

#前処理(生存曲線解析用の形式に変換)
head(data)                              #確認してるだけです(ヘッダー行を除いた最初の6行分を表示)
dim(data)                               #確認してるだけです(行数と列数を表示)
hoge <- Surv(time=time, event=cens)~treat#survival関数のformula部分を作成した結果をhogeに格納
sf <- survfit(formula=hoge, data=data) #survfit関数を用いてクラスオブジェクトを作成した結果をsfに格納
sf                                      #確認してるだけです
summary(sf)                             #確認してるだけです

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイルの各種パラメータを指定
plot(sf, xlab="Time(in weeks)", ylab="Survival rate")#生存曲線の描画
dev.off()                               #おまじない
```

⑦ ⑧

トップページへ

生存曲線の描画33

(Rで)塩基配列解析

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

1. サンプルデータ48のsample48.txtの場合：

```
in_f <- "sample48.tx                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.png"                    #出力ファイル名を指定してout_fに格
param_fig <- c(500, 400)                #ファイル出力時の横幅と縦幅を指定(単

#必要なパッケージをロード
library(survival)                       #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#in_fで指定したファイルの

#前処理(生存曲線解析用の形式に変換)
head(data)                              #確認してるだけです(ヘッダー行を除
dim(data)                               #確認してるだけです(行数と列数を表
hoge <- Surv(time=time, event=cens)~treat#survival関数のformula部分を作成
sf <- survfit(formula=hoge, data=data) #survfit関数を用いてクラスオブジェクトを作成した結果をsfに格納
sf                                      #確認してるだけです
summary(sf)                             #確認してるだけです

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイルの各種パラメータを指定
plot(sf, xl  "Time(in weeks)", ylab="Survival rate")#生存曲線の描画
dev.off()                               #おまじない
```

トップページへ

⑦ ⑧ ⑨ ⑩

さきほどまでは①Filesタブがアクティブな状態でしたが、②Plotsタブに自動で切り替わり、③生存曲線が描画されていることがわかります。ただ、②Plotsタブ上で③のように表示されているだけではうれしくないので、pngファイルで保存したい。④それをやるのがこれ。⑤リンク先の、⑥例題1。このスクリプトは、描画結果を⑦hoge1.pngというファイルに保存するものであり、そのサイズは⑧500×400ピクセルの大きさ。少し下に移動しただけ。⑨のpng関数のところで、⑦と⑧で指定した内容が利用されます。⑩dev.off()は、⑨で開いたpngファイルの作成が完了したことを宣言するものという理解でよい。

生存曲線の描画34

コピペ実行結果として得られた、⑦hoge1.png。確かに⑧横が500ピクセル、縦が400ピクセルっぽいので妥当ですね。

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#figure_survivalcurve_kiso2

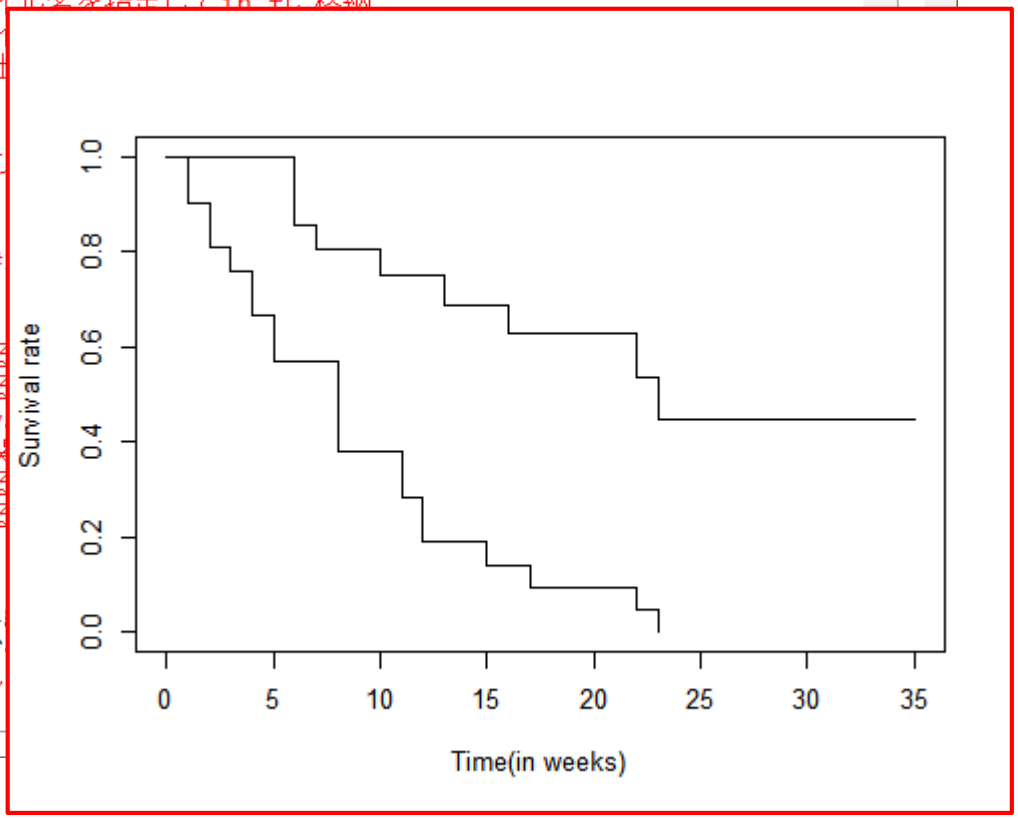
ゲスト

1. サンプルデータ48のsample48.txtの場合：

⑦

⑧

```
in_f <- "sample48.tx                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.png"                    #出力ファイ
param_fig <- c(500, 400)                #ファイル出

#必要なパッケージをロード
library(survival)                       #パッケージ

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                              #確認してる
dim(data)                               #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
sf <- survfit(formula=hoge, data=data) #survfit関
sf                                      #確認してる
summary(sf)                             #確認してる

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], hei
plot(sf, xlab="Time(in weeks)", ylab="Survival r
dev.off()                               #おまじない
```



生存曲線の描画35

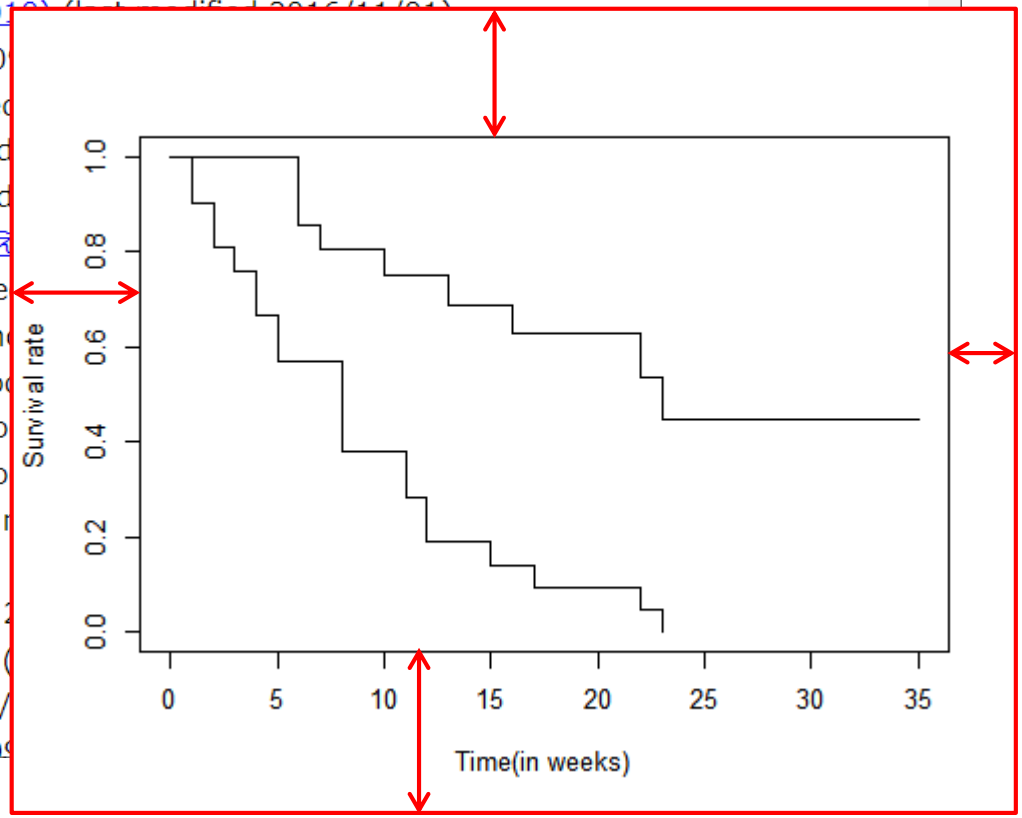
次は、赤矢印で示したデフォルトの余白を任意に変更するやり方。①をクリック。

(Rで)塩基配列解析

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

ゲスト

- 作図 | ヒートマップ(heatmap) | について (last modified 2019/09/18)
- 作図 | ヒートマップ(heatmap) | ComplexHeatmap(Gu_2016) (last modified 2018/06/27)
- 作図 | ヒートマップ(heatmap) | NeatMap(Rajaram_201
- 作図 | 生存曲線 | 基礎 | について (last modified 2019/0
- 作図 | 生存曲線 | 基礎 | 1. まずはプロット (last modifie
- 作図 | 生存曲線 | 基礎 | 2. pngファイルに保存 (
- 作図 | 生存曲線 | 基礎 | 3. 余白を変える(mar)
- 作図 | 生存曲線 | 基礎 | 4. 軸ラベルや数値の大き
- 作図 | 生存曲線 | 基礎 | 5. 色分けする(col) (last modifie
- 作図 | 生存曲線 | 基礎 | 6. グリッドを追加(grid) (last m
- 作図 | 生存曲線 | 基礎 | 7. 凡例を追加(legend) (last mo
- 作図 | 生存曲線 | 基礎 | 8. 95%信頼区間を追加 (last mo
- 作図 | 生存曲線 | 基礎 | 9. 80%信頼区間を追加 (last mo
- 作図 | 生存曲線 | 基礎 | 10. 信頼区間の凡例も追加 (last
- 作図 | M-A plot | について (last modified 2019/08/31)
- 作図 | M-A plot | 基礎 | 1. 感覚をつかむ (last modified
- 作図 | M-A plot | 基礎 | 2. 発現変動遺伝子を色分けする
- 作図 | M-A plot | 応用 | ggplot2編 (last modified 2018/
- 作図 | クラスタリング | について (last modified 2019/0



生存曲線の描画36

次は、赤矢印で示したデフォルトの余白を任意に変更するやり方。①をクリック。
②例題1のスクリプトは、③で余白をそれぞれ「下が4、左が5、上が1、右が0」となるように指定しています。

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#figure_survivalcurve_kis05

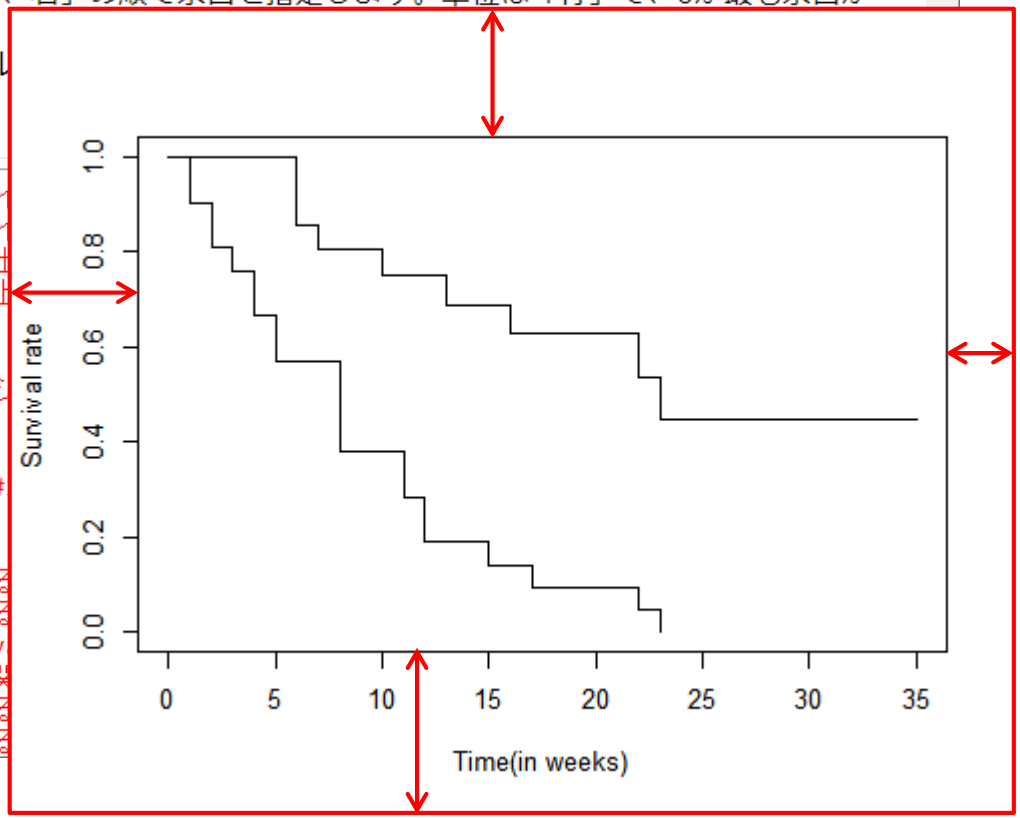
作図 | 生存曲線 | 基礎 | 3. 余白を変える NEW

この項目では、図の余白を任意に変更します。 「下、左、上、右」の順で余白を指定します。単位は「行」で、0が最も余白が小さく、大体5程度が最大です。
「ファイル」－「ディレクトリの変更」で解析したいファイル

1. サンプルデータ48のsample48.txtの場合： ②

