2019.03.01版

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/

予定講義内容です

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

門田　幸二のホームページ

名前　門田　幸二（かどた　こうじ）

所属　東京大学 大学院農学生命科学研究科 アグリバイオインフォマティクス教育研究ユニット
東京大学 微生物科学イノベーション連携研究機構

身分　准教授

研究分野　バイオインフォマティクス（トランスクリプトーム解析）

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

研究テーマ

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

①本日分はコチラ

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 ～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

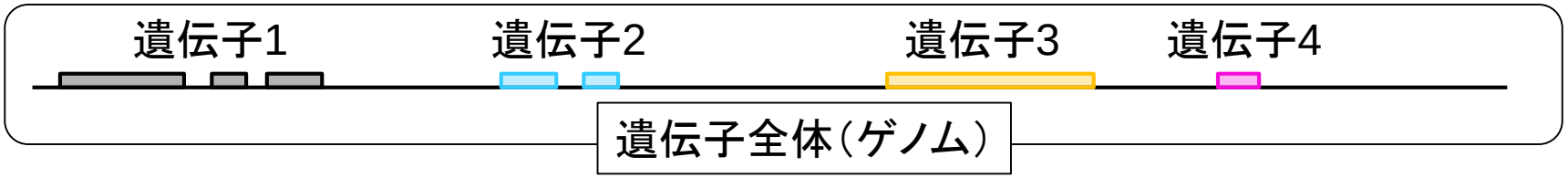
Contents

- 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

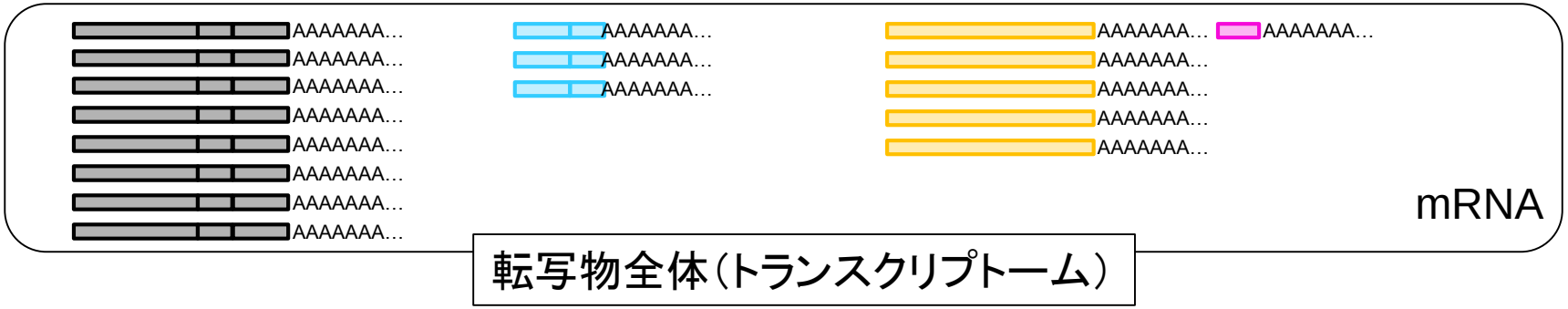
働いているRNAの種類や量を調べるのが目的

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
・…

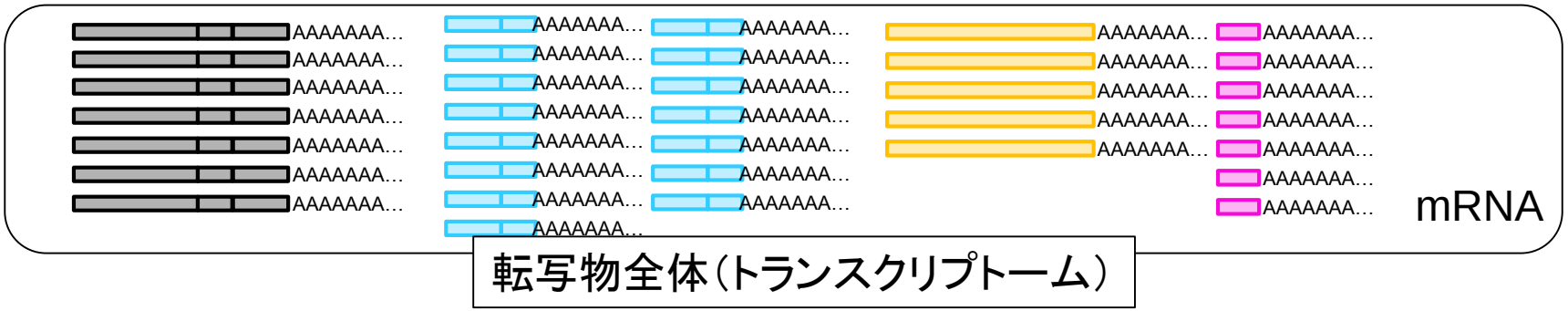
働いているRNAの種類や量を調べるのが目的

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

光刺激

・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…）

