

バイオインフォマティクス ～RNA-seq発現解析～

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/>

計3回の予定講義内容

■ バイオインフォマティクス: 概論とRの基礎(1/25)

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

■ バイオインフォマティクス: Rパッケージの話(2/22)

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お話。

■ バイオインフォマティクス: 解析結果の解釈など(3/15)

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。

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



①

門田 幸二のホームページ

②

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・論文など](#) (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**

③

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に分けていますが、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

講義 NEW

- [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、2014-2018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、分担)
- [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; 2016-2017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①](#)」 [C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)

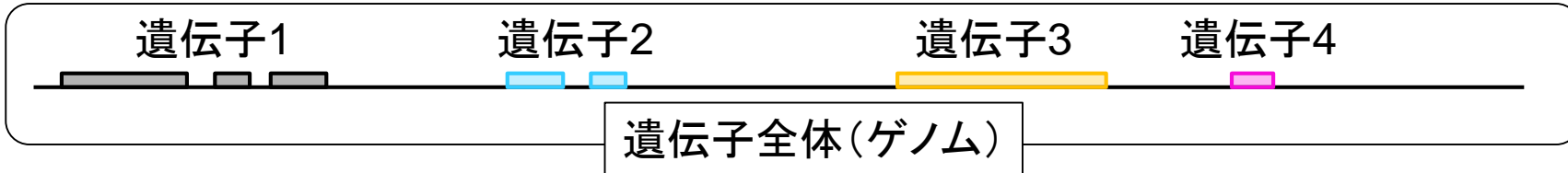
①

Contents

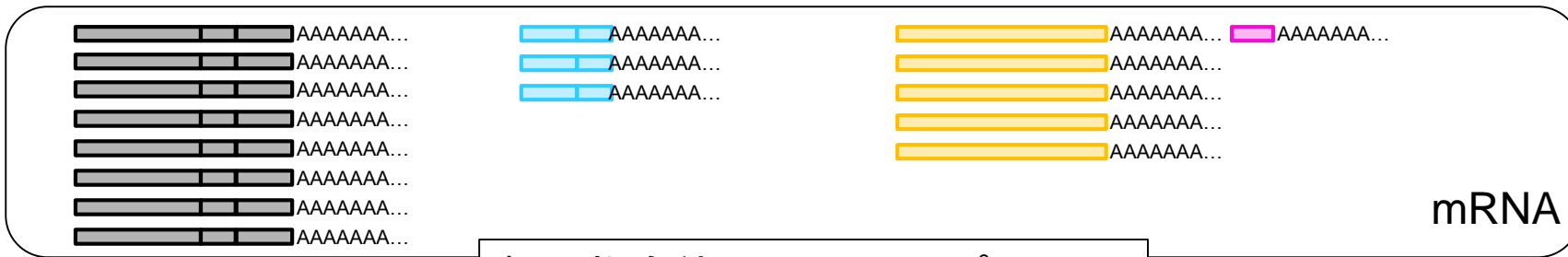
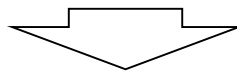
- トランスクリプトーム (RNA-seq) 解析の原理、データ解析戦略のイメージ
- 公共カウントデータの取得 (recount2)
- サンプル間クラスタリング
 - 手元のデータを実行、結果の解釈
 - グループ間の分離度を客観的に示すスコア
- TCC-GUI
 - ファイルのアップロード、グループラベル情報の付与、平均シルエットスコア (AS値)
 - 探索的解析 (Exploratory Analysis) : 階層的クラスタリングやPCAなど
 - 発現変動解析 (TCC Computation)
- すぐにDisconnectedとなる問題とその対策
 - RStudioのイン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

トランスクリプトーム解析とは

- ある状態のあるサンプル(例:目)のあるゲノムの領域



・どの染色体上のどの領域にどの遺伝子があるかは調べる個体(例:ヒト)が同じなら不変(目だろうが心臓だろうが...)



mRNA

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

働いているRNAの種類
や量を調べるのが目的

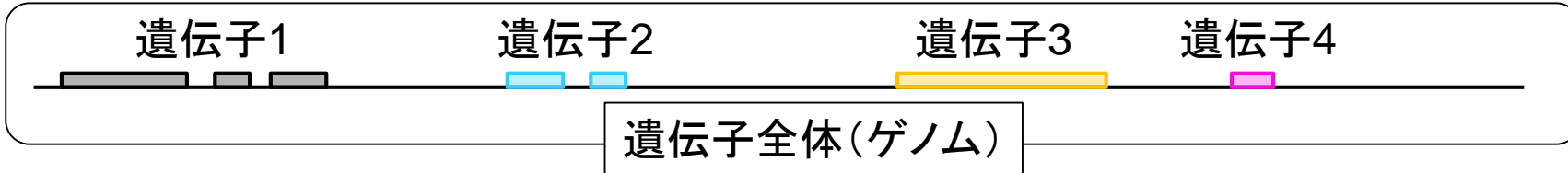
光刺激

ヒト

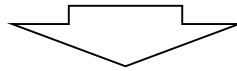


トランスクリプトーム解析とは

- ある状態のあるサンプル(例:目)のあるゲノムの領域



・どの染色体上のどの領域にどの遺伝子があるかは調べる個体(例:ヒト)が同じなら不変(目だろうが心臓だろうが...)

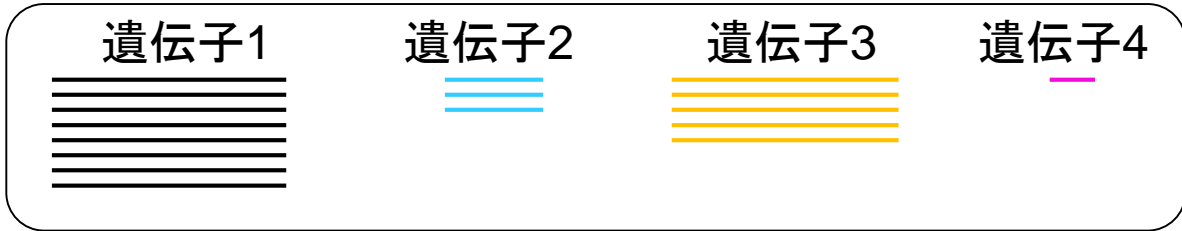


- ・遺伝子2は光刺激に应答して発現亢進
- ・遺伝子4も光刺激に应答して発現亢進

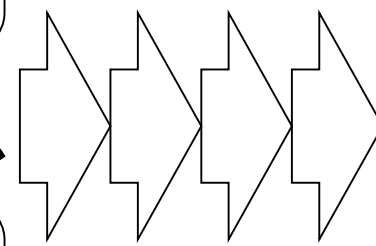
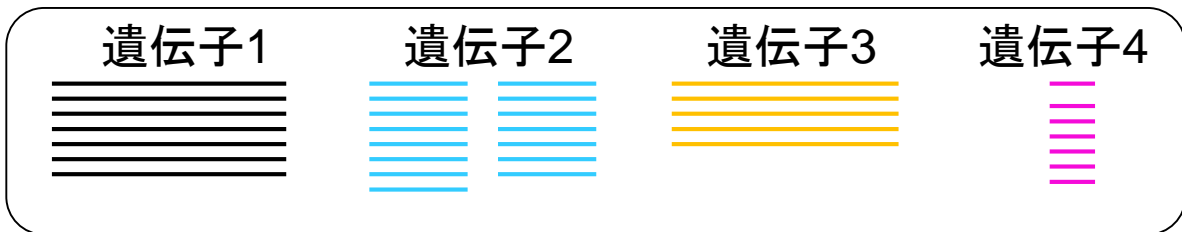
状態の異なる複数サンプルのデータを取得して解析するのが一般的。サンプル間比較

トランスクリプトーム解析

■ 光刺激前 (T1) の目のトランスクリプトーム



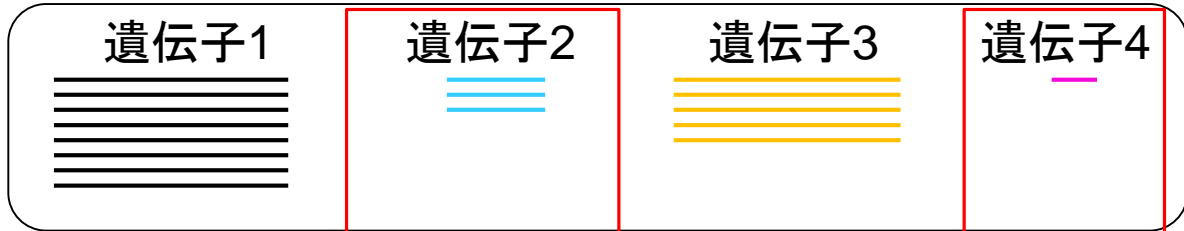
■ 光刺激後 (T2) の目のトランスクリプトーム



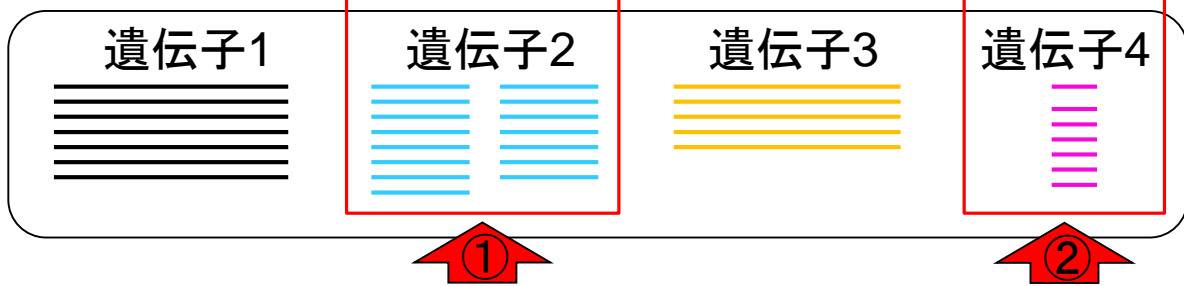
	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
...

トランスクリプトーム解析

■ 光刺激前 (T1) の目のトランスクリプトーム



■ 光刺激後 (T2) の目のトランスクリプトーム



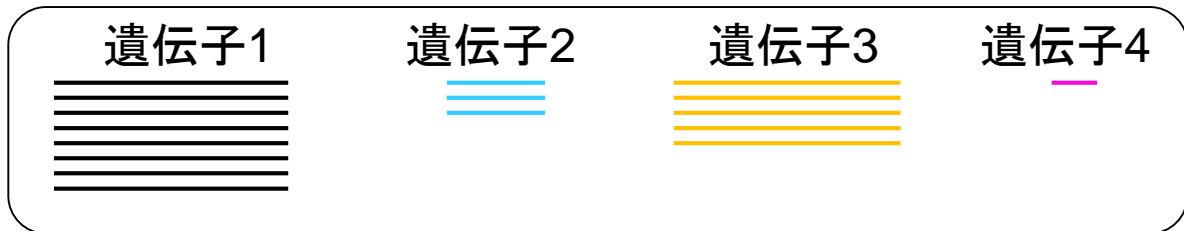
これがいわゆる
「遺伝子発現行列」

	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
...

トランスクリプトームデータ取得

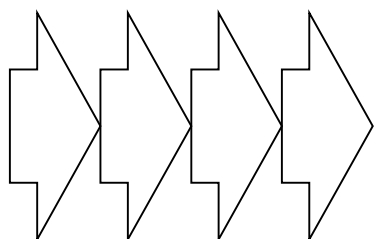
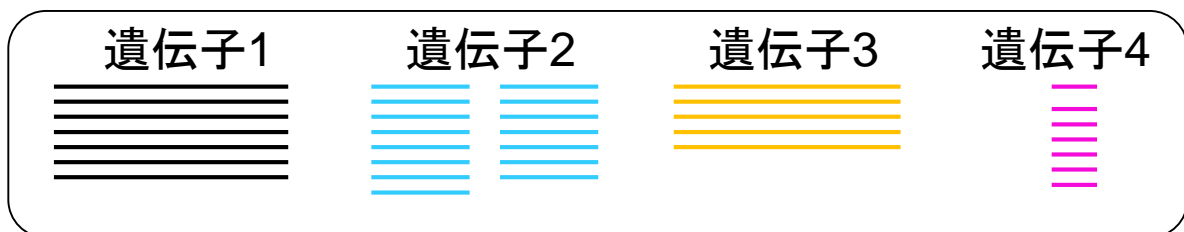
21世紀初頭によく利用されたのがマイクロアレイ。
今でも利用されています

■ 光刺激前 (T1) の目のトランスクリプトーム

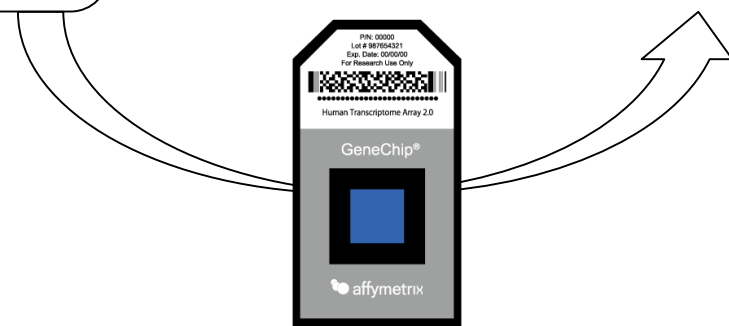


これがいわゆる
「遺伝子発現行列」

■ 光刺激後 (T2) の目のトランスクリプトーム



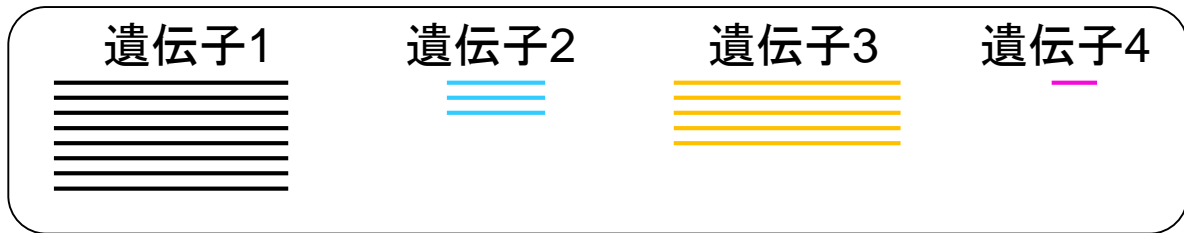
	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
...



現在はNGSの利用が主流。
NGSを用いたRNAの配列決定
(sequencing)なので、RNA-seq

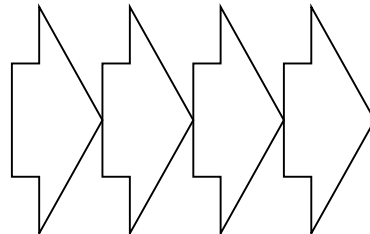
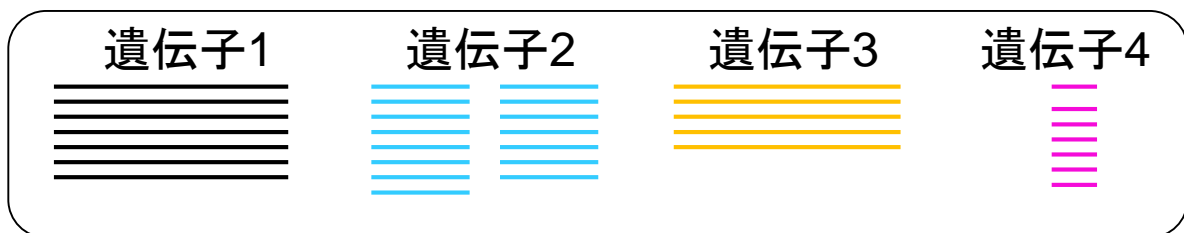
トランスクリプトームデータ取得

■ 光刺激前 (T1) の目のトランスクリプトーム

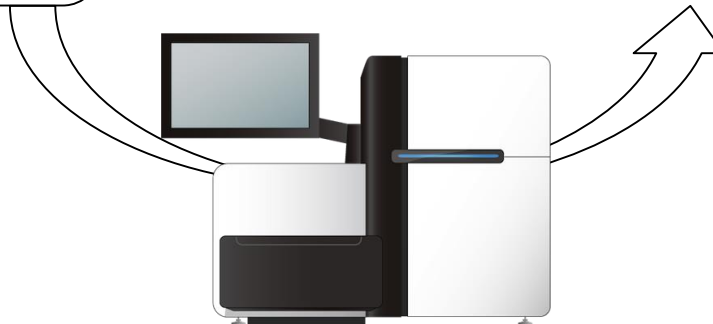


これがいわゆる
「遺伝子発現行列」

■ 光刺激後 (T2) の目のトランスクリプトーム



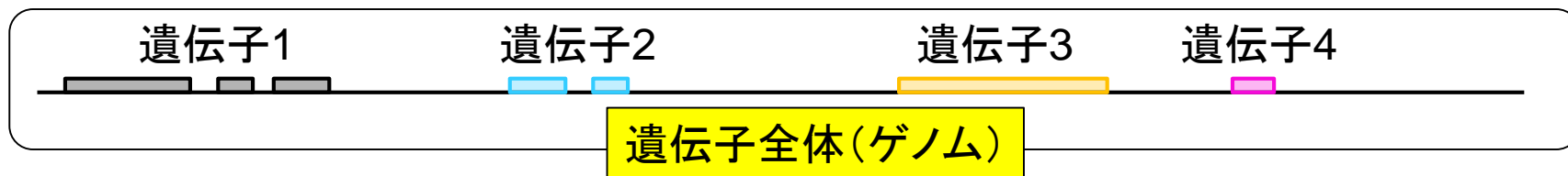
	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
...



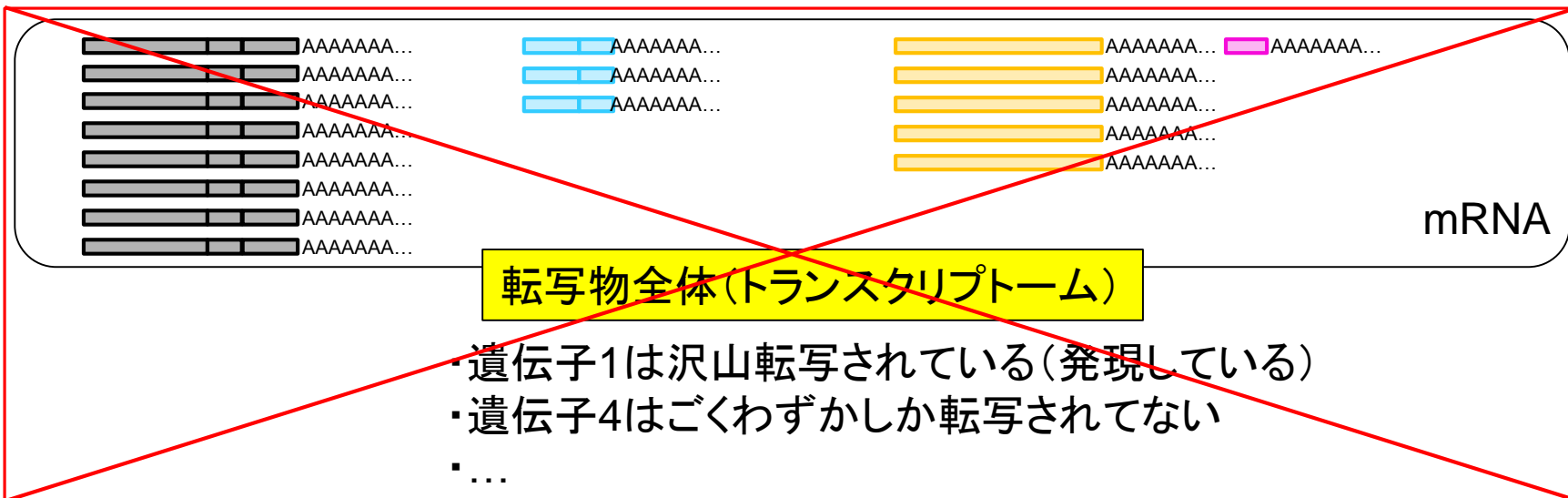
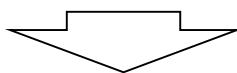
遺伝子 ≠ 転写物

赤枠部分の表現は、本当は不正確。昔は実験機器の解像度が事実上遺伝子レベルだった。遺伝子発現解析という表現はその名残り

- ある状態のあるサンプル(例:目)のあるゲノムの領域



- ・どの染色体上のどの領域にどの遺伝子があるかは調べる個体(例:ヒト)が同じなら不変(目だろうが心臓だろうが...)



ある遺伝子領域から転写 (transcription) されている転写物 (transcript) は、1種類とは限らない

遺伝子 ≠ 転写物

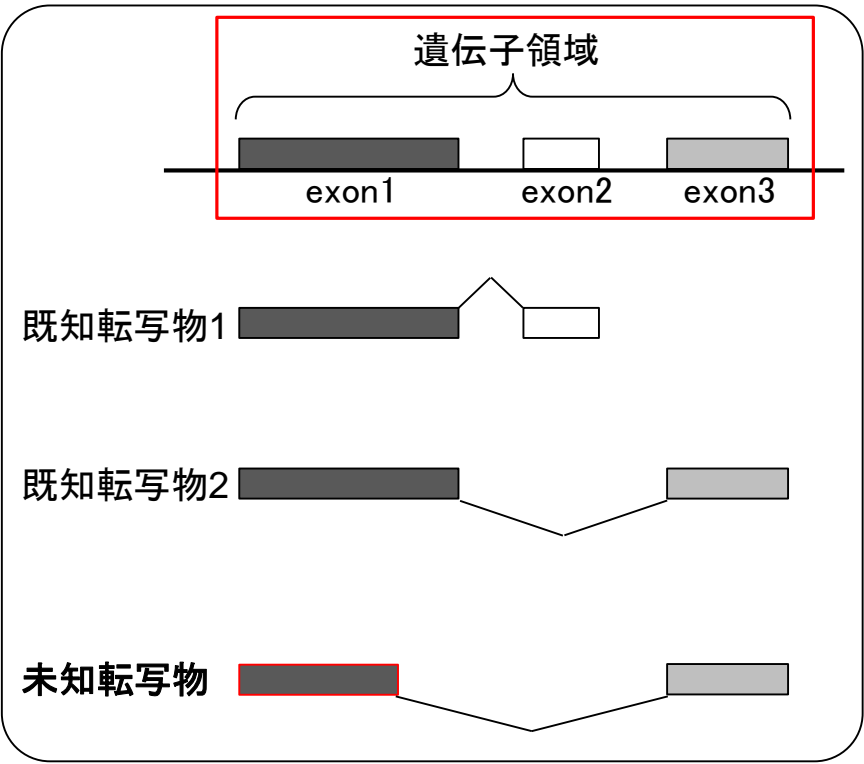
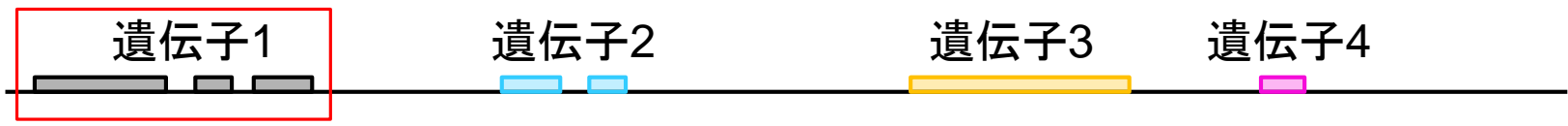
- ある状態のあるサンプル (例: 目) のあるゲノムの領域



例えば、遺伝子1の領域では、3種類の真の転写物が存在し、そのうち2種類は既知とする

遺伝子 ≠ 転写物

- ある状態のあるサンプル(例:目)のあるゲノムの領域

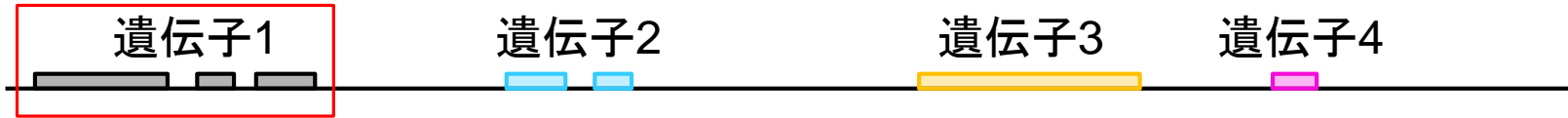


真の転写物情報

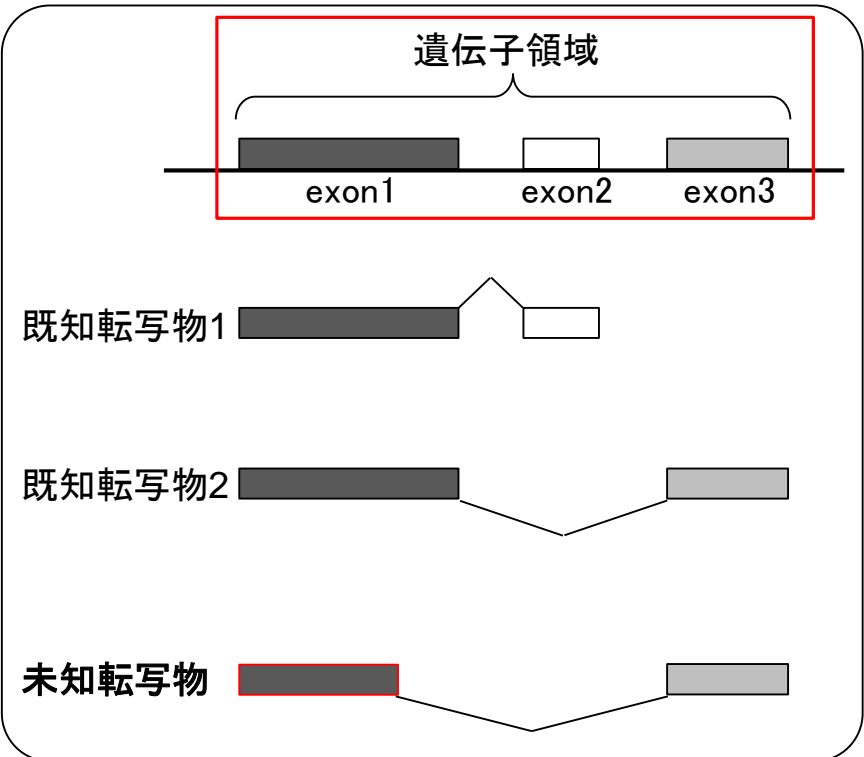
実際の細胞内(例:目のサンプル)での発現情報(働いている度合い)が①のような感じだったとする

遺伝子 ≠ 転写物

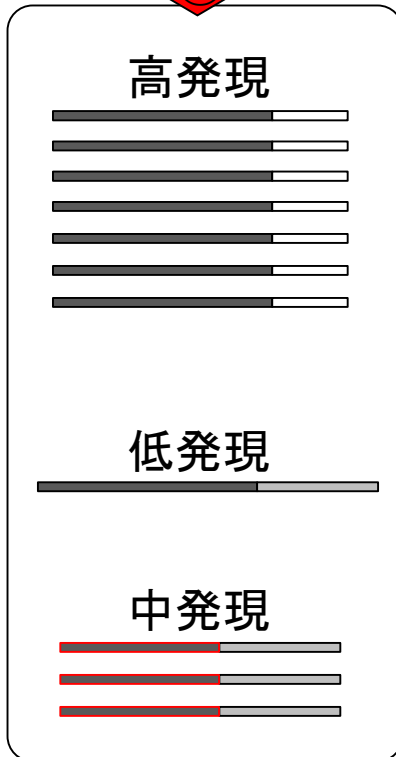
- ある状態のあるサンプル(例:目)のあるゲノムの領域



①



真の転写物情報



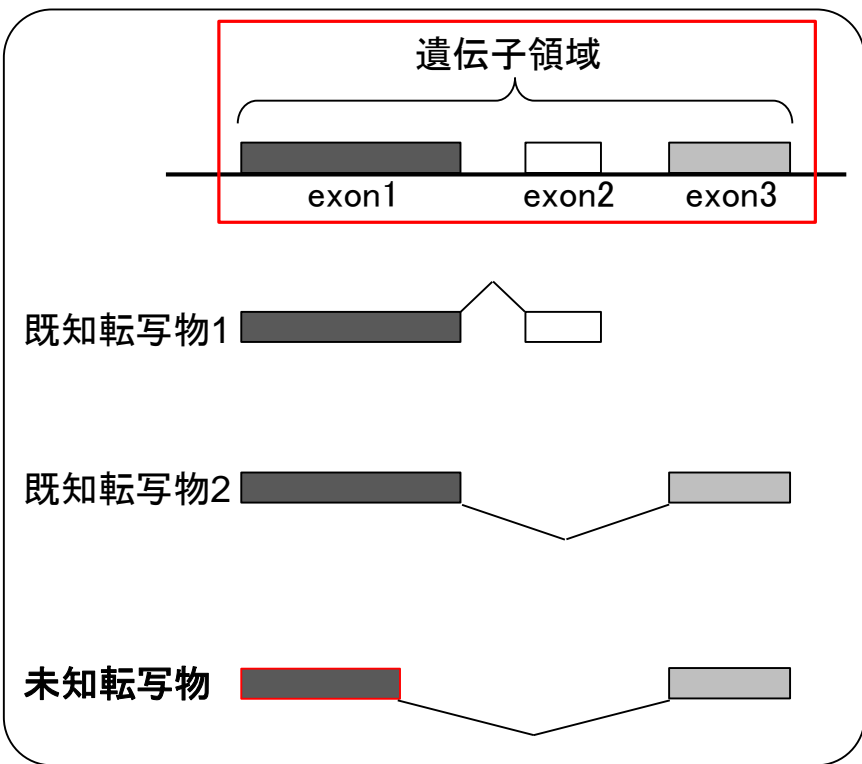
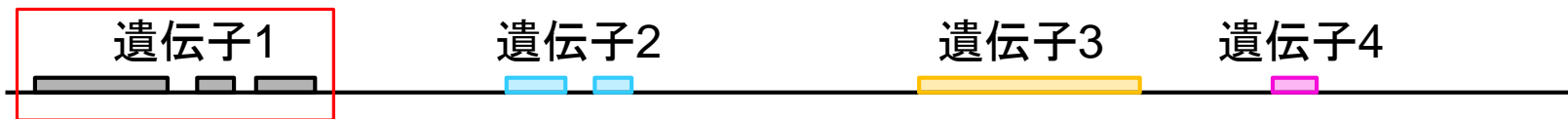
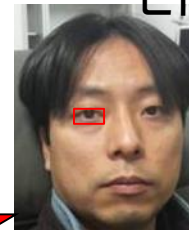
真の発現情報

遺伝子 ≠ 転写物

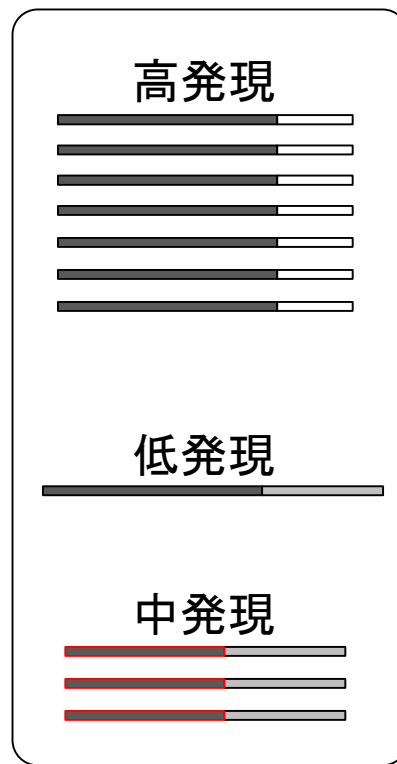
- ある状態のあるサンプル(例:目)のあるゲノムの領域

妄想

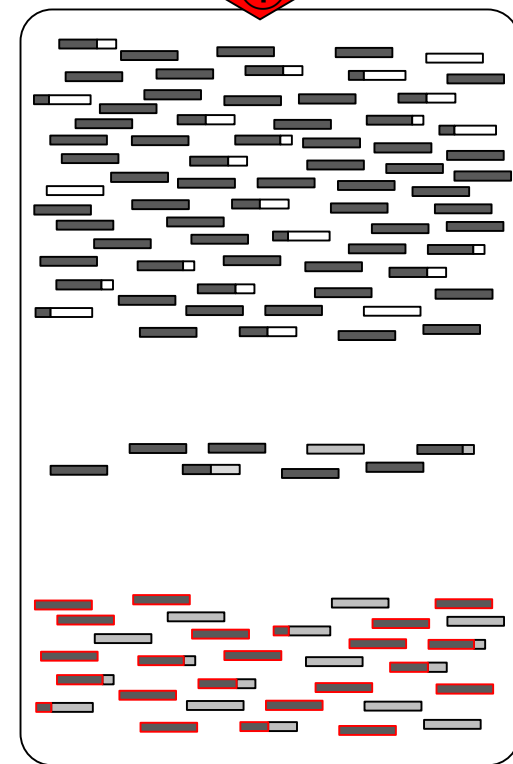
ヒト



真の転写物情報



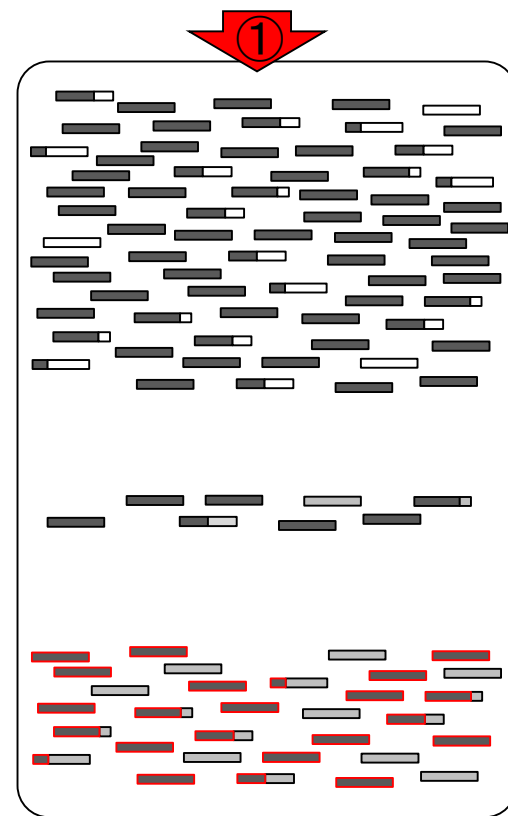
真の発現情報



RNA-seqで得られるリード情報 (色は不明)