```
in_f <- "sample48.txt"                    #入力ファイ
out_f <- "hoge1.png"                      #出力ファイ
param_fig <- c(500, 400)                  #ファイル出
param_mar <- c(4, 5, 1, 0)                #下、左、上

#必要なパッケージをロード
library(survival)                         #パッケージ

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                                #確認してる
dim(data)                                 #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
sf <- survfit(formula=hoge, data=data) #survfit関
sf                                        #確認してる
summary(sf)                               #確認してる
```

③



生存曲線の描画37

次は、赤矢印で示したデフォルトの余白を任意に変更するやり方。①をクリック。②例題1のスクリプトは、③で余白をそれぞれ「下が4、左が5、上が1、右が0」となるように指定しています。コピペ実行結果がこれ。妥当ですね。

(Rで)塩基配列解析

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

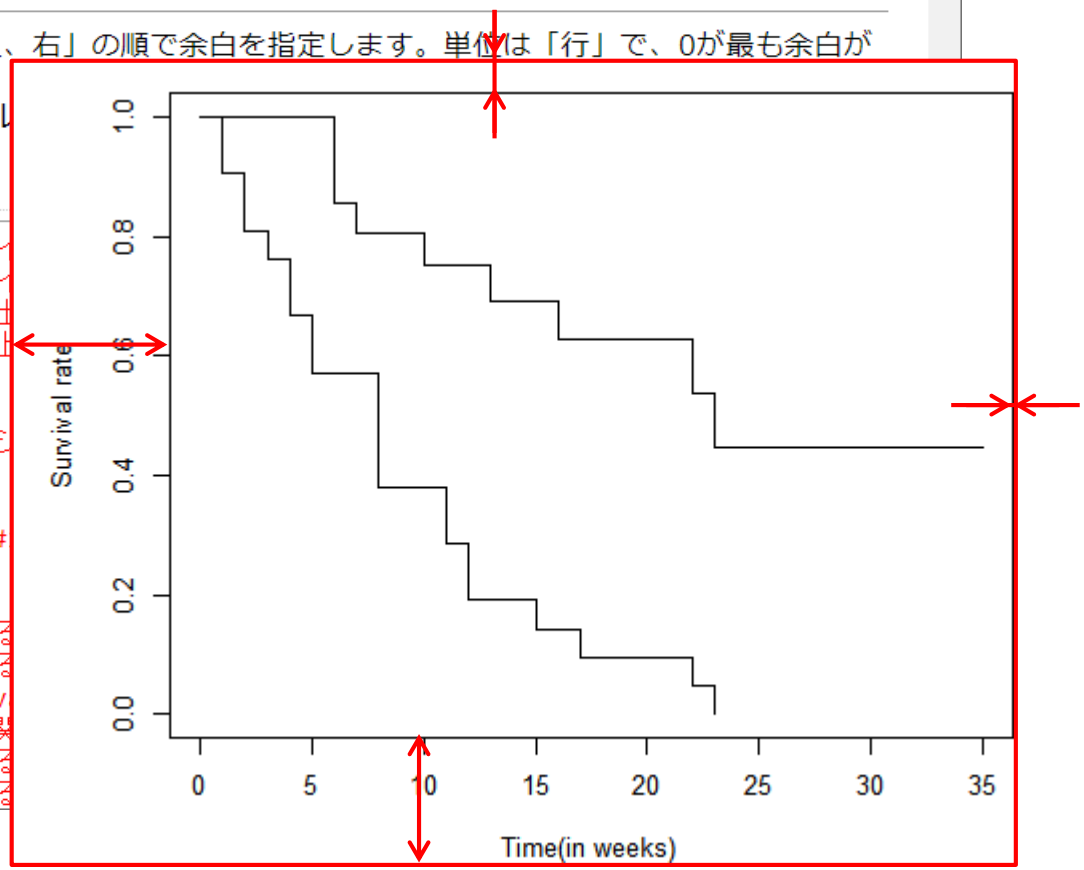
作図 | 生存曲線 | 基礎 | 3. 余白を変える NEW

この項目では、図の余白を任意に変更します。 「下、左、上、右」の順で余白を指定します。単位は「行」で、0が最も余白が小さく、大体5程度が最大です。
「ファイル」－「ディレクトリの変更」で解析したいファイル

1. サンプルデータ48のsample48.txtの場合：

```
in_f <- "sample48.txt"                    #入力ファイ
out_f <- "hoge1.png"                      #出力ファイ
param_fig <- c(500, 400)                  #ファイル出
param_mar <- c(4, 5, 1, 0)                #下、左、上

#必要なパッケージをロード
library(survival)                         #パッケージ

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                                #確認してる
dim(data)                                 #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
sf <- survfit(formula=hoge, data=data) #survfit関
sf                                        #確認してる
summary(sf)                               #確認してる
```

③



生存曲線の描画38

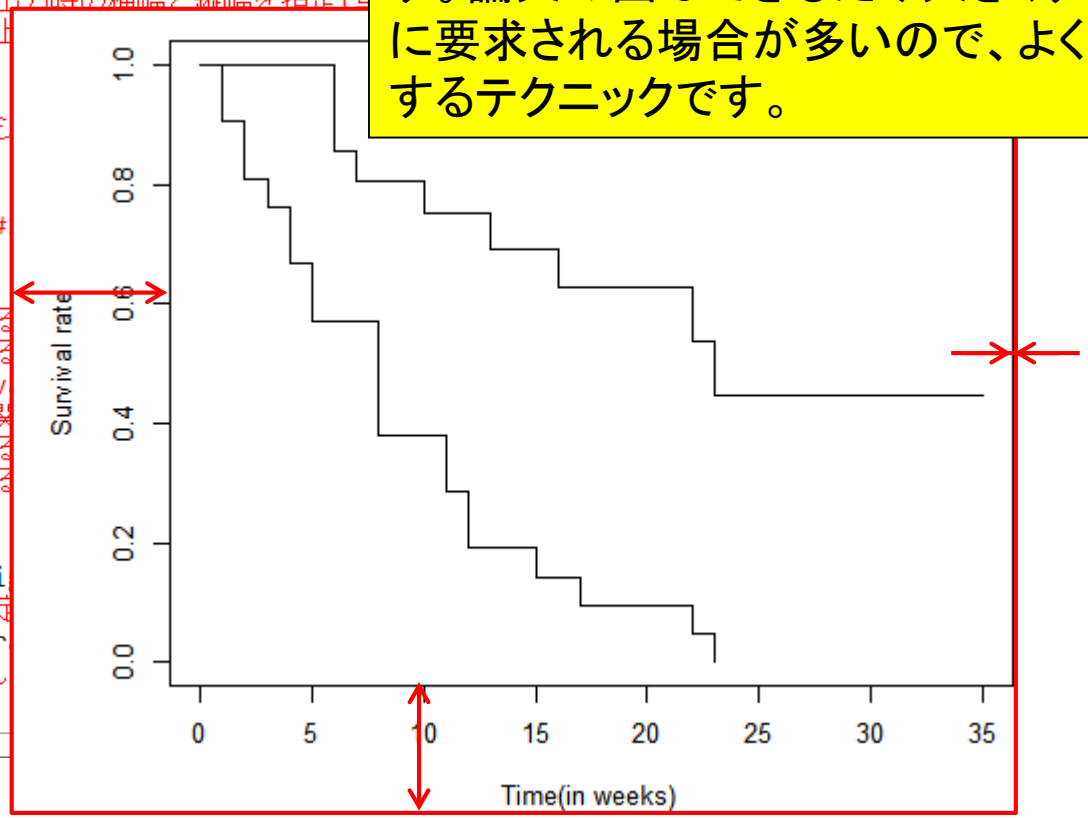
次は、赤矢印で示したデフォルトの余白を任意に変更するやり方。①をクリック。②例題1のスクリプトは、③で余白をそれぞれ「下が4、左が5、上が1、右が0」となるように指定しています。コピペ実行結果がこれ。妥当ですね。③で指定したパラメータは、④のところで利用されています。論文の図はできるだけ大きくするように要求される場合が多いので、よく利用するテクニックです。

(Rで)塩基配列解析

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

1. サンプルデータ48のsample48.txtの場合：