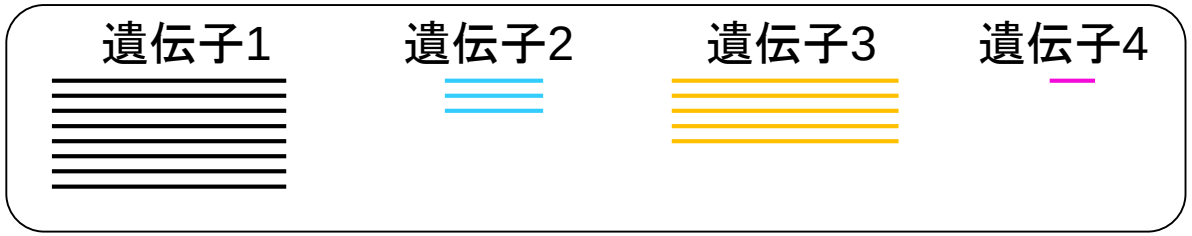


・遺伝子2は光刺激に応答して発現亢進
・遺伝子4も光刺激に応答して発現亢進

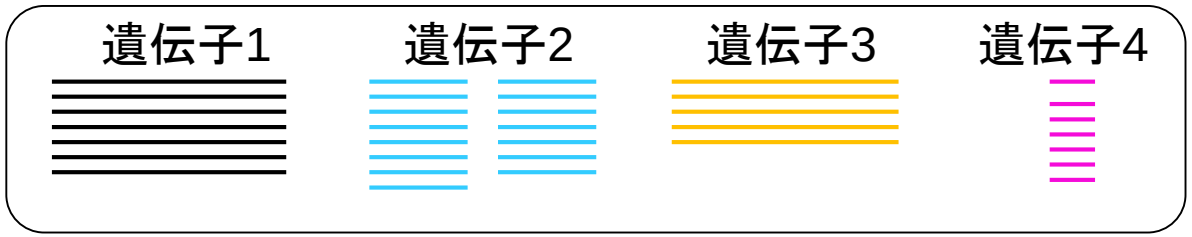
トランスクリプトーム解析

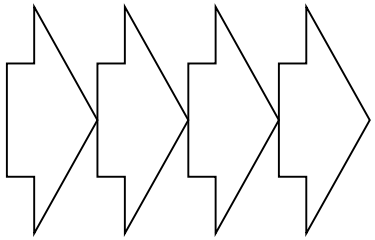
状態の異なる複数サンプルのデータを取得して解析するのが一般的。サンプル間比較

- 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



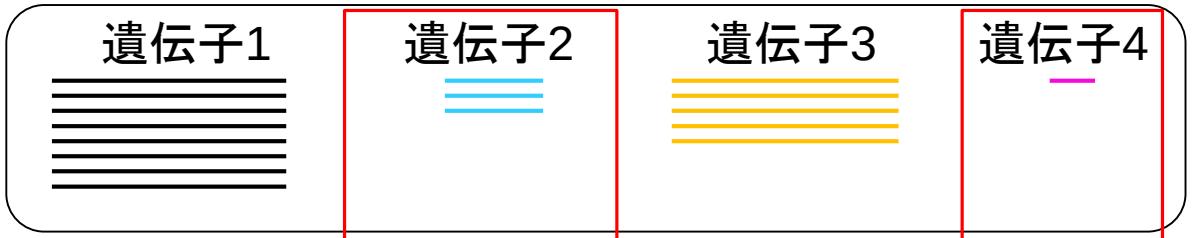


	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…

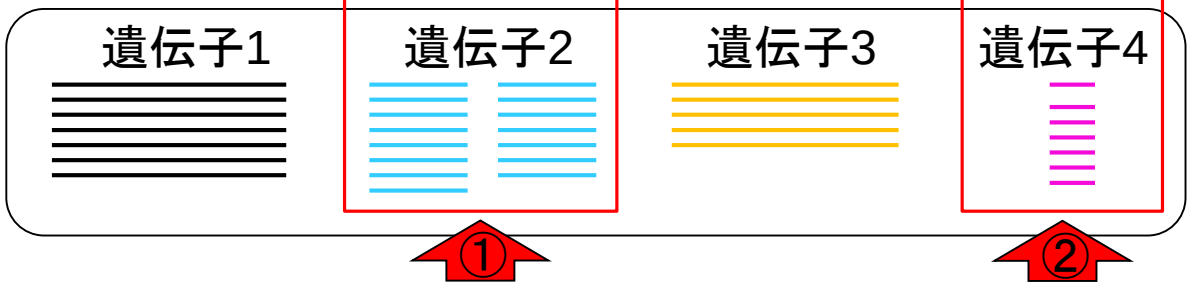
トランスクリプトーム解析

具体的な目的は、①や②の発現変動遺伝子同定など

- 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



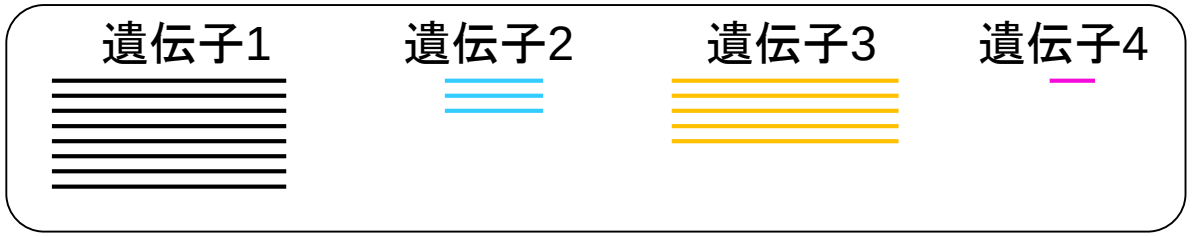
これがいわゆる「遺伝子発現行列」

	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
…	…	…

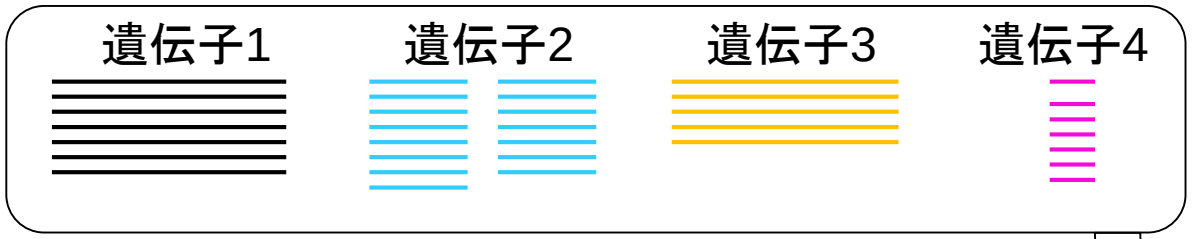
トランスクリプトームデータ取得

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

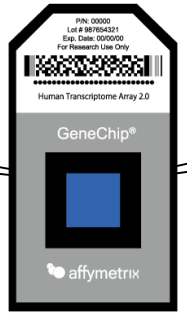


- 光刺激後（T2）の目のトランスクリプトーム



これがいわゆる
「遺伝子発現行列」

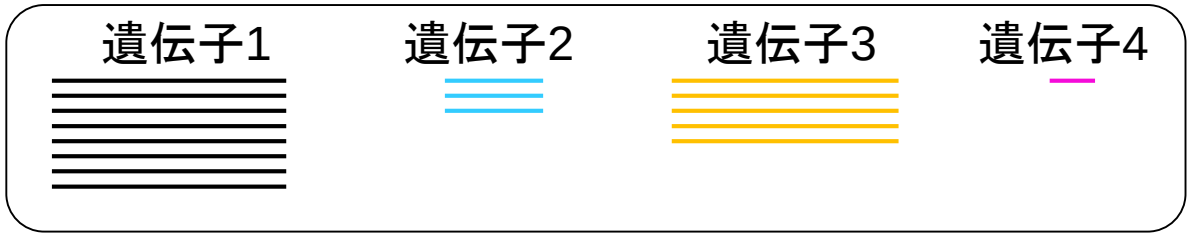
	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…