データ解析の出発点

トランスクリプトーム (RNA-seq) データ解析の出発点は、①RNA-seqデータファイル、



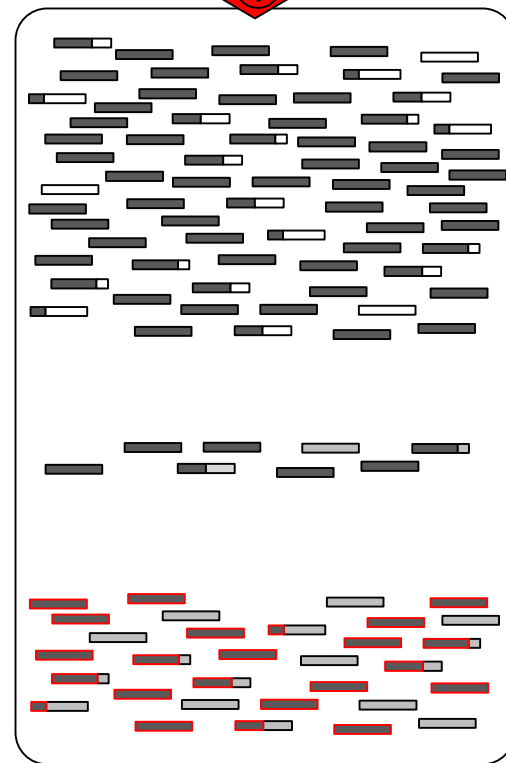
RNA-seqデータ

データ解析の出発点

トランスクリプトーム (RNA-seq) データ解析の出発点は、①RNA-seqデータファイル、②ゲノム配列情報、

②

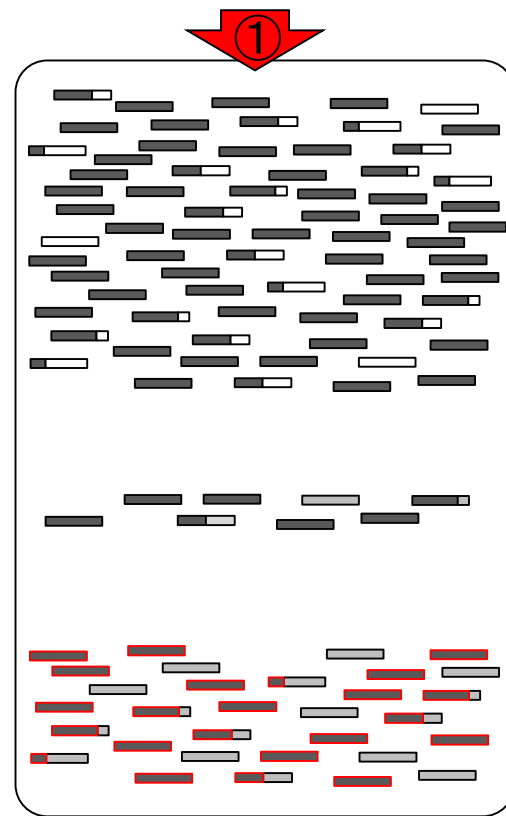
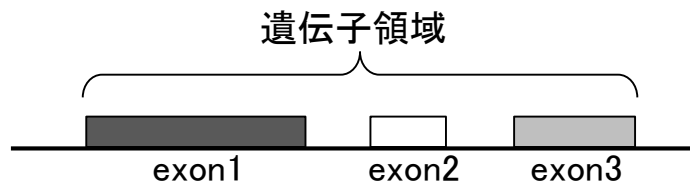
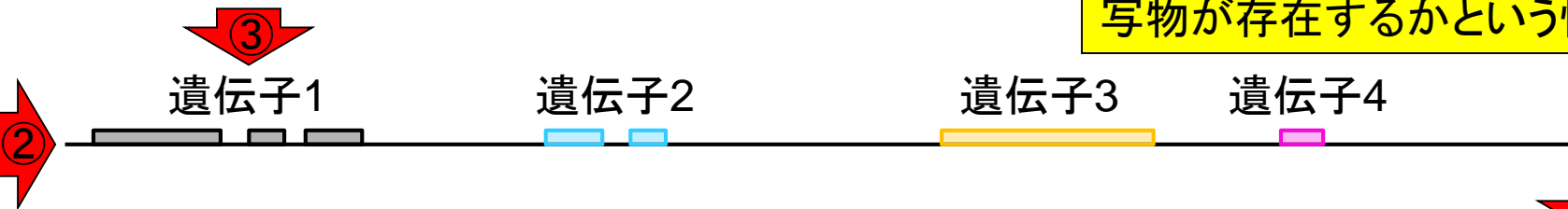
①



RNA-seqデータ

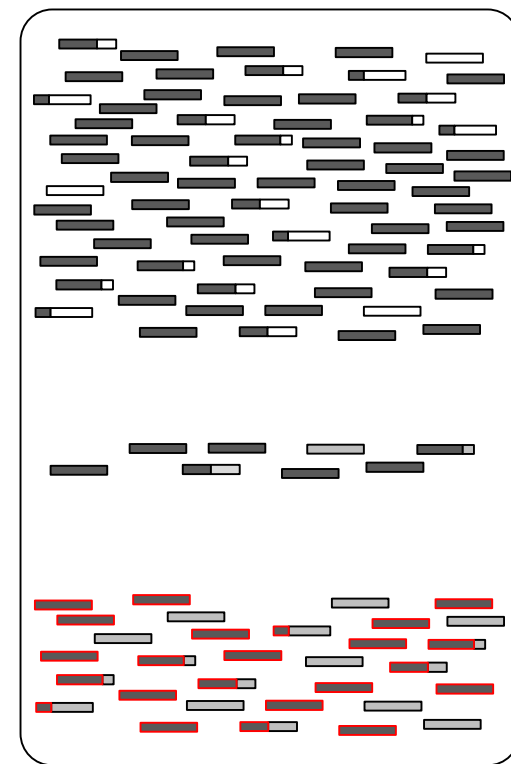
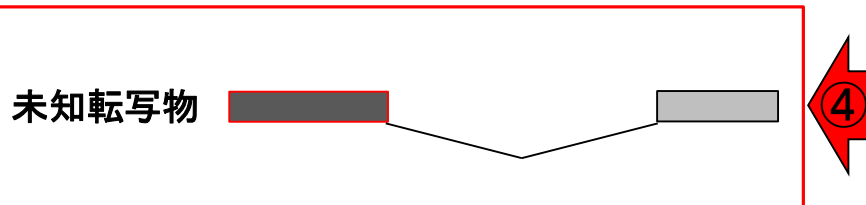
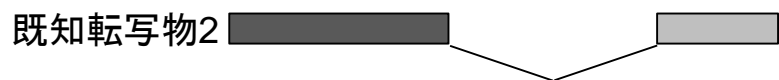
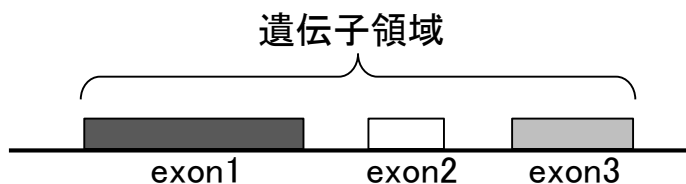
データ解析の出発点

トランスクリプトーム (RNA-seq) データ解析の出発点は、①RNA-seqデータファイル、②ゲノム配列情報、③アノテーション情報(ゲノム上のどこにどんな遺伝子、exon、転写物が存在するかという情報)



解析結果のイメージ

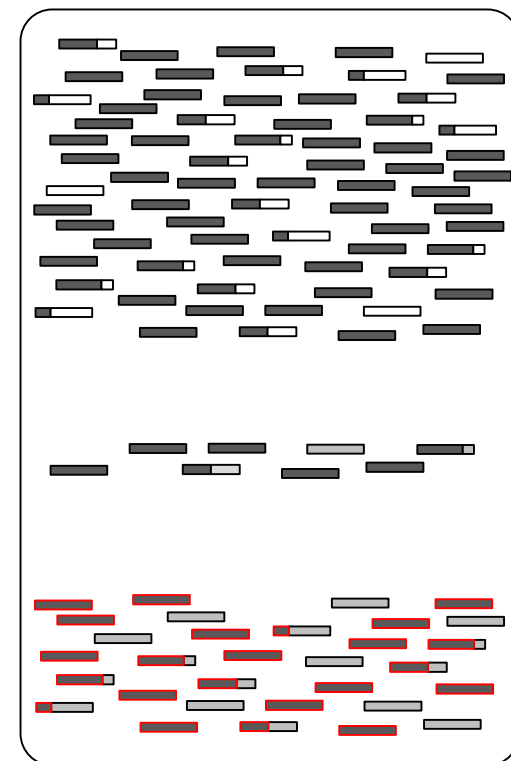
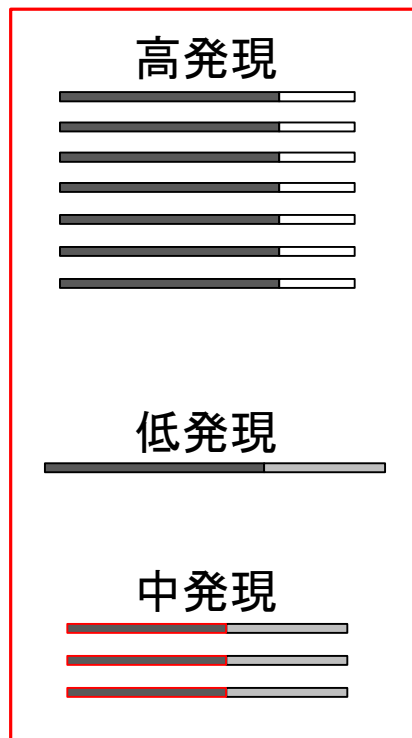
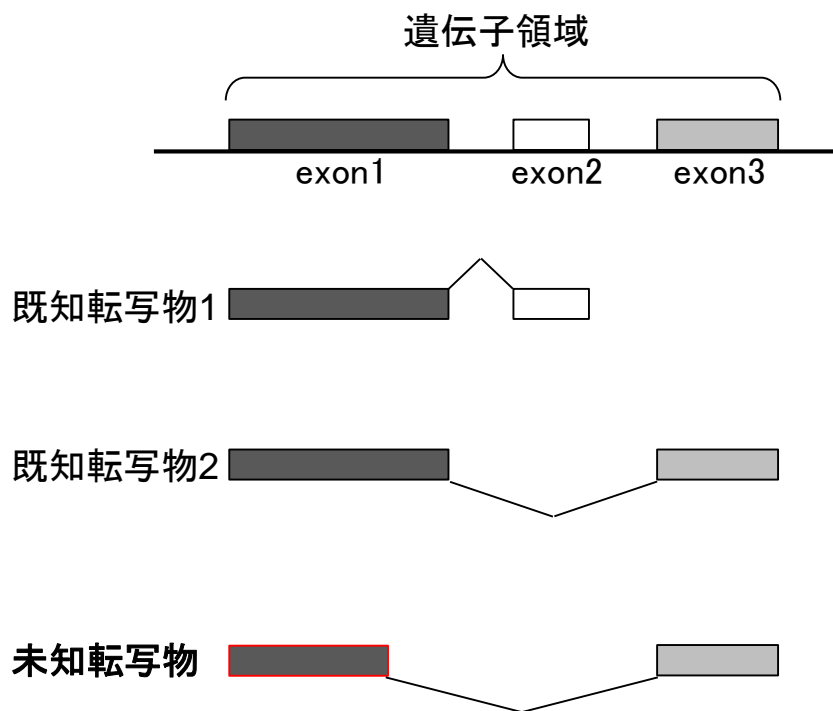
①RNA-seqデータ、②ゲノム配列情報、③アノテーション情報を利用して、④未知転写物(新規isoform)の同定ができる。



RNA-seqデータ

解析結果のイメージ

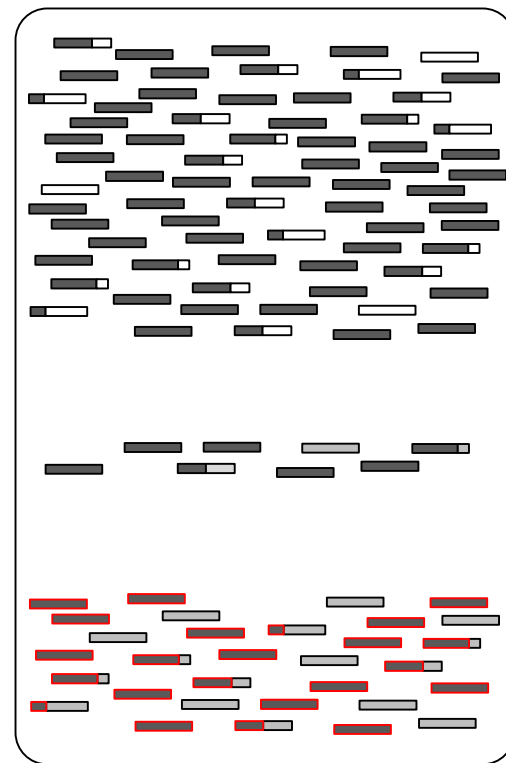
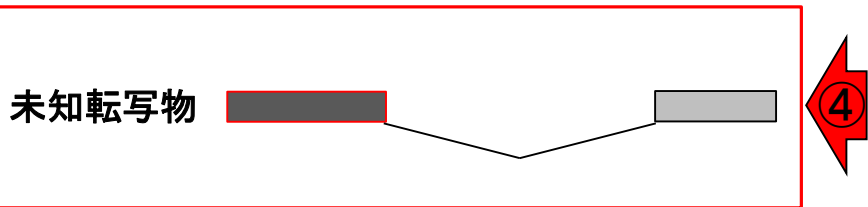
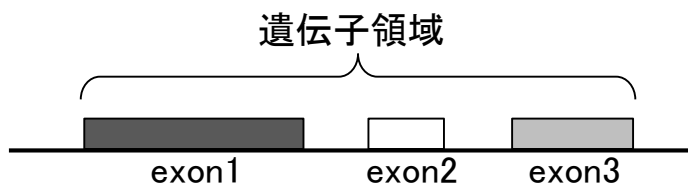
①RNA-seqデータ、②ゲノム配列情報、③アノテーション情報を利用して、④未知転写物(新規isoform)の同定ができる。⑤転写物の発現量(働いている度合い)推定も原理的に可能



RNA-seqデータ

具体的な戦略は？

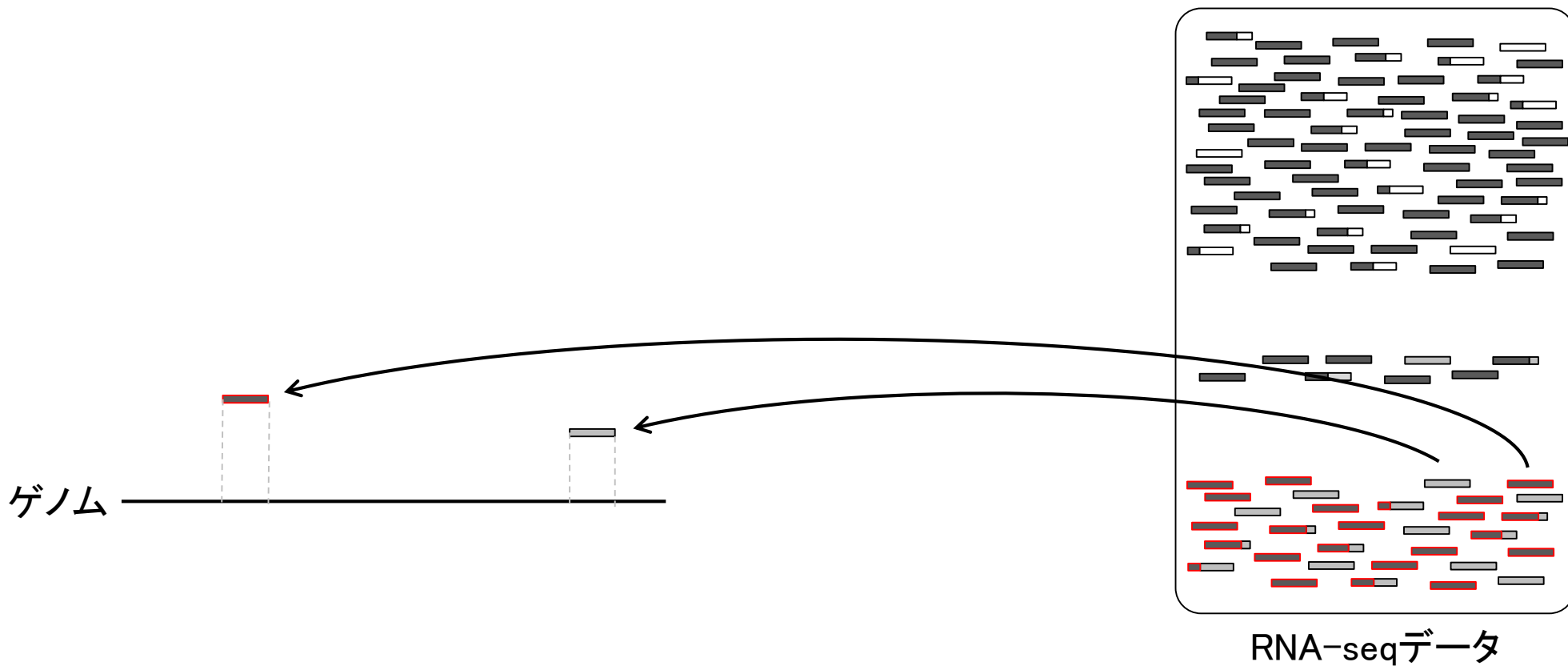
①RNA-seqデータ、②ゲノム配列情報、③アノテーション情報を利用して、④未知転写物(新規isoform)の同定ができる。



RNA-seqデータ

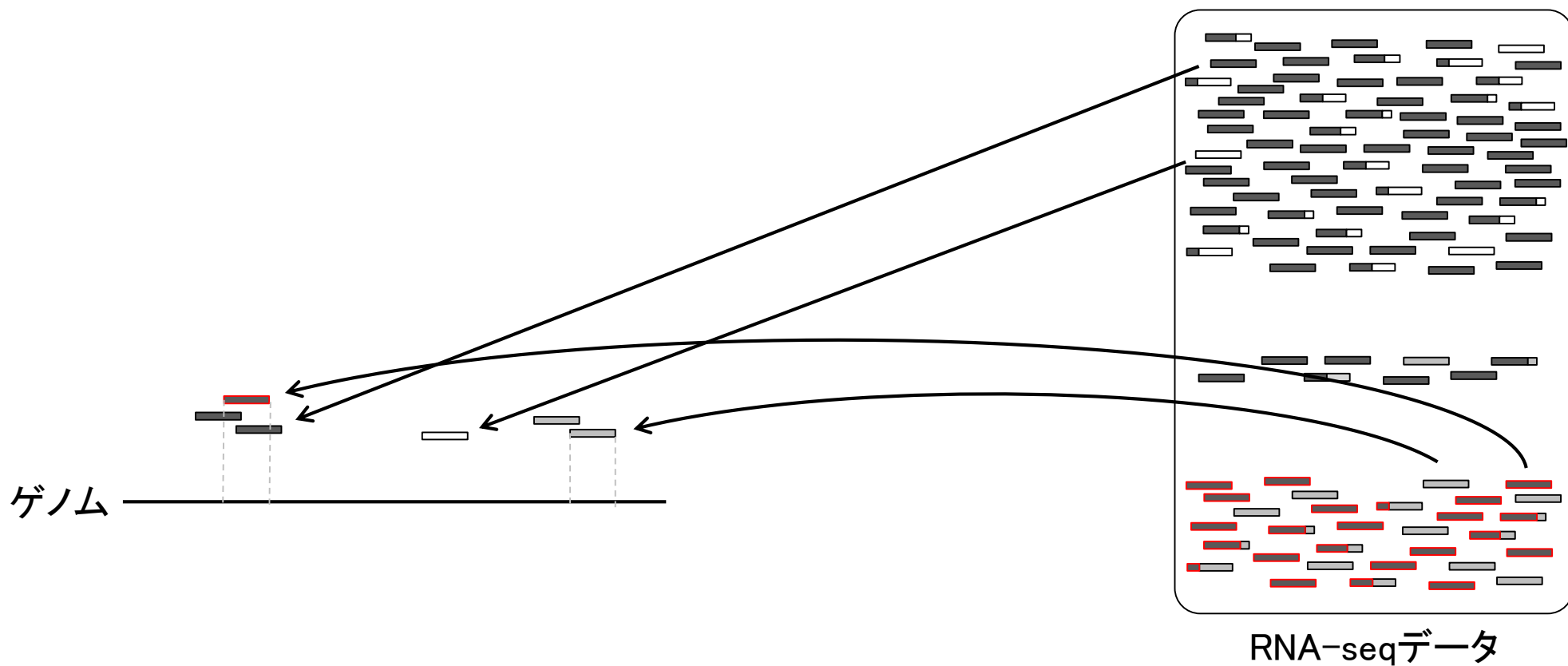
具体的な戦略

RNA-seqデータ中の1本1本のリード(横棒)がゲノム上のどの領域から転写されたのかを調べる。文字列検索と本質的に同じであり、これがマッピングという作業に相当する



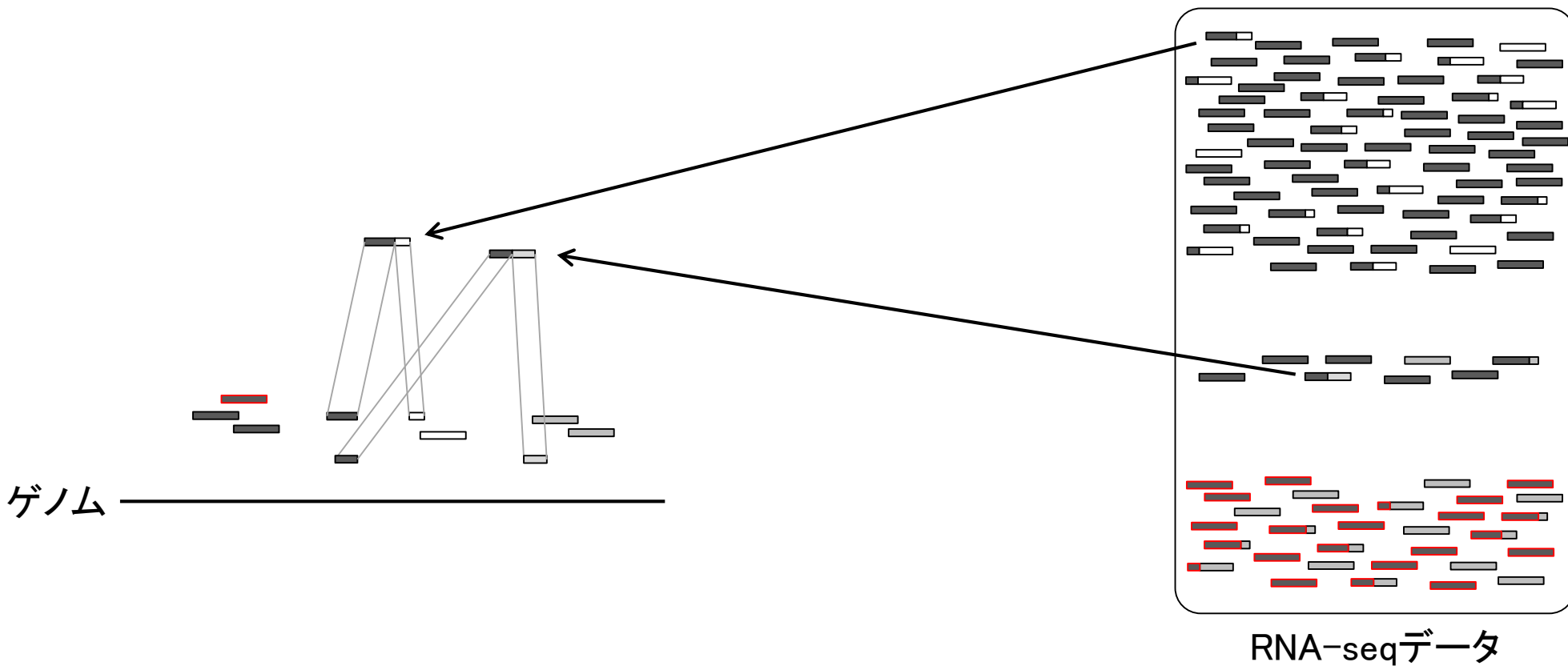
具体的な戦略

RNA-seqデータ中の1本1本のリード(横棒)がゲノム上のどの領域から転写されたのかを調べる。文字列検索と本質的に同じであり、これがマッピングという作業に相当する



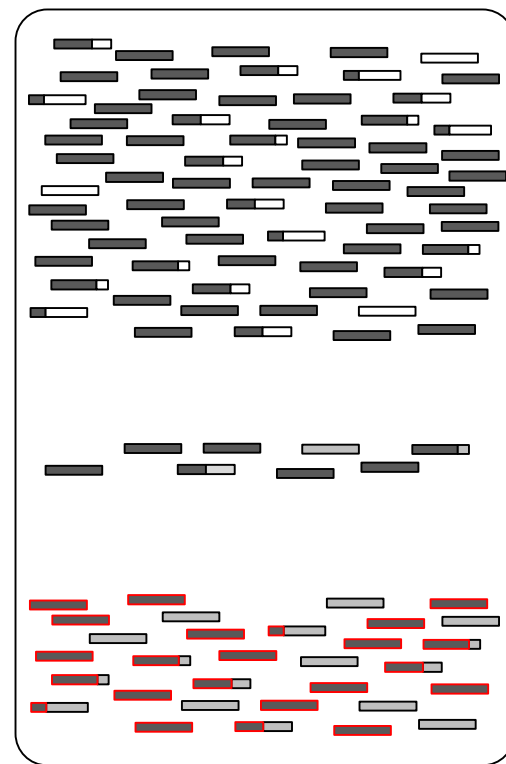
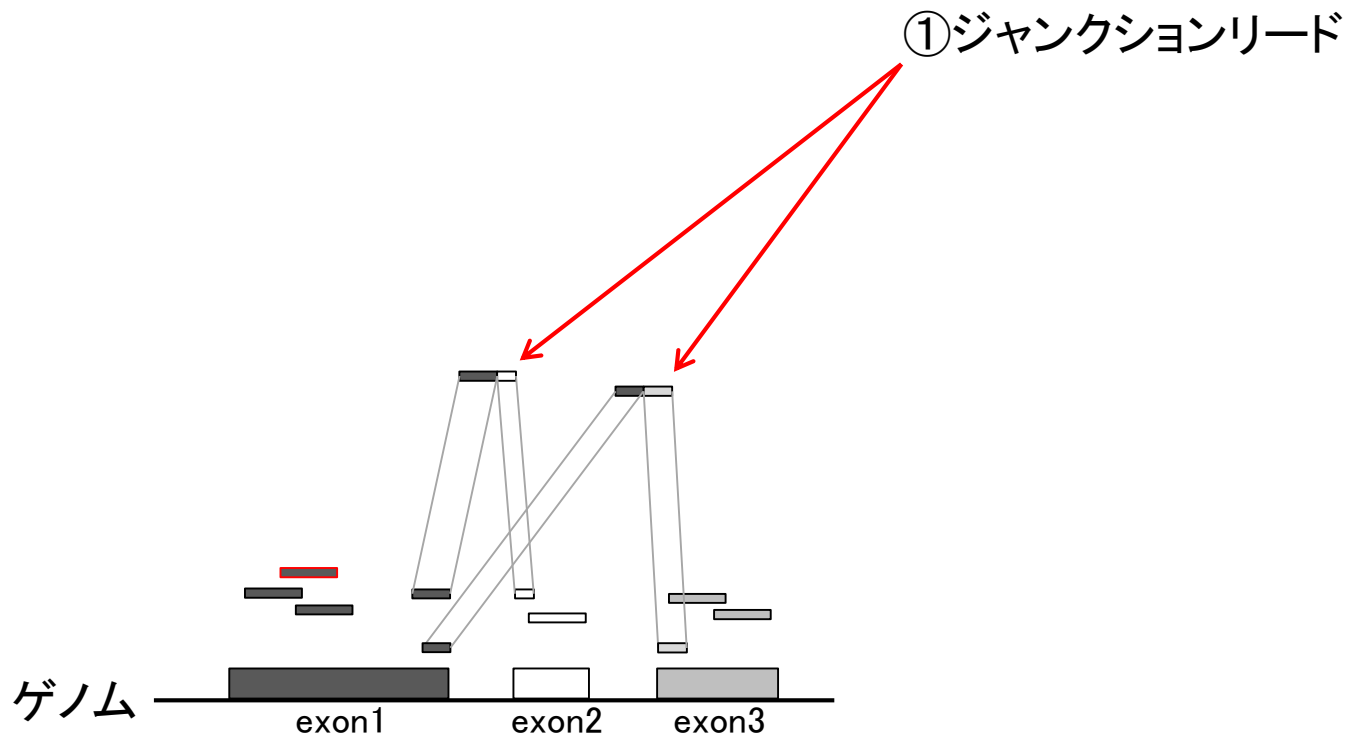
具体的な戦略

リードの長さが初期は35塩基程度だったが、現在は150塩基程度まで伸びている。そのおかげで、リードを分割してマップすることもできる



具体的な戦略

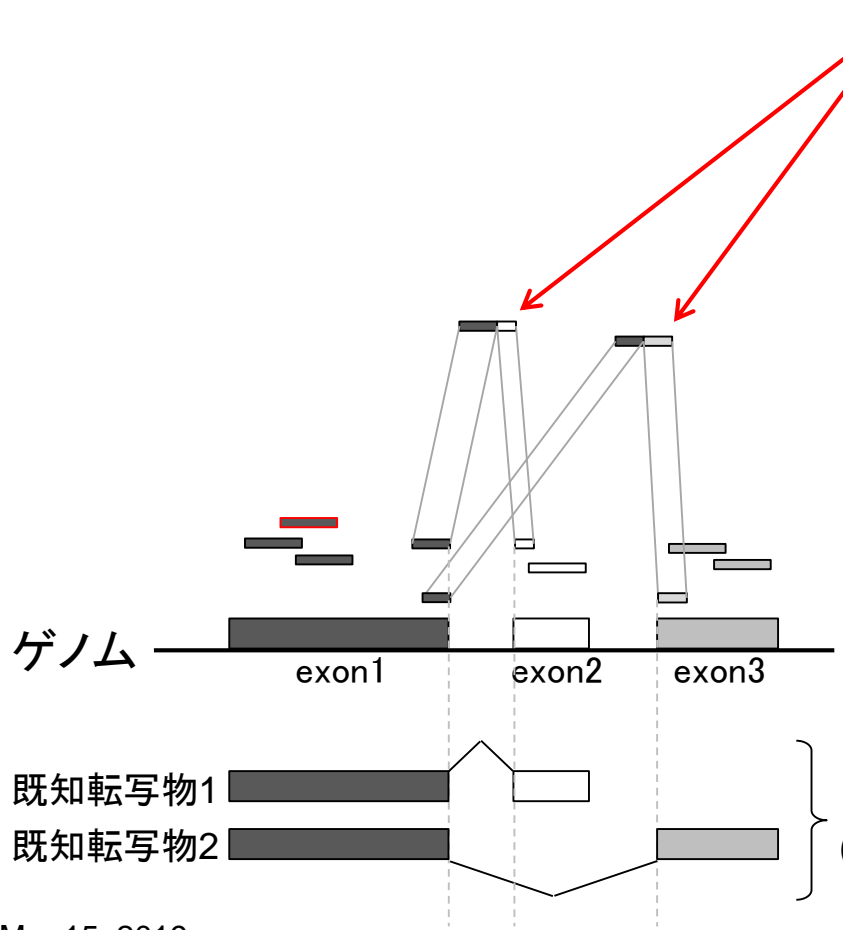
分割してマップされたリードは、大抵の場合複数のエクソン(exon)をまたぐリードであり、①ジャンクションリード(junction read)と呼ばれる



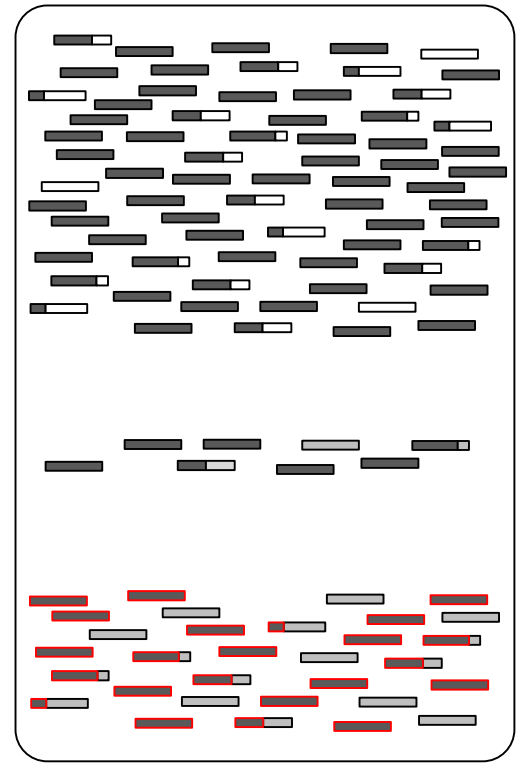
RNA-seqデータ

具体的な戦略

①ジャンクションリード



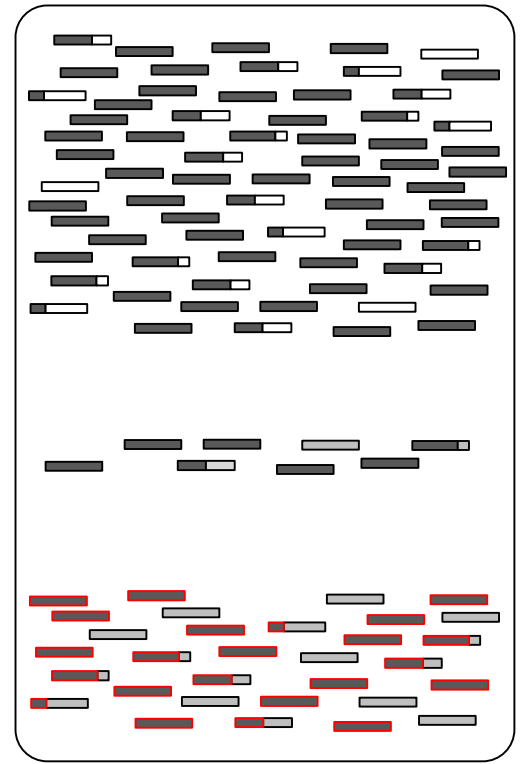
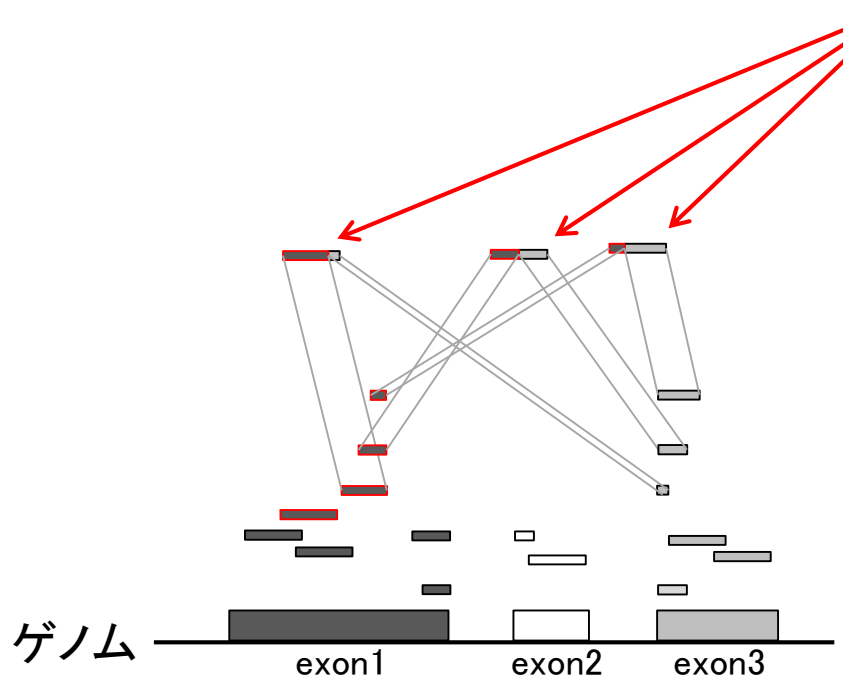
アノテーション情報
(既知遺伝子座標情報)