```
param_fig <- c(500, 400)                    #ファイル出力時の横幅と縦幅を指定(単
param_mar <- c(4, 5, 1, 0)                  #下、左、上

#必要なパッケージをロード
library(survival)                           #パッケージ

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                                  #確認してる
dim(data)                                   #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
sf <- survfit(formula=hoge, data=data) #survfit関
sf                                          #確認してる
summary(sf)                                 #確認してる

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], hei
par(mar=param_mar)                          #余白を指定
plot(sf, xlab="Time(in weeks)", ylab="Survival r
dev.off()                                   #おまじない
```



生存曲線の描画39

次は、①軸ラベルや数値の大きさを変更するやり方。②の部分は不変ですが…

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#figure_survivalcurve_kiso4

ゲスト

作図 | 生存曲線 | 基礎 | 4. 軸ラベルや数値の大きさを変える(cex.labとcex.axis) NEW

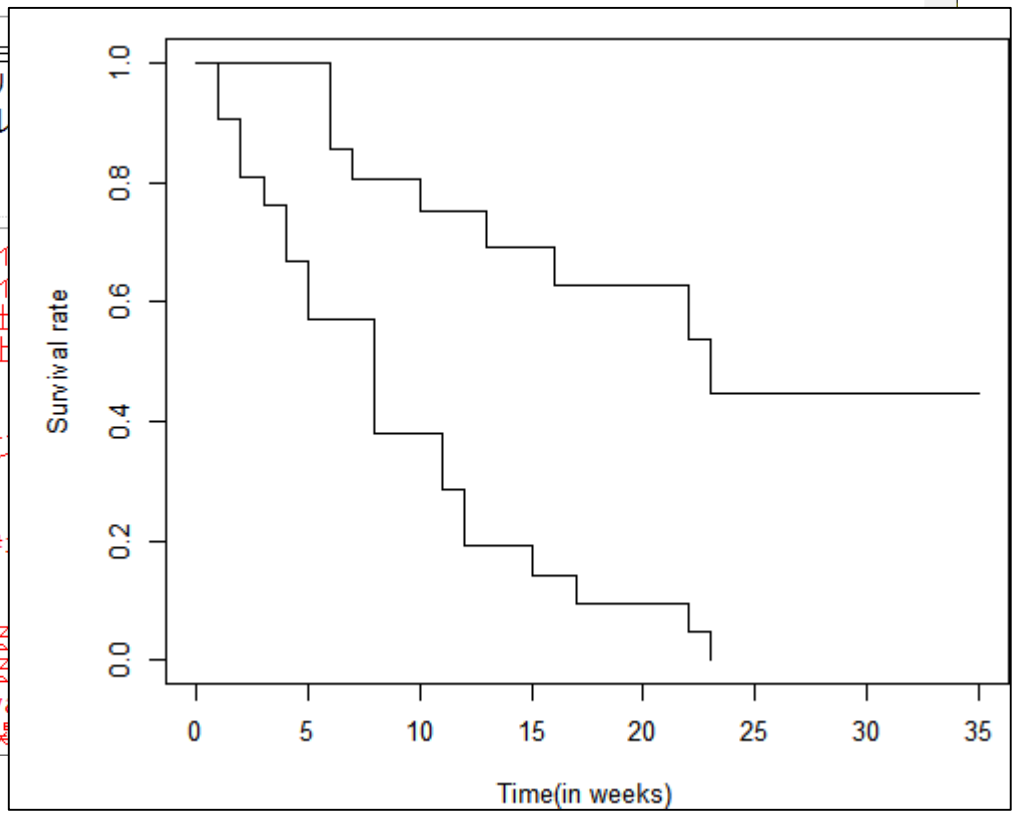
①

この項目では、plot関数実行時に、cex.labとcex.axisオプシ
という指定の仕方をします。 例えば、cex.lab=1.5は軸ラベ
「ファイル」－「ディレクトリの変更」で解析したいファイル

1. サンプルデータ48のsample48.txtの場合：

②

```
in_f <- "sample48.txt"                    #入力ファイ
out_f <- "hoge1.png"                      #出力ファイ
param_fig <- c(500, 400)                  #ファイル出
param_mar <- c(4, 5, 1, 0)                #下、左、上

#必要なパッケージをロード
library(survival)                         #パッケージ

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                                #確認してる
dim(data)                                 #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
sf <- survfit(formula=hoge, data=data) #survfit関
```



生存曲線の描画40

次は、①軸ラベルや数値の大きさを変更するやり方。②の部分は不変ですが、コード下部の③で変更を加えています。

(Rで)塩基配列解析

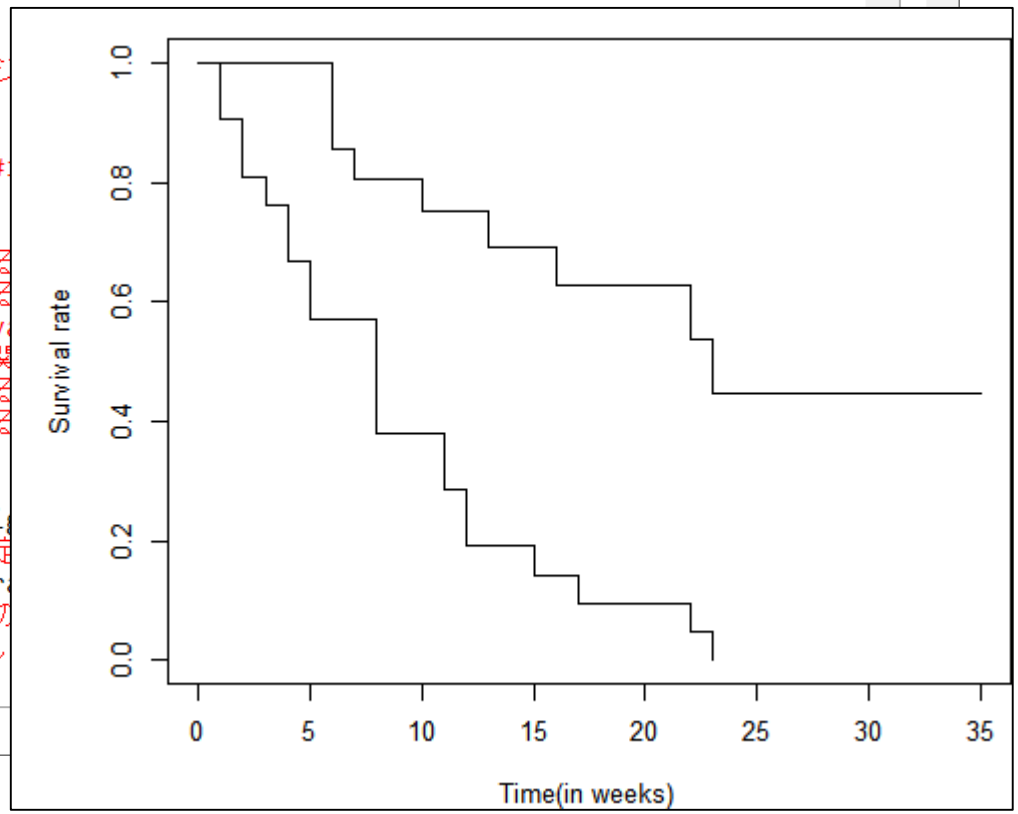
保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#figure_survivalcurve_kiso4

ゲスト

1. サンプルデータ48のsample48.txtの場合：

```
param_mar <- c(4, 5, 1, 0)              #下、左、上、右の順で余白を指定(単位は行)

#必要なパッケージをロード
library(survival)                       #パッケージ

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                              #確認してる
dim(data)                               #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
sf <- survfit(formula=hoge, data=data) #survfit関
sf                                      #確認してる
summary(sf)                             #確認してる

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], hei
par(mar=param_mar)                      #余白を指定
plot(sf, xlab="Time(in weeks)", ylab="Survival r
    cex.lab=1.5, cex.axis=1.5)          #生存曲線の
dev.off()                               #おまじない
```

③



生存曲線の描画41

次は、①軸ラベルや数値の大きさを変更するやり方。②の部分は不変ですが、コード下部の③で変更を加えています。
コピペ実行結果。④cex.lab=1.5は、⑤ラベル(label)情報の大きさを通常の1.5倍にするという指定。同様に⑥cex.axis=1.5は、⑦軸(axis)の数値の大きさを通常の1.5倍にするという指定。

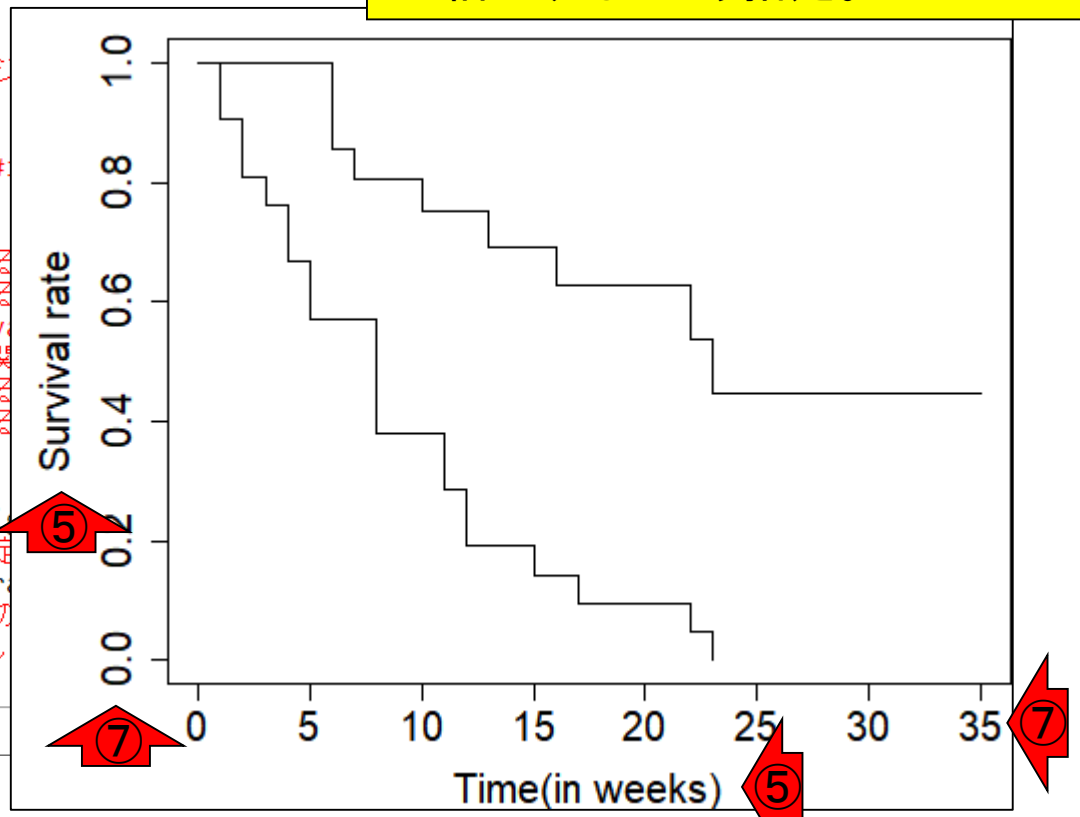
(Rで)塩基配列解析

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

1. サンプルデータ48のsample48.txtの場合：