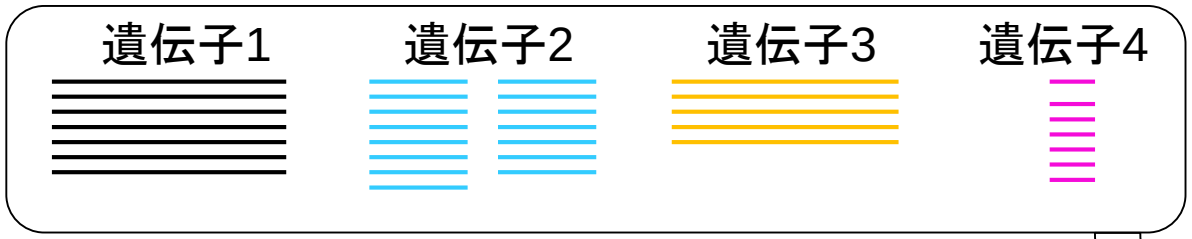
トランスクリプトームデータ取得

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

- 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」

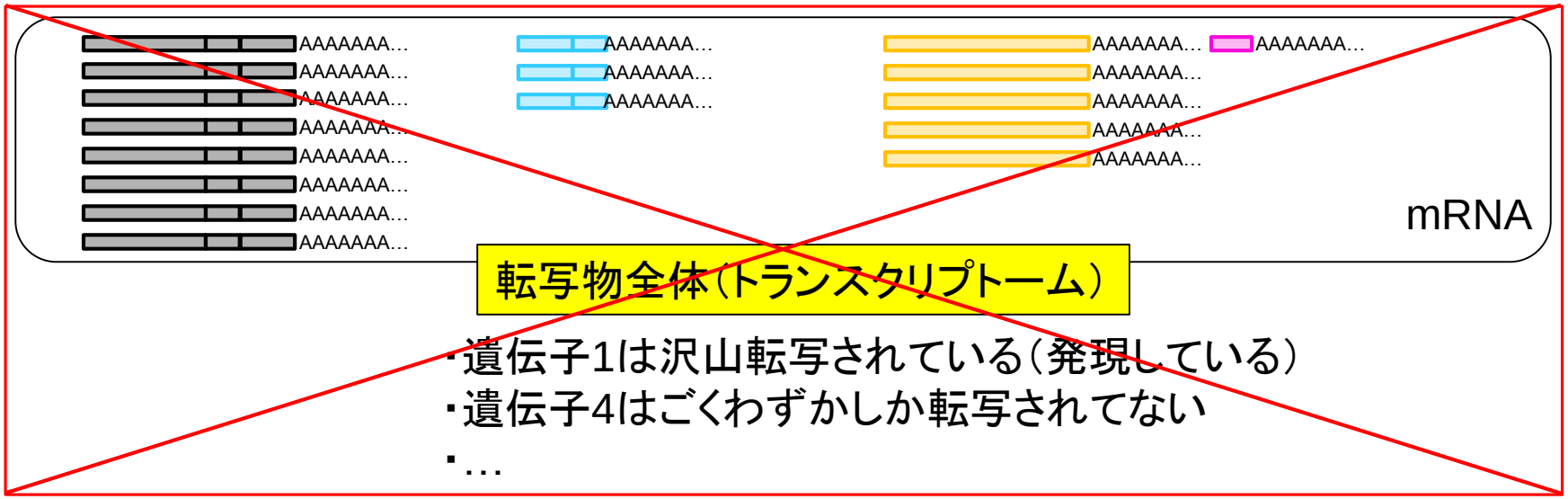
	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…