同様にして、他のジャンクションリードも既知転写物と比較することで…

具体的な戦略

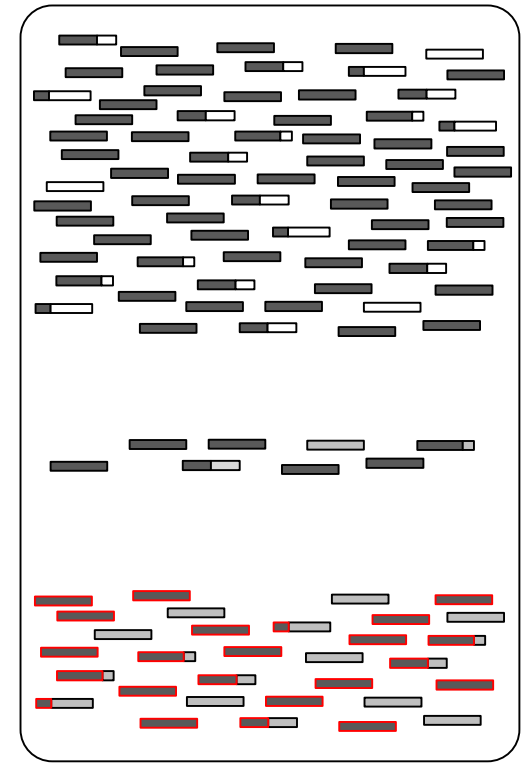
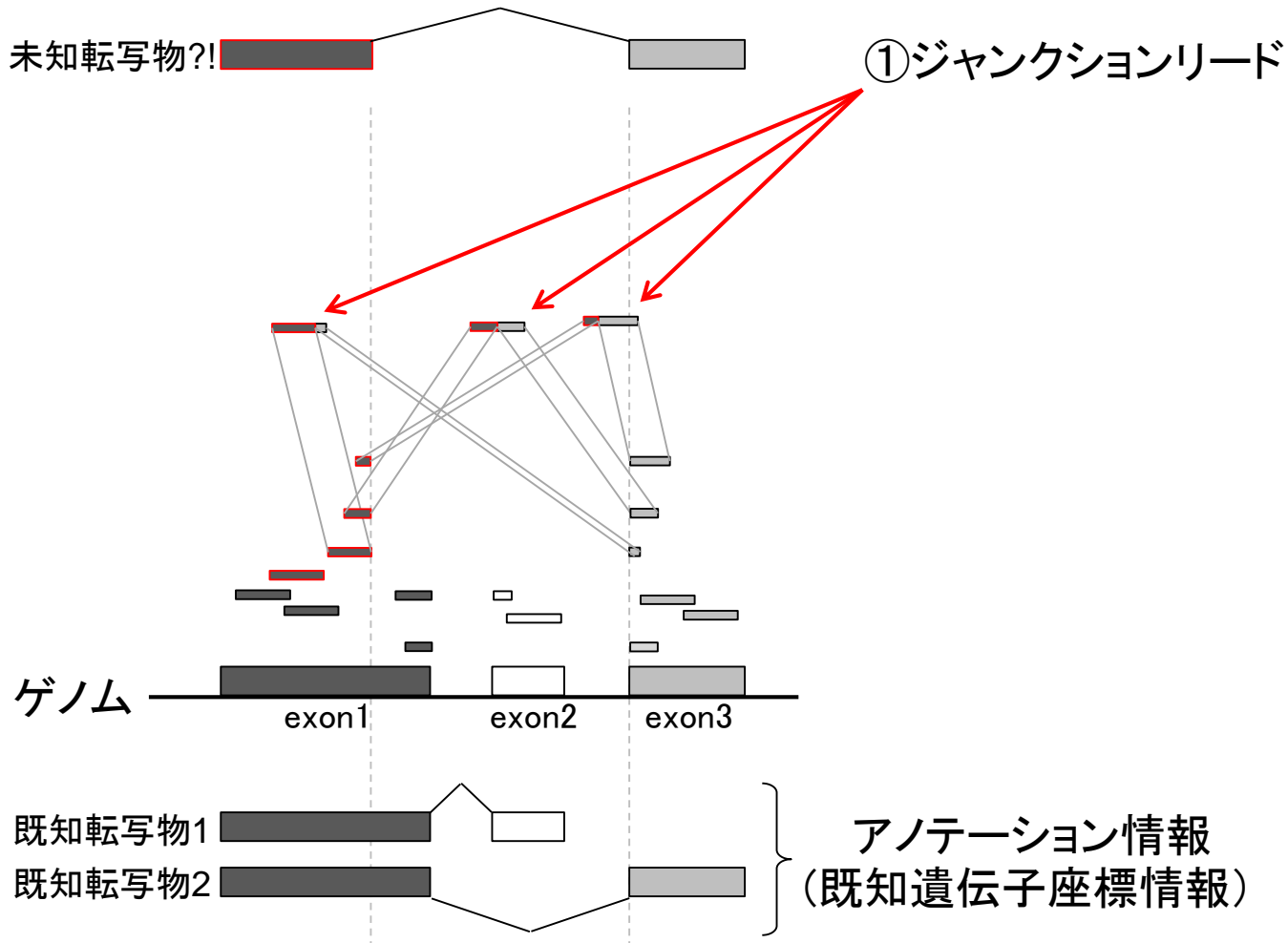
①ジャンクションリード



アノテーション情報
(既知遺伝子座標情報)

RNA-seqデータ

具体的な戦略



RNA-seqデータ

新規転写物同定の例

RNA-seq(トランスクリプトーム解析)は、癌でよくみられる融合遺伝子の検出などにも利用されます。理由:そこそこ発現している転写物は原理的に検出可能だから。肺がんでみられるALK融合遺伝子(fusion gene)は有名な例ですが、それ以外の①新たな融合遺伝子の発見などに役立っています。主に「トランスクリプトーム配列解析」の話

The screenshot shows a PubMed search result for the article: "Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data." The article is from Genome Biol. (2015) 16:7. The abstract describes a method called TRUP for identifying chimeric transcripts in cancer specimens using RNA-seq data. The page includes a "Full text links" section with a BioMed Central PMC Full text link, a "Save items" section with an "Add to Favorites" button, and a "Similar articles" section with three related articles. A red arrow with the number 1 points to the PMC Full text link.



qデータ

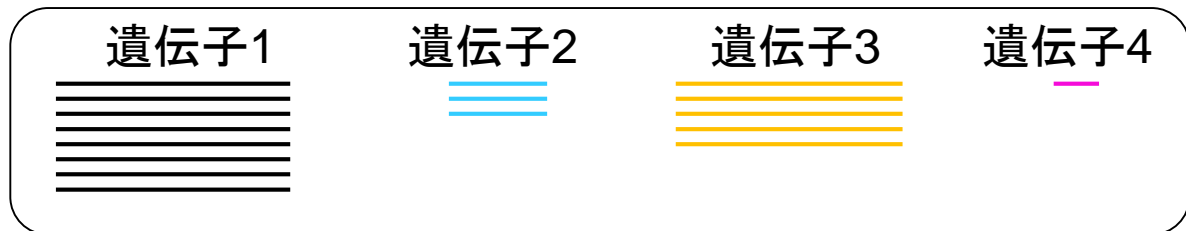
Contents

- 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

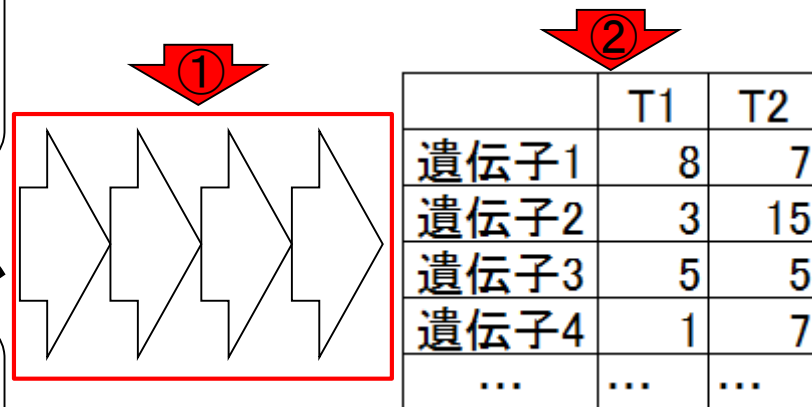
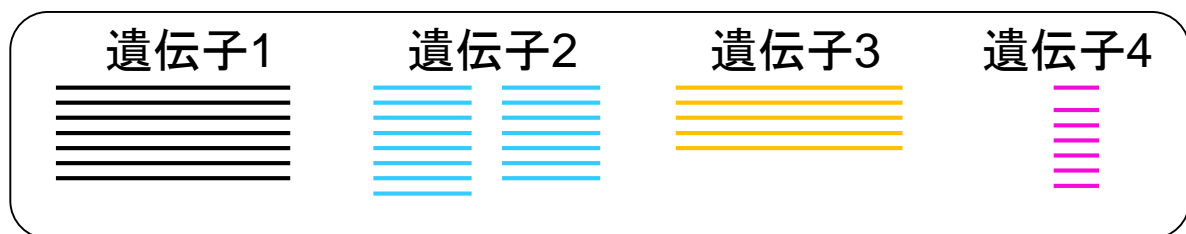
公共カウントデータ

現実問題として、①のところ(データ取得、マッピング、マップされたリード数のカウント)が大変。しかし、公共RNA-seq生データを②の状態にまで行ったものを提供してくれているサイトがあります。それが...

■ 光刺激前 (T1) の目のトランスクリプトーム



■ 光刺激後 (T2) の目のトランスクリプトーム



recount2

①recount2というウェブサイト。②原著論文はコチラ。③TCGAなど、ヒトのRNA-seqカウントデータに特化しているので、医学部のヒトにとって特に有用だと思われます。

recount2: analysis-ready RNA-seq × +
← → ↻ 🏠 ⓘ https://jhubiostatistics.shinyapps.io/recount/ ★ 人 ⓘ ⋮

recount2: analysis-ready RNA-seq gene and exon counts datasets

Datasets Popular datasets GTEx TCGA Documentation Download data with R Accessing recount2 via SciServer Contribute your data



Transcript counts are now available thanks to the work of Fu et al, bioRxiv, 2018. Exon counts are now from disjoint exons (v2) instead of reduced ones (v1). Check the Documentation tab for further information. ✕



A multi-experiment resource of analysis-ready RNA-seq gene and exon count datasets

recount2 is an online resource consisting of RNA-seq gene and exon counts as well as coverage bigWig files for 2041 different studies. It is the second generation of the [ReCount project](#). The raw sequencing data were processed with [Rail-RNA](#) as described in the [recount2 paper](#) and at [Nellore et al, Genome Biology, 2016](#) which created the coverage bigWig files. For ease of statistical analysis, for each study we created count tables at the gene and exon levels and extracted phenotype data, which we provide in their raw formats as well as in RangedSummarizedExperiment R objects (described in the [SummarizedExperiment](#) Bioconductor package). We also computed the mean coverage per study and provide it in a bigWig file, which can be used with the [derfinder](#) Bioconductor package to perform annotation-agnostic differential expression analysis at the expressed regions-level as described at [Collado-Torres et al, Nucleic Acids Research, 2017](#). The count tables, RangedSummarizeExperiment objects, phenotype tables, sample bigWigs, mean bigWigs, and file information tables are ready to use and freely available here. We also created the [recount](#) Bioconductor package which allows you to search and download the data for a specific study. By taking care of several preprocessing steps and combining many datasets into one easily-accessible website, we make finding and analyzing RNA-seq data considerably more straightforward.



Main publication

- [Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. *Reproducible RNA-seq analysis using recount2. Nature Biotechnology, 2017.* doi: 10.1038/nbt.3838.](#)

recount2

①recount2というウェブサイト。②原著論文はコチラ。③TCGAなど、ヒトのRNA-seqカウントデータに特化しているので、医学部のヒトにとって特に有用だと思われれます。まずは細かい説明はすっ飛ばして、すい臓のデータを探してみます。④少しページ下部に移動。

recount2: analysis-ready RNA-seq × +
← → ↻ 🏠 ⓘ https://jhubiostatistics.shinyapps.io/recount/

recount2: analysis-ready RNA-seq gene and exon counts datasets

Datasets Popular datasets GTEx TCGA Documentation Download data with R Accessing recount2 via SciServer Contribute your data

Transcript counts are now available thanks to the work of Fu et al, bioRxiv, 2018. Exon counts are now from disjoint exons (v2) instead of reduced ones (v1). Check the Documentation tab for further information. ×

A multi-experiment resource of analysis-ready RNA-seq gene and exon count datasets

recount2 is an online resource consisting of RNA-seq gene and exon counts as well as coverage bigWig files for 2041 different studies. It is the second generation of the [ReCount project](#). The raw sequencing data were processed with [Rail-RNA](#) as described in the [recount2 paper](#) and at [Nellore et al, Genome Biology, 2016](#) which created the coverage bigWig files. For ease of statistical analysis, for each study we created count tables at the gene and exon levels and extracted phenotype data, which we provide in their raw formats as well as in RangedSummarizedExperiment R objects (described in the [SummarizedExperiment](#) Bioconductor package). We also computed the mean coverage per study and provide it in a bigWig file, which can be used with the [derfinder](#) Bioconductor package to perform annotation-agnostic differential expression analysis at the expressed regions-level as described at [Collado-Torres et al, Nucleic Acids Research, 2017](#). The count tables, RangedSummarizeExperiment objects, phenotype tables, sample bigWigs, mean bigWigs, and file information tables are ready to use and freely available here. We also created the [recount](#) Bioconductor package which allows you to search and download the data for a specific study. By taking care of several preprocessing steps and combining many datasets into one easily-accessible website, we make finding and analyzing RNA-seq data considerably more straightforward.

Main publication

- **Collado-Torres L, Nellore A**, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. [Reproducible RNA-seq analysis using *recount2*](#). *Nature Biotechnology*, 2017. doi: 10.1038/nbt.3838.

recount2

①Searchという検索窓が見えたら、②pancreaticなどそれっぽいキーワードを打ち込んでみる。

The screenshot shows the recount2 web interface. At the top, there is a search bar with the text "pancreatic" entered. A red arrow labeled "1" points to the search bar, and another red arrow labeled "2" points to the search button. Below the search bar, there is a table of datasets. The table has columns for accession, number of samples, species, abstract, gene, exon, junctions, transcripts, phenotype, and files. The first row of the table is for accession SRP061240, which has 384 samples and is from a human. The abstract text describes the study of extracellular vesicles in RNA sequencing analysis.

accession	number of samples	species	abstract	gene	exon	junctions	transcripts	phenotype	files
All	All	All	All			All	All	All	
SRP061240	384	human	Extracellular vesicles such as exosomes are selectively enriched in RNA that has potential for use as disease biomarkers. To systemically characterize circulating extracellular RNA profiles, we performed RNA sequencing analysis on plasma extracellular vesicles derived from 192 individuals including 100 colon cancer, 36 prostate cancer and 6 pancreatic cancer patients along with 50 healthy individuals. Of ~12.6 million raw reads for each of these subjects, the number of mappable reads aligned to RNA references was ~5.4 million including microRNAs(miRNAs) (~40.4%), piwi-interacting RNAs(piwiRNAs) (~40.0%), pseudo-genes (~3.7%), long noncoding RNAs (lncRNAs) (~2.4%), tRNAs (~2.1%), and mRNAs (~2.1%). To select the best candidates for potential extracellular RNA reference controls, we performed abundant level stability testing and identified a set of miRNAs showing relatively consistent expression. To estimate biological variations, we performed association analysis of expression levels with age and sex in healthy individuals. This analysis showed	RSE v2 counts v1 counts v1	RSE v2 counts v1 counts v1	RSE jx_bed jx_cov counts	RSE v2 RSE v1	link	v2 v1

recount2

①Searchという検索窓が見えたら、②pancreaticなどそれっぽいキーワードを打ち込んでみる。そうすると、打ち込んでいるそばから赤枠内がごにょごにょ変化する。よく見ると、③打ち込んだキーワードを含むデータセットが表示されていることがわかる。

The screenshot shows the recount2 website interface. At the top, there is a search bar with the text 'pancreatic' entered. Below the search bar, there is a table of datasets. The first dataset, SRP061240, is highlighted with a red box. The abstract for this dataset contains the word 'pancreatic' underlined. Red arrows with numbers 1, 2, and 3 point to the search bar, the search input, and the highlighted dataset respectively.

accession	number of samples	species	abstract	gene	exon	junctions	transcripts	phenotype	files info
SRP061240	384	human	Extracellular vesicles such as exosomes are selectively enriched in RNA that has potential for use as disease biomarkers. To systemically characterize circulating extracellular RNA profiles, we performed RNA sequencing analysis on plasma extracellular vesicles derived from 192 individuals including 100 colon cancer, 36 prostate cancer and 6 <u>pancreatic</u> cancer patients along with 50 healthy individuals. Of ~12.6 million raw reads for each of these subjects, the number of mappable reads aligned to RNA references was ~5.4 million including microRNAs(miRNAs) (~40.4%), piwi-interacting RNAs(piwiRNAs) (~40.0%), pseudo-genes (~3.7%), long noncoding RNAs (lncRNAs) (~2.4%), tRNAs (~2.1%), and mRNAs (~2.1%). To select the best candidates for potential extracellular RNA reference controls, we performed abundant level stability testing and identified a set of miRNAs showing relatively consistent expression. To estimate biological variations, we performed association analysis of expression levels with age and sex in healthy individuals. This analysis showed	RSE v2 counts v2 RSE v1 counts v1	RSE v2 counts v2 RSE v1 counts v1	RSE jx_bed jx_cov counts	RSE v2 RSE v1	link	v2 v1

recount2

①Searchという検索窓が見えたら、②pancreaticなどそれっぽいキーワードを打ち込んでみる。そうすると、打ち込んでいるそばから赤枠内がごによごによ変化する。よく見ると、③打ち込んだキーワードを含むデータセットが表示されていることがわかる。④このデータセットのAccession番号。原著論文へのリンクもここから。

recount2: analysis-ready RNA-seq x +

← → ↻ 🏠 ⓘ https://jhubiostatistics.shinyapps.io/recount/

The Datasets

Show 10 entries

accession	number of samples	species	abstract	gene	exon	junctions	transcripts	phenotype	files info
All	All	All	All			All	All	All	
SRP061240	384	human	Extracellular vesicles such as exosomes are selectively enriched in RNA that has potential for use as disease biomarkers. To systemically characterize circulating extracellular RNA profiles, we performed RNA sequencing analysis on plasma extracellular vesicles derived from 192 individuals including 100 colon cancer, 36 prostate cancer and 6 <u>pancreatic</u> cancer patients along with 50 healthy individuals. Of ~12.6 million raw reads for each of these subjects, the number of mappable reads aligned to RNA references was ~5.4 million including microRNAs(miRNAs) (~40.4%), piwi-interacting RNAs(piwiRNAs) (~40.0%), pseudo-genes (~3.7%), long noncoding RNAs (lncRNAs) (~2.4%), tRNAs (~2.1%), and mRNAs (~2.1%). To select the best candidates for potential extracellular RNA reference controls, we performed abundant level stability testing and identified a set of miRNAs showing relatively consistent expression. To estimate biological variations, we performed association analysis of expression levels with age and sex in healthy individuals. This analysis showed	RSE v2 counts v2 RSE v1 counts v1	RSE v2 counts v2 RSE v1 counts v1	RSE jx_bed jx_cov counts	RSE v2 RSE v1	link	v2 v1



recount2

リードのマッピング後にどこの領域上にマップされたリードをカウントするかによって、①～④様々なバリエーションが存在する。通常利用は①geneだと思われれます。

The screenshot shows the recount2 web interface. The browser address bar displays 'https://jhubiostatistics.shinyapps.io/recount/'. The page title is 'The Datasets'. A search box contains the text 'pancreatic'. The table has columns for 'accession', 'number of samples', 'species', 'abstract', 'gene', 'exon', 'junctions', 'transcripts', 'phenotype', and 'files info'. Red arrows labeled 1, 2, 3, and 4 point to the 'gene', 'exon', 'junctions', and 'transcripts' columns respectively. The first row of data is for accession SRP061240, with 384 samples, human species, and a detailed abstract. The 'gene' column for this row contains 'RSE v2 counts v2 RSE v1 counts v1'. The 'exon' column contains 'RSE v2 counts v2 RSE v1 counts v1'. The 'junctions' column contains 'RSE jx_bed jx_cov counts'. The 'transcripts' column contains 'RSE v2 RSE v1'. The 'phenotype' column contains 'link'. The 'files info' column contains 'v2 v1'.

accession	number of samples	species	abstract	gene	exon	junctions	transcripts	phenotype	files info
SRP061240	384	human	Extracellular vesicles such as exosomes are selectively enriched in RNA that has potential for use as disease biomarkers. To systemically characterize circulating extracellular RNA profiles, we performed RNA sequencing analysis on plasma extracellular vesicles derived from 192 individuals including 100 colon cancer, 36 prostate cancer and 6 pancreatic cancer patients along with 50 healthy individuals. Of ~12.6 million raw reads for each of these subjects, the number of mappable reads aligned to RNA references was ~5.4 million including microRNAs(miRNAs) (~40.4%), piwi-interacting RNAs(piwiRNAs) (~40.0%), pseudo-genes (~3.7%), long noncoding RNAs (lncRNAs) (~2.4%), tRNAs (~2.1%), and mRNAs (~2.1%). To select the best candidates for potential extracellular RNA reference controls, we performed abundant level stability testing and identified a set of miRNAs showing relatively consistent expression. To estimate biological variations, we performed association analysis of expression levels with age and sex in healthy individuals. This analysis showed	RSE v2 counts v2 RSE v1 counts v1	RSE v2 counts v2 RSE v1 counts v1	RSE jx_bed jx_cov counts	RSE v2 RSE v1	link	v2 v1

recount2

リードのマッピング後にどこの領域上にマップされたリードをカウントするかによって、①～④様々なバリエーションが存在する。通常利用は①geneだと思われれます。赤枠内には、4つのリンク先(RSE v2, counts v2, RSE v1, and counts v1)が存在する。基本的にはversion 2に相当するv2のものを利用すればよい。また、RSEはRangedSummarizedExperimentというRの中で用いるデータの型の1つ。第2回(2/22)で教えたDNAStrngSetやAAStringSetと同じようなものです。②慣れないうちはcounts v2のほうを利用しましょう。

recount2: analysis-ready RNA-seq × +
https://jhubiostatistics

The Datasets

Show 10 entries

accession	number of samples	species	abstract	gene	exon	junctions	transcripts	phenotype	files info
All	All	All	All			All	All	All	
SRP061240	384	human	Extracellular vesicles such as exosomes are selectively enriched in RNA that has potential for use as disease biomarkers. To systemically characterize circulating extracellular RNA profiles, we performed RNA sequencing analysis on plasma extracellular vesicles derived from 192 individuals including 100 colon cancer, 36 prostate cancer and 6 pancreatic cancer patients along with 50 healthy individuals. Of ~12.6 million raw reads for each of these subjects, the number of mappable reads aligned to RNA references was ~5.4 million including microRNAs(miRNAs) (~40.4%), piwi-interacting RNAs(piwiRNAs) (~40.0%), pseudo-genes (~3.7%), long noncoding RNAs (lncRNAs) (~2.4%), tRNAs (~2.1%), and mRNAs (~2.1%). To select the best candidates for potential extracellular RNA reference controls, we performed abundant level stability testing and identified a set of miRNAs showing relatively consistent expression. To estimate biological variations, we performed association analysis of expression levels with age and sex in healthy individuals. This analysis showed	RSE v2 counts v2 RSE v1 counts v1	RSE v2 counts v2 RSE v1 counts v1	RSE jx_bed jx_cov counts	RSE v2 RSE v1	link	v2 v1



recount2

③384という数値は、②でダウンロードした数値行列の(行名部分を除く)列数に相当する。

The Datasets

Show 10 entries Search: pancreatic

accession	number of samples	species	abstract	gene	exon	junctions	transcripts	phenotype	files info
All	All	All	All			All	All	All	
SRP061240	384	human	Extracellular vesicles such as exosomes are selectively enriched in RNA that has potential for use as disease biomarkers. To systematically characterize circulating extracellular RNA profiles, we performed RNA sequencing analysis on plasma extracellular vesicles derived from 192 individuals including 100 colon cancer, 36 prostate cancer and 6 pancreatic cancer patients along with 50 healthy individuals. Of ~12.6 million raw reads for each of these subjects, the number of mappable reads aligned to RNA references was ~5.4 million including microRNAs(miRNAs) (~40.4%), piwi-interacting RNAs(piwiRNAs) (~40.0%), pseudo-genes (~3.7%), long noncoding RNAs (lncRNAs) (~2.4%), tRNAs (~2.1%), and mRNAs (~2.1%). To select the best candidates for potential extracellular RNA reference controls, we performed abundant level stability testing and identified a set of miRNAs showing relatively consistent expression. To estimate biological variations, we performed association analysis of expression levels with age and sex in healthy individuals. This analysis showed	RSE v2 counts v1	RSE v2 counts v1	RSE jx_bed jx_cov counts	RSE v2 RSE v1	link	v2 v1

recount2

③384という数値は、②でダウンロードした数値行列の(行名部分を除く)列数に相当する。④このあたりで、計192例のうち100例が結腸癌(colon cancer)みたいな情報を大まかに把握したうえで、⑤の情報と突き合わせて数値行列上のどの列が対応するのかを特定する。このデータセットはデカすぎるので、ページ下部に移動して、他の適度なサンプル数のデータセットを探すことにする。

recount2: analysis-ready RNA-seq × +
https://jhubiostatistics.shinyapps.io/recount2/

The Datasets

Show 10 entries

Search: pancreatic

accession	number of samples	species	abstract	gene	exon	junctions	transcripts	phenotype	files info
SRP061240	384	human	Extracellular vesicles such as exosomes are selectively enriched in RNA that has potential for use as disease biomarkers. To systematically characterize circulating extracellular RNA profiles, we performed RNA sequencing analysis on plasma extracellular vesicles derived from 192 individuals including 100 colon cancer, 36 prostate cancer and 6 pancreatic cancer patients along with 50 healthy individuals. Of ~12.6 million raw reads for each of these subjects, the number of mappable reads aligned to RNA references was ~5.4 million including microRNAs(miRNAs) (~40.4%), piwi-interacting RNAs(piwiRNAs) (~40.0%), pseudo-genes (~3.7%), long noncoding RNAs (lncRNAs) (~2.4%), tRNAs (~2.1%), and mRNAs (~2.1%). To select the best candidates for potential extracellular RNA reference controls, we performed abundant level stability testing and identified a set of miRNAs showing relatively consistent expression. To estimate biological variations, we performed association analysis of expression levels with age and sex in healthy individuals. This analysis showed	RSE v2 counts v1	RSE v2 counts v1	RSE jx_bed jx_cov counts	RSE v2 RSE v1	link	v2 v1



recount2

①このあたりまで下に移動。②まだ24サンプル分ありますね。③recount2は、全部で2,039エントリー(データセット数に相当)あることがわかる。このうち、④29エントリーがpancreaticというキーワードを含むものだということがわかる。⑤次のページに移動。

recount2: analysis-ready RNA-seq x +

← → ↻ 🏠 ⓘ https://jhubiostatistics.shinyapps.io/recount/ ★ 🗄️ 🔍 ⋮

SRP056835	24	human	Understanding distinct gene expression patterns of normal adult and developing fetal human pancreatic a and b cells is crucial for developing stem cell therapies, islet regeneration strategies, and therapies designed to increase b cell function in patients with diabetes (type 1 or 2). Toward that end, we have developed methods to highly purify a, b, and d cells from human fetal and adult pancreata by intracellular staining for the cell-specific hormone content, sorting the sub-populations by flow cytometry and, using next generation RNA sequencing, we report on the detailed transcriptomes of fetal and adult a and b cells. We observed that human islet composition was not influenced by age, gender, or body mass index and transcripts for inflammatory gene products were noted in fetal b cells. In addition, within highly purified adult glucagon-expressing a cells, we observed surprisingly high insulin mRNA expression, but not insulin protein expression. This transcriptome analysis from highly purified islet a and b cell subsets from fetal and adult pancreata offers clear implications for strategies that seek to increase insulin expression in type 1 and type 2 diabetes. Overall design: RNA-sequencing of highly purified human adult and fetal islet cell subset was performed using our newly developed method. Using this data, we can study and compare the detailed transcriptome or alpha and beta cells during development.	RSE v2 counts v1	RSE v2 counts v1	RSE jx_bed jx_cov counts	RSE v2 RSE v1	link	v2 v1
-----------	----	-------	---	------------------	------------------	--------------------------	---------------	------	-------



Showing 1 to 10 of 29 entries (filtered from 2,039 total entries)

Previous 1 2 3 Next

recount2

赤枠が20番目のエントリー。①手頃な計6サンプル。②最後の文章から、2群間比較(control vs. TCF7L2)なのだろうと読み取る。正確には、PANC1という膵臓がん細胞中のTCF7L2 (Wnt signaling pathway中の遺伝子でTCF4ともいう)をsiRNAsでknockdownした3例とcontrolの3例を比較したデータセット。③のリンク先から原著論文に辿れます。

TFE3

SRP050497	6	human	We have compared the genome-wide effects on the transcriptome after treatment with ICG-001 (the specific CBP inhibitor) versus C646, a compound that competes with acetyl-coA for the Lys-coA binding pocket of both CBP and p300. We found that both drugs cause large-scale changes in the transcriptome of HCT116 colon cancer cells and PANC1 pancreatic cancer cells, and reverse some tumor-specific changes in gene expression. Interestingly, although the epigenetic inhibitors affect cell cycle pathways in both the colon and pancreatic cancer cell lines, the WNT signaling pathway was affected only in the colon cancer cells. Notably, WNT target genes were similarly down-regulated after treatment of HCT116 with C646 as with ICG-001. Overall design: To identify genes affected by direct targeting of a component of the transcriptional complex implicated in WNT regulation, we used siRNAs to knockdown TCF7L2 in PANC1 cells. Cells were treated with control siRNAs or siRNAs specific for TCF7L2 and RNA was analyzed by RNA-seq.	RSE v2 counts v1	RSE v2 counts v1	RSE jx_bed jx_cov counts	RSE v2 RSE v1	link	v2 v1
-----------	---	-------	---	------------------	------------------	--------------------------	---------------	------	-------

Showing 11 to 20 of 29 entries (filtered from 2,039 total entries)

Previous 1 2 3 Next

Download list of studies matching search results

Note that GTEx is separated from this list.

recount2

Epigenetics Chromatin. 2015 Feb 24;8:9. doi: 10.1186/1756-8935-8-9. eCollection 2015.

Altering cancer transcriptomes using epigenomic inhibitors.

Gaddis M¹, Gerrard D², Fietze S², Farnham PJ¹.

+ Author information

Abstract

BACKGROUND: Due to the hyper-activation of WNT signaling in a variety of cancer types, there has been a strong drive to develop pathway-specific inhibitors with the eventual goal of providing a chemotherapeutic antagonist of WNT signaling to cancer patients. A new category of drugs, called epigenetic inhibitors, are being developed that hold high promise for inhibition of the WNT pathway. The canonical WNT signaling pathway initiates when WNT ligands bind to receptors, causing the nuclear localization of the co-activator β -catenin (CTNNB1), which leads to an association of β -catenin with a member of the TCF transcription factor family at regulatory regions of WNT-responsive genes. The TCF/ β -catenin complex then recruits CBP (CREBBP) or p300 (EP300), leading to histone acetylation and gene activation. A current model in the field is that CBP-driven expression of WNT target genes supports proliferation whereas p300-driven expression of WNT target genes supports differentiation. The small molecule inhibitor ICG-001 binds to CBP, but not to p300, and competitively inhibits the interaction of CBP with β -catenin. Upon treatment of cancer cells, this should reduce expression of CBP-regulated transcription, leading to reduced tumorigenicity and enhanced differentiation.

RESULTS: We have compared the genome-wide effects on the transcriptome after treatment with ICG-001 (the specific CBP inhibitor) versus C646, a compound that competes with acetyl-coA for the Lys-coA binding pocket of both CBP and p300. We found that both drugs cause large-scale changes in the transcriptome of HCT116 colon cancer cells and PANC1 pancreatic cancer cells and reverse some tumor-specific changes in gene expression. Interestingly, although the epigenetic inhibitors affect cell cycle pathways in both the colon and pancreatic cancer cell lines, the WNT signaling pathway was affected only in the colon cancer cells. Notably, WNT target genes were similarly downregulated after treatment of HCT116 with C646 as with ICG-001.

CONCLUSION: Our results suggest that treatment with a general HAT inhibitor causes similar effects on the transcriptome as does treatment with a CBP-specific inhibitor and that epigenetic inhibition affects the WNT pathway in HCT116 cells and the cholesterol biosynthesis pathway in PANC1 cells.

recount2

①をダウンロードして解凍すると、counts_gene.tsvが得られます。また、②からSRP050497.tsvが得られます。

recount2: analysis-ready RNA-seq x +

https://jhubiostatistics.shinyapps.io/recount/

TFE3

SRP050497	6	human	We have compared the genome-wide effects on the transcriptome after treatment with ICG-001 (the specific CBP inhibitor) versus C646, a compound that competes with acetyl-coA for the Lys-coA binding pocket of both CBP and p300. We found that both drugs cause large-scale changes in the transcriptome of HCT116 colon cancer cells and PANC1 pancreatic cancer cells, and reverse some tumor-specific changes in gene expression. Interestingly, although the epigenetic inhibitors affect cell cycle pathways in both the colon and pancreatic cancer cell lines, the WNT signaling pathway was affected only in the colon cancer cells. Notably, WNT target genes were similarly down-regulated after treatment of HCT116 with C646 as with ICG-001. Overall design: To identify genes affected by direct targeting of a component of the transcriptional complex implicated in WNT regulation, we used siRNAs to knockdown TCF7L2 in PANC1 cells. Cells were treated with control siRNAs or siRNAs specific for TCF7L2 and RNA was analyzed by RNA-seq.	RSE v2 counts v1	RSE v2 counts v1	RSE jx_bed jx_cov counts v1	RSE v2 RSE v1	link	v2 v1

Showing 11 to 20 of 29 entries (filtered from 2,039 total entries)

Previous 1 2 3 Next

Download list of studies matching search results

Note that GTEx is separated from this list.