```
param_mar <- c(4, 5, 1, 0)                #下、左、上、右の順で余白を指定(単位

#必要なパッケージをロード
library(survival)                         #パッケージ

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                                #確認してる
dim(data)                                 #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
sf <- survfit(formula=hoge, data=data) #survfit関
sf                                        #確認してる
summary(sf)                               #確認してる

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], hei
par(mar=param_mar)                        #余白を指定
plot(sf, xlab="Time(in weeks)", ylab="Survival r
    cex.lab=1.5, cex.axis=1.5)            #生存曲線の
dev.off()                                 #おまじない
```

③ ④ ⑥



生存曲線の描画42

次は、①色をを変更するやり方。②の部分で指定しています。

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#figure_survivalcurve_kiso5

ゲスト

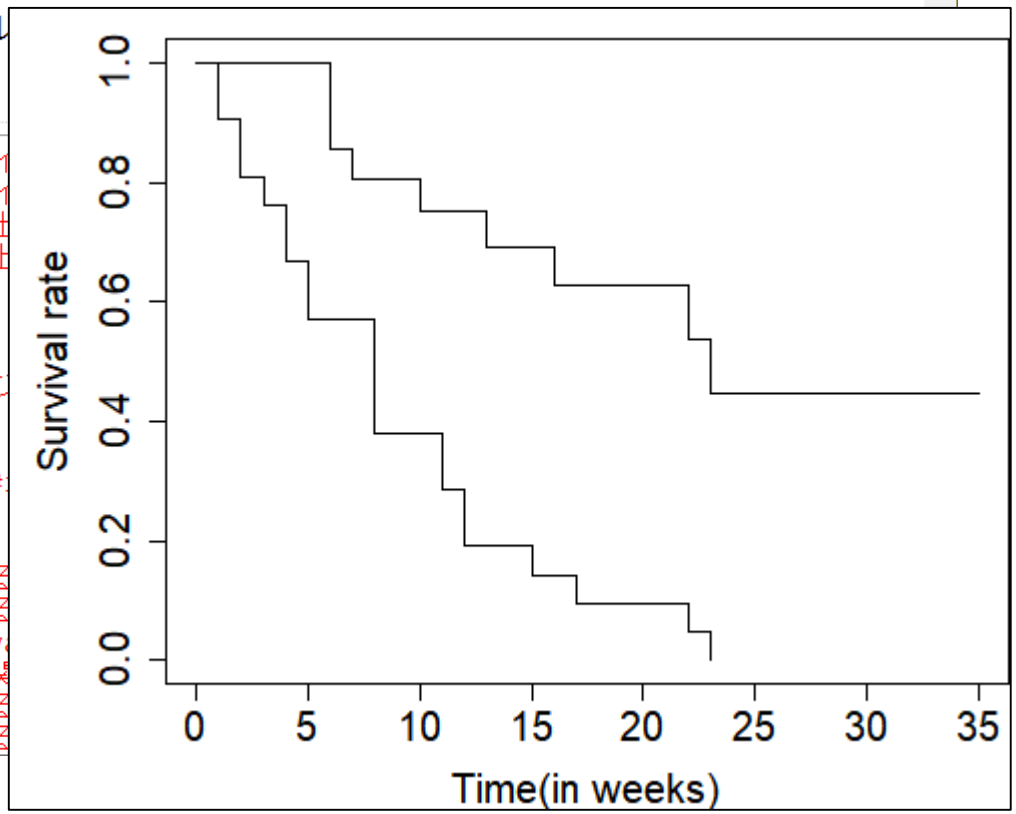
作図 | 生存曲線 | 基礎 | 5. 色分けする(col) ①NEW

この項目では、colオプションを追加して色分けします。
「ファイル」－「ディレクトリの変更」で解析したいファイル

1. サンプルデータ48のsample48.txtの場合：

```
in_f <- "sample48.txt"                    #入力ファイ
out_f <- "hoge1.png"                      #出力ファイ
param_fig <- c(500, 400)                  #ファイル出
param_mar <- c(4, 5, 1, 0)                #下、左、上
param_col <- c("red", "black")            #色を指定

#必要なパッケージをロード
library(survival)                         #パッケージ

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                                #確認してる
dim(data)                                 #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
sf <- survfit(formula=hoge, data=data) #survfit関
sf                                        #確認してる
summary(sf)                               #確認してる
```

②



生存曲線の描画43

次は、①色を変更するやり方。②の部分で指定しています。実行結果。②の指定内容は、コード下部の、③のところで反映されています。

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#figure_survivalcurve_kiso5

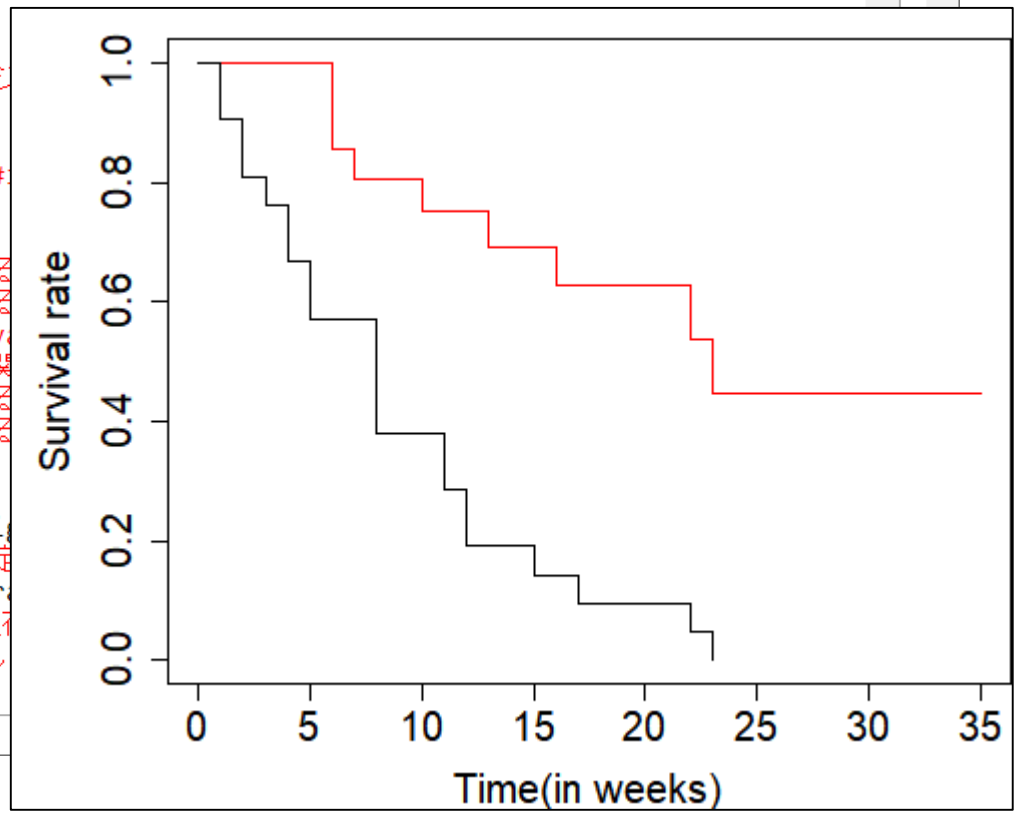
ゲスト

1. サンプルデータ48のsample48.txtの場合：

```
param_col <- c("red", "black")      #色を指定

#必要なパッケージをロード
library(survival)                   #パッケージ

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                          #確認してる
dim(data)                           #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
sf <- survfit(formula=hoge, data=data) #survfit関
sf                                  #確認してる
summary(sf)                         #確認してる

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], hei
par(mar=param_mar)                  #余白を指定
plot(sf, xlab="Time(in weeks)", ylab="Survival r
    cex.lab=1.5, cex.axis=1.5, col=param_col)#生存
dev.off()                           #おまじない
```

②

③



生存曲線の描画44

次は、①色を変更するやり方。②の部分で指定しています。実行結果。②の指定内容は、コード下部の、③のところで反映されています。今は明確にredが6-MP投与群で、blackがcontrol群だと分かります。しかし、3群以上の場合はどの順番で色を指定するのか難しい場合があります。その場合は、④の実行結果の並びを参考にすればよいです。

(Rで)塩基配列解析

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

1. サンプルデータ48

```
param_col <- c(

#必要なパッケージ
library(surviva

#入力ファイルの読
data <- read.ta

#前処理(生存曲線
head(data)
dim(data)
hoge <- Surv(ti
sf <- survfit(f
sf
summary(sf)

#ファイルに保存
png(out_f, poin
par(mar=param_m
plot(sf, xlab="
    cex.lab=1.5
dev.off()
```

④

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Console Terminal Jobs

C:/Users/kadota/Desktop/