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…）



・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
・…

遺伝子 ≠ 転写物

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

- ある状態のあるサンプル(例:目)のあるゲノムの領域

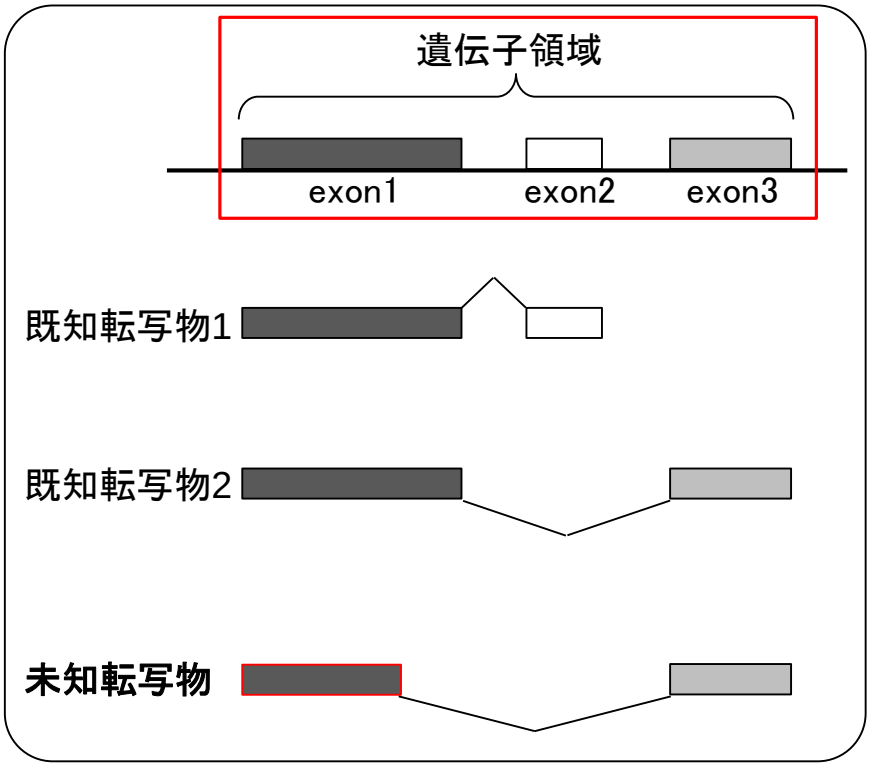
ヒト

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

ヒト

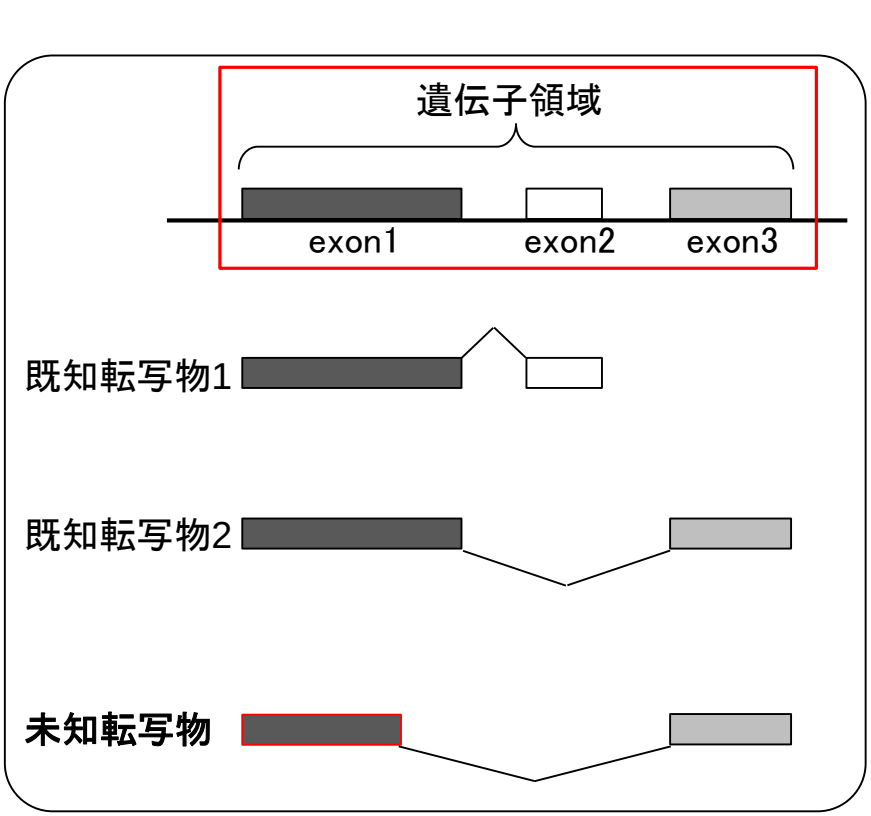


真の転写物情報

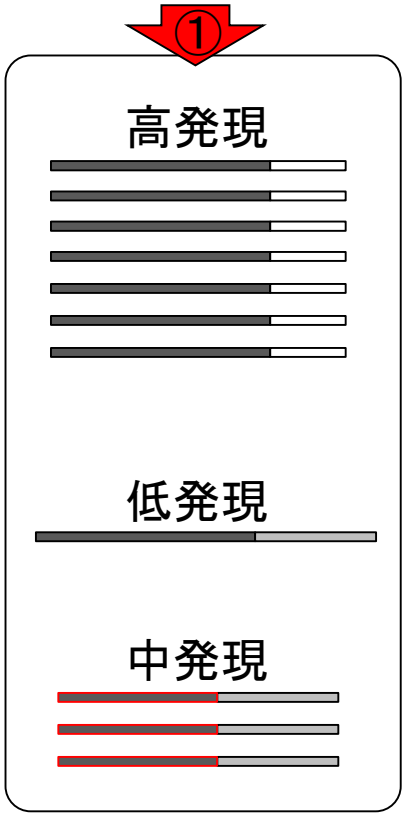
実際の細胞内（例：目のサンプル）での発現情報（働いている度合い）が①のような感じだったとする

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



真の転写物情報



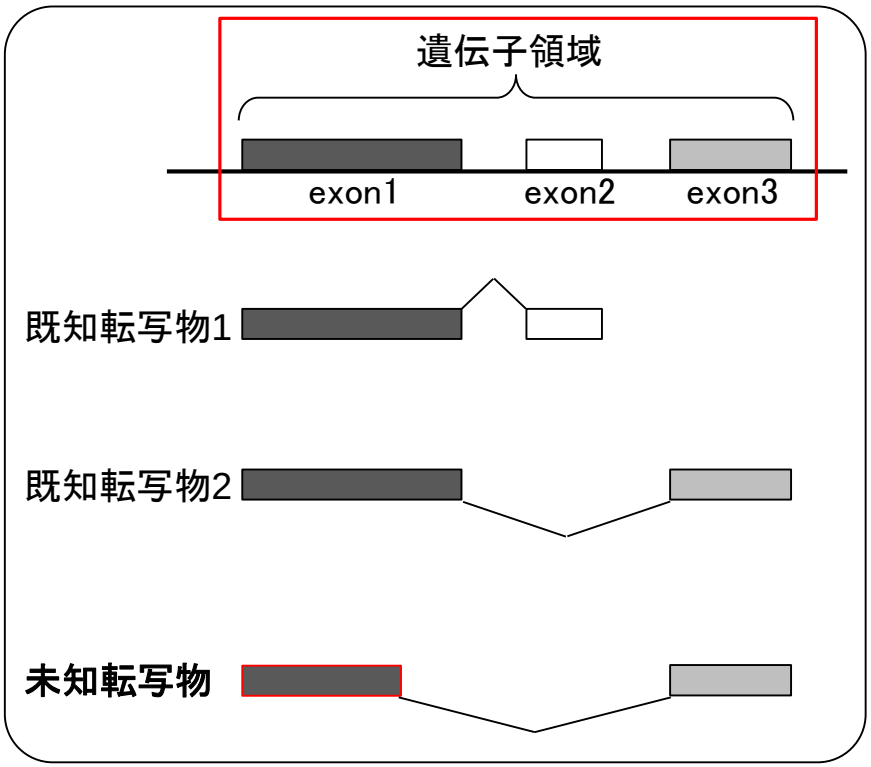
真の発現情報

遺伝子 ≠ 転写物

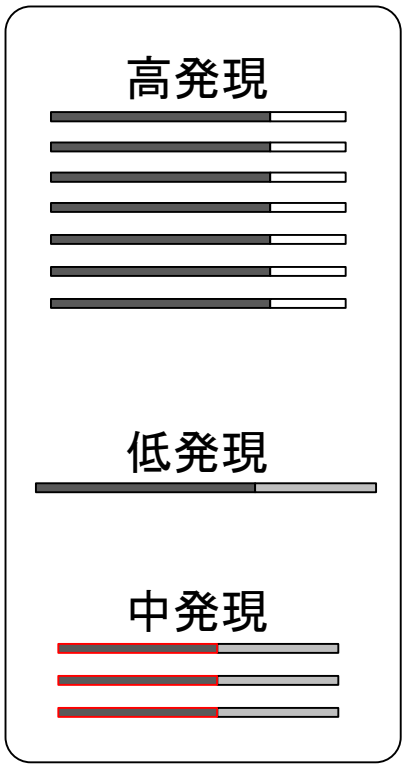
①NGS機器を用いて転写されているmRNA配列決定（RNA-seq）をした結果のイメージ