数値行列

①をダウンロードして解凍すると、counts_gene.tsvが得られます。また、②からSRP050497.tsvが得られます。counts_gene.tsvの中身。①行名情報の列が一番右側になっているので、一番左側に移動させます。



SRR1692137	SRR1692138	SRR1692139	SRR1692140	SRR1692141	SRR1692142	gene_id
11156	9957	14914	15968	15138	16747	ENSG00000000003.14
0	0	0	0	0	0	ENSG00000000005.5
64023	55036	88460	62819	67807	70965	ENSG000000000419.12
24666	22830	37528	31189	31249	36804	ENSG000000000457.13
42199	38508	58545	36169	39613	42044	ENSG000000000460.16
0	0	0	0	0	0	ENSG000000000938.12
647	628	794	450	650	446	ENSG000000000971.15
89682	81128	123915	121702	121412	143520	ENSG00000001036.13
86325	77847	120691	73226	76177	87010	ENSG00000001084.10
57229	49468	80253	89081	92579	109634	ENSG00000001167.14
16982	16121	25284	19119	18513	19189	ENSG00000001460.17
16155	14981	27084	22228	20813	22194	ENSG00000001461.16

数値行列

①をダウンロードして解凍すると、counts_gene.tsvが得られます。また、②からSRP050497.tsvが得られます。counts_gene.tsvの中身。①行名情報の列が一番右側になっているので、一番左側に移動させます。次に、赤枠内のサンプル名情報を理解しやすいものに変更します。

gene_id	SRR1692137	SRR1692138	SRR1692139	SRR1692140	SRR1692141	SRR1692142
ENSG00000000003.14	11156	9957	14914	15968	15138	16747
ENSG00000000005.5	0	0	0	0	0	0
ENSG00000000419.12	64023	55036	88460	62819	67807	70965
ENSG00000000457.13	24666	22830	37528	31189	31249	36804
ENSG00000000460.16	42199	38508	58545	36169	39613	42044
ENSG00000000938.12	0	0	0	0	0	0
ENSG00000000971.15	647	628	794	450	650	446
ENSG00000001036.13	89682	81128	123915	121702	121412	143520
ENSG00000001084.10	86325	77847	120691	73226	76177	87010
ENSG00000001167.14	57229	49468	80253	89081	92579	109634
ENSG00000001460.17	16982	16121	25284	19119	18513	19189
ENSG00000001461.16	16155	14981	27084	22228	20813	22194

phenotype

①をダウンロードして解凍すると、counts_gene.tsvが得られます。また、②からSRP050497.tsvが得られます。counts_gene.tsvの中身。①行名情報の列が一番右側になっているので、一番左側に移動させます。次に、赤枠内のサンプル名情報を理解しやすいものに変更します。その際に用いるのが、phenotypeのところからダウンロードした②のSRP050497.tsvです。赤枠内がその中身。

recount2: analysis-ready RNA-seq x +

← → ↻ 🏠 ⓘ https://jhubiostatistics.shinyapps.i

TFE3

SRP050497	6	human	We have compared the genome-wide effects on the transcriptome after treatment with ICG-001 (the specific CBP inhibitor) versus C646, a compound that competes with acetyl-coA for the Lys-coA binding pocket of both CBP and p300. We found that both drugs cause large-scale changes in the transcriptome of HCT116 colon cancer cells and PANC1 pancreatic cancer cells, and reverse some tumor-specific changes in gene expression. Interestingly, although the epigenetic inhibitors affect cell cycle pathways in both	RSE v2 counts v1	RSE v2 counts v1	RSE jx_bed jx_cov counts	RSE v2 RSE v1	link	v2 v1
-----------	---	-------	---	------------------	------------------	--------------------------	---------------	------	-------

sample	experiment	run	read_count	reads_down	proportion	paired_end	sra_misread	mapped_reads	auc	sharq_beta	sharq_beta	biosample	biosample	biosample	avg_read	geo_accession	bigwig_file	title	characteristics
SRS77852	SRX79502	SRR1692137	30007994	30007994	1	FALSE	FALSE	29958199	1.49E+09	NA	NA	2014-12-01	2015-01-01	2015-01-01	50	GSM1556/SRR16921	PANC1	counts	"cell line: pancreatic ductal carcinoma cell line", "siRNA: control siRNA"
SRS77852	SRX79502	SRR1692138	27252897	27252897	1	FALSE	FALSE	27208781	1.36E+09	NA	NA	2014-12-01	2015-01-01	2015-01-01	50	GSM1556/SRR16921	PANC1	counts	"cell line: pancreatic ductal carcinoma cell line", "siRNA: control siRNA"
SRS77852	SRX79502	SRR1692139	42212497	42212497	1	FALSE	FALSE	42146090	2.1E+09	NA	NA	2014-12-01	2015-01-01	2015-01-01	50	GSM1556/SRR16921	PANC1	counts	"cell line: pancreatic ductal carcinoma cell line", "siRNA: control siRNA"
SRS77852	SRX79502	SRR1692140	31456271	31456271	1	FALSE	FALSE	31384977	1.56E+09	NA	NA	2014-12-01	2015-01-01	2015-01-01	50	GSM1556/SRR16921	PANC1	counts	"cell line: pancreatic ductal carcinoma cell line", "siRNA: TCF7L2 siRNA"
SRS77852	SRX79502	SRR1692141	31569339	31569339	1	FALSE	FALSE	31516283	1.57E+09	NA	NA	2014-12-01	2015-01-01	2015-01-01	50	GSM1556/SRR16921	PANC1	counts	"cell line: pancreatic ductal carcinoma cell line", "siRNA: TCF7L2 siRNA"
SRS77853	SRX79502	SRR1692142	37477777	37477777	1	FALSE	FALSE	37411196	1.86E+09	NA	NA	2014-12-01	2015-01-01	2015-01-01	50	GSM1556/SRR16921	PANC1	counts	"cell line: pancreatic ductal carcinoma cell line", "siRNA: TCF7L2 siRNA"

affected by direct targeting of a component of the transcriptional complex implicated in WNT regulation, we used siRNAs to knockdown TCF7L2 in PANC1 cells. Cells were treated with control siRNAs or siRNAs specific for TCF7L2 and RNA was analyzed by RNA-seq.

Showing 11 to 20 of 29 entries (filtered from 2,039 total entries)

Previous 1 2 3 Next

Download list of studies matching search results

Note that GTEx is separated from this list.

phenotype

①をダウンロードして解凍すると、counts_gene.tsvが得られます。また、②からSRP050497.tsvが得られます。counts_gene.tsvの中身。①行名情報の列が一番右側になっているので、一番左側に移動させます。次に、赤枠内のサンプル名情報を理解しやすいものに変更します。その際に用いるのが、phenotypeのところからダウンロードした②のSRP050497.tsvです。赤枠内がその中身。この場合は③run列と④title列の情報を対応付けて、適当に名前を変更します。

recount2: analysis-ready RNA-seq

https://jhubiostatistics.shinyapps.io/recount2/

TFE3

SRP050497 6 human We have compared the transcriptome after treatment with TCF7L2 (a transcription factor-specific CBP inhibitor) versus C646, a compound that competes with acetyl-coA for the Lys-coA binding pocket of both CBP and p300. We found that both drugs cause large-scale changes in the transcriptome of HCT116 colon cancer cells and PANC1 pancreatic cancer cells, and reverse some tumor-specific changes in gene expression. Interestingly, although the epigenetic inhibitors affect cell cycle pathways in both

sample	experiment	run	read_count	reads_down	proportion	paired_end	sra_misread	mapped_reads	auc	sharq_beta	sharq_beta	biosample	biosample	biosample	avg_read	geo_accession	bigwig_file	title	Characteristics
SRS77852	SRX79502	SRR1692137	30007994	30007994	1	FALSE	FALSE	29958199	1.49E+09	NA	NA	2014-12-01	2015-01-01	2015-01-01	50	GSM1556/SRR1692137	PANC1	control siRNA 1	("cell line: pancreatic ductal carcinoma cell line", "siRNA: control siRNA")
SRS77852	SRX79502	SRR1692138	27252897	27252897	1	FALSE	FALSE	27208781	1.36E+09	NA	NA	2014-12-01	2015-01-01	2015-01-01	50	GSM1556/SRR1692138	PANC1	control siRNA 2	("cell line: pancreatic ductal carcinoma cell line", "siRNA: control siRNA")
SRS77852	SRX79502	SRR1692139	42212497	42212497	1	FALSE	FALSE	42146090	2.1E+09	NA	NA	2014-12-01	2015-01-01	2015-01-01	50	GSM1556/SRR1692139	PANC1	control siRNA 3	("cell line: pancreatic ductal carcinoma cell line", "siRNA: control siRNA")
SRS77852	SRX79502	SRR1692140	31456271	31456271	1	FALSE	FALSE	31384977	1.56E+09	NA	NA	2014-12-01	2015-01-01	2015-01-01	50	GSM1556/SRR1692140	PANC1	siTCF7L2 1	("cell line: pancreatic ductal carcinoma cell line", "siRNA: TCF7L2 siRNA")
SRS77852	SRX79502	SRR1692141	31569339	31569339	1	FALSE	FALSE	31516283	1.57E+09	NA	NA	2014-12-01	2015-01-01	2015-01-01	50	GSM1556/SRR1692141	PANC1	siTCF7L2 2	("cell line: pancreatic ductal carcinoma cell line", "siRNA: TCF7L2 siRNA")
SRS77852	SRX79502	SRR1692142	37477777	37477777	1	FALSE	FALSE	37411196	1.86E+09	NA	NA	2014-12-01	2015-01-01	2015-01-01	50	GSM1556/SRR1692142	PANC1	siTCF7L2 3	("cell line: pancreatic ductal carcinoma cell line", "siRNA: TCF7L2 siRNA")

run

SRR1692137

SRR1692138

SRR1692139

SRR1692140

SRR1692141

SRR1692142

Showing 11 from 2,039 total entries)

Download search results

Note that Gist.

title

PANC1 control siRNA 1

PANC1 control siRNA 2

PANC1 control siRNA 3

PANC1 siTCF7L2 1

PANC1 siTCF7L2 2

PANC1 siTCF7L2 3

1 2 3 Next

列名変更完了

ここでは、シンプルにcontrol vs. treatmentにしました。これで基本的にデータ解析の準備完了です。ウェブサイト(Rで)塩基配列解析ではタブ区切りテキストファイルを基本の入力フォーマットとしているので、ここではそのような形式でcounts_gene.txtというファイル名で保存しました。

gene_id	Control1	Control2	Control3	Treatment1	Treatment2	Treatment3
ENSG00000000003.14	11156	9957	14914	15968	15138	16747
ENSG00000000005.5	0	0	0	0	0	0
ENSG00000000419.12	64023	55036	88460	62819	67807	70965
ENSG00000000457.13	24666	22830	37528	31189	31249	36804
ENSG00000000460.16	42199	38508	58545	36169	39613	42044
ENSG00000000938.12	0	0	0	0	0	0
ENSG00000000971.15	647	628	794	450	650	446
ENSG00000001036.13	89682	81128	123915	121702	121412	143520
ENSG00000001084.10	86325	77847	120691	73226	76177	87010
ENSG00000001167.14	57229	49468	80253	89081	92579	109634
ENSG00000001460.17	16982	16121	25284	19119	18513	19189
ENSG00000001461.16	16155	14981	27084	22228	20813	22194

counts_gene.txt

Contents

- 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

クラスタリング

どのサンプルとどのサンプルが似ているかを眺め、全体としておかしくないか(外れサンプルがあるか)などをチェックする目的で行います。

gene_id	Control1	Control2	Control3	Treatment1	Treatment2	Treatment3
ENSG00000000003.14	11156	9957	14914	15968	15138	16747
ENSG00000000005.5	0	0	0	0	0	0
ENSG00000000419.12	64023	55036	88460	62819	67807	70965
ENSG00000000457.13	24666	22830	37528	31189	31249	36804
ENSG00000000460.16	42199	38508	58545	36169	39613	42044
ENSG00000000938.12	0	0	0	0	0	0
ENSG00000000971.15	647	628	794	450	650	446
ENSG00000001036.13	89682	81128	123915	121702	121412	143520
ENSG00000001084.10	86325	77847	120691	73226	76177	87010
ENSG00000001167.14	57229	49468	80253	89081	92579	109634
ENSG00000001460.17	16982	16121	25284	19119	18513	19189
ENSG00000001461.16	16155	14981	27084	22228	20813	22194

counts_gene.txt

①の項目群の中から、②を探す。多数の項目があるので最初のうちは見つけづらいですが、慣れです。

クラスタリング

(Rで)塩基配列解析

(last modified 2019/02/22, since 2010)

このウェブページのR関連部分は、[インストール](#) (last modified 2018/08/08) (Macintosh2018.11.27版)に従ってフリーウェアのインストールを前提で記述しています。初心者の方は[基礎](#) (last modified 2018/08/15) (Macintosh2019.01.15版)で自習してください。また、[書籍](#)・[学会誌](#)などを切り分けて[サブページ](#)に

What's new? (過去のお知らせはこちら)

- 「[生命科学データ解析を支える情報技術](#)」を含むかなり広範な内容を含んでいます。Homebrew, Docker, GitHub, EC2, AWSというメリットがあると思います。(2019/02/22)
- 「[イントロ](#) | [NGS](#) | [配列取得](#) | [FASTQ](#) | [シーケンシング](#)」

- 解析 | 前処理 | ID変換 | [Ensembl Gene ID中のバージョン情報を除去](#) (last modified 2018/08/08)
- 解析 | 前処理 | ID変換 | Ensembl Gene ID --> gene symbols | [基礎](#) (last modified 2018/08/15)
- 解析 | 前処理 | ID変換 | Ensembl Gene ID --> gene symbols | [RangedSummarizedExperiment](#) (last modified 2018/08/15)
- 解析 | [クラスタリング](#) | [について](#) (last modified 2018/07/17)
- 解析 | [クラスタリング](#) | サンプル間 | [hclust](#) (last modified 2015/02/26)
- 解析 | [クラスタリング](#) | サンプル間 | [TCC\(Sun_2013\)](#) (last modified 2018/08/06)
- 解析 | [クラスタリング](#) | 遺伝子間(基礎) | [MBCluster.Seq\(Si_2014\)](#) (last modified 2018/09/23)
- 解析 | [クラスタリング](#) | 遺伝子間(応用) | [TCC正規化\(Sun_2013\)+MBCluster.Seq\(Si_2014\)](#) (last modified 2018/09/23)
- 解析 | [外れサンプル検出](#) | [について](#) (last modified 2018/07/17)
- 解析 | [発現変動](#) | [について\(2013年頃の記載事項で記念に残しているだけ\)](#) (last modified 2014/07/10)
- 解析 | [発現変動](#) | [について](#) (last modified 2018/07/10)
- 解析 | [発現変動](#) | [2群間](#) | [対応なし](#) | [について](#) (last modified 2016/10/07)
- 解析 | [発現変動](#) | [2群間](#) | [対応なし](#) | [複製あり](#) | [DESeq2\(Love_2014\)](#) (last modified 2015/11/15)
- 解析 | [発現変動](#) | [2群間](#) | [対応なし](#) | [複製あり](#) | [TCC\(Sun_2013\)](#) (last modified 2015/07/07)推奨
- 解析 | [発現変動](#) | [2群間](#) | [対応なし](#) | [複製あり](#) | [Blekhmanデータ](#) | [TCC\(Sun_2013\)](#) (last modified 2015/07/07)

クラスタリング

①に飛びます。②の赤下線部分にも書かれていますが、これは、③TCCというRのパッケージ中の関数を用いてサンプル間クラスタリングを行う際に用います。

(Rで)塩基配列解析

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

解析 | クラスタリング | サンプル間 | TCC(Sun_2013)

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱論文 ([Tang et al., BMC Bioinformatics, 2015](#))中でもこの関数を用いています(2015/11/05追加)。xlsx形式ファイルを入力とするやり方も追加しました(2015/11/15)。

「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ー。

1. 59,857 genes×6 samplesのリアルデータ([srp017142_count_bowtie.txt](#))の場合 :

[Neyret-Kahn et al., Genome Res., 2013](#)の2群間比較用(3 proliferative samples vs. 3 Ras samples)ヒトRNA-seqカウントデータです。 [パイプライン | ゲノム | 発現変動 | 2群間 | 対応なし | 複製あり | SRP017142\(Neyret-Kahn 2013\)](#)から得られます。

```
in_f <- "srp017142_count_bowtie.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(TCC) #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を読み込み
dim(data) #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman", #クラスタリング実行結果をoutに格納
                    hclust.method="average", unique.pattern=TRUE) #クラスタリング実行結果をoutに格納

#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設定
plot(out) #樹形図(デンドログラム)の表示
dev.off() #おまじない
```

[トップページへ](#)

クラスタリング

①に飛びます。②の赤下線部分にも書かれていますが、これは、③TCCというRのパッケージ中の関数を用いてサンプル間クラスタリングを行う際に用います。④は、③TCCパッケージの原著論文の筆頭著者名と出版年です。原著論文がある場合は、このように記載する場合があります。

(Rで)塩基配列解析

保護されていない通信 | www.iu.a.u-to

解析 | クラスタリング | サンプル間 | TCC(Sun_2013)

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示し、clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドラインは、原著論文 (Tang et al., BMC Bioinformatics, 2015)中でもこの関数を用いています(2015/11/05追加)。xlsx形式ファイルを入力とするやり方も追加しました(2015/11/15)。

「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ー。

1. 59,857 genes×6 samplesのリアルデータ(srp017142_count_bowtie.txt)の場合 :

Neyret-Kahn et al., Genome Res., 2013の2群間比較用(3 proliferative samples vs. 3 Ras samples)ヒトRNA-seqカウントデータです。パイプライン | ゲノム | 発現変動 | 2群間 | 対応なし | 複製あり | SRP017142(Neyret-Kahn 2013)から得られます。

```
in_f <- "srp017142_count_bowtie.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(TCC) #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
dim(data) #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman", #クラスタリング実行結果をoutに格納
                    hclust.method="average", unique.pattern=TRUE) #クラスタリング実行結果をoutに格納

#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
plot(out) #樹形図(デンドログラム)の表示
dev.off() #おまじない
```

[トップページへ](#)

クラスタリング

①例題1は、②を入力ファイルとしてサンプル間クラスタリングを実行した結果を、③500×400ピクセルの大きさを `hoge1.png` というファイル名で保存するためのコード。

(Rで)塩基配列解析

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

解析 | クラスタリング | サンプル間 | TCC(Sun_2013)

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱論文 ([Tang et al., BMC Bioinformatics, 2015](#))中でもこの関数を用いています(2015/11/05追加)。xlsx形式ファイルを入力とするやり方も追加しました(2015/11/15)。

「ファイル」 - 「ディレクトリの変更」で解析したいファイルが②にあるディレクトリに移動し以下をコピー。

1. 59,857 genes×6 samplesのリアルデータ([srp017142_count_bowtie.txt](#))の場合 : ①

[Neyret-Kahn et al., Genome Res., 2013](#)の2群間比較用(3 proliferative samples vs. 3 Ras samples)ヒトRNA-seqカウントデータです。 [パイプライン | ゲノム | 発現変動 | 2群間 | 対応なし | 複製あり | SRP017142\(Neyret-Kahn 2013\)](#)から得られます。

```
in_f <- "srp017142_count_bowtie.txt" ② 入力ファイル名を指定してin_fに格納
out_f <- "hoge1.png"                ③ 出力ファイル名を指定してout_fに格納
param_fig <- c(500, 400)             ③ #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(TCC)                         #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
dim(data)                             #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman", #クラスタリング実行結果をoutに格納
                     hclust.method="average", unique.pattern=TRUE) #クラスタリング実行結果をoutに格納

#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
plot(out)                                                            #樹形図(デンドログラム)の表示
dev.off()                                                            #おまじない
```

[トップページへ](#)

Contents

- 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

クラスタリング

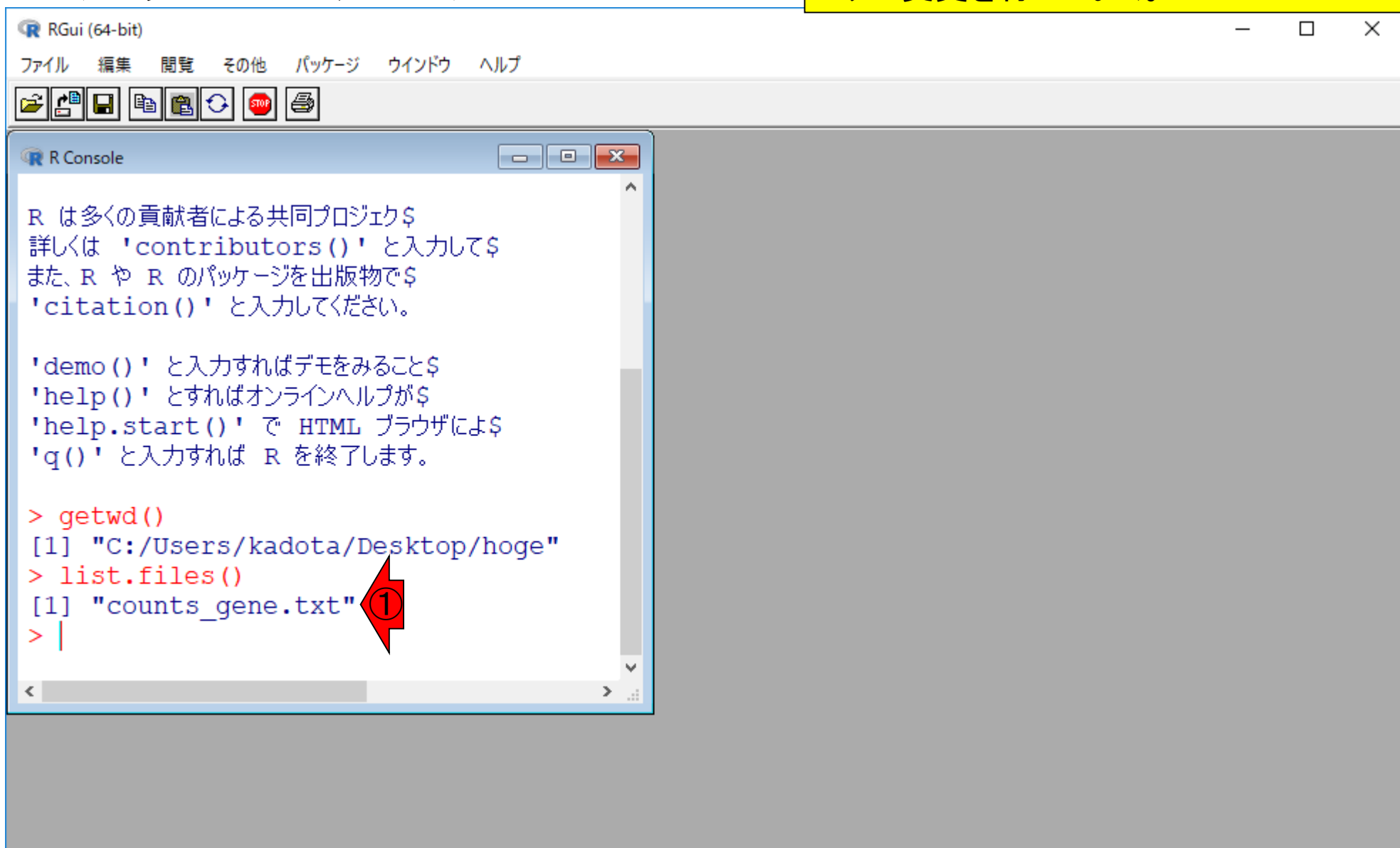
①手元にあるcounts_gene.txtを入力ファイルとして実行するにはどうすればよい?

gene_id	Control1	Control2	Control3	Treatment1	Treatment2	Treatment3
ENSG00000000003.14	11156	9957	14914	15968	15138	16747
ENSG00000000005.5	0	0	0	0	0	0
ENSG00000000419.12	64023	55036	88460	62819	67807	70965
ENSG00000000457.13	24666	22830	37528	31189	31249	36804
ENSG00000000460.16	42199	38508	58545	36169	39613	42044
ENSG00000000938.12	0	0	0	0	0	0
ENSG00000000971.15	647	628	794	450	650	446
ENSG00000001036.13	89682	81128	123915	121702	121412	143520
ENSG00000001084.10	86325	77847	120691	73226	76177	87010
ENSG00000001167.14	57229	49468	80253	89081	92579	109634
ENSG00000001460.17	16982	16121	25284	19119	18513	19189
ENSG00000001461.16	16155	14981	27084	22228	20813	22194

① counts_gene.txt

クラスタリング

①手元にあるcounts_gene.txtを入力ファイルとして実行できるように、Rを起動してディレクトリの変更を行っておく。



```
RGui (64-bit)
ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ

R Console
R は多くの貢献者による共同プロジェクトです。
詳しくは 'contributors()' と入力してください。
また、R や R のパッケージを出版物で
'citation()' と入力してください。

'demo()' と入力すればデモをみることも
'help()' とすればオンラインヘルプが
'help.start()' で HTML ブラウザによ
'q()' と入力すれば R を終了します。

> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files()
[1] "counts_gene.txt"
> |
```

クラスタリング

The screenshot shows the RStudio application window. The 'File' menu is open, and the 'New Script' option is highlighted with a red arrow labeled '2'. A red arrow labeled '1' points to the 'File' menu itself. In the background, a terminal window displays R code and its output:

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files()
[1] "counts_gene.txt"
> |
```

クラスタリング

①ファイル、②新しいスクリプト。こんな感じで③REディタが起動します。別にこれじゃなきゃいけないというわけではありませんが、前回述べた二重クォーテーション問題を回避可能なエディタです。

The screenshot shows the RGui (64-bit) interface. The top menu bar includes 'ファイル', '編集', 'パッケージ', 'ウインドウ', and 'ヘルプ'. Below the menu is a toolbar with icons for file operations. The main window is titled 'R Console' and contains the following text:

```
R は多くの貢献者による共同プロジェクトです。
詳しくは 'contributors()' と入力してください。
また、R や R のパッケージを出版物で
'citation()' と入力してください。

'demo()' と入力すればデモをみることも
'help()' とすればオンラインヘルプが
'help.start()' で HTML ブラウザによる
'q()' と入力すれば R を終了します。

> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files()
[1] "counts_gene.txt"
> |
```

A new window titled '無題 - RIディタ' is open in the foreground, indicated by a red arrow with the number '3' pointing to its title bar. The window is currently empty.

クラスタリング

①赤枠内のコードをテンプレートとして利用
すべくコピーして...

(Rで)塩基配列解析

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

解析 | クラスタリング | サンプル間 | TCC(Sun_2013)

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱論文 ([Tang et al., BMC Bioinformatics, 2015](#))中でもこの関数を用いています(2015/11/05追加)。xlsx形式ファイルを入力とするやり方も追加しました(2015/11/15)。

「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ー。

1. 59,857 genes×6 samplesのリアルデータ([srp017142_count_bowtie.txt](#))の場合 :

[Neyret-Kahn et al., Genome Res., 2013](#)の2群間比較用(3 proliferative samples vs. 3 Ras samples)ヒトRNA-seqカウントデータです。 [パイプライン | ゲノム | 発現変動 | 2群間 | 対応なし | 複製あり | SRP017142\(Neyret-Kahn 2013\)](#)から得られます。

```
in_f <- "srp017142_count_bowtie.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(TCC) #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
dim(data) #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman", #クラスタリング実行結果をoutに格納
                    hclust.method="average", unique.pattern=TRUE) #クラスタリング実行結果をoutに格納

#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
plot(out) #樹形図(デンドログラム)の表示
dev.off() #おまじない
```

[トップページへ](#)

クラスタリング

①赤枠内のコードをテンプレートとして利用
すべくコピーして、②Rエディタ上でペースト。

RGui (64-bit) window showing the R Console and an R Editor window.

R Console:

```
R は多くの貢献者による共同プロジェクト$
詳しくは 'contributors()' と入力して$
また、R や R のパッケージを出版物で$
'citation()' と入力してください。

'demo()' と入力すればデモをみること$
'help()' とすればオンラインヘルプが$
'help.start()' で HTML ブラウザによ$
'q()' と入力すれば R を終了します。

> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files()
[1] "counts_gene.txt"
> |
```

R Editor (無題 - Rエディタ):

```
in_f <- "src/017142_count_bowtie.txt" #入力ファイル名を^
out_f <- "hogel.png" #出力ファイル名を
param_fig <- c(500, 400) #ファイル出力時の

#必要なパッケージをロード
library(TCC) #パッケージの読み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1,
dim(data) #オブジェクトdata

#本番
out <- clusterSample(data, dist.method="spearman", #
hclust.method="average", unique.pattern

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height
plot(out) #樹形図(デンドロ
dev.off() #おまじない
```

クラスタリング

①赤枠内のコードをテンプレートとして利用
すべくコピーして、②Rエディタ上でペースト。
必要最小限の変更は、③ここを…

The screenshot shows the RGui (64-bit) interface. The R Console window displays the following text:

```
R は多くの貢献者による共同プロジェクト$
詳しくは 'contributors()' と入力して$
また、R や R のパッケージを出版物で$
'citation()' と入力してください。

'demo()' と入力すればデモをみること$
'help()' とすればオンラインヘルプが$
'help.start()' で HTML ブラウザによ$
'q()' と入力すれば R を終了します。

> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files()
[1] "counts_gene.txt"
> |
```

The R Editor window (無題 - RIデータ) contains the following code:

```
in_f <- "srp017142_count_bowtie.txt" #入力ファイル名を^
out_f <- "hogel.png" #出力ファイル名を
param_fig <- c(500, 400) #ファイル出力時の

#必要なパッケージをロード
library(TCC) #パッケージの読み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1,
dim(data) #オブジェクトdata

#本番
out <- clusterSample(data, dist.method="spearman", #
hclust.method="average", unique.pattern

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height
plot(out) #樹形図(デンドロ
dev.off() #おまじない
```

A red arrow with the number 3 points to the line `param_fig <- c(500, 400)` in the R Editor window.

クラスタリング

①赤枠内のコードをテンプレートとして利用
すべくコピーして、②Rエディタ上でペースト。
必要最小限の変更は、③ここをこんな感じで
変更するだけ。

RGui (64-bit)

ファイル 編集 パッケージ ウィンドウ ヘルプ



R Console

```
R は多くの貢献者による共同プロジェクトです。
詳しくは 'contributors()' と入力してください。
また、R や R のパッケージを出版物で
'citation()' と入力してください。

'demo()' と入力すればデモをみることも
'help()' とすればオンラインヘルプが
'help.start()' で HTML ブラウザによ
'q()' と入力すれば R を終了します。

> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files()
[1] "counts_gene.txt"
> |
```

無題 - Rエディタ

```
in_f <- "counts_gene.txt" #入力ファイル名を指定してin_fに
out_f <- "hogel.png" #出力ファイル名を
param_fig <- c(500, 400) #ファイル出力時の

#必要なパッケージをロード
library(TCC) #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,
dim(data) #オブジェクトdata

#本番
out <- clusterSample(data, dist.method="spearman", #
hclust.method="average", unique.pattern

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height
plot(out) #樹形図(デンドログラム)
dev.off() #おまじない
```

クラスタリング

①赤枠内のコードをテンプレートとして利用
すべくコピーして、②Rエディタ上でペースト。
必要最小限の変更は、③ここをこんな感じで
変更するだけ。後はコード全体をコピーしてR
Console画面上でペーストするだけ。Rエディ
タを使うメリットとしては、画面のように全選
択した後で (Windowsの場合は)CTRL + R
キーを押すだけで、コピー(CTRL + Cした後
にCTRL + Vすること)と同じ効果がある。

RGui (64-bit)

ファイル 編集 パッケージ ウィンドウ ヘルプ



R Console

```
R は多くの貢献者による共同プロジェクト$
詳しくは 'contributors()' と入力して$
また、R や R のパッケージを出版物で$
'citation()' と入力してください。

'demo()' と入力すればデモをみること$
'help()' とすればオンラインヘルプが$
'help.start()' で HTML ブラウザによ$
'q()' と入力すれば R を終了します。

> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files()
[1] "counts_gene.txt"
> |
```

無題 - Rエディタ

```
in_f <- "counts_gene.txt"
out_f <- "hogel.png" #出力ファイル名を指定
param_fig <- c(500, 400) #ファイル出力時の幅と高さ

#必要なパッケージをロード
library(TCC) #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")
dim(data) #オブジェクトdataの次元

#本番
out <- clusterSample(data, dist.method="spearman", #クラスタリング方法
                      hclust.method="average", unique.pattern=T)

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])
plot(out) #樹形図(デンドログラム)
dev.off() #おまじない
```

クラスタリング

①赤枠内のコードをテンプレートとして利用
すべくコピーして、②Rエディタ上でペースト。
必要最小限の変更は、③ここをこんな感じで
変更するだけ。後はコード全体をコピーしてR
Console画面上でペーストするだけ。Rエディ
タを使うメリットとしては、画面のように全選
択した後で (Windowsの場合は)CTRL + R
キーを押すだけで、コピー(CTRL + Cした後
にCTRL + Vすること)と同じ効果がある。右
クリックで④の部分を選択するのと同じです。

RGui (64-bit)

ファイル 編集 パッケージ ウィンドウ ヘルプ



R Console

```
R は多くの貢献者による共同プロジェクトです。
詳しくは 'contributors()' と入力して
また、R や R のパッケージを出版物で
'citation()' と入力してください。
```

```
'demo()' と入力すればデモをみることに
'help()' とすればオンラインヘルプが
'help.start()' で HTML ブラウザによ
'q()' と入力すれば R を終了します。
```

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files()
[1] "counts_gene.txt"
> |
```

無題 - Rエディタ

```
in_f <- "counts_gene.txt"
out_f <- "hoge1.png"
param_fig <- c(500, 100, 100)
```

```
#必要なパッケージをロード
library(TCC)
```

```
#入力ファイルの読み込み
data <- read.table(in_f, as.is=T)
dim(data)
```

```
#本番
out <- clusterSample(d, n=100, replace=T,
                      hclust.met
```

```
#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])
plot(out) #樹形図 (デンドログラム)
dev.off() #おまじない
```

カーソル行または選択中の R コードを実行 Ctrl+R

やり直し Ctrl+Z

カット Ctrl+X

コピー Ctrl+C

ペースト Ctrl+V

消去

全て選択 Ctrl+A

④

クラスタリング

実行中…。①で見えているのは、入力ファイルの数値行列部分が58,037行×6列だということを表しています。確かに計6サンプルのデータを読み込ませているので妥当ですね。

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ Vignettes

R Console

```
以下のオブジェクトは 'package:edge$  
  
calcNormFactors  
  
>  
> #入力ファイルの読み込み  
> data <- read.table(in_f, header=TRUE)  
> dim(data) [1] 58037 6  
>  
> #本番  
> out <- clusterSample(data, dist.method="spearman",  
+ hclust.method="average")
```

R Editor

```
in_f <- "counts_gene.txt" #入力ファイル名を指定してin_fに本  
out_f <- "hogel.png" #出力ファイル名を指  
param_fig <- c(500, 400) #ファイル出力時の横  
  
#パッケージをロード  
library(TCC) #パッケージの読み込  
  
#データの読み込み  
data <- read.table(in_f, header=TRUE, row.names=1, se  
ca) #オブジェクトdataの  
  
#クラスタリング  
out <- clusterSample(data, dist.method="spearman", #ク  
hclust.method="average", unique.pattern=1  
  
#結果を保存  
png(out_f, pointsize=13, width=param_fig[1], height=pa  
plot(out) #樹形図(デンドログ  
dev.off() #おまじない
```