```
4     2     7     1     6-MP
5     3     3     1  control
6     3    32     0     6-MP
> dim(data)                                  #確認してるだけです(行数と列数を表示)
[1] 42  4
> hoge <- Surv(time=time, event=cens)~treat#survival関数のformula部分を作成した結果をhogeに格納
> sf <- survfit(formula=hoge, data=data) #survfit関数を用いてクラスオブジェクトを作成した結果をsfに格納
> sf                                         #確認してるだけです
Call: survfit(formula = hoge, data = data)

                n events median 0.95LCL 0.95UCL
treat=6-MP     21      9     23      16      NA
treat=control  21     21      8       4      12
> summary(sf)                                #確認してるだけです
Call: survfit(formula = hoge, data = data)

                treat=6-MP
 time n.risk n.event survival std.err lower 95% CI
    6     21       3    0.857  0.0764        0.720
```

④

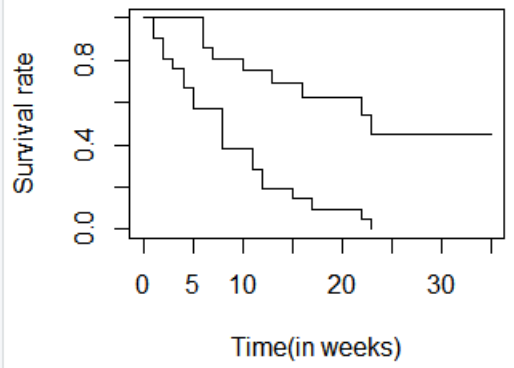
Environment History Connections

To Console To Source

```
par(mar=param_mar) #余白を指定
plot(sf, xlab="Time(in weeks…
cex.lab=1.5, cex.axis=1.5, c…
dev.off() #おまじない
```

Files Plots Packages Help Viewer

Zoom Export



次は、①グリッド線を追加するやり方。

生存曲線の描画45

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#figure_survivalcurve_kiso6

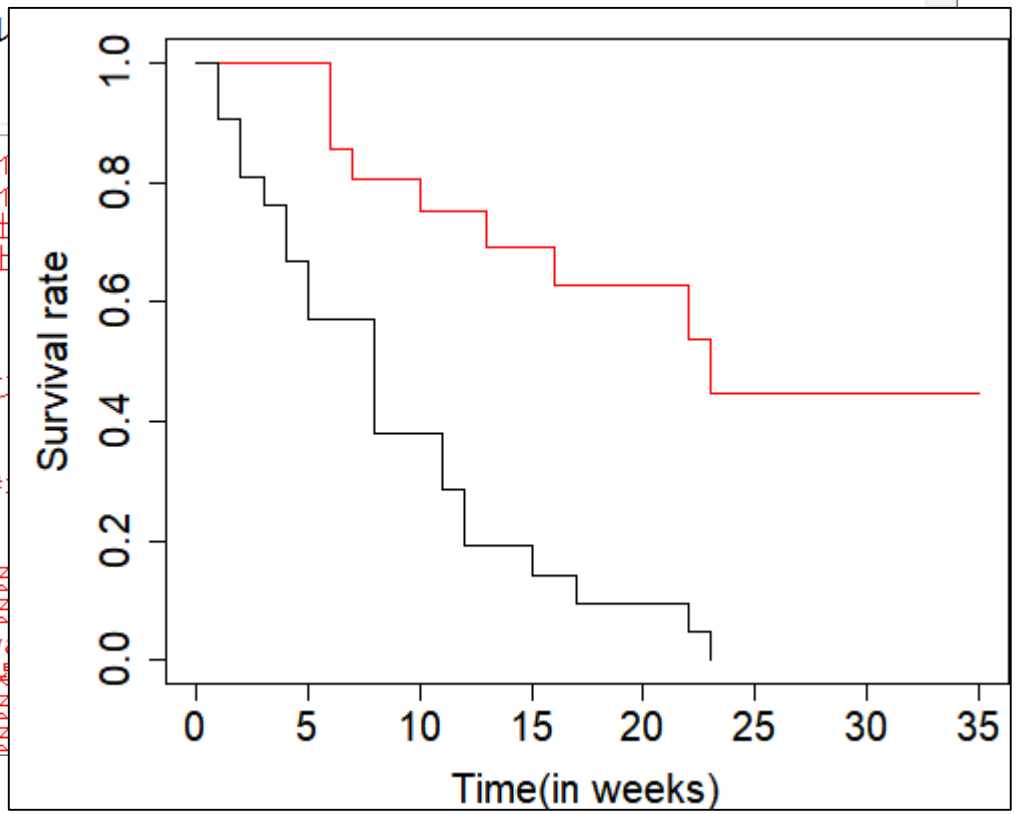
ゲスト

作図 | 生存曲線 | 基礎 | 6. グリッドを追加(grid) ①NEW

この項目では、grid関数を用いてグリッドを追加します。
「ファイル」－「ディレクトリの変更」で解析したいファイル

1. サンプルデータ48のsample48.txtの場合：

```
in_f <- "sample48.txt"                  #入力ファイ
out_f <- "hoge1.png"                    #出力ファイ
param_fig <- c(500, 400)                #ファイル出
param_mar <- c(4, 5, 1, 0)              #下、左、上
param_col <- c("red", "black")          #色を指定

#必要なパッケージをロード
library(survival)                       #パッケージ

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                              #確認してる
dim(data)                               #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
sf <- survfit(formula=hoge, data=data) #survfit関
sf                                      #確認してる
summary(sf)                             #確認してる
```



生存曲線の描画46

次は、①グリッド線を追加するやり方。実行結果。コード下部に、②grid関数が追加されています。colオプションやltyオプションの意味が実行結果と比較するとよくわかると思います。例えばltyでは、dotted（点線）の代わりにdashed（破線）やsolid（実線）なども指定できます。

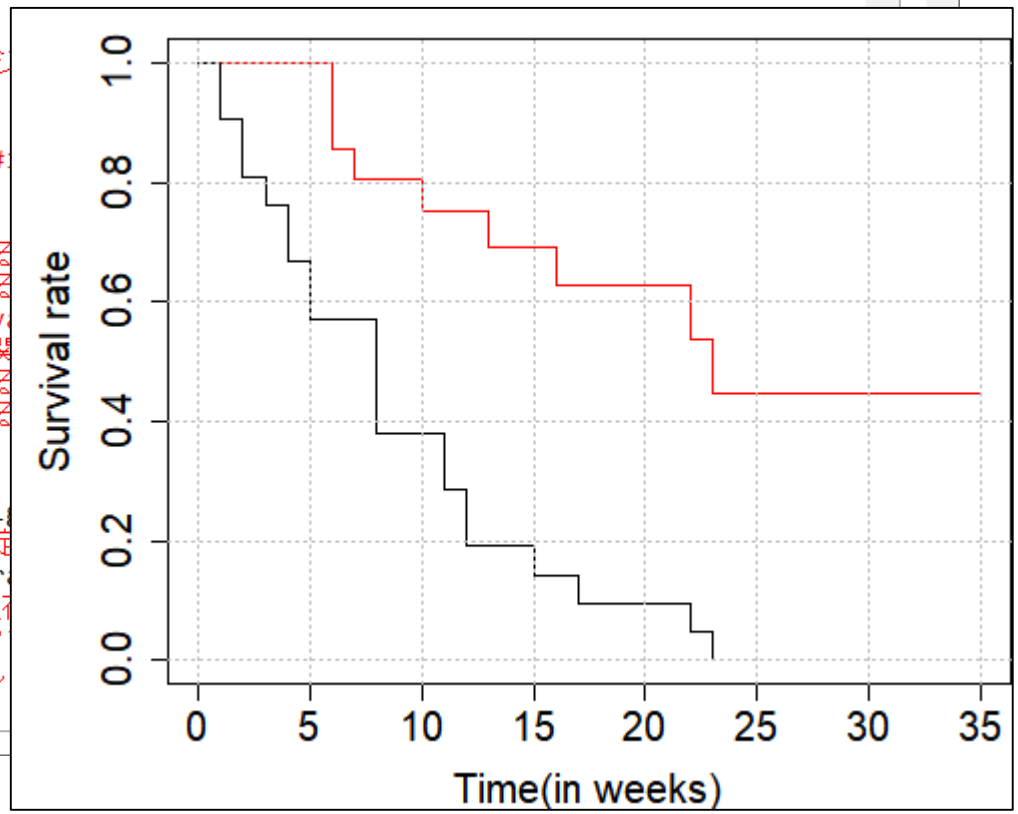
(Rで)塩基配列解析

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

1. サンプルデータ48のsample48.txtの場合：

```
#必要なパッケージをロード
library(survival)                          #パッケージ

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                                 #確認してる
dim(data)                                  #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
sf <- survfit(formula=hoge, data=data) #survfit関
sf                                         #確認してる
summary(sf)                                #確認してる

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], hei
par(mar=param_mar)                         #余白を指定
plot(sf, xlab="Time(in weeks)", ylab="Survival r
    cex.lab=1.5, cex.axis=1.5, col=param_col)#生存
grid(col="gray", lty="dotted")             #指定したパ
dev.off()                                  #おまじない
```

②



生存曲線の描画47

次は、①凡例を追加するやり方。②の部分で指定しています。順番は、③と同じです。

(Rで)塩基配列解析

保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#figure_survivalcurve_kiso7

ゲスト

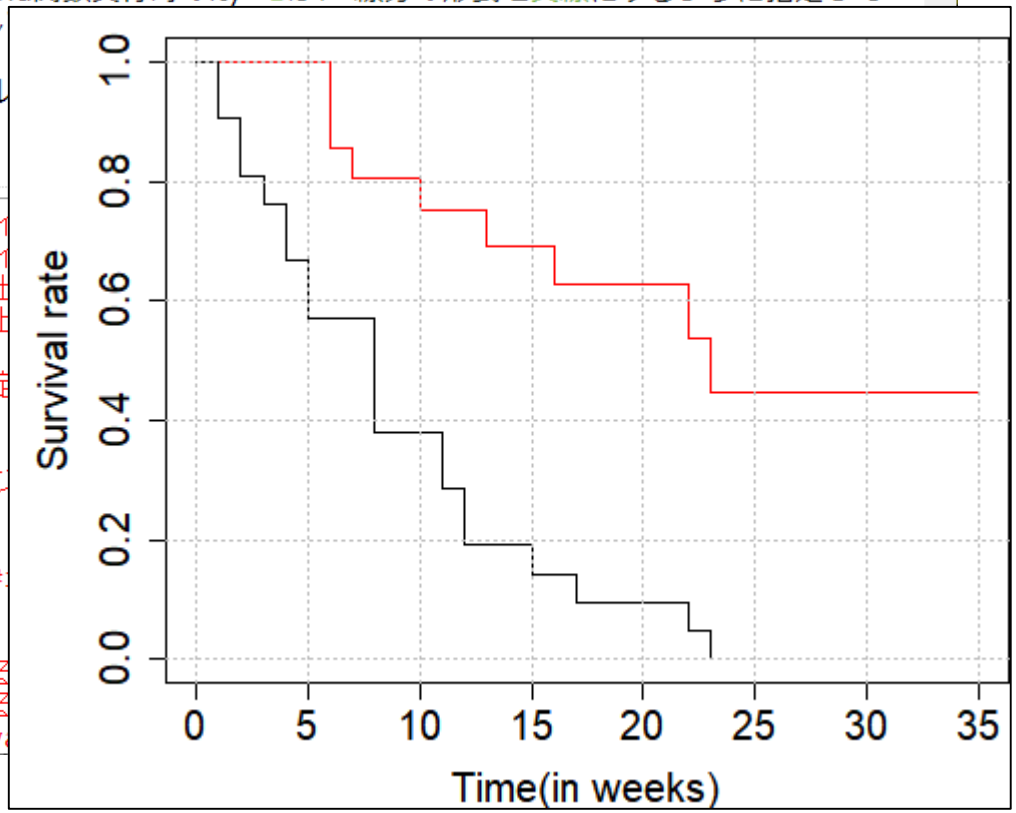
作図 | 生存曲線 | 基礎 | 7. 凡例を追加(legend) ①NEW

この項目では、legend関数を用いて凡例を追加します。legend関数実行時のlty=1は、 線分の形式を実線にするように指定していることに相当します。 lty=2は破線に相当します。他にもいです。この場合は通常の1.5倍に相当します。
「ファイル」－「ディレクトリの変更」で解析したいファイル

1. サンプルデータ48のsample48.txtの場合：