妄想

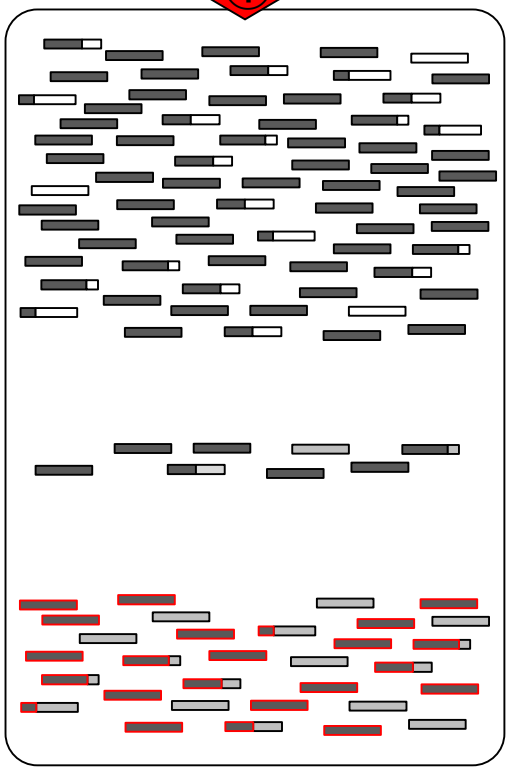
- ある状態のあるサンプル（例：目）のあるゲノムの領域



真の転写物情報



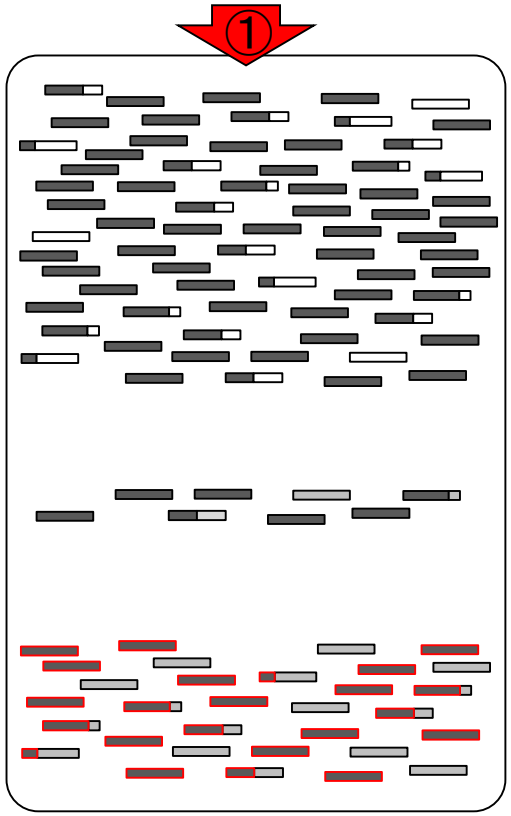
真の発現情報



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、

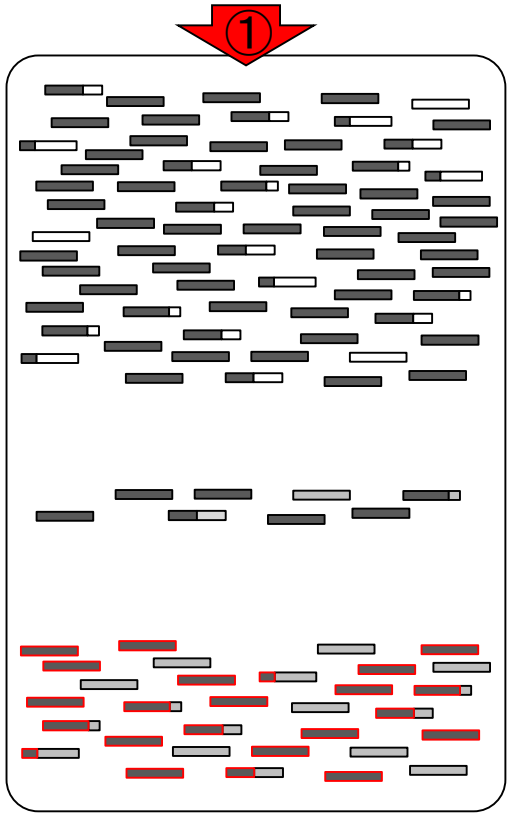


RNA-seqデータ

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、