クラスタリング

実行中…。①で見えているのは、入力ファイルの数値行列部分が58,037行×6列だということを表しています。確かに計6サンプルのデータを読み込ませているので妥当ですね。実行完了。特にエラーなく終わったら、こんな感じに見えます。

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ Vignettes



R Console

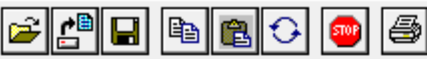
```
> dim(data) $
[1] 58037 6 ①
>
> #本番
> out <- clusterSample(data, dist.me$
+ hclust.method="average"$
>
> #ファイルに保存
> png(out_f, pointsize=13, width=par$
> plot(out) $
> dev.off() $
null device
      1
>
> |
```

```
エディタ
- "counts_gene.txt" #入力ファイル名を指定してin_fに本
<- "hogel.png" #出力ファイル名を指
fig <- c(500, 400) #ファイル出力時の横
パッケージをロード
y(TCC) #パッケージの読み込
ファイルの読み込み
- read.table(in_f, header=TRUE, row.names=1, se
ca) #オブジェクトdataの
clusterSample(data, dist.method="spearman", #ク
hclust.method="average", unique.pattern=1
保存
png(out_f, pointsize=13, width=param_fig[1], height=p
plot(out) #樹形図(デンドログ
dev.off() #おまじない
```

クラスタリング

実行中…。①で見えているのは、入力ファイルの数値行列部分が58,037行×6列だということを表しています。確かに計6サンプルのデータを読み込ませているので妥当ですね。実行完了。特にエラーなく終わったら、こんな感じに見えます。①list.files()で確認。確かに②出力ファイル名として指定した、③hoge1.pngが見えていますね。

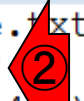
RGui (64-bit)
ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ Vignettes



```
> #本番
> out <- clusterSample(data, dist.me$
+ hclust.method="average$
>
> #ファイルに保存
> png(out_f, pointsize=13, width=par$
> plot(out)
> dev.off()
null device
      1
>
> list.files()
[1] "counts_gene.txt"
[2] "hoge1.png"
> |
```



```
counts_gene.txt" #入力ファイル名を指定してin_fに本
"hogel.png" #出力ファイル名を指
fig <- c(500, 400) #ファイル出力時の横
パッケージをロード
y(TCC) #パッケージの読み込
ファイルの読み込み
read.table(in_f, header=TRUE, row.names=1, se
ca) #オブジェクトdataの
clusterSample(data, dist.method="spearman", #クラ
hclust.method="average", unique.pattern=1
保存
png(out_f, pointsize=13, width=param_fig[1], height=p
plot(out) #樹形図(デンドログ
dev.off() #おまじない
```



クラスタリング

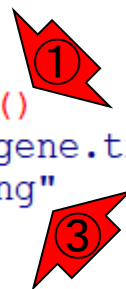
実行中…。①で見えているのは、入力ファイルの数値行列部分が58,037行×6列だということを表しています。確かに計6サンプルのデータを読み込ませているので妥当ですね。実行完了。特にエラーなく終わったら、こんな感じに見えます。①list.files()で確認。確かに②出力ファイル名として指定した、③hoge1.pngが見えていますね。④実物。

RGui (64-bit)
ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ Vignettes

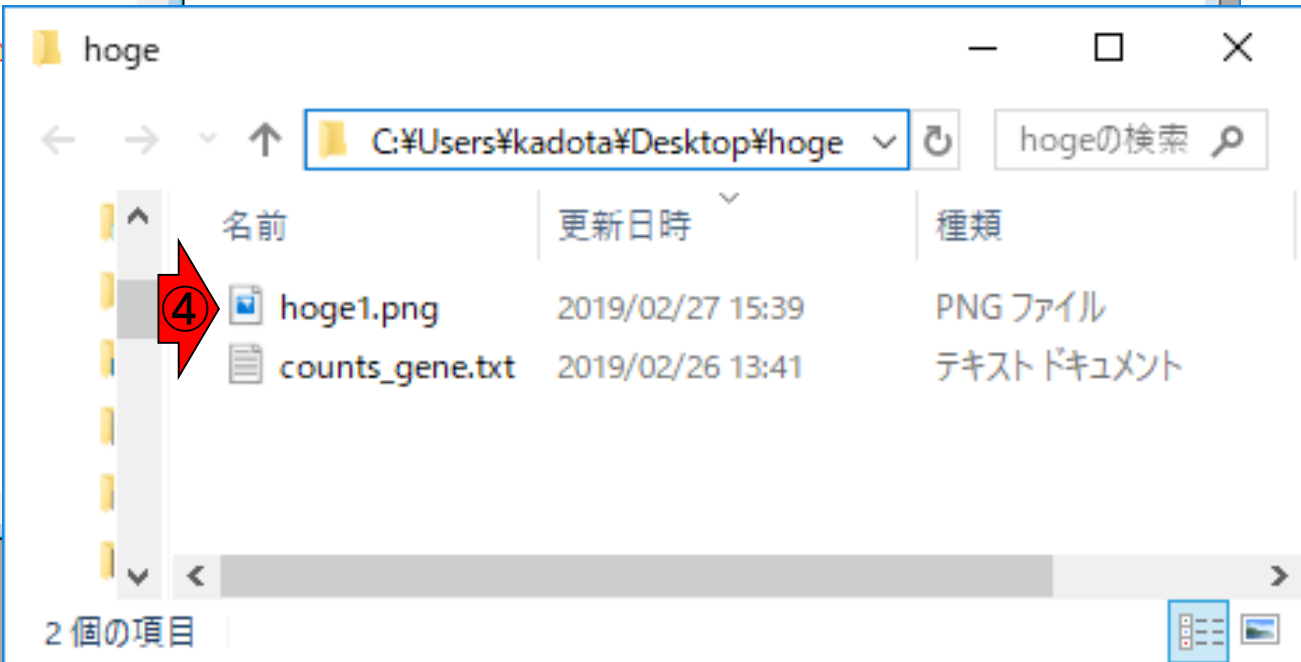


R Console

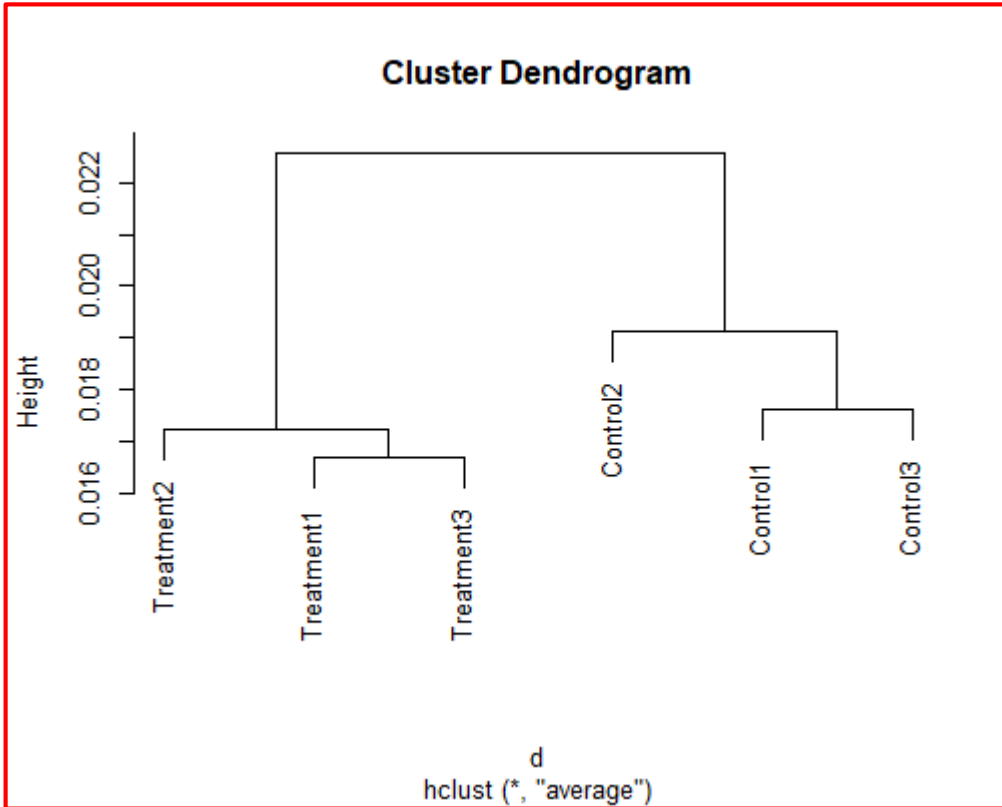
```
> #本番
> out <- clusterSample(data, dist.me$
+                       hclust.method="average$
>
> #ファイルに保存
> png(out_f, pointsize=13, width
> plot(out)
> dev.off()
null device
      1
>
> list.files()
[1] "counts_gene.txt"
[2] "hoge1.png"
> |
```



```
in_f <- "counts_gene.txt" #入力ファイル名を指定してin_fに本
out_f <- "hoge1.png" #出力ファイル名を指
fig <- c(500, 400) #ファイル出力時の横
```



クラスタリング



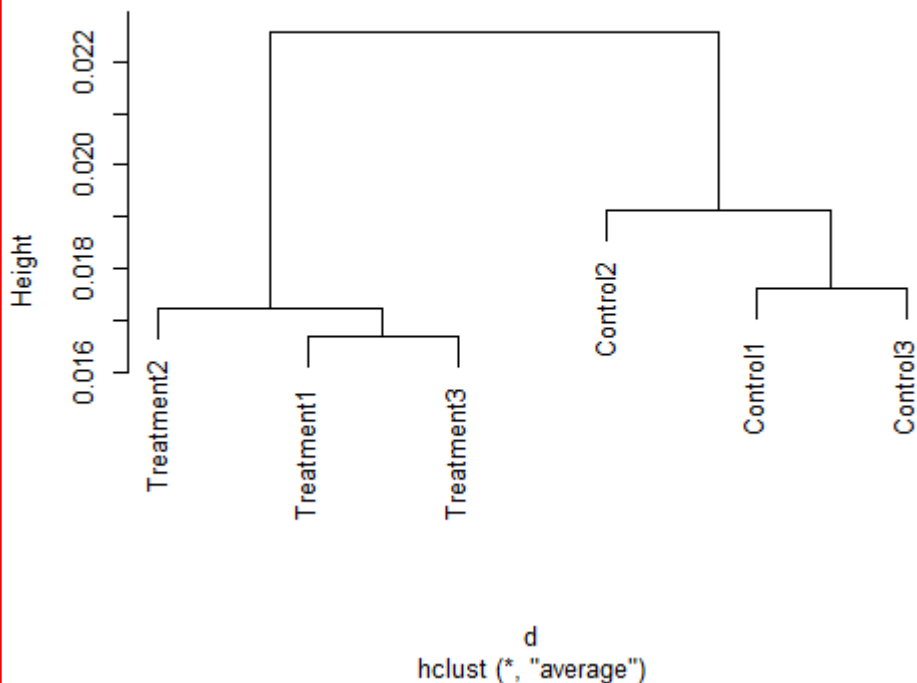
```
counts_gene.txt" #入力ファイル名を指定してin_fに本
hoge1.png" #出力ファイル名を指
k- c(500, 400) #ファイル出力時の様
-ジをロード
c) #パッケージの読み込
読み込み
ad.table(in_f, header=TRUE, row.names=1, se
#オブジェクトdataの
out <- clusterSample(data, dist.method="spearman", #ク
hclust.method="average", unique.pattern=1
#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=p
plot(out) #樹形図(デンドログ
dev.off() #おまじない
```



クラスタリング

赤枠が出力ファイル①hoge1.pngの中身。②横が500ピクセル、③縦が400ピクセルっぽい感じの縦横比になっていますね。

Cluster Dendrogram



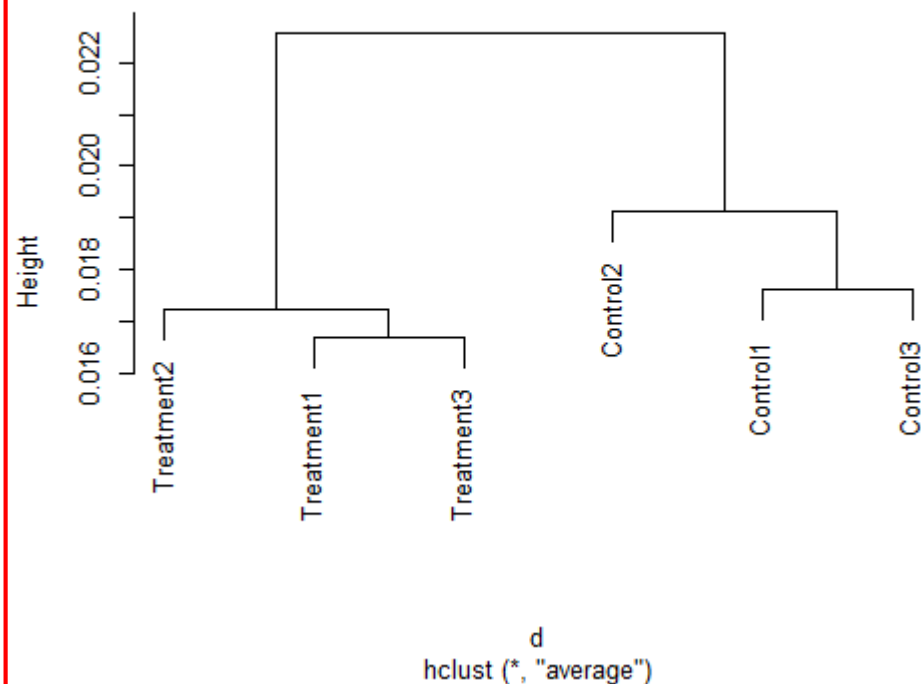
```
counts_gene.txt" #入力ファイル名を指定してin_fに本  
hoge1.png" #出力ファイル名を指  
c(500, 400) #ファイル出力時の横  
#パッケージを読み込  
読み込み  
read.table(in_f, header=TRUE, row.names=1, se  
#オブジェクトdataの
```

```
out <- clusterSample(data, dist.method="spearman", #クラ  
hclust.method="average", unique.pattern=1  
#ファイルに保存  
png(out_f, pointsize=13, width=param_fig[1], height=p  
plot(out) #樹形図(デンドログ  
dev.off() #おまじない
```

クラスタリング

赤枠が出力ファイル①hoge1.pngの中身。②横が500ピクセル、③縦が400ピクセルっぽい感じの縦横比になっていますね。④をいじれば多数のサンプルの場合でも横幅を広げるなどの対策が可能です。

Cluster Dendrogram



```
counts_gene.txt" #入力ファイル名を指定してin_fに本  
"hoge1.png" #出力ファイル名を指  
c(500, 400) #ファイル出力時の横  
#パッケージの読み込  
読み込み  
read.table(in_f, header=TRUE, row.names=1, se  
#オブジェクトdataの
```

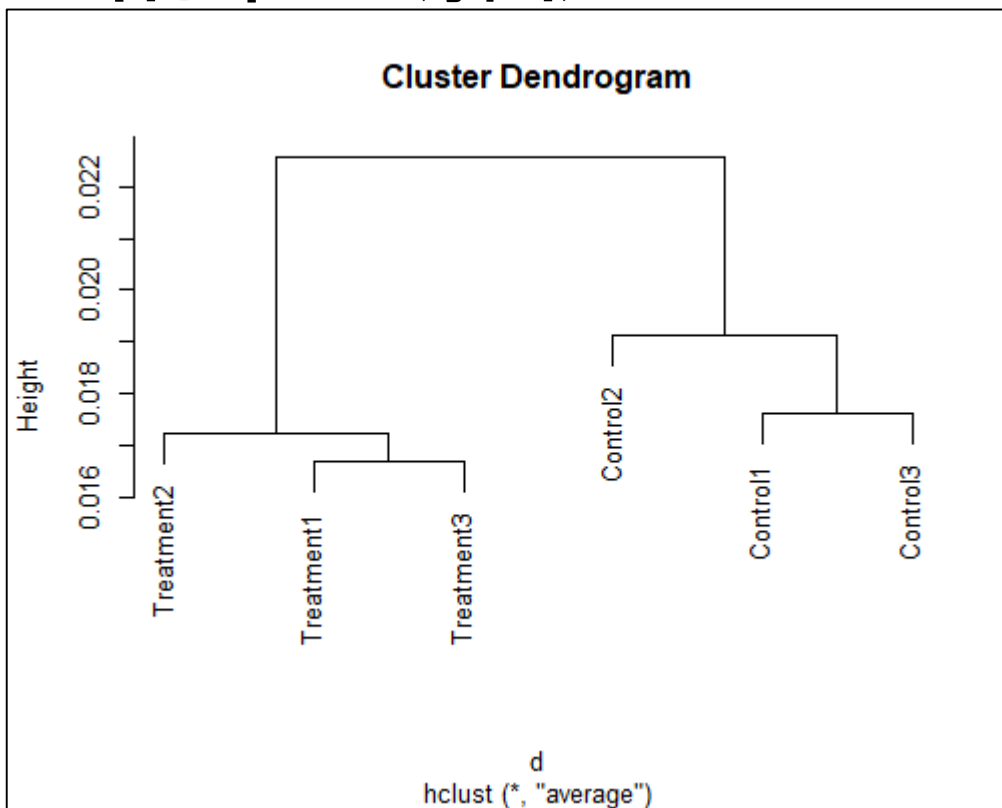
```
out <- clusterSample(data, dist.method="spearman", #ク  
hclust.method="average", unique.pattern=1  
#ファイルに保存  
png(out_f, pointsize=13, width=param_fig[1], height=p  
plot(out) #樹形図(デンドログ  
dev.off() #おまじない
```

Contents

- 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

結果の解釈

サンプル間クラスタリング結果の合理的な解釈について説明します。一定の割合で脱落者がいますが、まずは入力データについてのおさらい。



おさらい

入力データは計6サンプル。②最後の文章から、2群間比較(control vs. TCF7L2)なのだろうと読み取る。正確には、PANC1という膵臓がん細胞中のTCF7L2 (Wnt signaling pathway中の遺伝子でTCF4ともいう)をsiRNAsでknockdownした3例とcontrolの3例を比較したデータセット。これをTreatment vs. Controlとみなしてクラスタリングした結果を眺めているのです。

SRP050497 6 human

We have compared the genome-wide effects on the transcriptome after treatment with ICG-001 (the specific CBP inhibitor) versus C646, a compound that competes with acetyl-coA for the Lys-coA binding pocket of both CBP and p300. We found that both drugs cause large-scale changes in the transcriptome of HCT116 colon cancer cells and PANC1 pancreatic cancer cells, and reverse some tumor-specific changes in gene expression. Interestingly, although the epigenetic inhibitors affect cell cycle pathways in both the colon and pancreatic cancer cell lines, the WNT signaling pathway was affected only in the colon cancer cells. Notably, WNT target genes were similarly down-regulated after treatment of HCT116 with C646 as with ICG-001. Overall design: To identify genes affected by direct targeting of a component of the transcriptional complex implicated in WNT regulation, we used siRNAs to knockdown TCF7L2 in PANC1 cells. Cells were treated with control siRNAs or siRNAs specific for TCF7L2 and RNA was analyzed by RNA-seq.

RSE v2 counts v2 RSE counts v1 RSE v2 counts v2 RSE counts v1 RSE jx_bed jx_cov counts RSE v2 RSE v1 link v2 v1

Showing 11 to 20 of 29 entries (filtered from 2,039 total entries)

Previous 1 2 3 Next

Download list of studies matching search results

Note that GTEx is separated from this list.

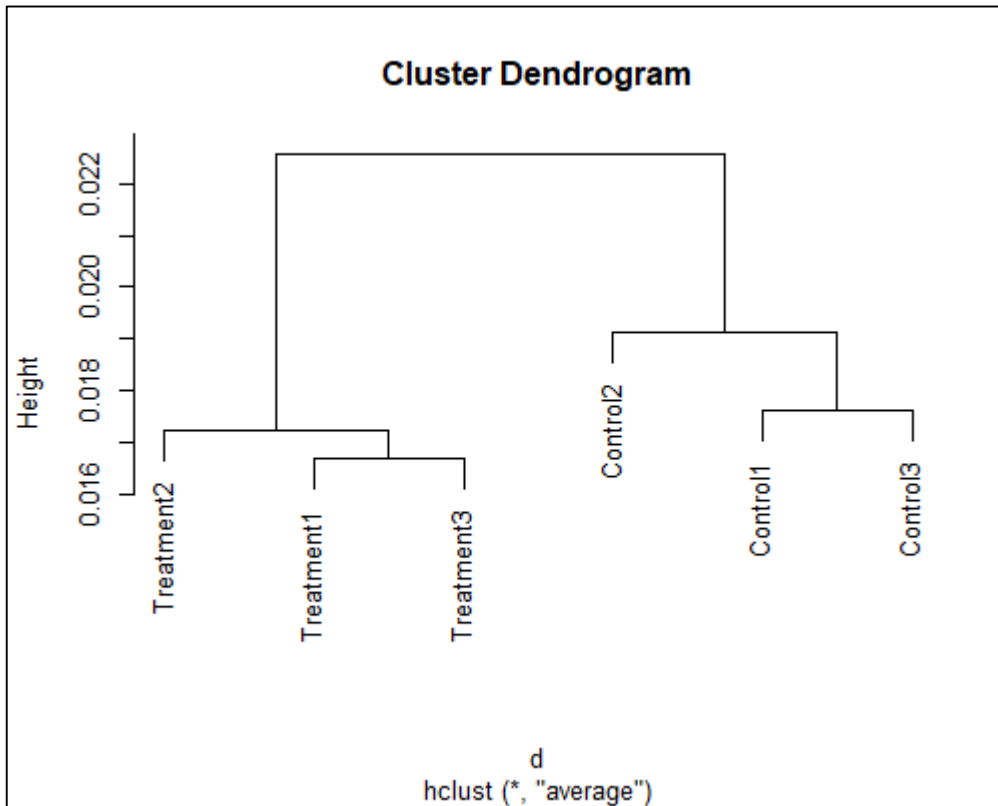
おさらい

入力データは58,037遺伝子×6サンプル。どのサンプルとどのサンプルが似ているかを眺め、全体としておかしくないか(外れサンプルがあるか)などをチェックする目的で、サンプル間クラスタリングを行いました。

gene_id	Control1	Control2	Control3	Treatment1	Treatment2	Treatment3
ENSG00000000003.14	11156	9957	14914	15968	15138	16747
ENSG00000000005.5	0	0	0	0	0	0
ENSG00000000419.12	64023	55036	88460	62819	67807	70965
ENSG00000000457.13	24666	22830	37528	31189	31249	36804
ENSG00000000460.16	42199	38508	58545	36169	39613	42044
ENSG00000000938.12	0	0	0	0	0	0
ENSG00000000971.15	647	628	794	450	650	446
ENSG00000001036.13	89682	81128	123915	121702	121412	143520
ENSG00000001084.10	86325	77847	120691	73226	76177	87010
ENSG00000001167.14	57229	49468	80253	89081	92579	109634
ENSG00000001460.17	16982	16121	25284	19119	18513	19189
ENSG00000001461.16	16155	14981	27084	22228	20813	22194

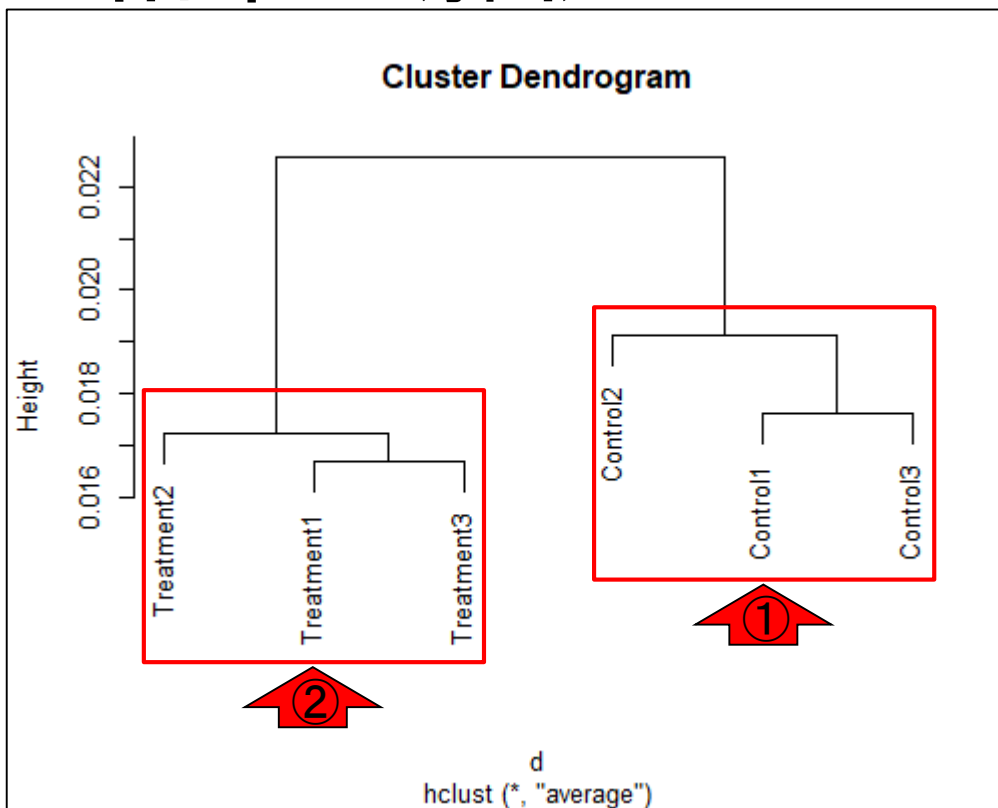
counts_gene.txt

結果の解釈

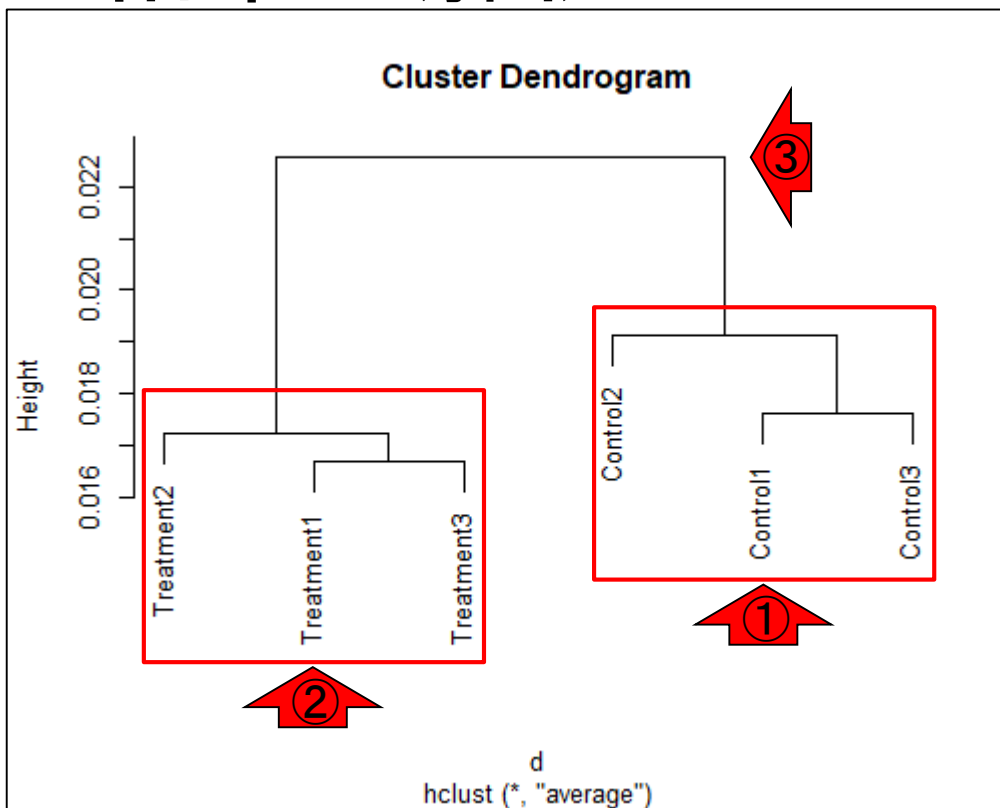


結果の解釈

私はこの結果をみて「大丈夫」と判断します。理由は、①Control群、②Treatment群ともに、外れサンプルがないように見えるからです。

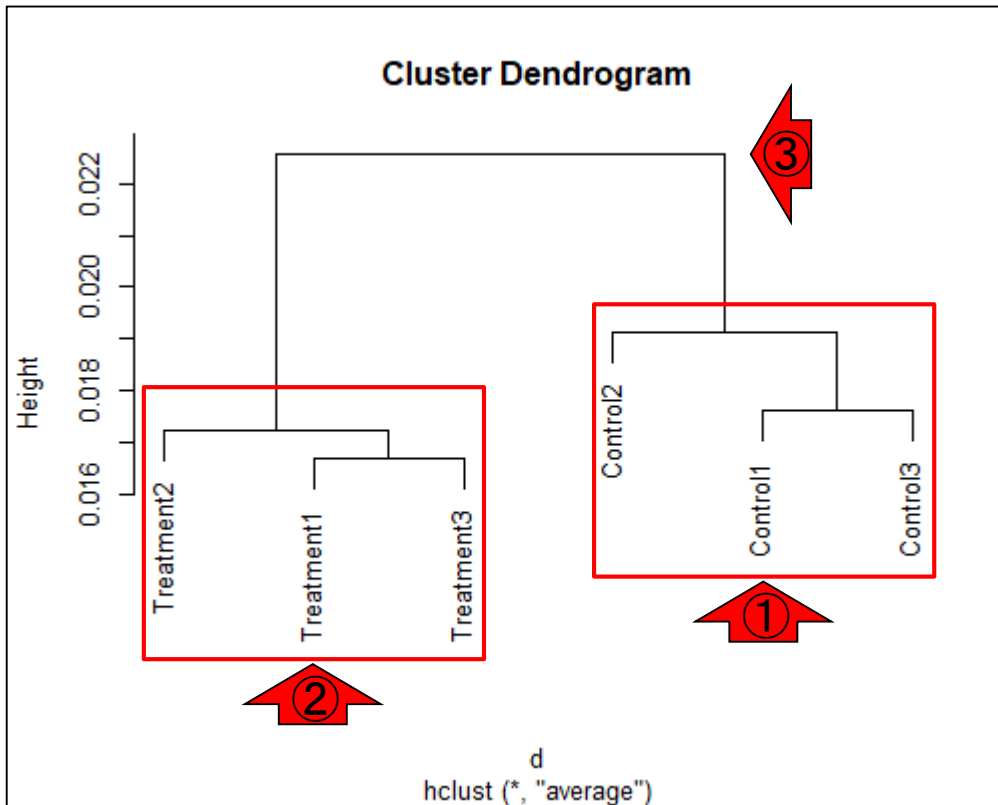


結果の解釈



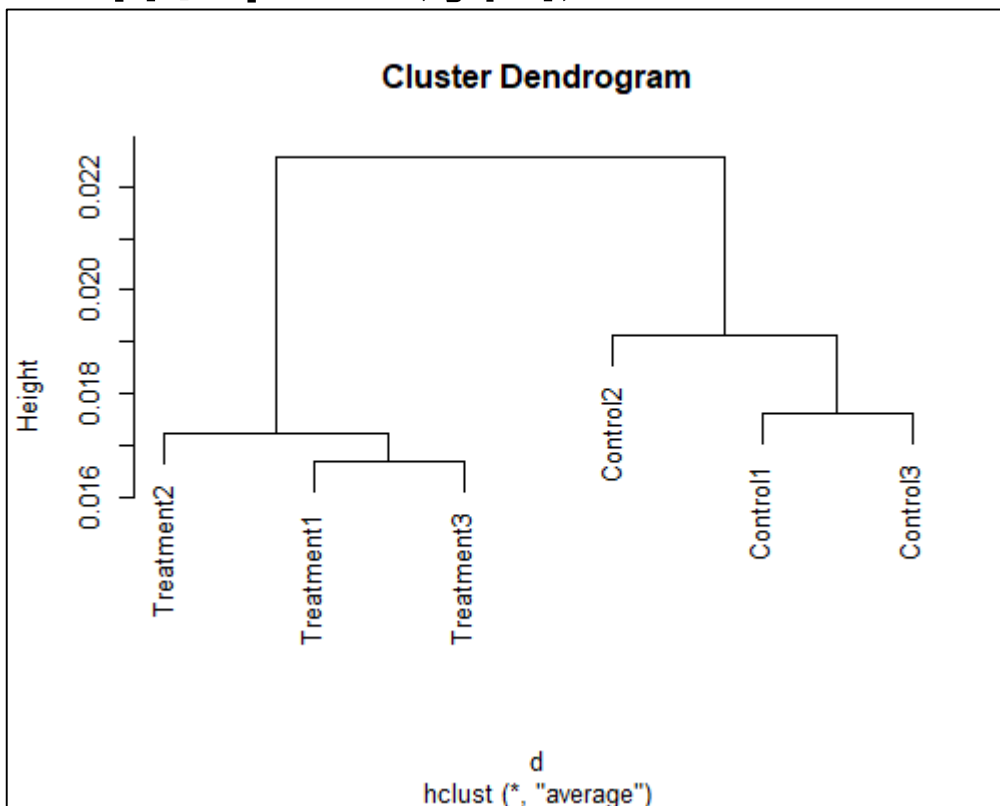
私はこの結果をみて「大丈夫」と判断します。理由は、①Control群、②Treatment群ともに、外れサンプルがないように見えるからです。また、①と②でみられる群内の平均類似度が、③群間の平均類似度よりも高いからです。

結果の解釈



私はこの結果をみて「大丈夫」と判断します。理由は、①Control群、②Treatment群ともに、外れサンプルがないように見えるからです。また、①と②でみられる群内の平均類似度が、③群間の平均類似度よりも高いからです。この実験デザインは、siRNAsでknockdown処理前後の発現パターンに違いがあるかどうかを調べるためのものです。比較する群間で発現に差がありそう（つまり発現変動遺伝子がありそう）だと判断します。決して科学的な表現ではありませんが、「そうあってほしい」という結論が「得られそうなクラスタリング結果」なので「大丈夫」なわけです。

結果の解釈



私はこの結果をみて「大丈夫」と判断します。理由は、①Control群、②Treatment群ともに、外れサンプルがないように見えるからです。また、①と②でみられる群内の平均類似度が、③群間の平均類似度よりも高いからです。この実験デザインは、siRNAsでknockdown処理前後の発現パターンに違いがあるかどうかを調べるためのものです。比較する群間で発現に差がありそう（つまり発現変動遺伝子がありそう）だと判断します。決して科学的な表現ではありませんが、「そうあってほしい」という結論が「得られそうなクラスタリング結果」なので「大丈夫」なわけです。実際には、この後「比較する群間には差がない」という帰無仮説に基づく統計解析を行います。もし本当に差がなければ、クラスタリング結果におけるControlとTreatmentのサンプルが入り混じった感じになるのです。

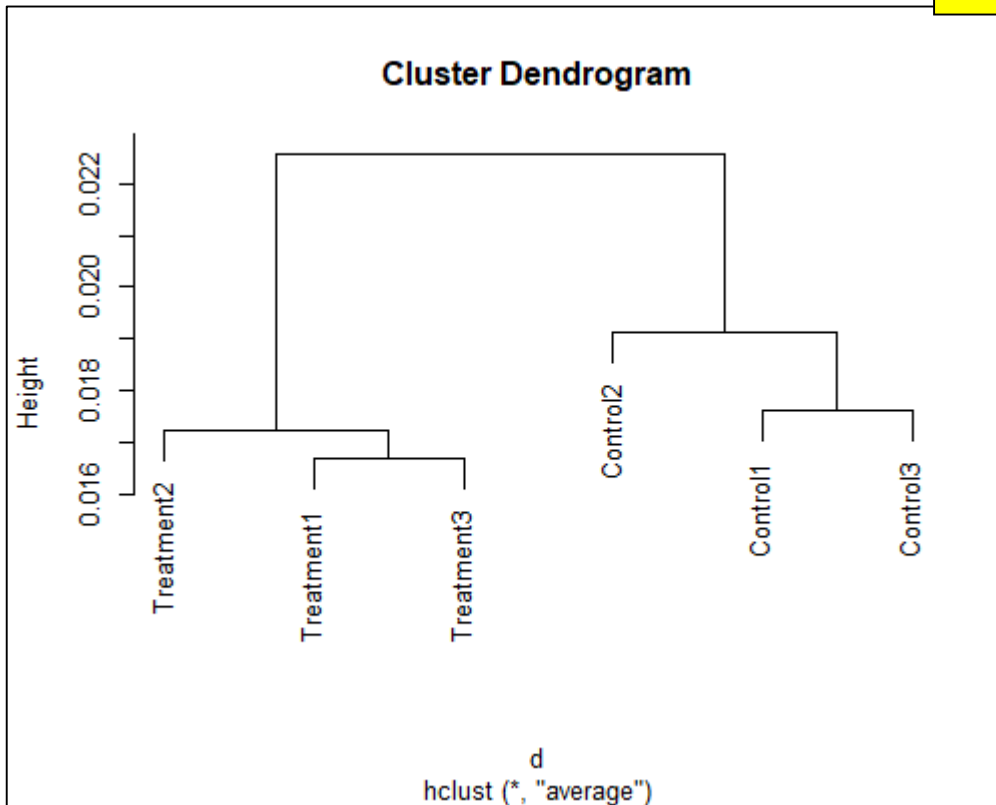


Contents

- 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

クラスタリング結果の解釈は主観的になりがちですが、比較したい任意の群間(例: ControlとTreatment間)がどれだけ似ているかを客観的に示す指標も存在します。

客観的な指標



Biol Proced Online. 2018 Mar 1;20:5. doi: 10.1186/s12575-018-0067-8. eCollection 2018.

Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results.

Zhao S¹, Sun J¹, Shimizu K¹, Kadota K¹.

Author information

Abstract

BACKGROUND: Hierarchical Sample clustering (HSC) is widely performed to examine associations within expression data obtained from microarrays and RNA sequencing (RNA-seq). Researchers have investigated the HSC results with several possible criteria for grouping (e.g., sex, age, and disease types). However, the evaluation of arbitrary defined groups still counts in subjective visual inspection.

RESULTS: To objectively evaluate the degree of separation between groups of interest in the HSC dendrogram, we propose to use *Silhouette* scores. Silhouettes was originally developed as a graphical aid for the validation of data clusters. It provides a measure of how well a sample is classified when it was assigned to a cluster by according to both the tightness of the clusters and the separation between them. It ranges from 1.0 to -1.0, and a larger value for the average silhouette (AS) over all samples to be analyzed indicates a higher degree of *cluster* separation. The basic idea to use an AS is to replace the term *cluster* by *group* when calculating the scores. We investigated the validity of this score using simulated and real data designed for differential expression (DE) analysis. We found that larger (or smaller) AS values agreed well with both higher (or lower) degrees of separation between different groups and higher percentages of differentially expressed genes (P_{DEG}). We also found that the AS values were generally independent on the number of replicates (N_{rep}). Although the P_{DEG} values depended on N_{rep} , we confirmed that both AS and P_{DEG} values were close to zero when samples in the data showed an intermingled nature between the groups in the HSC dendrogram.

CONCLUSION: Silhouettes is useful for exploring data with predefined group labels. It would help provide both an objective evaluation of HSC dendrograms and insights into the DE results with regard to the compared groups.

客観的な指標

平均シルエットスコア (Average Silhouette score; AS) の①取りうる範囲は $[-1, 1]$ であり、②値が大きいほど比較する群間の分離度が高いことを示す。AS ≈ 0 が分離度がほぼゼロ(入り混じっている状態)に相当する。

Biol Proced Online. 2018 Mar 1;20:5. doi: 10.1186/s12575-018-0

Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results.

Zhao S¹, Sun J¹, Shimizu K¹, Kadota K¹.

Author information

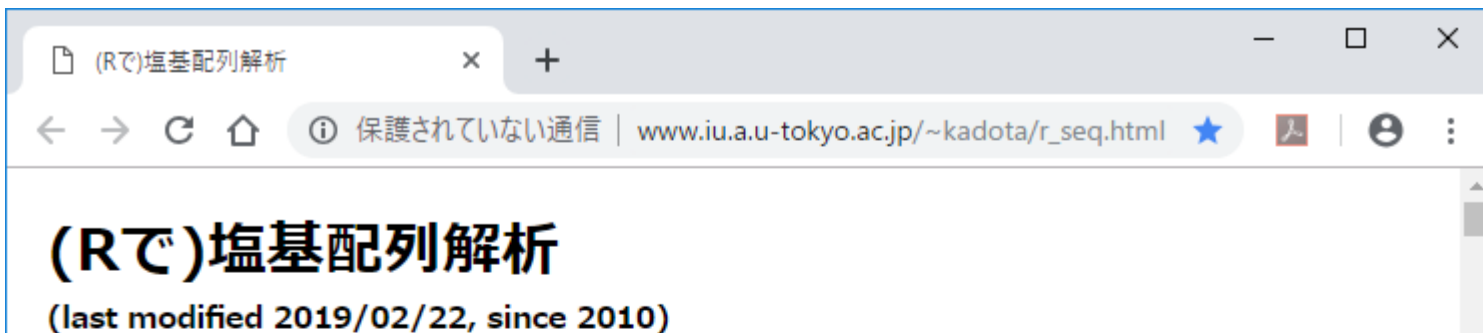
Abstract

BACKGROUND: Hierarchical Sample clustering (HSC) is widely performed to examine associations within expression data obtained from microarrays and RNA sequencing (RNA-seq). Researchers have investigated the HSC results with several possible criteria for grouping (e.g., sex, age, and disease types). However, the evaluation of arbitrary defined groups still counts in subjective visual inspection.

RESULTS: To objectively evaluate the degree of separation between groups of interest in the HSC dendrogram, we propose to use *Silhouette* scores. Silhouettes was originally developed as a graphical aid for the validation of data clusters. It provides a measure of how well a sample is classified when it was assigned to a cluster by according to both the tightness of the clusters and the separation between them. It ranges from 1.0 to - 1.0, and a larger value for the average silhouette (AS) over all samples to be analyzed indicates a higher degree of cluster separation. The basic idea to use an AS is to replace the term *cluster* by *group* when calculating the scores. We investigated the validity of this score using simulated and real data designed for differential expression (DE) analysis. We found that larger (or smaller) AS values agreed well with both higher (or lower) degrees of separation between different groups and higher percentages of differentially expressed genes (P_{DEG}). We also found that the AS values were generally independent on the number of replicates (N_{rep}). Although the P_{DEG} values depended on N_{rep} , we confirmed that both AS and P_{DEG} values were close to zero when samples in the data showed an intermingled nature between the groups in the HSC dendrogram.

CONCLUSION: Silhouettes is useful for exploring data with predefined group labels. It would help provide both an objective evaluation of HSC dendrograms and insights into the DE results with regard to the compared groups.

シルエットスコア



このウェブページのR関連部分は、[Macintosh2018.11.27版](#)に従って、[Macintosh2019.01.15版](#)で自習し、[書籍・学会誌など](#)を切り分けてサ

- 解析 | 一般 | アラインメント | ペアワイズ | 応用 | [Biostrings](#) (last modified 2016/12/2)
- 解析 | 一般 | アラインメント | マルチプル | [DECIPHER\(Wright_2015\)](#) (last modified 2016/12/2)
- 解析 | 一般 | アラインメント | マルチプル | [msa\(Bordenhofer_2015\)](#) (last modified 2016/12/2)
- 解析 | 一般 | [Silhouette scores\(シルエットスコア\)](#) (last modified 2018/06/16)
- 解析 | 一般 | [パターンマッチング](#) (last modified 2018/06/19)
- 解析 | 一般 | [GC含量\(GC contents\)](#) (last modified 2015/09/12)
- 解析 | 一般 | [Sequence logos | について](#) (last modified 2018/06/29)
- 解析 | 一般 | Sequence logos | [seqLogo](#) (last modified 2018/06/29)
- 解析 | 一般 | Sequence logos | [ggseqlogo\(Wagih_2017\)](#) (last modified 2018/06/29)
- 解析 | 一般 | 上流配列解析 | [LDSS\(Yamamoto_2007\)](#) (last modified 2015/02/19)
- 解析 | 一般 | 上流配列解析 | [Relative Appearance Ratio\(Yamamoto_2011\)](#) (last modified 2015/02/19)
- 解析 | 基礎 | k-mer | ゲノムサイズ推定(基礎) | [qrqc](#) (last modified 2016/01/06)
- 解析 | 基礎 | [平均-分散プロット | について](#) (last modified 2015/11/11)
- 解析 | 基礎 | 平均-分散プロット | [Technical replicates](#) (last modified 2014/02/18)



シルエットスコア

①得られた平均シルエットスコアの解釈について、ごちゃごちゃ書いています。

(Rで)塩基配列解析

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

解析 | 一般 | Silhouette scores(シルエットスコア) NEW

Silhouetteスコアの新たな使い道提唱論文(Zhao et al., Biol. Proc. Online, 2018)の利用法を説明します。入力
は「解析 | 発現変動 | 2群間 | 対応なし | 複製あり | ICC(Sun 2013)」などと同じく、遺伝子発現行列データ
と比較したいグループラベル情報 (Group1が1、Group2が2みたいなやつ) です。出力は、Average
Silhouette(AS値)というスカラー情報 (1つの数値) です。AS値の取り得る範囲は $[-1, 1]$ で、数値が大きいほど
指定したグループ間の類似度が低いことを意味し、発現変動解析結果としてDifferentially Expressed Genes
(DEGs)が沢山得られる傾向にあります。逆に、AS値が低い (通常は-1に近い値になることはほぼ皆無で、相関
係数と同じく0に近い) ほど指定したグループ間の類似度が高いことを意味し、DEGがほとんど得られない傾向
にあります。論文中で提案している使い道としては、「発現変動解析を行ってDEGがほとんど得られなかった場
合に、サンプル間クラスタリング(SC)結果とAS値を提示して、(客観的な数値情報である) AS値が0に近い値
だったのでDEGがないのは妥当だね」みたいなdiscussionに使ってもらえればと思っています。RNA-seqカウ
ントデータでもマイクロアレイデータでも使えます。

例題の多くは、[サンプルデータ42](#)の20,689 genes×18 samplesのリアルカウントデータ
([sample blekhman 18.txt](#))を入力としています。ヒトHomo sapiens; HS)のメス3サンプル(HSF1-3)とオス
3サンプル(HSM1-3)、チンパンジー(Pan troglodytes; PT)のメス3サンプル(PTF1-3)とオス3サンプル(PTM1-
3)、アカゲザル(Rhesus macaque; RM)のメス3サンプル(RMF1-3)とオス3サンプル(RMM1-3)の並びになっ
ています。つまり、以下のような感じです。FはFemale(メス)、MはMale(オス)を表します。

ヒト(1-6列目): HSF1, HSF2, HSF3, HSM1, HSM2, and HSM3

チンパンジー(7-12列目): PTF1, PTF2, PTF3, PTM1, PTM2, and PTM3

アカゲザル(13-18列目): RMF1, RMF2, RMF3, RMM1, RMM2, and RMM3

「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. HSF vs. PTFの場合 :

[トップページ](#)へ

HSF (ヒトメス) データが存在する1-3列目と、PTF (チンパンジーマス) データが存在する 7-9 列目のデー

シルエットスコア

①の最初のほうでも書いていますが、例題1-3は、②原著論文と全く同じ結果が得られます。例題4-5は、原著論文と同じ入力データですが、3群間比較でも対応可能だという実例を示しています。例題6が理解しやすい形式になっているので、そちらをテンプレートとして利用します。

(Rで)塩基配列解析 × +
保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_s

解析 | 一般 | Silhouette scores(シルエットスコア) NEW

Silhouetteスコアの新たな使い道提唱論文([Zhao et al., Biol. Proc. Online, 2018](#))の利用法を説明します。入力は「解析 | 発現変動 | 2群間 | 対応なし | 複製あり | [TCC\(Sun 2013\)](#)」などと同じく、遺伝子発現行列データと比較したいグループラベル情報 (Group1が1、Group2が2みたいなやつ) です。出力は、Average Silhouette(AS値)というスカラー情報 (1つの数値) です。AS値の取り得る範囲は[-1, 1]で、数値が大きいほど指定したグループ間の類似度が低いことを意味し、発現変動解析結果としてDifferentially Expressed Genes (DEGs)が沢山得られる傾向にあります。逆に、AS値が低い (通常は-1に近い値になることはほぼ皆無で、相関係数と同じく0に近い) ほど指定したグループ間の類似度が高いことを意味し、DEGがほとんど得られない傾向にあります。論文中で提案している使い道としては、「発現変動解析を行ってDEGがほとんど得られなかった場合に、サンプル間クラスタリング(SC)結果とAS値を提示して、(客観的な数値情報である) AS値が0に近い値だったのでDEGがないのは妥当だね」みたいなdiscussionに使ってもらえればと思っています。RNA-seqカウントデータでもマイクロアレイデータでも使えます。

例題の多くは、[サンプルデータ42](#)の20,689 genes×18 samplesのリアルカウントデータ ([sample blekhman 18.txt](#))を入力としています。ヒトHomo sapiens; HS)のメス3サンプル(HSF1-3)とオス3サンプル(HSM1-3)、チンパンジー(Pan troglodytes; PT)のメス3サンプル(PTF1-3)とオス3サンプル(PTM1-3)、アカゲザル(Rhesus macaque; RM)のメス3サンプル(RMF1-3)とオス3サンプル(RMM1-3)の並びになっています。つまり、以下のような感じです。FはFemale(メス)、MはMale(オス)を表します。

ヒト(1-6列目): HSF1, HSF2, HSF3, HSM1, HSM2, and HSM3

チンパンジー(7-12列目): PTF1, PTF2, PTF3, PTM1, PTM2, and PTM3

アカゲザル(13-18列目): RMF1, RMF2, RMF3, RMM1, RMM2, and RMM3

「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ー。

1. HSF vs. PTFの場合 :

[トップページ](#) ▼

HSF (ヒトメス) データが存在する1-3列目と、PTF (チンパンジーメス) データが存在する 7-9 列目のデー

シルエットスコア

①の最初のほうでも書いていますが、例題1-3は、②原著論文と全く同じ結果が得られます。例題4-5は、原著論文と同じ入力データですが、3群間比較でも対応可能だという実例を示しています。例題6が理解しやすい形式になっているので、そちらをテンプレートとして利用します。③例題6。④入力ファイル名と、⑤比較する群ごとのサンプル数を指定するだけ。出力は⑥AS値。

(Rで)塩基配列解析

x +

③ 保護されていない通信 | www.iu.a.u-tokyo.ac.jp/~kadota/r_s

6. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data)

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～gene_2000までがDEG(最初の1800個がG1群で高発現、残りの200個がgene_10000までがnon-DEGであることが既知です。

```

in_f <- "data_hyp/data_3vs3.txt" ④ #入力ファイル名を指定してin_fに格納
param_G1 <- 3 } ⑤ #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

#必要なパッケージをロード
library(cluster) #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", quote="")#in_fで指定したファイルを読み込み
data.cl <- c(rep(1, param_G1), rep(2, param_G2))#G1群を1、G2群を2としたベクトルdata.clを作成
dim(data) #オブジェクトdataの行数と列数を表示

#前処理(フィルタリング)
obj <- as.logical(rowSums(data) > 0) #条件を満たすかどうかを判定した結果をobjに格納
data <- unique(data[obj,]) #objがTRUEとなる行のみ抽出し、ユニークパターンのみにし
dim(data) #オブジェクトdataの行数と列数を表示

#本番(AS値の計算)
d <- as.dist(1 - cor(data, method="spearman"))#サンプル間の距離を計算し、結果をdに格納
AS <- mean(silhouette(data.cl, d)[, "sil_width"])#サンプルごとのSilhouette scoreを計算し、
AS #AS値を表示

```

[トップページ](#)▲

⑥

おさらい

①手元ファイルの入カデータは、58,037遺伝子×6サンプル。②最初のControl群が3サンプル、③残りのTreatment群が3サンプルなので…

gene_id	Control1	Control2	Control3	Treatment1	Treatment2	Treatment3
ENSG00000000003.14	11156	9957	14914	15968	15138	16747
ENSG00000000005.5	0	0	0	0	0	0
ENSG00000000419.12	64023	55036	88460	62819	67807	70965
ENSG00000000457.13	24666	22830	37528	31189	31249	36804
ENSG00000000460.16	42199	38508	58545	36169	39613	42044
ENSG00000000938.12	0	0	0	0	0	0
ENSG00000000971.15	647	628	794	450	650	446
ENSG00000001036.13	89682	81128	123915	121702	121412	143520
ENSG00000001084.10	86325	77847	120691	73226	76177	87010
ENSG00000001167.14	57229	49468	80253	89081	92579	109634
ENSG00000001460.17	16982	16121	25284	19119	18513	19189
ENSG00000001461.16	16155	14981	27084	22228	20813	22194

① counts_gene.txt

シルエットスコア

(Rで塩基配列解析

x +

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

6. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata_3vs3.txt)の場合:

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)です。gene_1～gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001～gene_10000までがnon-DEGであることが既知です。

```

in_f <- "data_hypodata_3vs3.txt" ① #入力ファイル名を指定してin_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

#必要なパッケージをロード
library(cluster) #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", quote="")#in_fで指定したファイルを読み込み
data.cl <- c(rep(1, param_G1), rep(2, param_G2))#G1群を1、G2群を2としたベクトルdata.clを作成
dim(data) #オブジェクトdataの行数と列数を表示

#前処理(フィルタリング)
obj <- as.logical(rowSums(data) > 0) #条件を満たすかどうかを判定した結果をobjに格納
data <- unique(data[obj,]) #objがTRUEとなる行のみ抽出し、ユニークパターンのみにし
dim(data) #オブジェクトdataの行数と列数を表示

#本番(AS値の計算)
d <- as.dist(1 - cor(data, method="spearman"))#サンプル間の距離を計算し、結果をdに格納
AS <- mean(silhouette(data.cl, d)[, "sil_width"])#サンプルごとのSilhouette scoreを計算し、
AS #AS値を表示

```

[トップページ](#)⤴

シルエットスコア

```
in_f <- "data_hypodata_3vs3.txt" #入力ファイル名を^
param_G1 <- 3 #G1群のサンプル:
param_G2 <- 3 #G2群のサンプル:

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

#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,
data.cl <- c(rep(1, param_G1), rep(2, param_G2)) #G1
dim(data) #オブジェクトdata

#前処理 (フィルタリング)
obj <- as.logical(rowSums(data) > 0) #条件を満たすか
data <- unique(data[obj,]) #objがTRUEとな
dim(data) #オブジェクトdata

#本番 (AS値の計算)
d <- as.dist(1 - cor(data, method="spearman")) #サンブ
AS <- mean(silhouette(data.cl, d)[, "sil_width"]) #
```


シルエットスコア

テンプレートのRコードをコピーして、①入力ファイル名部分のみ変更して、コピー実行した結果。

```

RGui (64-bit)
ファイル 編集 パッケージ ウィンドウ ヘルプ

R Console
[1] 58037      6
>
> #前処理 (フィルタリング)
> obj <- as.logical(rowSums(data)
> data <- unique(data[obj,])
> dim(data)
[1] 34114      6
>
> #本番 (AS値の計算)
> d <- as.dist(1 - cor(data, meth
> AS <- mean(silhouette(data.cl,
> AS
[1] 0.2083361
>
> |

無題 - RIデータ
in_f <- "counts_gene.txt" #入力ファイル名を指定して:
param_G1 <- 3 #G1群のサンプル:
param_G2 <- 3 #G2群のサンプル:

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

#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,
data.cl <- c(rep(1, param_G1), rep(2, param_G2))#G1
dim(data) #オブジェクトdata

#前処理 (フィルタリング)
obj <- as.logical(rowSums(data) > 0) #条件を満たすか
data <- unique(data[obj,]) #objがTRUEとな
dim(data) #オブジェクトdata

#本番 (AS値の計算)
d <- as.dist(1 - cor(data, method="spearman"))#サンパ
AS <- mean(silhouette(data.cl, d)[, "sil_width"])#

```


シルエットスコア

テンプレートのRコードをコピーして、①入力ファイル名部分のみ変更して、コピー実行した結果。
②平均シルエットスコア(AS値)は0.2083361であり、0よりも大きい。これくらいの数値のときに...

RGui (64-bit)

ファイル 編集 パッケージ ウィンドウ ヘルプ

R Console

```
[1] 58037      6
>
> #前処理 (フィルタリング)
> obj <- as.logical(rowSums(data)
> data <- unique(data[obj,])
> dim(data)
[1] 34114      6
>
> #本番 (AS値の計算)
> d <- as.dist(1 - cor(data, meth
> AS <- mean(silhouette(data.cl,
> AS
[1] 0.2083361
>
> |
```

無題 - RIデータ

```
in_f <- "counts_gene.txt"
param_G1 <- 3
param_G2 <- 3

#必要なパッケージをロード
library(cluster)

#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,
data.cl <- c(rep(1, param_G1), rep(2, param_G2))#G1
dim(data)

#前処理 (フィルタリング)
obj <- as.logical(rowSums(data) > 0)
data <- unique(data[obj,])
dim(data)

#本番 (AS値の計算)
d <- as.dist(1 - cor(data, method="spearman"))#サンブ
AS <- mean(silhouette(data.cl, d)[, "sil_width"])
```

①

②

シルエットスコア

テンプレートのRコードをコピーして、①入力ファイル名部分のみ変更して、コピー実行した結果。
 ②平均シルエットスコア(AS値)は0.2083361であり、0よりも大きい。これくらいの数値のときに、
 ③これくらいの分離度。ですので、ASが0.05くらいでも発現変動遺伝子はそこそこ得られます。

RGui (64-bit)

ファイル 編集 パッケージ ウィンドウ ヘルプ

R Console

```
[1] 58037      6
>
> #前処理 (フィルタリング)
> obj <- as.logical(...)
> data <- unique(data)
> dim(data)
[1] 34114      6
>
> #本番 (AS値の計算)
> d <- as.dist(1 - cor(...))
> AS <- mean(silhouette(data.cl, d)[, "sil_width"])#
> AS
[1] 0.2083361
>
> |
```

Cluster Dendrogram

```
#本番 (AS値の計算)
d <- as.dist(1 - cor(data, method="spearman"))#サンプル間相関
AS <- mean(silhouette(data.cl, d)[, "sil_width"])#平均シルエットスコア
```

②

③

Contents

- 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

TCCパッケージ

①さきほどのサンプル間クラスタリングの結果は、②TCCというパッケージが提供する、③clusterSampleという関数を利用しています。

(Rで)塩基配列解析

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

解析 | クラスタリング | サンプル間 | TCC(Sun_2013)

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。clusterSample関数を利用した頑健なクラスタリング結果を返します。多群間比較用の推奨ガイドライン提唱論文 ([Tang et al., BMC Bioinformatics, 2015](#))中でもこの関数を用いています(2015/11/05追加)。xlsx形式ファイルを入力とするやり方も追加しました(2015/11/15)。

「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ー。

1. 59,857 genes×6 samplesのリアルデータ([srp017142_count_bowtie.txt](#))の場合 :

[Neyret-Kahn et al., Genome Res., 2013](#)の2群間比較用(3 proliferative samples vs. 3 Ras samples)ヒトRNA-seqカウントデータです。 [パイプライン | ゲノム | 発現変動 | 2群間 | 対応なし | 複製あり | SRP017142\(Neyret-Kahn 2013\)](#)から得られます。

```
in_f <- "srp017142_count_bowtie.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(TCC) #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
dim(data) #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman", #クラスタリング実行結果をoutに格納
                    hclust.method="average", unique.pattern=TRUE) #クラスタリング実行結果をoutに格納

#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
plot(out) #樹形図(デンドログラム)の表示
dev.off() #おまじない
```

[トップページへ](#)

TCCパッケージ

①TCCの原著論文。これは本来は発現変動遺伝子検出部分を売りにしているパッケージであり、クラスタリングの関数はむしろおまけという位置づけです。このGUI(グラフィカルユーザーインターフェース)版がTCC-GUI。

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

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

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

Author information



Abstract

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

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

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

PMC Full text

Save items

Similar articles

Evaluation of methods for differential expression [BMC Bioinformatics. 2015]

A normalization strategy for comparing tag count data [Algorithms Mol Biol. 2012]

A comparison of per sample global scaling and per gene [PLoS One. 2017]

Review Statistical detection of differentially expressed genes [Brief Bioinform. 2016]

Review A statistical framework for applying RNA-seq [Chemosphere. 2017]

See reviews...

See all...

Cited by 75 PubMed Central articles

Combined treatment with HMGN1 and anti-CD4 [J Immunother Cancer. 2019]

Susceptibility of Escherichia coli O157:H7 grown at low [Sci Rep. 2019]

TCC-GUI

手元の入力ファイル(counts_gene.txt)をアップロードして、どうやってさきほどの結果が得られるのかまでをざっくりと解説します。① Step1のあたりをクリック。

TCC-GUI: Graphical User Interface for TCC package

Welcome to TCC-GUI

Data Simulation Exploratory Analysis TCC Computation MA Plot Volcano Plot

Heatmap Expression Level Plot Analysis Report

What's TCC?

TCC^[1] is a [R/Bioconductor](#) package provides a series of functions for performing differential expression (DE) analysis from RNA-seq count data using a robust normalization strategy (called DEGES).

The basic idea of DEGES is that potential differentially expressed genes (DEGs) among compared samples should be removed before data normalization to obtain a well-ranked gene list where true DEGs are top-ranked and non-DEGs are bottom ranked. This can be done by performing the multi-step normalization procedures based on DEGES (DEG elimination strategy) implemented in TCC.

TCC internally uses functions provided by [edgeR](#)^[2], [DESeq](#)^[3], [DESeq2](#)^[4], and [baySeq](#)^[5]. The multi-step normalization of TCC can be done by using functions in the four packages.

TCC-GUI: Graphical User Interface for TCC package

TCC-GUI

TCC-GUI: Graphical User Interface for TCC package

Documentation

Data Simulation **Step 0**

Exploratory Analysis **Step 1**

Sample Upload

Select Sample Data

hypoData (sample dataset)

1. The 10000 genes dataset **hypoData** is simulation data generated by `TCC::simulateReadCounts` function

2. After performing simulation in the **Step0**, **Simulation Data** can be selected and it's referring the latest simulation result.

1. Import Count Data

Group Assignment

Input your group info

Please input group information at here.
Here is a example format:

Read Count Table

No data to show. Click **Sample** or **Upload** your own dataset.

Count Distribution Filtering Threshold Density Plot MDS Plot PCA

Hierarchical Clustering

No data for plotting. Please import dataset and assign group information first.

TCC-GUI

こんな感じの画面に飛びます。①Upload。ここでさらに②Upload。

TCC-GUI: Graphical User Interface for TCC package

Documentation

Data Simulation **Step 0**

Exploratory Analysis **Step 1**

Sample Upload

1

Upload Count Data

Upload... No file has been uploaded

2

Text file in .tsv/.csv format, and the first column should be genes' name.

Group Assignment

Input your group info

Please input group information at here. Here is an example format:

G1_rep1,Group1

▶ 2. Assign Group Label

TCC-GUI expects first label should be Group1 (G1) and the next be Group2 (G2), and so on.

Read Count Table

No data to show. Click [Sample](#) or [Upload](#) your own dataset.

Count Distribution Filtering Threshold Density Plot MDS Plot PCA

Hierarchical Clustering

No data for plotting. Please import dataset and assign group information first.

TCC-GUI

こんな感じの画面に飛びます。①Upload。ここでさらに②Upload。③解析したいファイルを選択して、④開く。

The screenshot shows the TCC-GUI web interface. The main content area has a blue header with the title 'TCC-GUI: Graphical User Interface for TCC package'. Below the header, there are several sections: 'Documentation', 'Data Simulation' (Step 0), and 'Exploratory Analysis' (Step 1). The 'Read Count Table' section is highlighted in blue and contains the text 'No data to show. Click [Sample](#) or [Upload](#) your own dataset.' Below this, there is an 'Upload Count Data' section with an 'Upload...' button and the text 'No file has been uploaded yet'. A text box below explains: 'Text file in .tsv/.csv format, and the first column should be genes' name.' The 'Group Assignment' section is also visible, with a blue button labeled '2. Assign Group Label'. A file explorer window is overlaid on the interface, showing the path 'PC > デスクトップ > hoge'. The file 'counts_gene.txt' is selected, and a red arrow with the number 3 points to it. The file name 'counts_gene.txt' is also entered in the 'ファイル名(N):' field, with another red arrow and the number 3 pointing to it. The file type is set to 'テキストドキュメント'. The '開く(O)' button is highlighted with a red arrow and the number 4.

TCC-GUI

こんな感じの画面に飛びます。①Upload。ここでさらに②Upload。③解析したいファイルを選択して、④開く。⑤アップロード中…

TCC-GUI

こんな感じの画面に飛びます。①Upload。ここでさらに②Upload。③解析したいファイルを選択して、④開く。⑤アップロード中、⑥アップロード完了後の状態。⑦赤枠内の数値情報は…

TCC-GUI: Graphical User Interface × +
https://infinityloop.shinyapps.io/TCC-GUI/

TCC-GUI: Graphical User Interface for TCC package

Documentation

Data Simulation **Step 0**

Exploratory Analysis **Step 1**

Sample Upload

Upload Count Data

Upload... counts_gene

Upload complete

Text file in .tsv/.csv format, and the first column should be genes' name.

Group Assignment

Input your group info

Please input group information at here. Here is a example format:

G1_rep1,Group1

2. Assign Group Label

TCC-GUI expect first label should be Group1 (G1) and the next be Group2 (G2), and so on.

Read Count Table

Search:

Gene Name	Control1	Control2	Control3	Treatment1	Treatment2
ENSG00000000003.14	11156	9957	14914	15968	
ENSG00000000005.5	0	0	0	0	
ENSG000000000419.12	64023	55036	88460	62819	
ENSG000000000457.13	24666	22830	37528	31189	
ENSG000000000460.16	42199	38508	58545	36169	
ENSG000000000938.12	0	0	0	0	
ENSG000000000971.15	647	628	794	450	
ENSG000000001036.13	89682	81128	123915	121702	
ENSG000000001084.10	86325	77847	120691	73226	
ENSG000000001167.14	57229	49468	80253	89081	
ENSG000000001460.17	16982	16121	25284	19119	

Showing 1 to 12 of 58,037 entries

Count Distribution

Filtering Threshold

Density Plot

MDS Plot

PCA

おさらい

こんな感じの画面に飛びます。①Upload。ここでさらに②Upload。③解析したいファイルを選択して、④開く。⑤アップロード中、⑥アップロード完了後の状態。⑦赤枠内の数値情報は、アップロードしたファイルと確かに同じ

gene_id	Control1	Control2	Control3	Treatment1	Treatment2	Treatment3
ENSG00000000003.14	11156	9957	14914	15968	15138	16747
ENSG00000000005.5	0	0	0	0	0	0
ENSG00000000419.12	64023	55036	88460	62819	67807	70965
ENSG00000000457.13	24666	22830	37528	31189	31249	36804
ENSG00000000460.16	42199	38508	58545	36169	39613	42044
ENSG00000000938.12	0	0	0	0	0	0
ENSG00000000971.15	647	628	794	450	650	446
ENSG00000001036.13	89682	81128	123915	121702	121412	143520
ENSG00000001084.10	86325	77847	120691	73226	76177	87010
ENSG00000001167.14	57229	49468	80253	89081	92579	109634
ENSG00000001460.17	16982	16121	25284	19119	18513	19189
ENSG00000001461.16	16155	14981	27084	22228	20813	22194

counts_gene.txt

Contents

- 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

Group Assignment

次の作業は、①どのサンプル名のものがどの群に属するかを指定する、②Group Assignment。具体的には③の枠内で指定するのだが、これを行わないと④ページ下部のほうの情報がスカスカなままです。

TCC-GUI: Graphical User Interface for TCC package
https://infinityloop.shinyapps.io/TCC-GUI/

TCC-GUI: Graphical User Interface for TCC package

Documentation

Data Simulation **Step 0**

Exploratory Analysis **Step 1**

Sample Upload

Upload Count Data

Upload... counts_gene

Upload complete

Text file in .tsv/.csv format, and the first column should be genes' name.

Group Assignment

Input your group info

Please input group information at here. Here is a example format:

G1_rep1,Group1

2. Assign Group Label

TCC-GUI expect first label should be Group1 (G1) and the next be Group2 (G2), and so on.

Read Count Table

Search:

Gene Name	Control1	Control2	Control3	Treatment1	Treatment2
ENSG00000000003.14	11156	9957	14914	15968	
ENSG00000000005.5	0	0	0	0	
ENSG000000000419.12	64023	55036	88460	62819	
ENSG000000000457.13	24666	22830	37528	31189	
ENSG000000000460.16	42199	38508	58545	36169	
ENSG000000000938.12	0	0	0	0	
ENSG000000000971.15	647	628	794	450	
ENSG000000001036.13	89682	81128	123915	121702	
ENSG000000001084.10	86325	77847	120691	73226	
ENSG000000001167.14	57229	49468	80253	89081	
ENSG000000001460.17	16982	16121	25284	19119	

Showing 1 to 12 of 58,037 entries

Count Distribution

Filtering Threshold

Density Plot

MDS Plot

PCA

Group Assignment

次の作業は、①どのサンプル名のものがどの群に属するかを指定する、②Group Assignment。具体的には③の枠内で指定するのだが、これを行わないと④ページ下部のほうの情報がスカスカなままです。ページ下部に移動。⑤にも群情報の指定が必要だと書かれてますね。

TCC-GUI: Graphical User Interface

https://infinityloop.shinyapps.io/TCC-GUI/

Please input group information at here. Here is an example format:

G1_rep1,Group1

▶ 2. Assign Group Label

TCC-GUI expect first label should be Group1 (G1) and the next be Group2 (G2), and so on.

Summary

N_{gene} : 58037

N_{group} : 0

NR :

Assign group information needed.