```
in_f <- "sample48.txt"                    #入力ファイ
out_f <- "hoge1.png"                      #出力ファイ
param_fig <- c(500, 400)                  #ファイル出
param_mar <- c(4, 5, 1, 0)                #下、左、上
param_col <- c("red", "black")            #色を指定
param_legend <- c("6-MP", "Control")      #凡例を指定

#必要なパッケージをロード
library(survival)                         #パッケージ

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                                #確認してる
dim(data)                                 #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
```

③ ②



生存曲線の描画48

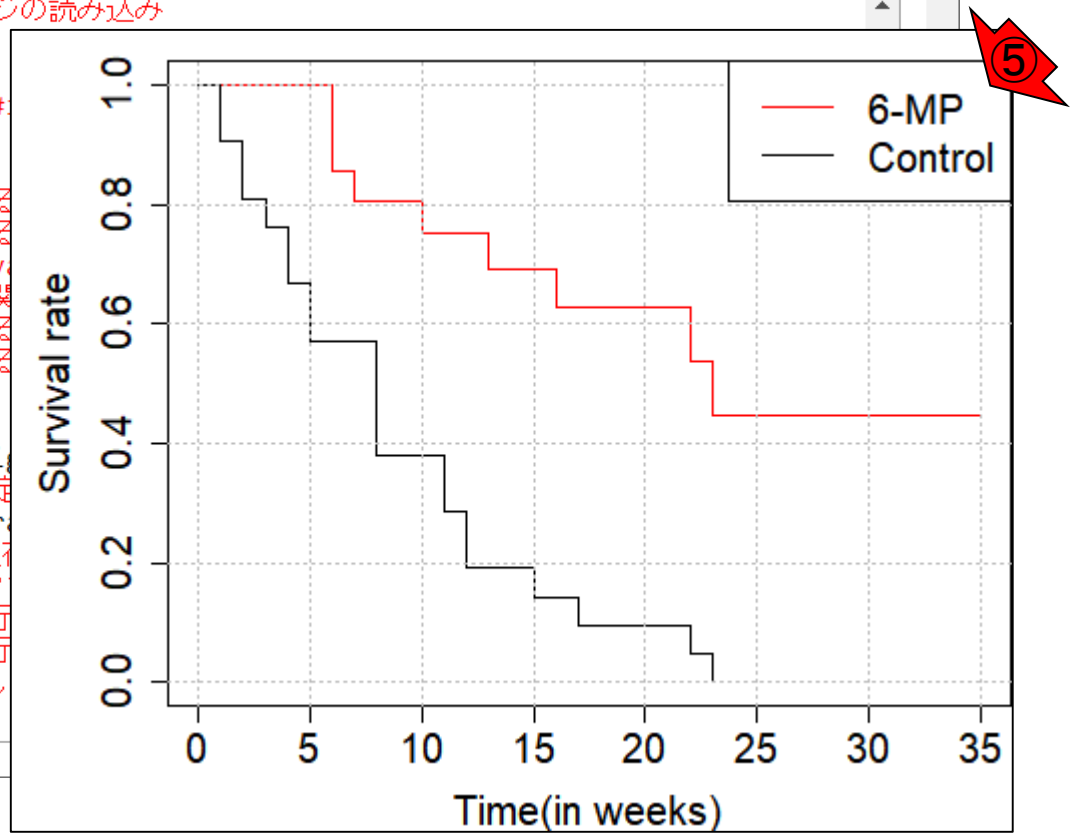
次は、①凡例を追加するやり方。②の部分で指定しています。順番は、③と同じです。実行結果。コード下部に、④legend関数が追加されています。⑤なぜ凡例が右上（topright）にあるかの理由もわかるでしょう。場合によっては左下（bottomleft）でもよいかもしれません。

(Rで)塩基配列解析

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

1. サンプルデータ48のsample48.txtの場合：

```
library(survival)                          #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, sep="\t")#

#前処理(生存曲線解析用の形式に変換)
head(data)                                 #確認してる
dim(data)                                  #確認してる
hoge <- Surv(time=time, event=cens)~treat#surviv
sf <- survfit(formula=hoge, data=data) #survfit関
sf                                         #確認してる
summary(sf)                                #確認してる