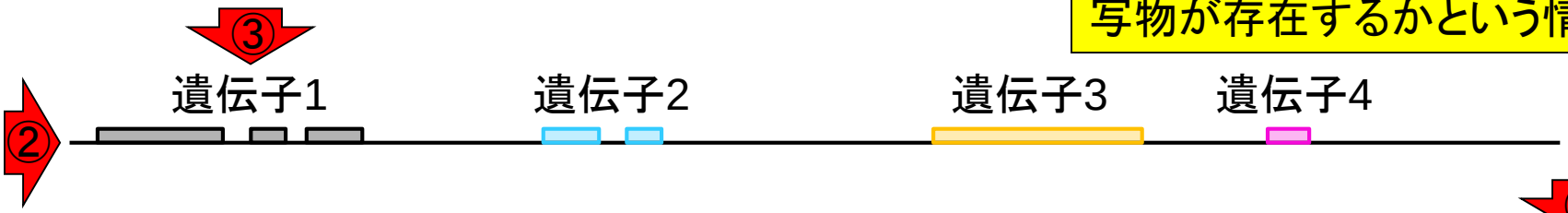
②

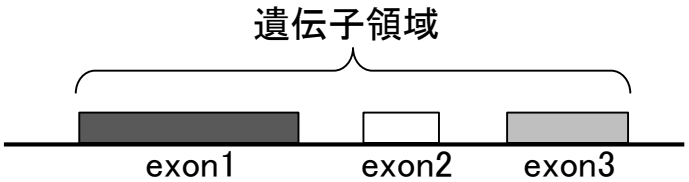


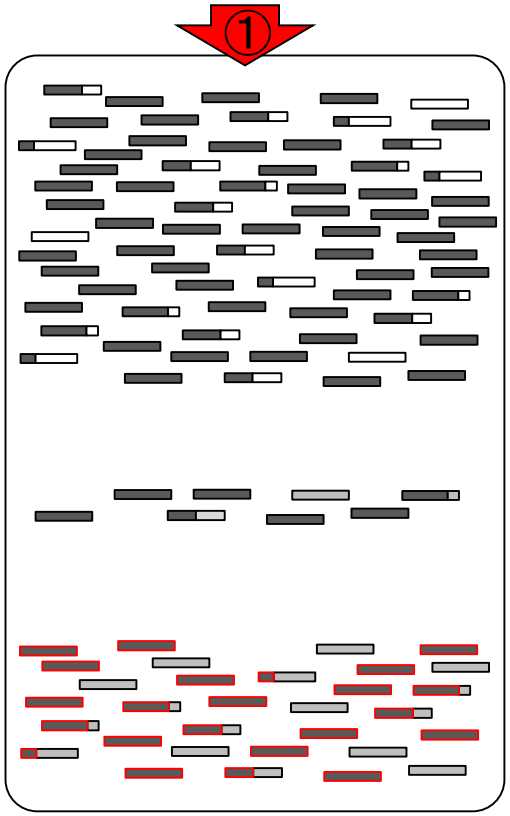
RNA-seqデータ

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）



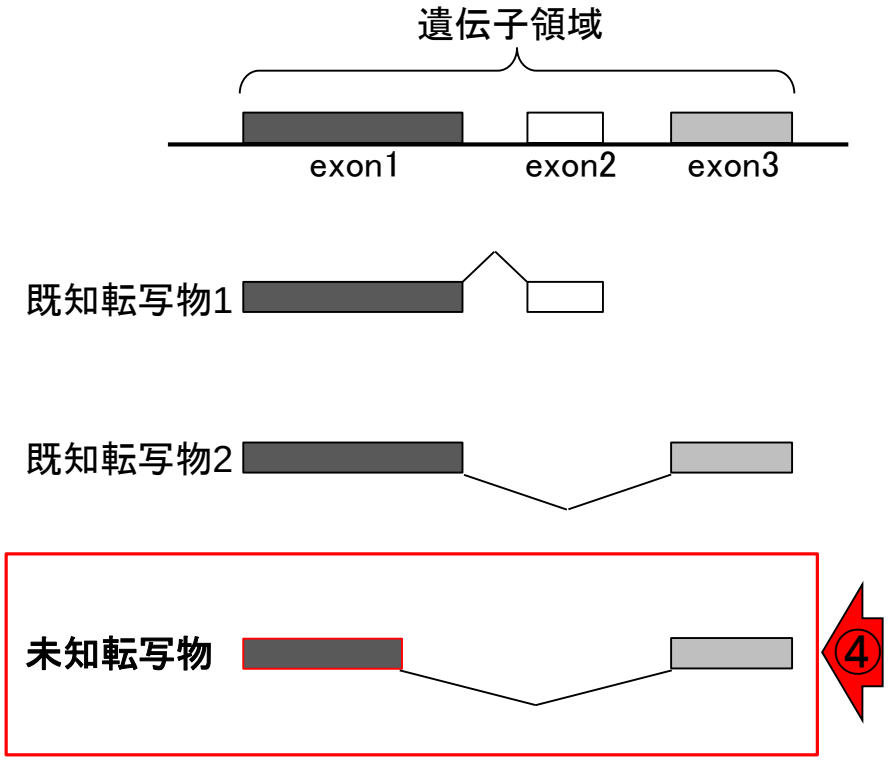


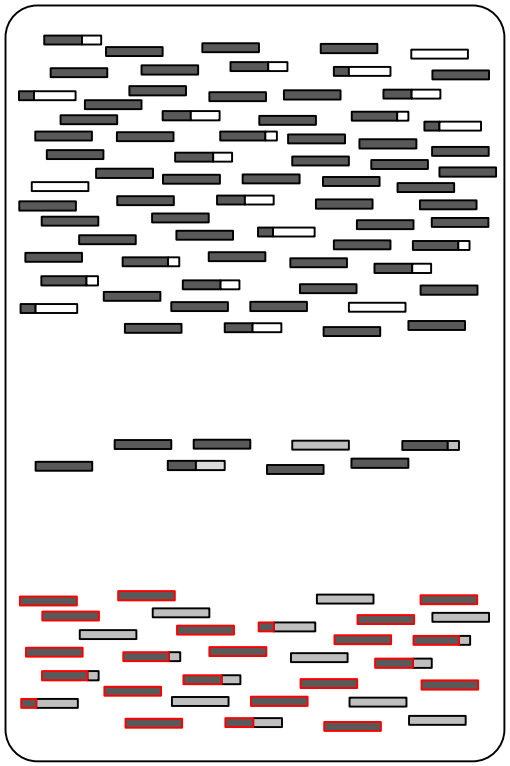


RNA-seqデータ

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。