ENSG00000000938.12				
ENSG00000000971.15				
ENSG00000001036.13	89682	81128	123915	121702
ENSG00000001084.10	86325	77847	120691	73226
ENSG00000001167.14	57229	49468	80253	89081
ENSG00000001460.17	16982	16121	25284	19119

Showing 1 to 12 of 58,037 entries

Count Distribution

Filtering Threshold

Density Plot

MDS Plot

PCA

Hierarchical Clustering

No data for plotting. Please import dataset and assign group information first.

Group Assignment

①群情報の指定はこんな感じです。コンマ(,)の左側がサンプル名で、右側が群情報。ここではContとTreatにしたが、別にG1とG2や、1と2など、識別さえできればなんでもよい。入力が完了したら、②を押す。サンプル名で自動判定してほしいという要望はあるだろうが、このあたりはエンドユーザが明確に指定することを要求するポリシーです。

Control1,Cont
Control2,Cont
Control3,Cont
Treatment1,Treat
Treatment2,Treat
Treatment3,Treat



2. Assign Group Label



TCC-GUI expect first label should be Group1 (G1) and the next be Group2 (G2), and so on.

Summary

N_{gene} : 58037
N_{group} : 0
NR :
Assign group information needed.

ENSG00000000938.12				
ENSG00000000971.15				
ENSG00000001036.13				
ENSG00000001084.10	86325	77847	120691	73226
ENSG00000001167.14	57229	49468	80253	89081
ENSG00000001460.17	16982	16121	25284	19119

Showing 1 to 12 of 58,037 entries

Control1,Cont
Control2,Cont
Control3,Cont
Treatment1,Treat
Treatment2,Treat
Treatment3,Treat

Group Assignment後

①このあたりがごによごによな、②無事群情報の指定が終了したことを示すメッセージが出ます。

The screenshot shows the TCC-GUI interface with a modal dialog box in the center. The dialog box contains a green checkmark icon, the word 'DONE', and the text 'Group labels were successfully assigned.' Below this text is an 'Ok' button. A red arrow with the number '2' points to the 'Ok' button. Another red arrow with the number '1' points to a bar chart icon in the background interface.

The background interface includes a table of gene expression data:

Gene ID	Sample 1	Sample 2	Sample 3	Sample 4
ENSG00000000938.12				
ENSG00000000971.15	647	628	794	450
ENSG00000001036.13	89682	81128	123915	121702
ENSG00000001084.10	86325	77847	120691	73226
ENSG00000001167.14	57229	49468	80253	89081
ENSG00000001460.17	16982	16121	25284	19119

Other interface elements include a sidebar with a list of controls and treatments, a '2. Assign Group' button, and a summary panel showing statistics like N_{gene}: 58037, N_{group}: 2, NR: Cont: 3 Tr, and AS: 0.208. The bottom of the interface has input fields for 'X label' (Sample) and 'Y label' (log₂(C)).

Group Assignment後

①このあたりがごによごによなつて、②無事群情報の指定が終了したことを示すメッセージが出ます。さらに数秒後にはごによごによしていた作業が終了し、箱ひげ図(box plot)が見られるようになります。①Okを押すと…

TCC-GUI: Graphical User Interface x +
https://infinityloop.shinyapps.io/TCC-GUI/

Control1,Cont
Control2,Cont
Control3,Cont
Treatment1,Treat
Treatment2,Treat
Treatment3,Treat


▶ 2. Assign Group

TCC-GUI expects
should be Group
next be Group
on.

Summary

N_{gene} : 58037
N_{group} : 2
NR : Cont: 3 Treat: 3
AS : 0.208

ENSG00000000936.12				
ENSG00000000971.15	647	628	794	450
ENSG00000001036.13	89682	81128	123915	121702
ENSG00000001084.10	86325	77847	120691	73226
ENSG00000001167.14	57229	49468	80253	89081
ENSG00000001460.17	16982	16121	25284	19119



DONE

Group labels were successfully assigned.

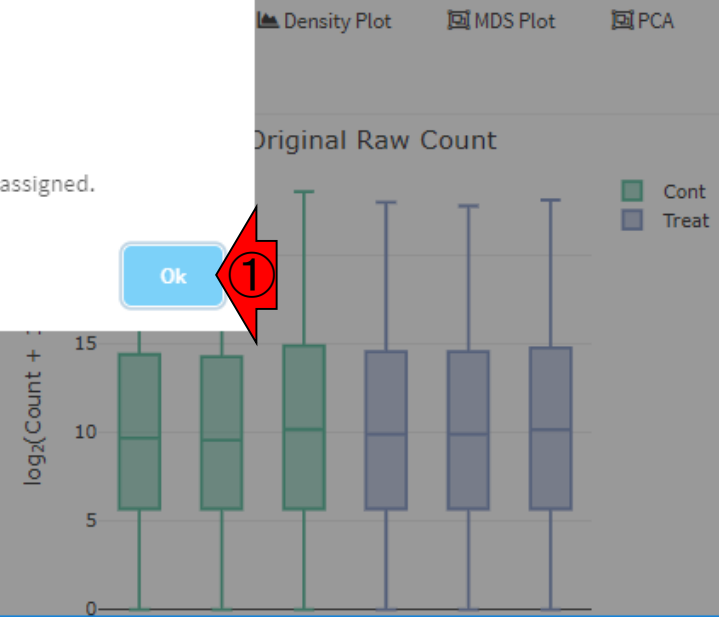
Ok



Original Raw Count

X label
Sample

Y label
log₂(Count + 1)



こんな感じになります。

Group Assignment後

The screenshot shows the TCC-GUI web interface. On the left, there is a sidebar with a list of labels: Control1, Control2, Control3, Treatment1, Treatment2, and Treatment3. Below this is a blue button labeled "2. Assign Group Label". A text box explains: "TCC-GUI expect first label should be Group1 (G1) and the next be Group2 (G2), and so on." Below that is a "Summary" section with the following information: N_{gene}: 58037, N_{group}: 2, NR: Cont: 3 Treat: 3, AS: 0.208.

The main area displays a table of gene expression data. The table has 5 columns: Gene ID, and four numerical values. The first row is partially visible: ENSG00000000936.12. The second row is: ENSG00000000971.15 with values 647, 628, 794, 450. The third row is: ENSG00000001036.13 with values 89682, 81128, 123915, 121702. The fourth row is: ENSG00000001084.10 with values 86325, 77847, 120691, 73226. The fifth row is: ENSG00000001167.14 with values 57229, 49468, 80253, 89081. The sixth row is: ENSG00000001460.17 with values 16982, 16121, 25284, 19119. Below the table, it says "Showing 1 to 12 of 58,037 entries".

Below the table, there are several analysis options: "Count Distribution", "Filtering Threshold", "Density Plot", "MDS Plot", and "PCA". Under "Filtering Threshold", there is a "Hierarchical Clustering" section with a "Filter low genes" slider set to 0. Below that are input fields for "Title" (Original Raw Count), "X label" (Sample), and "Y label" (log₂(Count + 1)).

On the right, there is a box plot titled "Original Raw Count". The y-axis is labeled "log₂(Count + 1)" and ranges from 0 to 20. The x-axis has six categories. The first three categories are green and represent "Cont" (Control) samples. The last three categories are blue and represent "Treat" (Treatment) samples. The box plots show the distribution of log₂(Count + 1) for each sample, with whiskers extending to the minimum and maximum values.

Group Assignment後

こんな感じになります。①Summary部分に注目！②遺伝子数はGroup Assignment前と同じですが、

TCC-GUI: Graphical User Interface

https://infinityloop.shinyapps.io/TCC-GUI/

Control1,Cont
Control2,Cont
Control3,Cont
Treatment1,Treat
Treatment2,Treat
Treatment3,Treat

▶ 2. Assign Group Label

TCC-GUI expect first label should be Group1 (G1) and the next be Group2 (G2), and so on.

①

Summary

N_{gene} : 58037

N_{group} : 2

NR : Cont: 3 Treat: 3

AS : 0.208

②

ENSG00000000936.12				
ENSG00000000971.15	647	628	794	450
ENSG00000001036.13	89682	81128	123915	121702
ENSG00000001084.10	86325	77847	120691	73226
ENSG00000001167.14	57229	49468	80253	89081
ENSG00000001460.17	16982	16121	25284	19119

Showing 1 to 12 of 58,037 entries

Count Distribution

Filtering Threshold

Density Plot

MDS Plot

PCA

Hierarchical Clustering

Filter low genes



Title

Original Raw Count

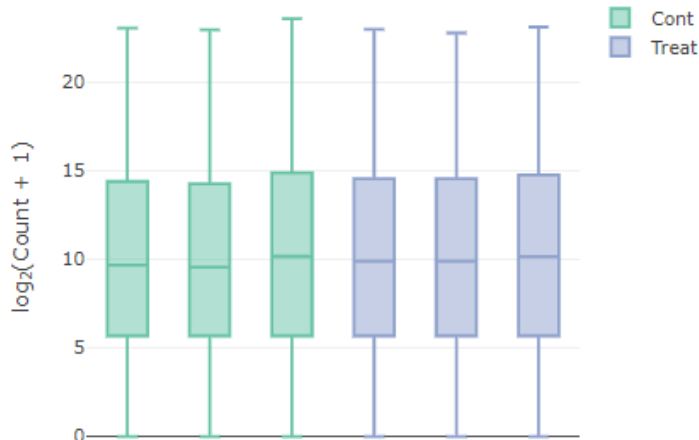
X label

Sample

Y label

log₂(Count + 1)

Original Raw Count



Group Assignment後

こんな感じになります。①Summary部分に注目！②遺伝子数はGroup Assignment前と同じですが、③グループ数が2、④Contが3サンプルTreatが3サンプル、そして⑤平均シルエットスコア(AS)が0.208であることが示されました。

TCC-GUI: Graphical User Interface

https://infinityloop.shinyapps.io/TCC-GUI/

Control1,Cont
Control2,Cont
Control3,Cont
Treatment1,Treat
Treatment2,Treat
Treatment3,Treat

▶ 2. Assign Group Label

TCC-GUI expect first label should be Group1 (G1) and the next be Group2 (G2), and so on.

①

Summary

N_{gene} : 58037

N_{group} : 2

NR : Cont: 3 Treat: 3

AS : 0.208

②

③

④

⑤

ENSG00000000936.12				
ENSG00000000971.15	647	628	794	450
ENSG00000001036.13	89682	81128	123915	121702
ENSG00000001084.10	86325	77847	120691	73226
ENSG00000001167.14	57229	49468	80253	89081
ENSG00000001460.17	16982	16121	25284	19119

Showing 1 to 12 of 58,037 entries

Count Distribution

Filtering Threshold

Density Plot

MDS Plot

PCA

Hierarchical Clustering

Filter low genes



Title

Original Raw Count

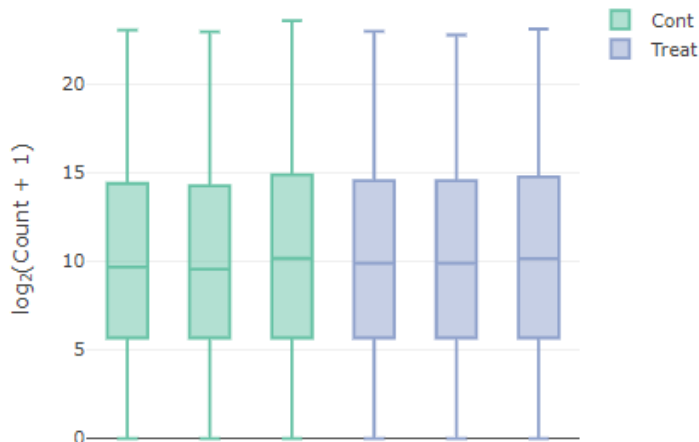
X label

Sample

Y label

log₂(Count + 1)

Original Raw Count



シルエットスコア

おさらい。①counts_gene.txtを入力としてR上で計算した結果(②AS = 0.2083361)と同じですね。

```

RGui (64-bit)
ファイル 編集 パッケージ ウィンドウ ヘルプ

R Console
[1] 58037      6
>
> #前処理 (フィルタリング)
> obj <- as.logical(rowSums(data))
> data <- unique(data[obj,])
> dim(data)
[1] 34114      6
>
> #本番 (AS値の計算)
> d <- as.dist(1 - cor(data, method="spearman"))
> AS <- mean(silhouette(data.cl, d), "sil_width")
> AS
[1] 0.2083361
>
> |

無題 - RIデータ
in_f <- "counts_gene.txt"
param_G1 <- 3
param_G2 <- 3

#必要なパッケージをロード
library(cluster)

#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,
data.cl <- c(rep(1, param_G1), rep(2, param_G2))#G1
dim(data)

#前処理 (フィルタリング)
obj <- as.logical(rowSums(data) > 0)
data <- unique(data[obj,])
dim(data)

#本番 (AS値の計算)
d <- as.dist(1 - cor(data, method="spearman"))#サンブ
AS <- mean(silhouette(data.cl, d), "sil_width")#
  
```

Contents

- 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

Count Distribution

赤枠で見ているのは、①Count Distributionです。
下のほうが切れているのでページ下部に移動。

The screenshot shows the TCC-GUI web interface. On the left, there are controls for assigning group labels and a summary section. The main area displays a table of gene counts and a 'Count Distribution' plot. A red arrow points to the 'Count Distribution' tab, and a red box highlights the 'Original Raw Count' box plot.

Table Data:

Gene ID	Sample 1	Sample 2	Sample 3	Sample 4
ENSG00000000936.12				
ENSG00000000971.15	647	628	794	450
ENSG00000001036.13	89682	81128	123915	121702
ENSG00000001084.10	86325	77847	120691	73226
ENSG00000001167.14	57229	49468	80253	89081
ENSG00000001460.17	16982	16121	25284	19119

Summary:

- N_{gene} : 58037
- N_{group} : 2
- NR : Cont: 3 Treat: 3
- AS : 0.208

Count Distribution Plot:

Original Raw Count

Y label: $\log_2(\text{Count} + 1)$

Legend: Cont (green), Treat (blue)

Count Distribution

TCC-GUI: Graphical User Interface

https://infinityloop.shinyapps.io/TCC-GUI/

Treatment3,Treat

▶ 2. Assign Group Label

TCC-GUI expect first label should be Group1 (G1) and the next be Group2 (G2), and so on.

Summary

N_{gene} : 58037
 N_{group} : 2
 NR : Cont: 3 Treat: 3
 AS : 0.208

ENSG00000001167.14	57229	49468	80253	89081
ENSG00000001460.17	16982	16121	25284	19119

Showing 1 to 12 of 58,037 entries

Count Distribution Filtering Threshold Density Plot MDS Plot PCA

Hierarchical Clustering

Filter low genes

0 20

-1 2 5 8 11 14 17 20

Title
Original Raw Count

X label
Sample

Y label
log₂(Count + 1)

Original Raw Count

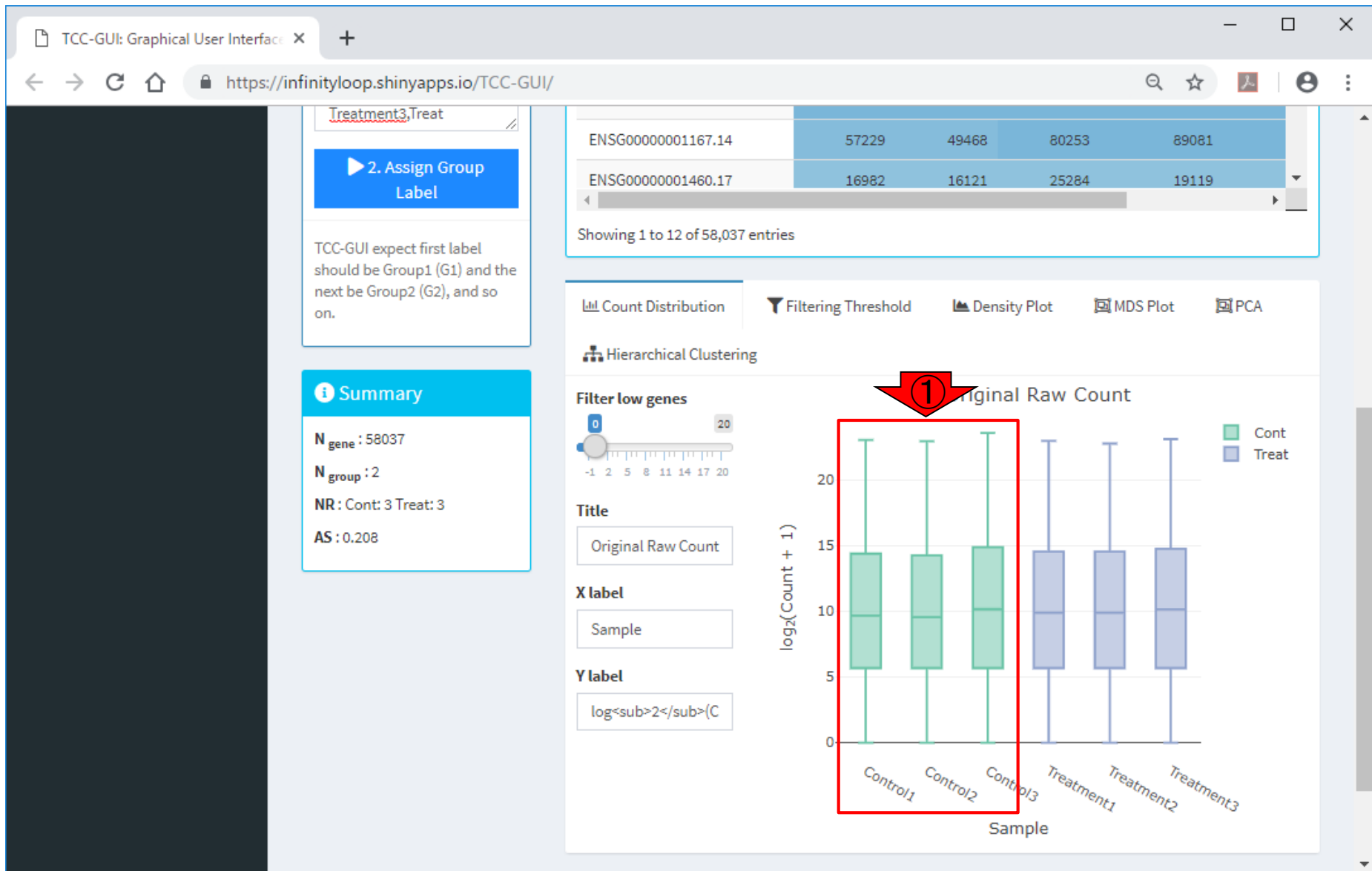
Cont Treat

Control1 Control2 Control3 Treatment1 Treatment2 Treatment3

Sample

Count Distribution

こんな感じになります。①最初の3列分がControl群3サンプルの、列ごとの数値分布。



Count Distribution

こんな感じになります。①最初の3列分がControl群3サンプルの、列ごとの数値分布。②残りの3列分がTreatment群3サンプルの分布。Treatment群のほうが若干平均全体的に値が大きいような気がする、という程度の考察は可能。

TCC-GUI: Graphical User Interface x +

https://infinityloop.shinyapps.io/TCC-GUI/

Treatment3,Treat

▶ 2. Assign Group Label

TCC-GUI expect first label should be Group1 (G1) and the next be Group2 (G2), and so on.

Summary

N_{gene} : 58037

N_{group} : 2

NR : Cont: 3 Treat: 3

AS : 0.208

ENSG00000001167.14	57229	49468	80253	89081
ENSG00000001460.17	16982	16121	25284	19119

Showing 1 to 12 of 58,037 entries

Count Distribution

Filtering Threshold

Density Plot

MDS Plot

PCA

Hierarchical Clustering

Filter low genes



Title

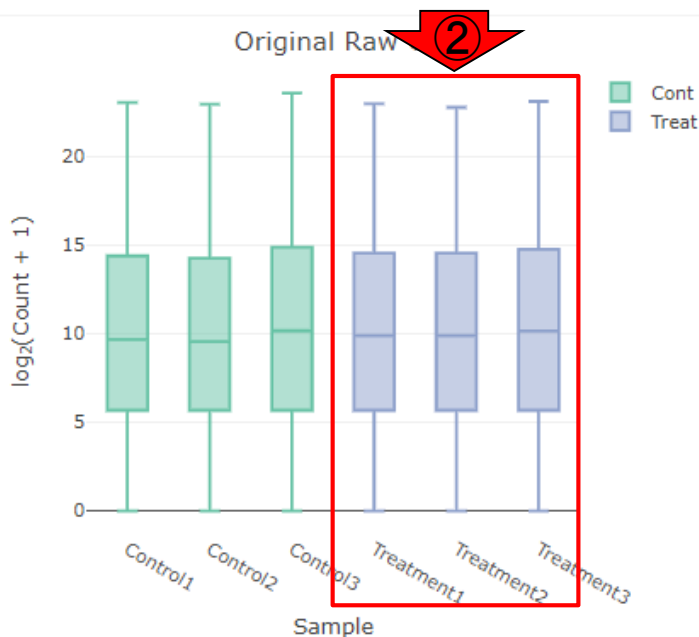
Original Raw Count

X label

Sample

Y label

$\log_2(\text{Count} + 1)$



階層的クラスタリング

① Hierarchical Clusteringタブをクリックした結果。②こちらの樹形図(デンドログラム)でも Control群とTreatment群でくっきり分かれていますね。

The screenshot shows the TCC-GUI web interface. At the top, there's a browser tab for 'TCC-GUI: Graphical User Interface' and the URL 'https://infinityloop.shinyapps.io/TCC-GUI/'. Below the browser, there's a data table with columns for gene IDs and four numerical values. A summary panel on the left provides statistics: N_{gene}: 58037, N_{group}: 2, NR: Cont: 3 Treat: 3, AS: 0.208. The main area features a 'Hierarchical Clustering' tab (indicated by red arrow ①) with settings for 'Complete' agglomeration and 'Spearman' distance measure. Below these settings is a dendrogram (indicated by red arrow ②) and a heatmap. The dendrogram shows two main clusters: Control (Control2, Control1, Control3) and Treatment (Treatment3, Treatment1, Treatment2). The heatmap uses a color scale from 0.00 (dark blue) to 0.03 (light green) to represent distance values.

Gene ID	Value 1	Value 2	Value 3	Value 4
ENSG00000001167.14	57229	49468	80253	89081
ENSG00000001460.17	16982	16121	25284	19119

階層的クラスタリング

① Hierarchical Clusteringタブをクリックした結果。②こちらの樹形図(デンドログラム)でもControl群とTreatment群でくっきり分かれていますね。③このあたりにマウスポインタを合わせると、画像をpng形式で保存することができます。

TCC-GUI: Graphical User Interface

https://infinityloop.shinyapps.io/TCC-GUI/

Treatment3,Treat

▶ 2. Assign Group Label

Label

TCC-GUI expect first label should be Group1 (G1) and the next be Group2 (G2), and so on.

Summary

N_{gene} : 58037

N_{group} : 2

NR : Cont: 3 Treat: 3

AS : 0.208

ENSG00000001167.14

ENSG00000001460.17

Showing 1 to 12 of 58,037 entries

Count Distribution Filtering Threshold Density Plot MDS Plot PCA

Hierarchical Clustering

Agglomeration Method: Complete

Distance Measure: Spearman

Download plot as a png

Control2

Control1

Control3

Treatment3

Treatment1

Treatment2

Treatment2 Treatment1 Treatment3 Control3 Control1 Control2

0.03

0.02

0.01

0.00

階層的クラスタリング

① Hierarchical Clusteringタブをクリックした結果。②こちらの樹形図(デンドログラム)でもControl群とTreatment群でくっきり分かれていますね。③このあたりにマウスポインタを合わせると、画像をpng形式で保存することができます。④こんな感じのファイル名で保存されます。⑤現状では縦横比が変わってしまっているのが少し残念。

The screenshot shows the TCC-GUI web application interface. The main content area displays a heatmap and a dendrogram. The heatmap has rows labeled Control2, Control1, Control3, Treatment3, Treatment1, and Treatment2, and columns labeled Treatment2, Treatment1, Treatment3, Control3, Control1, and Control2. A color scale on the right ranges from 0.00 (dark blue) to 0.03 (light green). The dendrogram on the left shows hierarchical clustering of the samples. A red box highlights the dendrogram, and a red arrow points to a file save dialog box. The dialog box shows the file name 'newplot.png' and the file type 'PNG Image (*.png)'. The dialog box also shows the current directory as 'PC > デスクトップ' and a search bar for 'デスクトップの検索'.

PCA

①PCA (Principal Component Analysis; 主成分分析)も実行可能です。

The screenshot shows the TCC-GUI web interface. On the left, there is a sidebar with a text input field containing "Treatment3,Treat", a blue button labeled "2. Assign Group Label", and a summary box. The summary box contains the following information: **N_{gene}** : 58037, **N_{group}** : 2, **NR** : Cont: 3 Treat: 3, and **AS** : 0.208. Below the summary box are three toggle switches for "Log(x+1) transform", "Center", and "Scale", all of which are currently turned on.

The main content area displays a table of gene expression data with columns for gene IDs and four numerical values. The first two rows are highlighted in blue. Below the table, it says "Showing 1 to 12 of 58,037 entries". A red arrow with the number "1" points to the "PCA" button in the analysis menu.

The analysis menu includes "Count Distribution", "Filtering Threshold", "Density Plot", "MDS Plot", and "PCA". Below the menu, there is a "Hierarchical Clustering" section and a "Summary Table" tab. The "Summary Table" displays the following data:

	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC1	5.922	0.351	0.351
PC2	4.873	0.237	0.588
PC3	4.125	0.170	0.758
PC4	3.543	0.126	0.884
PC5	3.408	0.116	1.000
PC6	0.000	0.000	1.000

PCA

①PCA (Principal Component Analysis; 主成分分析)も実行可能です。②2D Plot (2次元プロット)がおそらくPCAの一般的なイメージでしょう。

Treatment3,Treat

▶ 2. Assign Group Label

TCC-GUI expect first label should be Group1 (G1) and the next be Group2 (G2), and so on.

Summary

N_{gene} : 58037

N_{group} : 2

NR : Cont: 3 Treat: 3

AS : 0.208

ENSG00000001167.14	57229	49468	80253	89081
ENSG00000001460.17	16982	16121	25284	19119

Showing 1 to 12 of 58,037 entries

Count Distribution Filtering Threshold Density Plot MDS Plot **PCA**

Hierarchical Clustering

PCA is performed on the all genes (or top n gene) selected by non-zero row variance (or the most variable genes).

Top Gene

100

Log(x+1) transform

Center

Scale

Summary Table Scree Plot 3D Plot **2D Plot**

PCA Plot (2D)

PC2

PC1

Control1 Control2 Control3 Treatment1 Treatment2 Treatment3

Legend: Aa Cont, Aa Treat

Contents

- 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

TCC Computation

① ページ上部へ。TCC-GUIのメインは、② Step2のTCC Computation(発現変動遺伝子同定)です。

The screenshot shows the TCC-GUI web interface. The sidebar on the left has four items: 'Documentation', 'Data Simulation' (Step 0), 'Exploratory Analysis' (Step 1), and 'TCC Computation' (Step 2). A red arrow labeled '2' points to the 'TCC Computation' item. The main content area is divided into three sections: 'Upload Count Data', 'Group Assignment', and 'Read Count Table'. The 'Upload Count Data' section has an 'Upload...' button and a file named 'counts_gene'. Below it, it says 'Upload complete' and provides instructions: 'Text file in .tsv/.csv format, and the first column should be genes' name.' The 'Group Assignment' section has a '2. Assign Group Label' button and a list of group labels: 'Control1,Cont', 'Control2,Cont', 'Control3,Cont', 'Treatment1,Treat', 'Treatment2,Treat', and 'Treatment3,Treat'. The 'Read Count Table' section shows a table with columns for 'Gene Name', 'Control1', 'Control2', 'Control3', 'Treatment1', and 'Treatment2'. A search bar is located above the table. A red arrow labeled '1' points to the top right corner of the table area. The table contains 12 rows of data. Below the table, it says 'Showing 1 to 12 of 58,037 entries'. At the bottom of the interface, there are several tabs: 'Count Distribution', 'Filtering Threshold', 'Density Plot', 'MDS Plot', and 'PCA'.

Gene Name	Control1	Control2	Control3	Treatment1	Treatment2
ENSG00000000003.14	11156	9957	14914	15968	
ENSG00000000005.5	0	0	0	0	
ENSG000000000419.12	64023	55036	88460	62819	
ENSG000000000457.13	24666	22830	37528	31189	
ENSG000000000460.16	42199	38508	58545	36169	
ENSG000000000938.12	0	0	0	0	
ENSG000000000971.15	647	628	794	450	
ENSG000000001036.13	89682	81128	123915	121702	
ENSG000000001084.10	86325	77847	120691	73226	
ENSG000000001167.14	57229	49468	80253	89081	
ENSG000000001460.17	16982	16121	25284	19119	

TCC Computation

The screenshot displays the TCC-GUI web application interface. The browser address bar shows the URL <https://infinityloop.shinyapps.io/TCC-GUI/>. The page title is "TCC-GUI: Graphical User Interface for TCC package".

The interface is divided into several sections:

- Left Sidebar:** Contains navigation links for "Documentation", "Data Simulation" (Step 0), "Exploratory Analysis" (Step 1), and "TCC Computation" (Step 2).
- TCC Computation Parameters:** A central panel with the following settings:
 - Normalization Method:** TMM
 - DEG Identification Method:** edgeR
 - Filtering Threshold for Low Count Genes:** A slider set to "Do not filter" (value 30). Below it, it states "0 genes (0%) will be filtered out."
 - Number of Iteration:** A slider set to 3 (range 0 to 30).
 - FDR Cut-off:** A slider set to 0.1 (range 0 to 1).
 - Elimination of Potential DEGs:** A slider set to 0.05 (range 0 to 1).
- Result Table:** A section with the instruction: "Click [Run TCC Computation] to obtain Result Table."
- Summary of TCC Normalization:** A section with the instruction: "Summary of TCC normalization will be shown after TCC computation."

TCC Computation

こんな感じになります。①沢山選択可能なパラメータがありますが、とりあえず無視でよいです。②ページ下部に移動。

The screenshot shows the TCC-GUI web interface. The browser address bar displays `https://infinityloop.shinyapps.io/TCC-GUI/`. The page title is "TCC-GUI: Graphical User Interface for TCC package". A sidebar on the left contains navigation links: "Documentation", "Data Simulation" (Step 0), "Exploratory Analysis" (Step 1), and "TCC Computation" (Step 2). The main content area is divided into two columns. The left column, titled "TCC Computation Parameters", is enclosed in a red box with a red arrow labeled "1" pointing to it. This section includes several adjustable parameters: "Normalization Method" (set to TMM), "DEG Identification Method" (set to edgeR), "Filtering Threshold for Low Count Genes" (set to "Do not filter" with a slider at 30), "Number of Iteration" (set to 3 with a slider from 0 to 30), "FDR Cut-off" (set to 0.1 with a slider from 0 to 1), and "Elimination of Potential DEGs" (set to 0.05 with a slider from 0 to 1). The right column contains two informational boxes: "Result Table" (with the instruction "Click [Run TCC Computation] to obtain Result Table.") and "Summary of TCC Normalization" (with the instruction "Summary of TCC normalization will be shown after TCC computation."). A red arrow labeled "2" points to the bottom of the right column.

TCC Computation

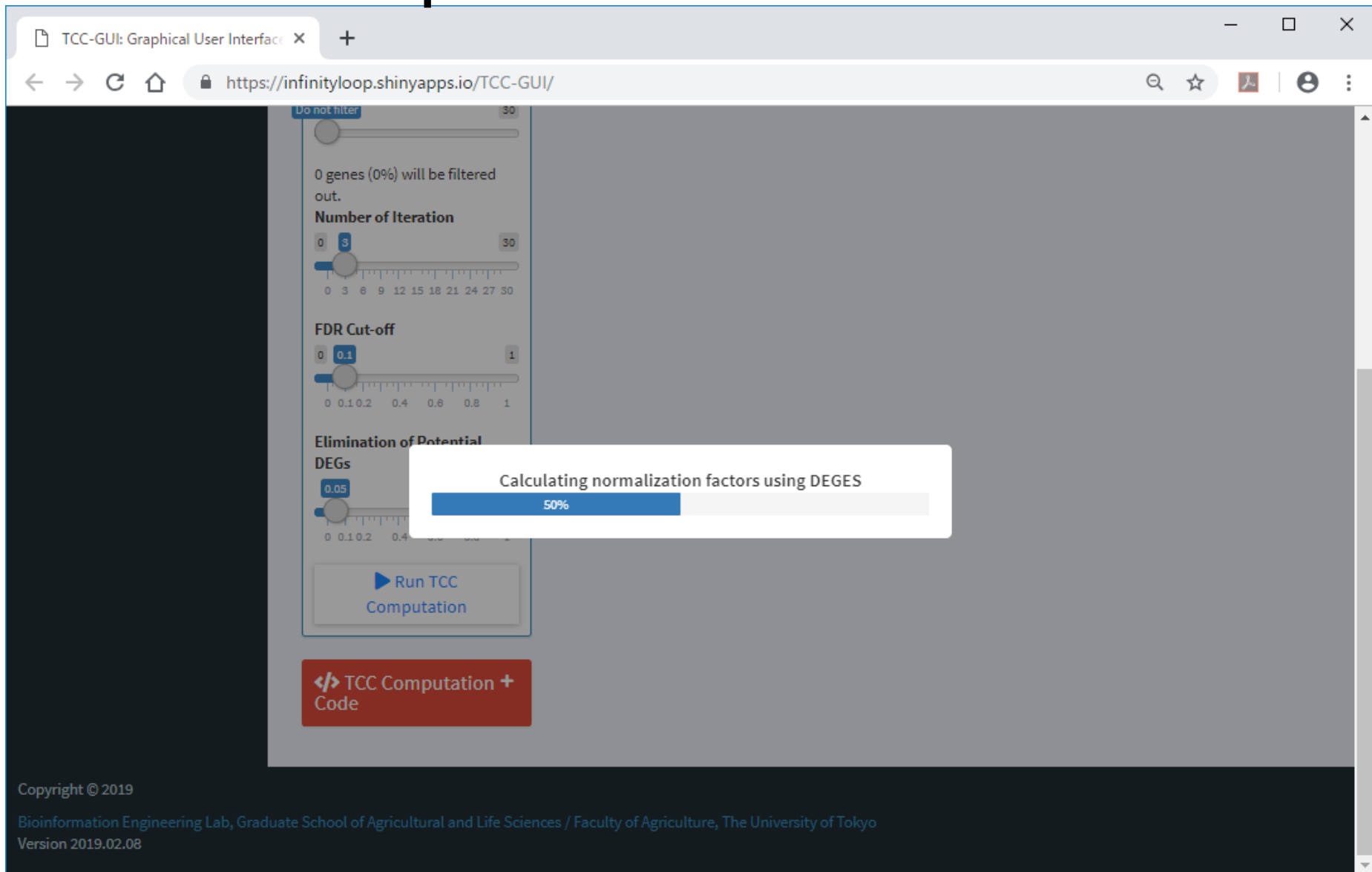
こんな感じになります。①沢山選択可能なパラメータがありますが、とりあえず無視でよいです。②ページ下部に移動後。③実行。

The screenshot shows the TCC-GUI web interface in a browser. The address bar displays <https://infinityloop.shinyapps.io/TCC-GUI/>. The interface includes several sliders for parameter adjustment:

- Do not filter:** A slider set to 30.
- Number of Iteration:** A slider set to 3.
- FDR Cut-off:** A slider set to 0.1.
- Elimination of Potential DEGs:** A slider set to 0.05.