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], hei
par(mar=param_mar)                         #余白を指定
plot(sf, xlab="Time(in weeks)", ylab="Survival r
    cex.lab=1.5, cex.axis=1.5, col=param_col)#生存
grid(col="gray", lty="dotted")             #指定したパ
④ legend("topright", legend=param_legend,#凡例を表示
       col=param_col, lty=1, cex=1.5)      #凡例を表示
dev.off()                                  #おまじない
```



生存曲線の描画49

①残りは割愛しますが、②80%信頼区間とその凡例なども追加することができます。

(Rで)塩基配列解析

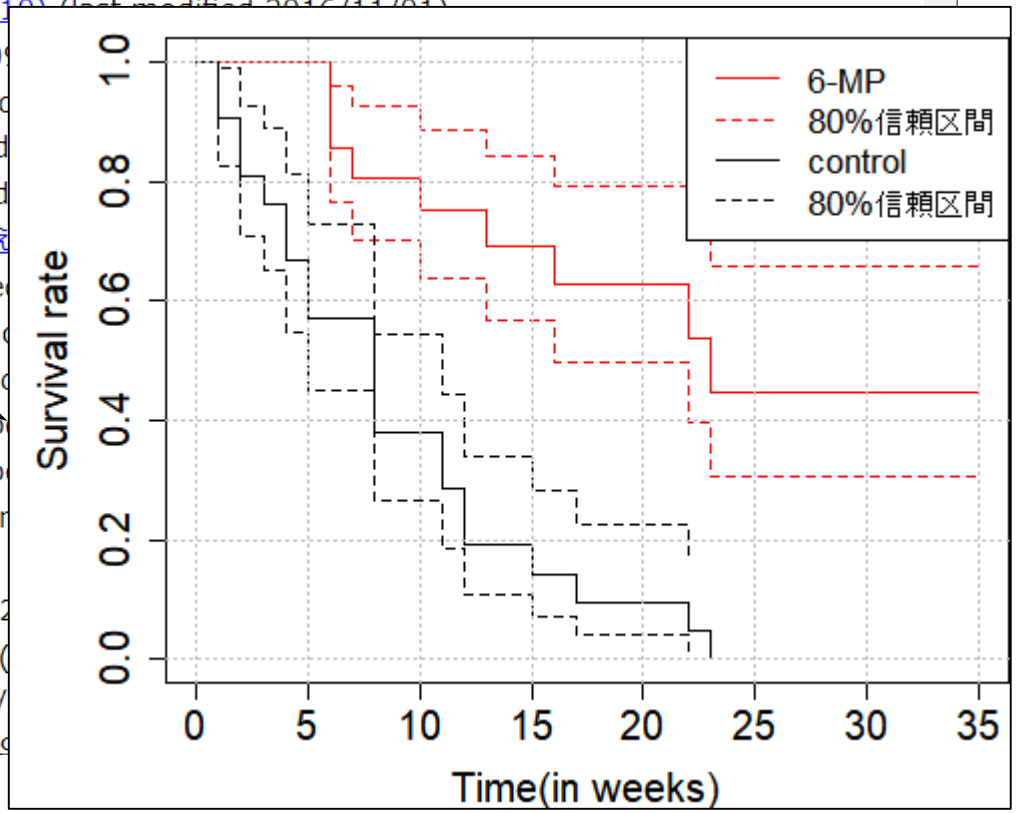
保護されていない通信 | iu.a.u-tokyo.ac.jp/~kadota/r_seq.html

ゲスト

- 作図 | ヒートマップ(heatmap) | について (last modified 2019/09/18)
- 作図 | ヒートマップ(heatmap) | ComplexHeatmap(Gu_2016) (last modified 2018/06/27)
- 作図 | ヒートマップ(heatmap) | NeatMap(Rajaram_201
- 作図 | 生存曲線 | 基礎 | について (last modified 2019/0
- 作図 | 生存曲線 | 基礎 | 1. まずはプロット (last modifie
- 作図 | 生存曲線 | 基礎 | 2. pngファイルに保存 (last mod
- 作図 | 生存曲線 | 基礎 | 3. 余白を変える(mar) (last mod
- 作図 | 生存曲線 | 基礎 | 4. 軸ラベルや数値の大きさを変え
- 作図 | 生存曲線 | 基礎 | 5. 色分けする(col) (last modifie
- 作図 | 生存曲線 | 基礎 | 6. グリッドを追加(grid) (last mo
- 作図 | 生存曲線 | 基礎 | 7. 凡例を追加(legend) (la
- 作図 | 生存曲線 | 基礎 | 8. 95%信頼区間を追加 (last mo
- 作図 | 生存曲線 | 基礎 | 9. 80%信頼区間を追加 (last mo
- 作図 | 生存曲線 | 基礎 | 10. 信頼区間の凡例も追加
- 作図 | M-A plot | について (last modified 2019/08/11)
- 作図 | M-A plot | 基礎 | 1. 感覚をつかむ (last modified
- 作図 | M-A plot | 基礎 | 2. 発現変動遺伝子を色分けする
- 作図 | M-A plot | 応用 | ggplot2編 (last modified 2018/
- 作図 | クラスタリング | について (last modified 2019/0

①

②

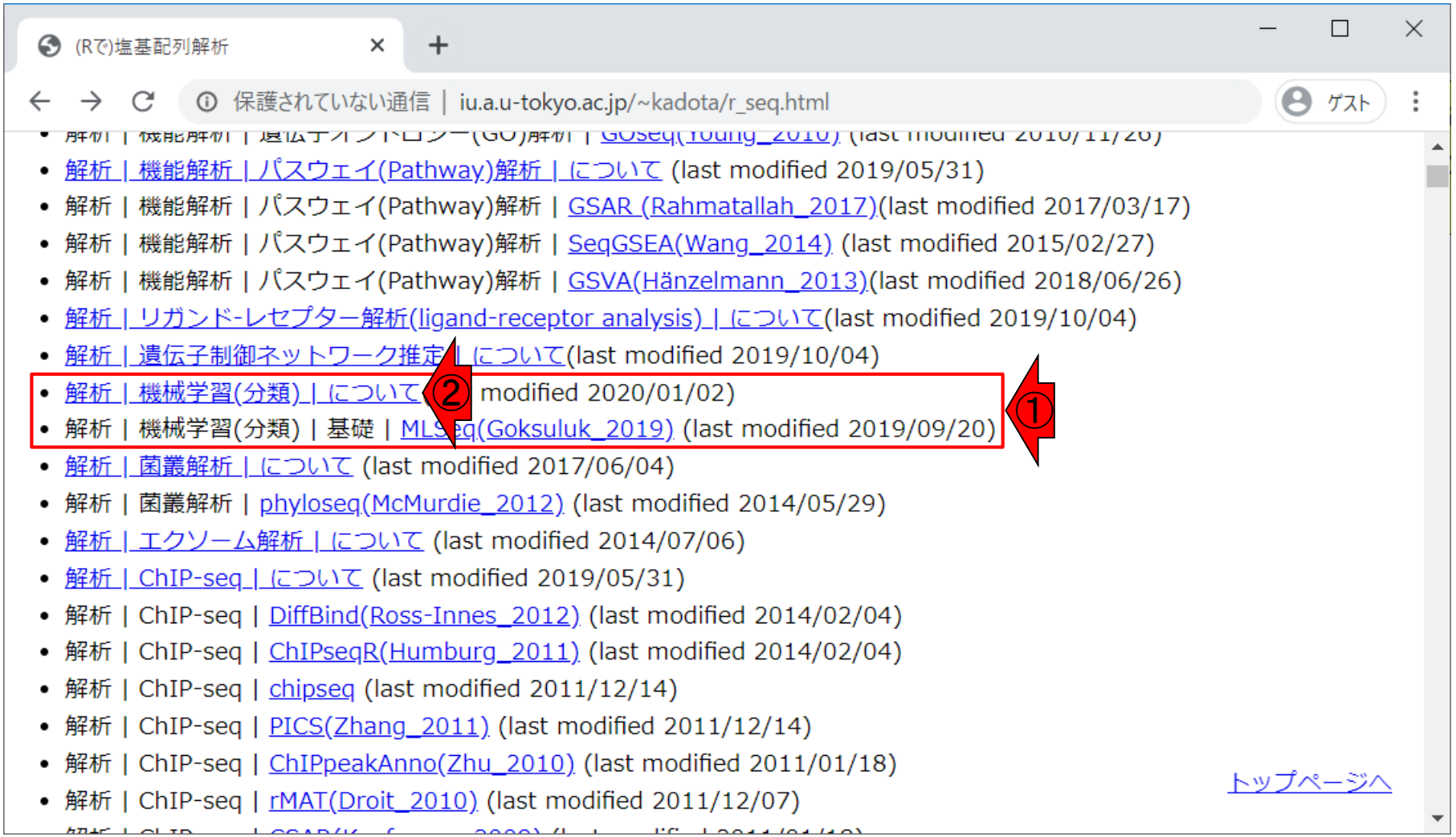


Contents（今回）

- 以前の講義内容のおさらい
- single-cell RNA-seqの論文：DECENT論文、NBID論文、sc性能評価論文
- symmetric（対称）とasymmetric（非対称）
- scRNA-seq性能評価論文再訪
- bulk RNA-seqの論文：DEGES論文、TCC論文、bulk性能評価論文
- 何故次から次へと新規手法論文が出るのか?
- Zero-inflated negative binomial (ZINB)モデルはscRNA-seq用なのか?
- 「反復数増やすとDEG増える」を細胞数で読み替えると…
- ここまでのまとめ
- GLMベースの3群間比較、Osabe法で発現パターンまで得る
- Osabe法の実践、生存曲線の描画、分類（診断）

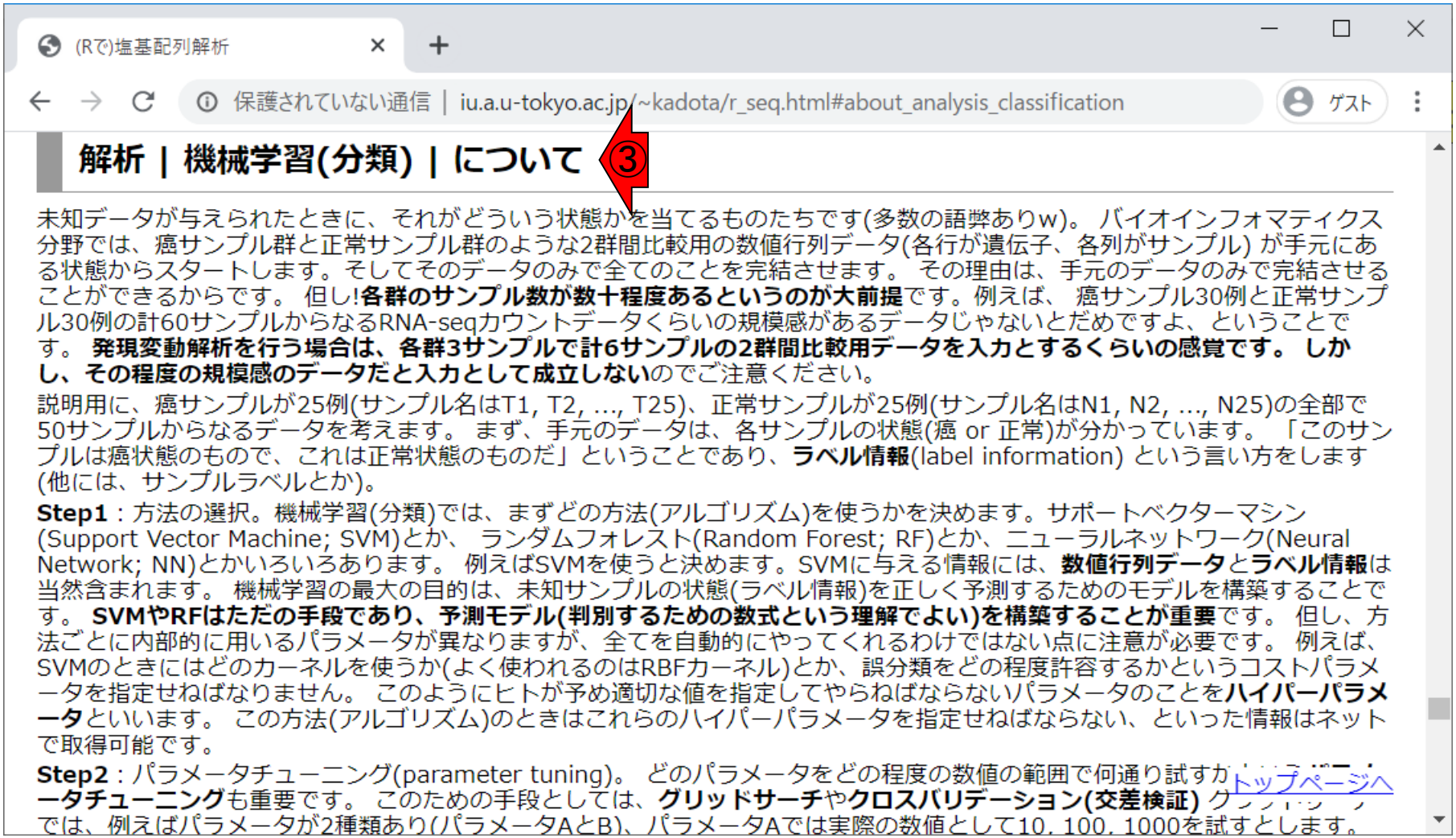
分類(診断)1

①機械学習(分類)はこのあたり。②のリンク先が…



分類(診断)2

①機械学習(分類)はこのあたり。②のリンク先が、③こんな感じです。



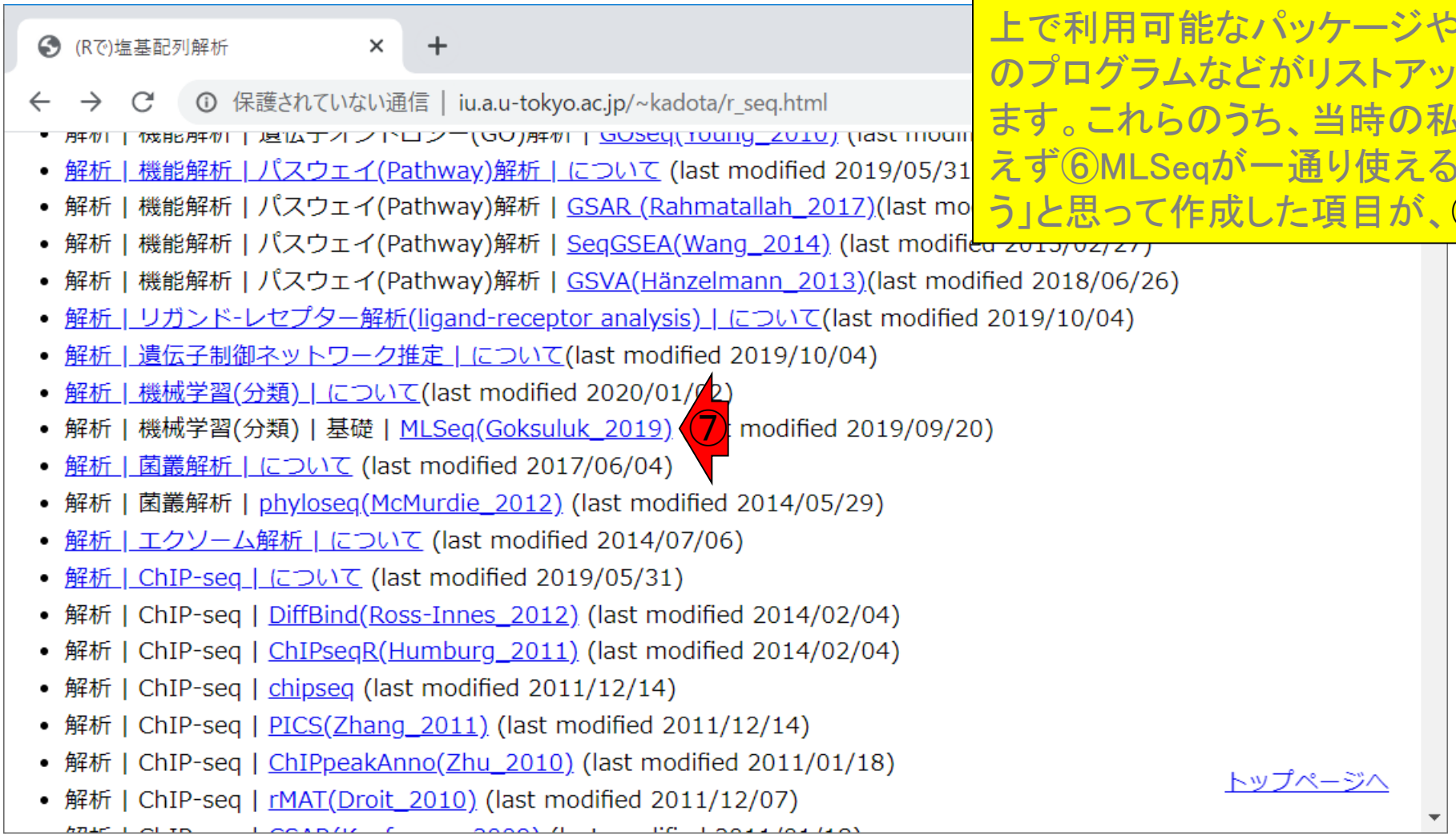
分類(診断)3

①機械学習(分類)はこのあたり。②のリンク先が、③こんな感じです。ページ下部に移動していくと、こんな感じで、④R上で利用可能なパッケージや、⑤R以外のプログラムなどがリストアップされています。

分類(診断)4

①機械学習(分類)はこのあたり。②のリンク先が、③こんな感じです。ページ下部に移動していくと、こんな感じで、④R上で利用可能なパッケージや、⑤R以外のプログラムなどがリストアップされています。これらのうち、当時の私が「とりあえず⑥MLSeqが一通り使えるようにしよう」と思って作成した項目が…

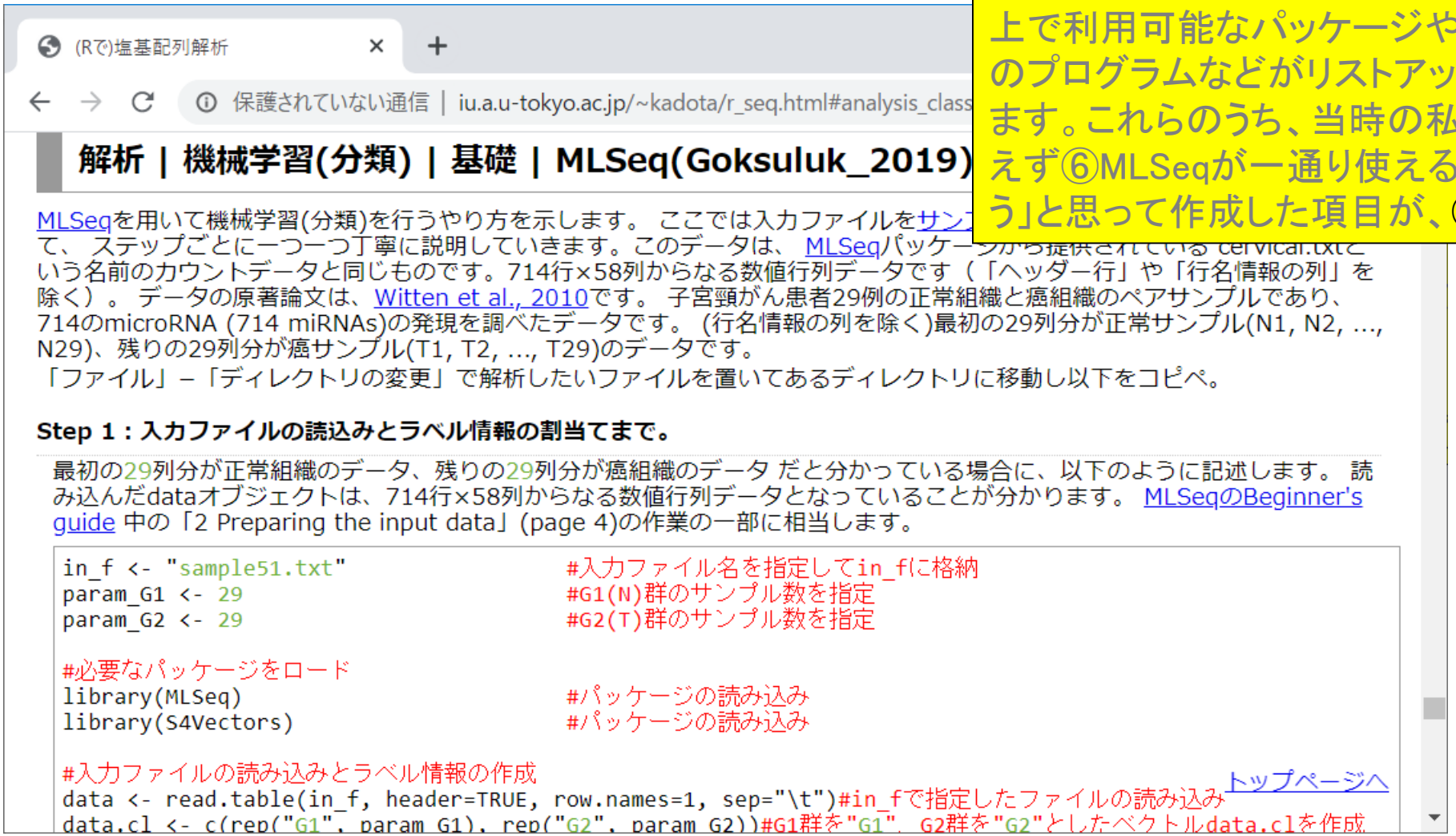
分類(診断)5



①機械学習(分類)はこのあたり。②のリンク先が、③こんな感じです。ページ下部に移動していくと、こんな感じで、④R上で利用可能なパッケージや、⑤R以外のプログラムなどがリストアップされています。これらのうち、当時の私が「とりあえず⑥MLSeqが一通り使えるようにしよう」と思って作成した項目が、⑦これ。

分類(診断)6

①機械学習(分類)はこのあたり。②のリンク先が、③こんな感じです。ページ下部に移動していくと、こんな感じで、④R上で利用可能なパッケージや、⑤R以外のプログラムなどがリストアップされています。これらのうち、当時の私が「とりあえず⑥MLSeqが一通り使えるようにしよう」と思って作成した項目が、⑦これ。



解析 | 機械学習(分類) | 基礎 | MLSeq(Goksuluk_2019)

MLSeqを用いて機械学習(分類)を行うやり方を示します。 ここでは入力ファイルをサン
て、 ステップごとに一つ一つ丁寧に説明していきます。このデータは、 MLSeqパッケージから提供されている cervical.txtと
いう名前のカウントデータと同じものです。714行×58列からなる数値行列データです（「ヘッダー行」や「行名情報の列」を
除く）。 データの原著論文は、Witten et al., 2010です。 子宮頸がん患者29例の正常組織と癌組織のペアサンプルであり、
714のmicroRNA (714 miRNAs)の発現を調べたデータです。 (行名情報の列を除く)最初の29列分が正常サンプル(N1, N2, ..., N29)、残りの29列分が癌サンプル(T1, T2, ..., T29)のデータです。
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

Step 1：入力ファイルの読込みとラベル情報の割当てまで。

最初の29列分が正常組織のデータ、残りの29列分が癌組織のデータ だと分かっている場合に、以下のように記述します。 読み込んだdataオブジェクトは、714行×58列からなる数値行列データとなっていることが分かります。 MLSeqのBeginner's guide 中の「2 Preparing the input data」(page 4)の作業の一部に相当します。