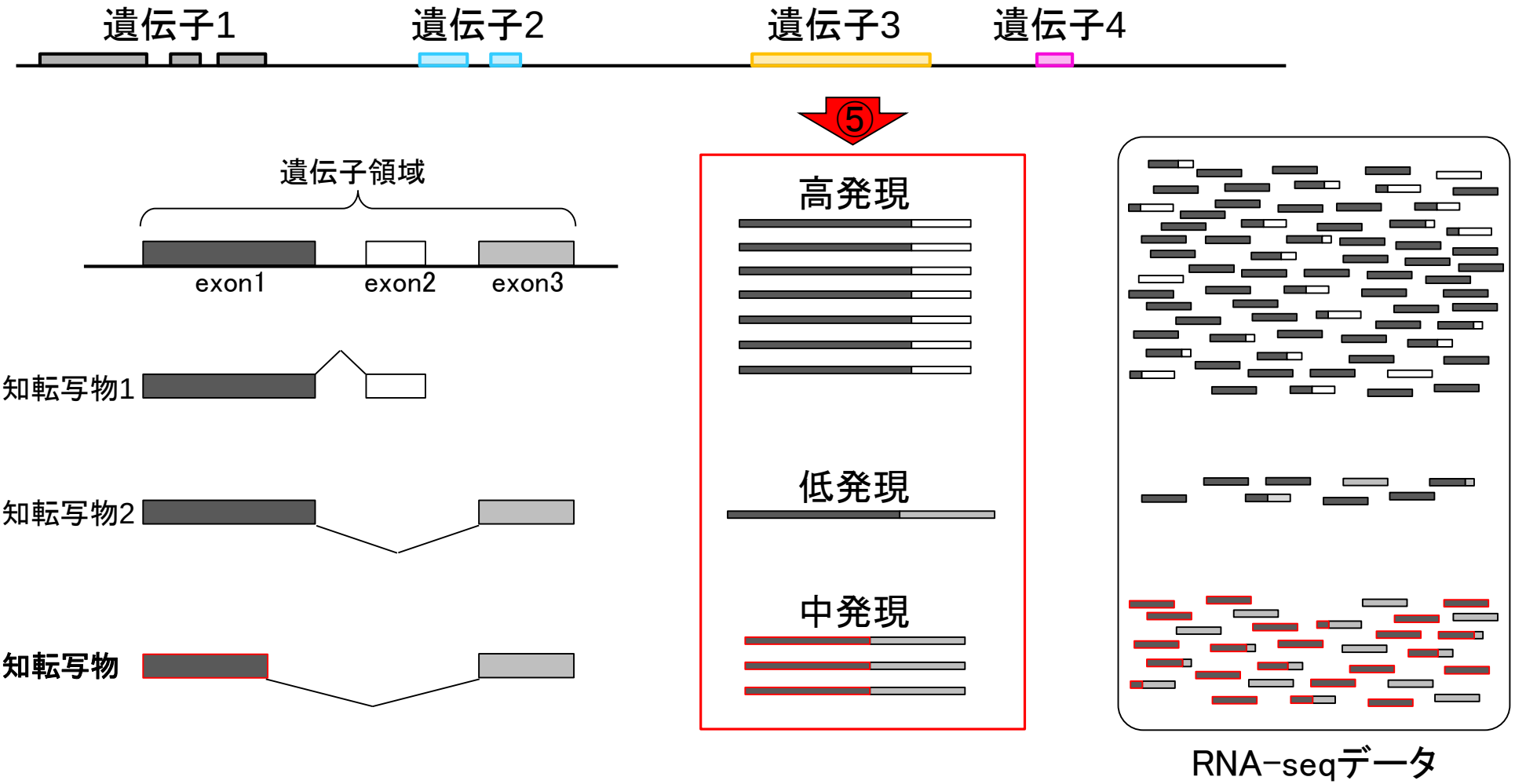




RNA-seqデータ

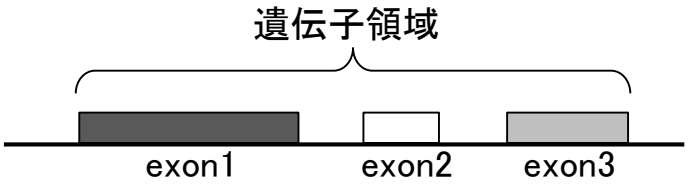
解析結果のイメージ

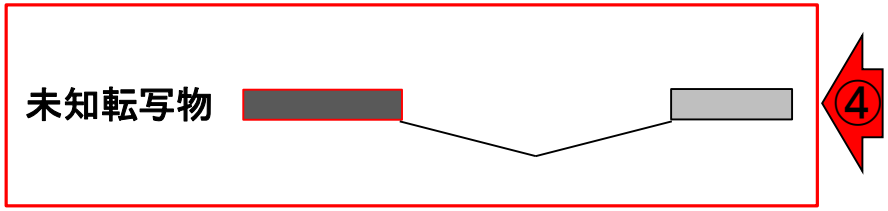
①RNA-seqデータ、②ゲノム配列情報、③アノテーション情報を利用して、④未知転写物（新規isoform）の同定ができる。**⑤転写物の発現量（働いている度合い）推定も原理的に可能**

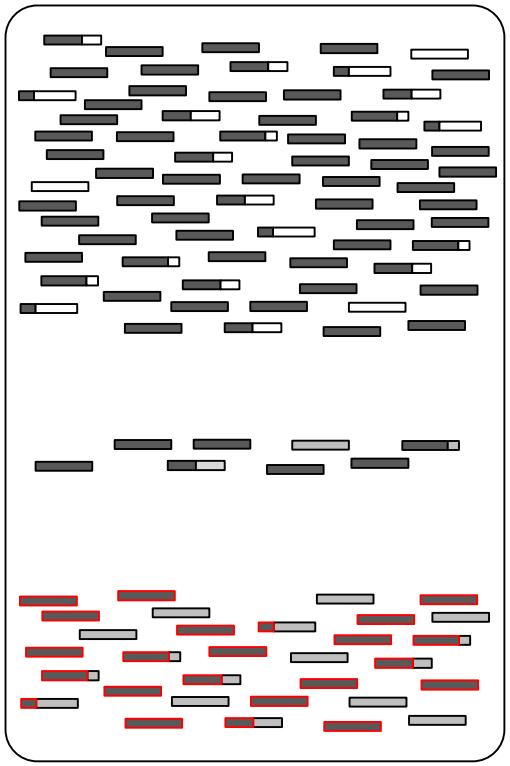


具体的な戦略は?