Below the sliders is a blue button labeled "Run TCC Computation" with a play icon. A red arrow with the number 3 points to this button. At the bottom of the page, there is a red button labeled "TCC Computation + Code" with a code icon. A red arrow with the number 2 points to the bottom of the page area. The footer contains the text: "Copyright © 2019 Bioinformation Engineering Lab, Graduate School of Agricultural and Life Sciences / Faculty of Agriculture, The University of Tokyo Version 2019.02.08".

TCC Computation



The screenshot shows a web browser window with the URL <https://infinityloop.shinyapps.io/TCC-GUI/>. The interface includes several sliders for configuration: 'Do not filter' (set to 30), 'Number of Iteration' (set to 3), 'FDR Cut-off' (set to 0.1), and 'Elimination of Potential DEGs' (set to 0.05). A 'Run TCC Computation' button is visible. A modal dialog box is overlaid on the interface, displaying the text 'Calculating normalization factors using DEGES' and a progress bar that is 50% filled. At the bottom of the page, there is a red button labeled 'TCC Computation + Code'.

Copyright © 2019
Bioinformatics Engineering Lab, Graduate School of Agricultural and Life Sciences / Faculty of Agriculture, The University of Tokyo
Version 2019.02.08

TCC Computation

The screenshot shows a web browser window with the URL <https://infinityloop.shinyapps.io/TCC-GUI/>. The interface includes several sliders for parameter adjustment:

- Filtering Threshold for Low Count Genes:** Set to "Do not filter" (value 30).
- Number of Iteration:** Set to 3.
- FDR Cut-off:** Set to 0.1.
- Elimination of DEGs:** Set to 0.05.

A central progress bar indicates the current status: "Identifying DE genes" is 70% complete. Below the sliders is a "Run TCC Computation" button, and at the bottom is a red button labeled "TCC Computation + Code".

Copyright © 2019
 Bioinformation Engineering Lab, Graduate School of Agricultural and Life Sciences / Faculty of Agriculture, The University of Tokyo

TCC Computation

The screenshot shows a web browser window with the URL <https://infinityloop.shinyapps.io/TCC-GUI/>. The interface includes a sidebar with navigation options: Heatmap (Step 3), Expression Level (Step 3), and Report (Step 4). The main panel displays several sliders for parameter adjustment: 'Filtering Threshold for Low Count Genes' (set to 'Do not filter'), 'Number of Iteration' (set to 3), 'FDR Cut-off' (set to 0.1), and 'Elimination of DEGs' (set to 0.05). A modal dialog box is centered on the screen, featuring a green checkmark icon and the text 'DONE' and 'TCC was successfully performed.' with an 'Ok' button. At the bottom of the interface, there is a red button labeled 'TCC Computation + Code'. The footer contains the text 'Copyright © 2019 Bioinformation Engineering Lab, Graduate School of Agricultural and Life Sciences / Faculty of Agriculture, The University of Tokyo'.

TCC Computation

Filtering Threshold for Low Count Genes: Do not filter (30)

Number of Iteration: 3 (30)

FDR Cut-off: 0.1 (0.4)

Elimination of DEGs: 0.05 (0.4)

Showing 1 to 12 of 58,037 entries

	Gene Name	A Value	M Value	P Value	FDR
1	ENSG00000000003.14	13.761	0.369	0.009	0.126
2	ENSG00000000005.5	-2.831	0.148	1.000	1.000
3	ENSG000000000419.12	16.058	-0.076	0.591	1.000
4	ENSG000000000457.13	14.897	0.191	0.176	0.948
			-0.281	0.047	0.418
			0.148	1.000	1.000
			-0.465	0.467	1.000
			0.349	0.014	0.168
			-0.309	0.029	0.296
			0.602	0.000	0.001
			-0.069	0.627	1.000

Summary of TCC Normalization

- Library Size ¹ = Sum of Raw Count.
- Effective Library Size ² = Library Size × Normalization Factor.

TCC Computation

The screenshot shows the TCC-GUI web interface. On the left, there is a sidebar with navigation options: Heatmap (Step 3), Expression Level (Step 3), and Report (Step 4). The main area is divided into a control panel on the left and a data table on the right.

Control Panel:

- Filtering Threshold for Low Count Genes:** Slider set to 30. Text: "Do not filter".
- Number of Iteration:** Slider set to 3.
- FDR Cut-off:** Slider set to 0.1.
- Elimination of Potential DEGs:** Slider set to 0.05.
- Run TCC Computation:** Blue button.
- TCC Computation + Code:** Red button.

Data Table:

	Gene Name	A Value	M Value	P Value	FDR (FDR)
1	ENSG000000000003.14	13.761	0.369	0.009	0.126
2	ENSG000000000005.5	-2.831	0.148	1.000	1.000
3	ENSG0000000000419.12	16.058	-0.076	0.591	1.000
4	ENSG0000000000457.13	14.897	0.191	0.176	0.948
5	ENSG0000000000460.16	15.384	-0.281	0.047	0.418
6	ENSG0000000000938.12	-2.831	0.148	1.000	1.000
7	ENSG0000000000971.15	9.240	-0.465	0.467	1.000
8	ENSG000000001036.13	16.780	0.349	0.014	0.168
9	ENSG000000001084.10	16.400	-0.309	0.029	0.296
10	ENSG000000001167.14	16.244	0.602	0.000	0.001
11	ENSG000000001460.17	14.230	-0.069	0.627	1.000

Showing 1 to 12 of 58,037 entries

Summary of TCC Normalization:

- Copy
- Print
- Download
- Library Size ¹ = Sum of Raw Count.
- Effective Library Size ² = Library Size × Normalization Factor.



計算結果の保存

① ページ上部に移動。② を押せば、③ 赤枠の中身をCSV形式ファイルでダウンロードできます。

TCC-GUI: Graphical User Interface for TCC package

TCC Computation Parameters

Normalization Method: TMM

DEG Identification Method: edgeR

Filtering Threshold for Low Count Genes: Do not filter (30)

Number of Iteration: 3 (30)

FDR Cut-off: 0.1 (1)

Elimination of Potential DEGs: 0.05 (1)

Result Table

Download All Result (CSV) | Download Normalized Data (CSV)

Copy | Print | Download

Search:

- Filter genes by typing conditions (such as 2...5) in the filter boxes to filter numeric columns. Copy, Print or Download the filtered result for further analysis.
- Gene Name is colored according to FDR cut-off.

	Gene Name	A Value	M Value	P Value	Q Value (FDR)
1	ENSG000000000003.14	13.761	0.369	0.009	0.126
2	ENSG000000000005.5	-2.831	0.148	1.000	1.000
3	ENSG0000000000419.12	16.058	-0.076	0.591	1.000
4	ENSG0000000000457.13	14.897	0.191	0.176	0.948
5	ENSG0000000000460.16	15.384	-0.281	0.047	0.418
6	ENSG0000000000938.12	-2.831	0.148	1.000	1.000
7	ENSG0000000000971.15	9.240	-0.465	0.467	1.000
8	ENSG000000001036.13	16.780	0.349	0.014	0.168
9	ENSG000000001084.10	16.400	-0.309	0.029	0.296
10	ENSG000000001167.14	16.244	0.602	0.000	0.001

計算結果の保存

①ページ上部に移動。②を押せば、③赤枠の中身をCSV形式ファイルでダウンロードできます。④このときはこんな感じのファイル名になりました。

The screenshot shows the TCC-GUI web interface. On the left is a navigation menu with steps 0-4. The main panel is divided into 'TCC Computation Parameters' and 'Result Table'. The 'Result Table' shows a list of genes with columns for gene ID, log2 fold change, and p-value. A file save dialog box is open over the table, showing the file name '2019-03-01_tmm_edger_3_0.1_0.05_TCC.csv' and the file type 'Microsoft Excel Comma Separated Values File (*.csv)'. A red arrow with the number 4 points to the file name field.

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

名	更新日時	種類
---	------	----

MA Plot

RNA-seq発現変動解析結果を描画する目的でよく利用される①MA Plot。②まずはデフォルトで実行。

The screenshot shows the TCC-GUI web interface. On the left is a navigation menu with steps 0 through 4. The main area is divided into three sections: MA Plot Parameters, MA Plot, and Result Table.

MA Plot Parameters:

- Point Size: Slider set to 3 (range 1 to 5).
- FDR Cut-off: Slider set to 0.1 (range 0.01 to 1).
- DEGs Color: Dropdown menu set to red, with 4006 genes listed below.
- Generate MA Plot: A blue button with a play icon.

MA Plot:

Please click [Generate MA Plot] first.

Result Table:

Buttons: Copy, Print, Download. Search:

- Above buttons only deal with loaded part of the whole table (max to 99 rows).
- Gene Name was colored according to FDR cut-off.

	Gene Name	A Value	M Value	P Value	Q Value (FDR)
1	ENSG000000000003.14	13.761	0.369	0.009	0.126
2	ENSG000000000005.5	-2.831	0.148	1.000	1.000
3	ENSG0000000000419.12	16.058	-0.076	0.591	1.000
4	ENSG0000000000457.13	14.897	0.191	0.176	0.948
5	ENSG0000000000460.16	15.384	-0.281	0.047	0.418
6	ENSG0000000000938.12	-2.831	0.148	1.000	1.000
7	ENSG0000000000971.15	9.240	-0.465	0.467	1.000
8	ENSG000000001036.13	16.780	0.349	0.014	0.168

FDR vs DEGs:

Number (#) and Percentage (%) of DEGs satisfying different FDR cut-off.

Cut-off	DEGs(#)	DEGs(%)
---------	---------	---------

MA Plot

RNA-seq発現変動解析結果を描画する目的でよく利用される①MA Plot。②まずはデフォルトで実行。描画完了。全部で58,037遺伝子あるので、(低発現遺伝子のフィルタリングをしなければ)その分だけプロットがあります。①赤っぽい点が発現変動遺伝子(Differentially Expressed Genes; DEGs)のもので、②全部で4,006個あります。

The screenshot shows the TCC-GUI web interface. The browser address bar displays <https://infinityloop.shinyapps.io/TCC-GUI/>. The main navigation sidebar on the left lists various analysis tools, with 'MA Plot' highlighted as 'Step 3'. The central panel, titled 'MA Plot Parameters', includes sliders for 'Point Size' (set to 3), 'FDR Cut-off' (set to 0.1), and a dropdown for 'DEGs Color' showing '4006 genes' in red. A 'Generate MA Plot' button is visible. Below the parameters is a table view for 'FDR vs DEGs'. The right panel, titled 'MA Plot', shows a scatter plot with the y-axis labeled $M = \log_2(G2) - \log_2(G1)$ and the x-axis labeled $A = (\log_2(G2) + \log_2(G1)) / 2$. The plot title is 'MA Plot with q-value < 0.1 (10% FDR)'. Red points represent DEGs, and black points represent non-DEGs. Two red arrows labeled '1' point to specific red points in the plot. A text box above the plot says 'Hover over the point to show gene's expression level of interest.' Below the plot is a 'Result Table' section with 'Copy', 'Print', and 'Download' buttons, a search bar, and a note: 'Above buttons only deal with loaded part of the whole table (max to 99 rows). Gene Name was colored according to FDR cut-off.'

MA Plot

①色は自由に変更可能。②4,006遺伝子は、発現変動解析結果として、③q-value < 0.1という条件を満たすものたちです。この0.1というのは、「本当はDEGではないものが混入している割合 (False Discovery Rate; FDR)」に相当します。この閾値 (cut-off) の場合は $4,006 \times 0.1 = 400.6$ 個が偽物だということです。

TCC-GUI: Graphical User Interface for TCC package

Documentation

Data Simulation Step 0

Exploratory Analysis Step 1

TCC Computation Step 2

MA Plot Step 3

Volcano Plot Step 3

Heatmap Step 3

Expression Level Step 3

Report Step 4

MA Plot Parameters

Point Size

FDR Cut-off

DEGs Color

4006 genes

Generate MA Plot

MA Plot

MA Plot with q-value < 0.1 (10% FDR)

M = $\log_2(G2) - \log_2(G1)$

A = $(\log_2(G2) + \log_2(G1)) / 2$

DEG

non-DEG

Hover over the point to show gene's expression level of interest.

Result Table

Copy Print Download

Search:

Number (#) and Percentage (%) of DEGs satisfying different FDR cut-off.

Cut-off	DEGs(#)	DEGs(%)
---------	---------	---------

• Above buttons only deal with loaded part of the whole table (max to 99 rows).
• Gene Name was colored according to FDR cut-off.

MA Plot

GUIの一番のメリットは、③の閾値を変更すれば、その閾値を満たす遺伝子数が即座にわかることです。例えば、FDR = 0.05を満たす遺伝子数は④3,389個だというのがすぐにわかります。これは、偽物混入割合を低めに設定し直したことを意味するため、条件を満たす遺伝子数は当然少なくなります。この閾値 (cut-off) の場合は、 $3,389 \times 0.05 = 169.45$ 個が偽物だということです。⑤を押して、MA Plotを再描画すると…

TCC-GUI: Graphical User Interface for TCC package

Documentation

Data Simulation Step 0

Exploratory Analysis Step 1

TCC Computation Step 2

MA Plot Step 3

Volcano Plot Step 3

Heatmap Step 3

Expression Level Step 3

Report Step 4

MA Plot Parameters

Point Size

FDR Cut-off

DEGs Color

3389 genes

Generate MA Plot

MA Plot

MA Plot with q-value < 0.1 (10% FDR)

DEG non-DEG

Hover over the point to show gene's expression level of interest.

M = $\log_2(G2) - \log_2(G1)$

A = $(\log_2(G2) + \log_2(G1)) / 2$

Table Plot

FDR vs DEGs

Number (#) and Percentage (%) of DEGs satisfying different FDR cut-off.

Result Table

Copy Print Download

Search:

- Above buttons only deal with loaded part of the whole table (max to 99 rows).
- Gene Name was colored according to FDR cut-off.

MA Plot

こんな感じになります。①の部分も反映され、②のDEG数も4,006個から3,389個になったのだからということが黒のプロット領域の増加から想像できます。

TCC-GUI: Graphical User Interface for TCC package

MA Plot Parameters

Point Size: 1 1.4 1.8 2.6 3 3.4 4.2 5

FDR Cut-off: 0.05

DEGs Color: 3389 genes

Generate MA Plot

MA Plot

MA Plot with q-value < 0.05 (5% FDR)

M = $\log_2(G_2) - \log_2(G_1)$

A = $(\log_2(G_2) + \log_2(G_1))/2$

DEG non-DEG

Hover over the point to show gene's expression level of interest.

Result Table

Copy Print Download

Search:

• Above buttons only deal with loaded part of the whole table (max to 99 rows).
• Gene Name was colored according to FDR cut-off.

Contents

- 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

Disconnected...

recount2やTCC-GUIのウェブ版は、接続から5分程度で①Disconnectedとなり、最初からやり直す必要が出てきます。これらのウェブツールに慣れてくればくるほど、イラッと感が募ってきます。

TCC-GUI: Graphical User Interface for TCC package

Documentation

Data Simulation Step 0

Exploratory Analysis Step 1

TCC Computation Step 2

MA Plot Step 3

Volcano Plot Step 3

Heatmap Step 3

Expression Level Step 3

Report Step 4

MA Plot Parameters

Point Size

FDR Cut-off

DEGs Color

4007 genes

Generate MA Plot

MA Plot

MA Plot with $q\text{-value} < 0.1$ (10% FDR)

$M = \log_2(G_2) - \log_2(G_1)$

$A = (\log_2(G_2) + \log_2(G_1)) / 2$

DEG non-DEG

Hover over the point to show gene's expression level of interest.

Result Table

Copy Print Download

Search:

Disconnected from the server. Reload

1

https://infinityloop.shinyapps.io/TCC-GUI/w_e99b7dde/#

Disconnected...

Disconnected from the server.
Reload

https://infinityloop.shinyapps.io/TCC-GUI/

TCC-GUI: Graphical User Interface for TCC package

Documentation

Data Simulation Step 0

Exploratory Analysis Step 1

TCC Computation Step 2

MA Plot Step 3

Volcano Plot Step 3

Heatmap Step 3

Expression Level Step 3

Report Step 4

MA Plot Parameters

Point Size

FDR Cut-off

DEGs Color

4007 genes

Generate MA Plot

MA Plot

MA Plot with q -value

$M = \log_2(G_2) - \log_2(G_1)$

$A = (\log_2(G_2) + \log_2(G_1)) / 2$

Result Table

Copy Print Download

Search:

- Above buttons only deal with loaded part of the whole table (max to 99 rows).
- Gene Name was colored according to FDR cut-off.

recount2やTCC-GUIのウェブ版は、接続から5分程度で①Disconnectedとなり、最初からやり直す必要が出てきます。これらのウェブツールに慣れてくればくるほど、イラッと感が募ってきます。この原因は、②の場所を使っているためです。一般にウェブツールを提供する場合は、自分たちのサーバ上に置きますが、recount2やTCC-GUIは②の場所に置かせてもらっているのです。TCC-GUIの場合は、たしか月20時間まで無料というプランを選択させてもらっています。それゆえ一定の時間が経過すると接続を切る設定にしないと、すぐに無料時間を超えてしまうのです。ユーザが増えれば増えるほど、月20時間なんてあっという間に過ぎてしまい、結果として(接続およびユーザがw)ブチブチ切れるという状況になるのです。その対策としては、TCC-GUIの場合はローカル版を利用することです。

Contents

- 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

RStudio

最初に行うのは、①RStudioのインストール。②
Download Rstudio。

The image shows a browser window displaying the RStudio website. The browser's address bar shows the URL `https://www.rstudio.com`. The website header includes the RStudio logo and navigation links for Products, Resources, Pricing, About Us, and Blogs. The main content area features a large blue banner with the text "RStudio" and "Open source and enterprise-ready professional software for R". A red arrow with the number "1" points to the "RStudio" text. To the right of the banner, there are four buttons: "Download RStudio", "Discover Shiny", "Discover RStudio Package Manager", and "Discover RStudio Connect". A red arrow with the number "2" points to the "Download RStudio" button. At the bottom of the page, there are icons for RStudio, Shiny, and tidyR.

RStudio

最初に行うのは、①RStudioのインストール。②
Download Rstudio。③DOWNLOAD。

Download RStudio - RStudio


https://www.rstudio.com/products/rstudio/download/

R Studio

Products Resources Pricing About Us Blogs

Choose Your Version of RStudio

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace. [Learn More about RStudio features.](#)



RStudio Desktop Open Source License	RStudio Desktop Commercial License	RStudio Server Open Source License	RStudio Server Pro Commercial License	RStudio Server RStudio Co Commercial License
FREE	\$995 per year	FREE	\$9,995 per year	\$29,995 year
DOWNLOAD	BUY	DOWNLOAD	DOWNLOAD	TALK
Learn More	Learn More	Learn More	Learn More	Learn More

RStudio

最初に行うのは、①RStudioのインストール。②Download Rstudio。③DOWNLOAD。④の中の、上のやつがWindows版、下のやつがMac版のインストーラ。基本的に言われるがままにインストールを進めればよい。

Download RStudio - RStudio

https://www.rstudio.com/products/rstudio/download/#download

RStudio

Products Resources Pricing About Us Blogs

RStudio Desktop 1.1.463 — Release Notes

RStudio requires R 3.0.1+. If you don't already have R, download it here.

Linux users may need to import RStudio's public code-signing key prior to installation, depending on the operating system's security policy.

Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 1.1.463 - Windows Vista/7/8/10	85.8 MB	2018-10-29	58b3d796d8cf96fb8580c62f46ab64d4
RStudio 1.1.463 - Mac OS X 10.6+ (64-bit)	74.5 MB	2018-10-29	a79032ba4d7daaa86a8da01948278d94
RStudio 1.1.463 - Ubuntu 12.04-15.10/Debian 8 (32-bit)	89.3 MB	2018-10-29	8a8755fa9fae2bafce289df3358aaf63
RStudio 1.1.463 - Ubuntu 12.04-15.10/Debian 8 (64-bit)	97.4 MB	2018-10-29	bc50d6bd34926c1cc3ae4a209d67d649
RStudio 1.1.463 - Ubuntu 16.04+/Debian 9+ (64-bit)	65 MB	2018-10-29	cf659db18619cc78d1592fefaa7c753
RStudio 1.1.463 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	88.1 MB	2018-10-29	742f0bad60dfeaa3281576e14ad6699e
RStudio 1.1.463 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	90.6 MB	2018-10-29	c7303067a0ca99deea7e427b956059d1

RStudio起動

インストールが無事完了して、起動するとこんな感じになります。

The screenshot shows the RStudio application window. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for file operations and a search bar. The main interface is divided into several panes:

- Console:** Displays the R version (3.5.1) and copyright information. It contains the following text:

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[workspace loaded from ~/.RData]

> |
```
- Environment:** Shows the Global Environment with a search bar. It lists variables in the Data and Functions sections:

Variable	Type
actionBttnP...	List of 4
arab	Large matrix (157332 elements, ...)
hypoData	num [1:1000, 1:6] 34 358 1144 ...
convert2c1	function (x, df)
make_summar...	function (df)
- Files:** A file explorer showing the contents of the Home directory. It lists files and folders such as .RData, .Rhistory, 2017, 2018, 2019, html, identity, Officeのカスタム テンプレート, Outlook ファイル, paper, public_html, R, and その他.

作業ディレクトリの変更

今は特に必要ないですが、RStudioの場合は、
①Session、②Set Working Directory、③
Choose Directory、のようにすればよいです。

The screenshot shows the RStudio interface. The 'Session' menu is open, with 'Set Working Directory' highlighted. A red arrow labeled '1' points to the 'Session' menu. A second red arrow labeled '2' points to 'Set Working Directory'. A third red arrow labeled '3' points to 'Choose Directory...' in a sub-menu. The sub-menu also shows 'To Source File Location' and 'To Files Pane Location'. The file explorer window shows the 'Home' directory with various folders and files.

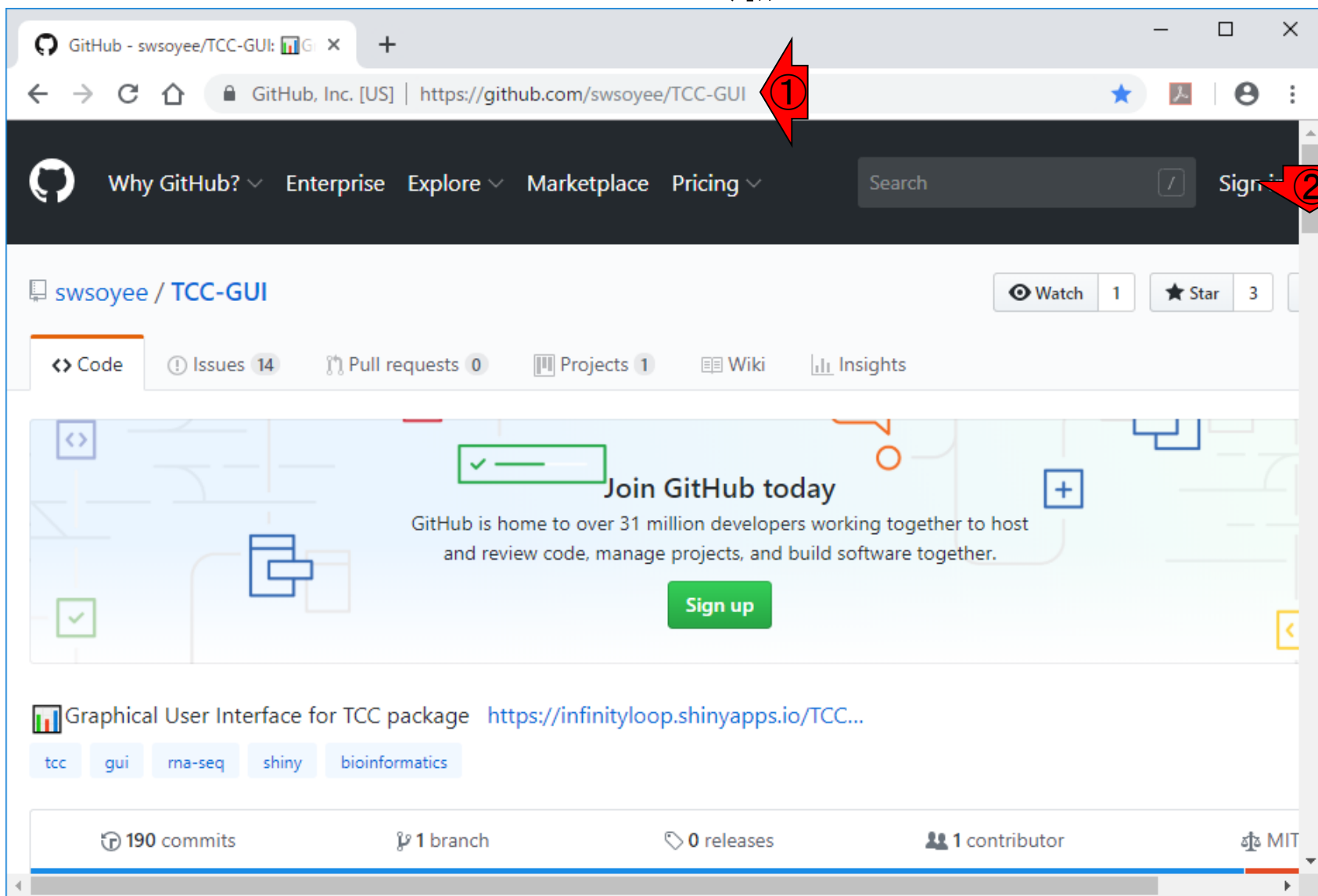
Name	Size	Modified
.RData	293.6 KB	Dec 12, 2018
.Rhistory	168 B	Feb 14, 2019
2017		
2018		
2019		
html		
identity	528 B	Sep 16, 2005
Officeのカスタム テンプレート		
Outlook ファイル		
paper		
public_html		
R		
その他		

Contents

- 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

TCC-GUIローカル版

「TCC-GUI」でググれば、①のサイトに辿り着けます。②ページ下部に移動。



TCC-GUIローカル版

「TCC-GUI」でググれば、①のサイトに辿り着けます。②ページ下部に移動。さきほどまで利用していた③オンライン版です。④Launchをクリック。

The screenshot shows the GitHub repository page for TCC-GUI. The browser address bar displays the URL <https://github.com/swsoyee/TCC-GUI>. The page content includes a navigation menu with the following items: Usage, Online version, Standalone version, Installation, and Launch. Below the menu is a 'References' section containing four citations. Red arrows with numbers 2, 3, and 4 point to specific elements: arrow 2 points to the bottom of the page, arrow 3 points to the 'TCC-GUI' link in the 'Go to' section, and arrow 4 points to the 'Launch' link in the navigation menu.

Usage

Online version

Go to [TCC-GUI](#)

Standalone version

Installation

Launch

References

[1] Sun J, Nishiyama T, Shimizu K, et al. TCC: an R package for comparing tag count data with robust normalization strategies. *BMC bioinformatics*, 2013, 14(1): 219.

[2] Robinson M D, McCarthy D J, Smyth G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 2010, 26(1): 139-140.

[3] Anders S, Huber W. Differential expression analysis for sequence count data. *Genome biology*, 2010, 11(10): R106.

[4] Love M I, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 2014, 15(12): 550.

TCC-GUIローカル版

「TCC-GUI」でググれば、①のサイトに辿り着けます。②ページ下部に移動。さきほどまで利用していた③オンライン版です。④Launchをクリック。こんな感じになります。基本的に⑤のコマンドを、RStudio上でコピー実行すれば、TCC-GUIのローカル版を起動することができます。

GitHub - swsoyee/TCC-GUI: TCC-GUI

GitHub, Inc. [US] | https://github.com/swsoyee/TCC-GUI

Usage

Online version

Go to [TCC-GUI](#).

Standalone version

▶ [Installation](#)

▼ [Launch](#)

Run the following command to launch `TCC-GUI` in your local environment, then it will download `TCC-GUI` automatically from github and launch.

Method 1

```
shiny::runGitHub("TCC-GUI", "swsoyee", subdir = "TCC-GUI", launch.browser = TRUE)
```

This method always download the source code from github before launching, so maybe you can try to download all the source code by yourself and launch it.

⑤

TCC-GUIローカル版

「TCC-GUI」でググれば、①のサイトに辿り着けます。②ページ下部に移動。さきほどまで利用していた③オンライン版です。④Launchをクリック。こんな感じになります。基本的に⑤のコマンドを、RStudio上でコピー実行すれば、TCC-GUIのローカル版を起動することができます。コマンドを反転させて、⑥コピー。

GitHub - swsoyee/TCC-GUI: TCC-GUI

GitHub, Inc. [US] | https://github.com/swsoyee/TCC-GUI

Usage

Online version

Go to [TCC-GUI](#).

Standalone version

▶ [Installation](#)

▼ [Launch](#)

Run the following command to launch `TCC-GUI` in your local environment, then it will download `TCC-GUI` automatically from github and launch.

Method 1

```
shiny::runGitHub("TCC-GUI", "swsoyee", subdir = "TCC-GUI", launch.browser = FALSE)
```

This method always download the source code from github before launch. You can also download the source code by yourself and launch it.

コピー(C)	Ctrl+C
Google で「shiny::runGitHub("TCC-GUI", "swsoyee", subdir = "...」を検索(S)	
印刷(P)...	Ctrl+P
検証(I)	Ctrl+Shift+I

TCC-GUIローカル版

「TCC-GUI」でググれば、①のサイトに辿り着けます。②ページ下部に移動。さきほどまで利用していた③オンライン版です。④Launchをクリック。こんな感じになります。基本的に⑤のコマンドを、RStudio上でコピー実行すれば、TCC-GUIのローカル版を起動することができます。コマンドを反転させて、⑥コピー。⑦Edit、⑧Paste。

The screenshot shows the RStudio interface. The 'Edit' menu is open, and the 'Paste' option is highlighted with a red arrow labeled '8'. A red arrow labeled '7' points to the top-left corner of the RStudio window. The console shows the following text:

```
-- "Feather Spray"
ation for Statistical Computing
x64 (64-bit)

with ABSOLUTELY NO WARRANTY.
te it under certain conditions.
)' for distribution details.

with many contributors.
e information and
or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

The file explorer window shows the 'Home' directory with the following files and folders:

Name	Size	Modified
2017		
2018		
2019		
html		
identity	528 B	Sep 16, 2005
Officeのカスタム テンプレート		
Outlook ファイル		
paper		
public_html		
R		
その他		

TCC-GUIローカル版

「TCC-GUI」でググれば、①のサイトに辿り着けます。②ページ下部に移動。さきほどまで利用していた③オンライン版です。④Launchをクリック。こんな感じになります。基本的に⑤のコマンドを、RStudio上でコピー実行すれば、TCC-GUIのローカル版を起動することができます。コマンドを反転させて、⑥コピー。⑦Edit、⑧Paste。⑨こんな感じになったことを確認できたらリターンキーを押す。

The image shows a screenshot of the RStudio interface with the terminal window open. The terminal displays the R version information and a command to run the TCC-GUI locally. A red arrow with the number 9 points to the end of the command line. To the right, a Windows File Explorer window shows the 'Home' directory with various folders and files.

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> shiny::runGitHub("TCC-GUI", "swoyee", subdir = "TCC-GUI", launch
  .browser = TRUE)|
```

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

Home

	Name	Size	Modified
<input type="checkbox"/>	2017		
<input type="checkbox"/>	2018		
<input type="checkbox"/>	2019		
<input type="checkbox"/>	html		
<input type="checkbox"/>	identity	528 B	Sep 16, 2005
<input type="checkbox"/>	Officeのカスタム テンプレート		
<input type="checkbox"/>	Outlook ファイル		
<input type="checkbox"/>	paper		
<input type="checkbox"/>	public_html		
<input type="checkbox"/>	R		
<input type="checkbox"/>	その他		

TCC-GUIローカル版

最初はこんな感じでTCC-GUIのプログラムをダウンロードしています。

The screenshot shows the RStudio interface. The console window displays the output of the `shiny::runGitHub` command, which has successfully downloaded the TCC-GUI archive from GitHub. The file explorer window shows the local file system structure, including folders for years (2017, 2018, 2019), document types (html, paper, public_html, R), and system files (identity, Office templates, Outlook files, その他).

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> shiny::runGitHub("TCC-GUI", "swsoyee", subdir = "TCC-GUI", launch.brow
ser = TRUE)
Downloading https://github.com/swsoyee/TCC-GUI/archive/master.tar.gz
|
```

Name	Size	Modified
2017		
2018		
2019		
html		
identity	528 B	Sep 16, 2005
Officeのカスタム テンプレート		
Outlook ファイル		
paper		
public_html		
R		
その他		

TCC-GUIローカル版

しばらくすると、画面がざ~っと流れて、ウェブブラウザが起動します。

The screenshot shows the RStudio interface. The left pane is the Terminal, displaying the following output:

```
~/
The following object is masked from 'package:genomickranges':
  shift
The following object is masked from 'package:IRanges':
  shift
The following objects are masked from 'package:S4vectors':
  first, second
The following objects are masked from 'package:dplyr':
  between, first, last
Attaching package: 'tidyr'
The following object is masked from 'package:S4vectors':
  expand
Attaching package: 'MASS'
The following object is masked from 'package:dplyr':
  select
The following object is masked from 'package:plotly':
  select
Listening on http://127.0.0.1:3730
```

The right pane shows the Environment pane, which is currently empty, displaying "Environment is empty". Below it is the Files pane, showing a file explorer view of the Home directory with the following files and folders:

Name	Size	Modified
2017		
2018		
2019		
html		
identity	528 B	Sep 16, 2005
Officeのカスタム テンプレート		
Outlook ファイル		
paper		
public_html		
R		
その他		

TCC-GUI ローカル版

無事TCC-GUIのローカル版が起動しました。
①の部分が異なっていることがわかります。
これでブチブチ切れる問題から解放されます。



TCC-GUI: Graphical User Interface for TCC package

Documentation
Data Simulation **Step 0**
Exploratory Analysis **Step 1**

Welcome to TCC-GUI | Data Simulation | Exploratory Analysis | TCC Computation

MA Plot | Volcano Plot | Heatmap | Expression Level Plot | Analysis Report

What's TCC?

TCC^[1] is a R/Bioconductor package provides a series of functions for performing differential expression (DE) analysis from RNA-seq count data using a robust normalization strategy (called DEGES).

The basic idea of DEGES is that potential differentially expressed genes (DEGs) among compared samples should be removed before data normalization to obtain a well-ranked gene list where true DEGs are top-ranked and non-DEGs are bottom ranked. This can be done by performing the multi-step normalization procedures based on DEGES (DEG elimination strategy) implemented in TCC.

TCC internally uses functions provided by edgeR^[2], DESeq^[3], DESeq2^[4], and baySeq^[5]. The multi-step