```
in_f <- "sample51.txt"                  #入力ファイル名を指定してin_fに格納
param_G1 <- 29                          #G1(N)群のサンプル数を指定
param_G2 <- 29                          #G2(T)群のサンプル数を指定

#必要なパッケージをロード
library(MLSeq)                          #パッケージの読み込み
library(S4Vectors)                      #パッケージの読み込み

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t")#in_fで指定したファイルの読み込み
data.cl <- c(rep("G1", param_G1), rep("G2", param_G2))#G1群を"G1"、G2群を"G2"としたベクトルdata.clを作成
```

トップページへ

分類(診断)7

①この項目を、2020年2月17-18日に丸2日間かけてやりました。子宮頸がん患者29例の正常組織とガン組織のペアサンプルデータを用いています。

(Rで)塩基配列解析

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

ゲスト

解析 | 機械学習(分類) | 基礎 | MLSeq(Goksuluk_2019)

MLSeqを用いて機械学習(分類)を行うやり方を示します。 ここでは入力ファイルをサンプルデータ51のsample51.txtに限定して、 ステップごとに一つ一つ丁寧に説明していきます。このデータは、 MLSeqパッケージから提供されている cervical.txtという名前のカウントデータと同じものです。714行×58列からなる数値行列データです（「ヘッダー行」や「行名情報の列」を除く）。 データの原著論文は、Witten et al., 2010です。 子宮頸がん患者29例の正常組織と癌組織のペアサンプルであり、714のmicroRNA (714 miRNAs)の発現を調べたデータです。 (行名情報の列を除く)最初の29列分が正常サンプル(N1, N2, ..., N29)、残りの29列分が癌サンプル(T1, T2, ..., T29)のデータです。
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

Step 1：入力ファイルの読込みとラベル情報の割当てまで。

最初の29列分が正常組織のデータ、残りの29列分が癌組織のデータ だと分かっている場合に、以下のように記述します。 読み込んだdataオブジェクトは、714行×58列からなる数値行列データとなっていることが分かります。 MLSeqのBeginner's guide 中の「2 Preparing the input data」(page 4)の作業の一部に相当します。

```
in_f <- "sample51.txt"                  #入力ファイル名を指定してin_fに格納
param_G1 <- 29                          #G1(N)群のサンプル数を指定
param_G2 <- 29                          #G2(T)群のサンプル数を指定

#必要なパッケージをロード
library(MLSeq)                          #パッケージの読み込み
library(S4Vectors)                      #パッケージの読み込み

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t")#in_fで指定したファイルの読み込み
data.cl <- c(rep("G1", param_G1), rep("G2", param_G2))#G1群を"G1"、G2群を"G2"としたベクトルdata.clを作成
```

トップページへ