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。



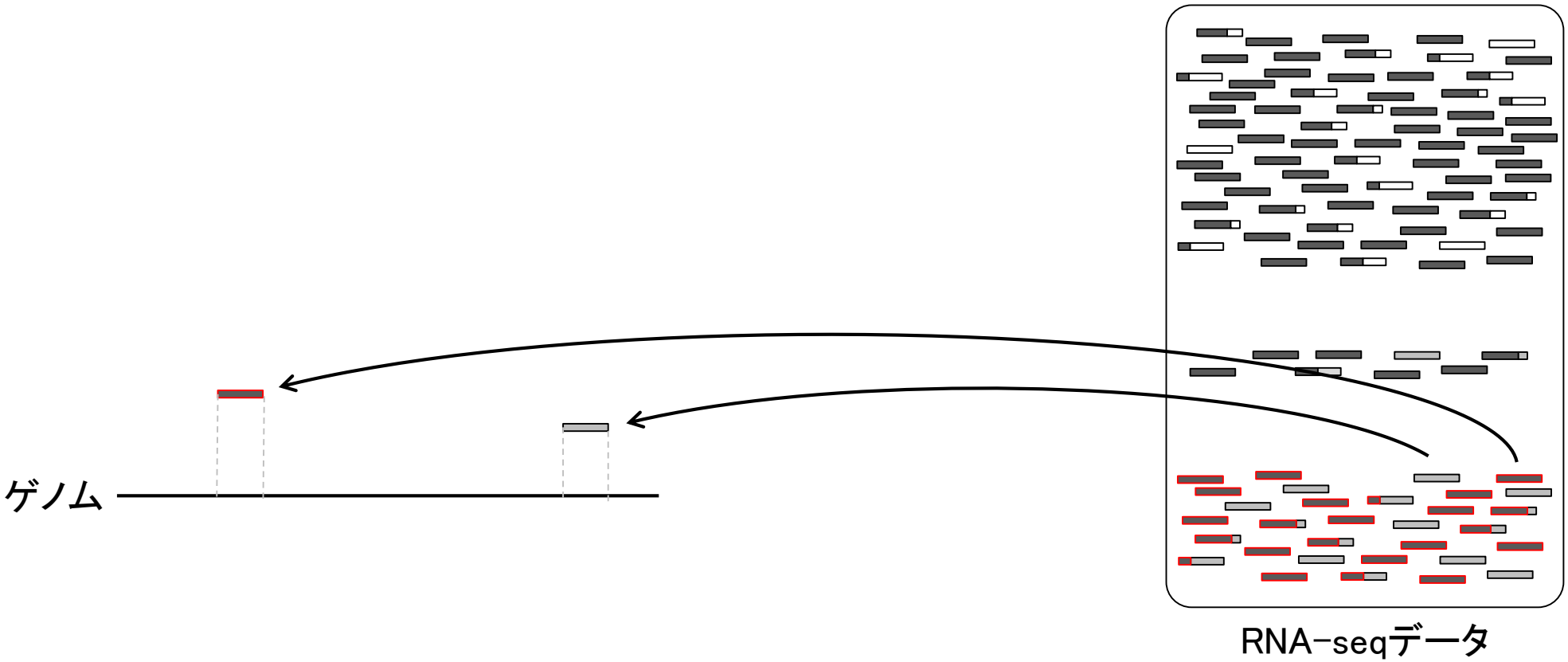




RNA-seqデータ

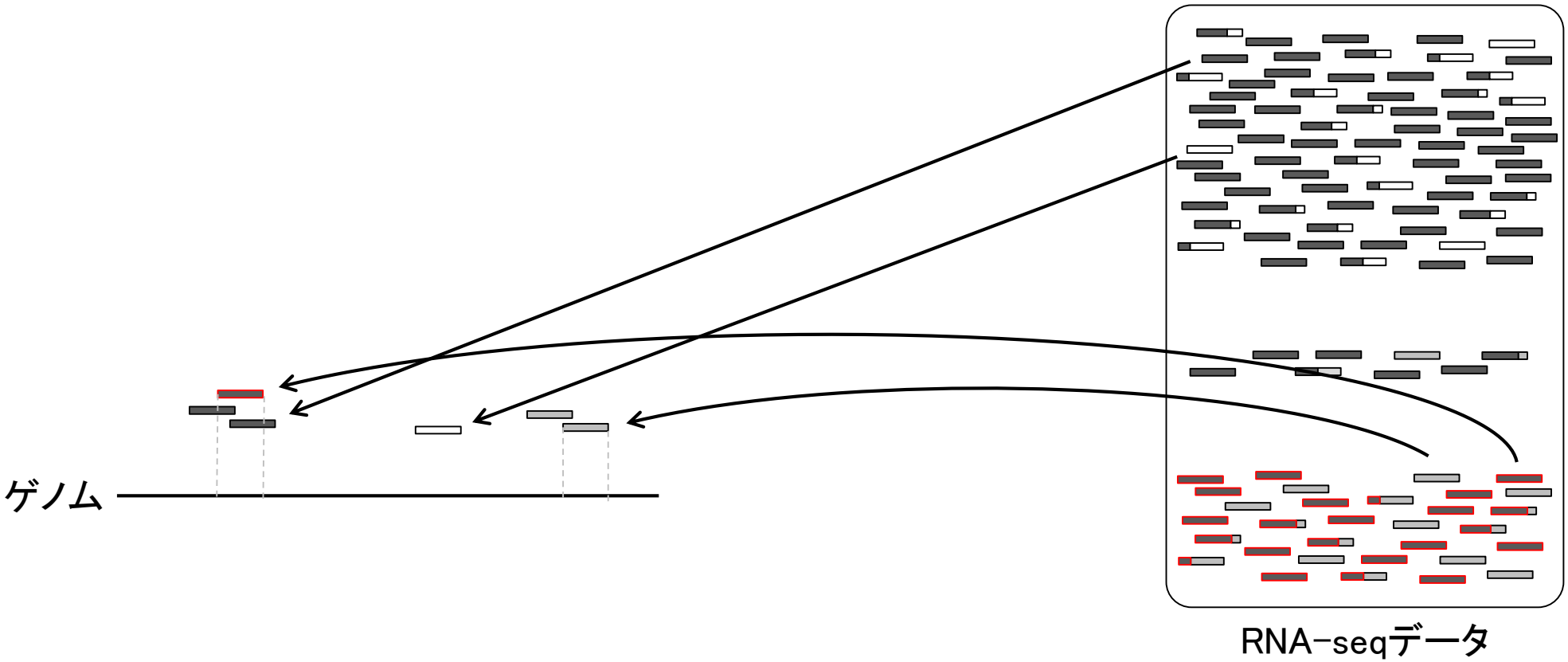
具体的な戦略

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



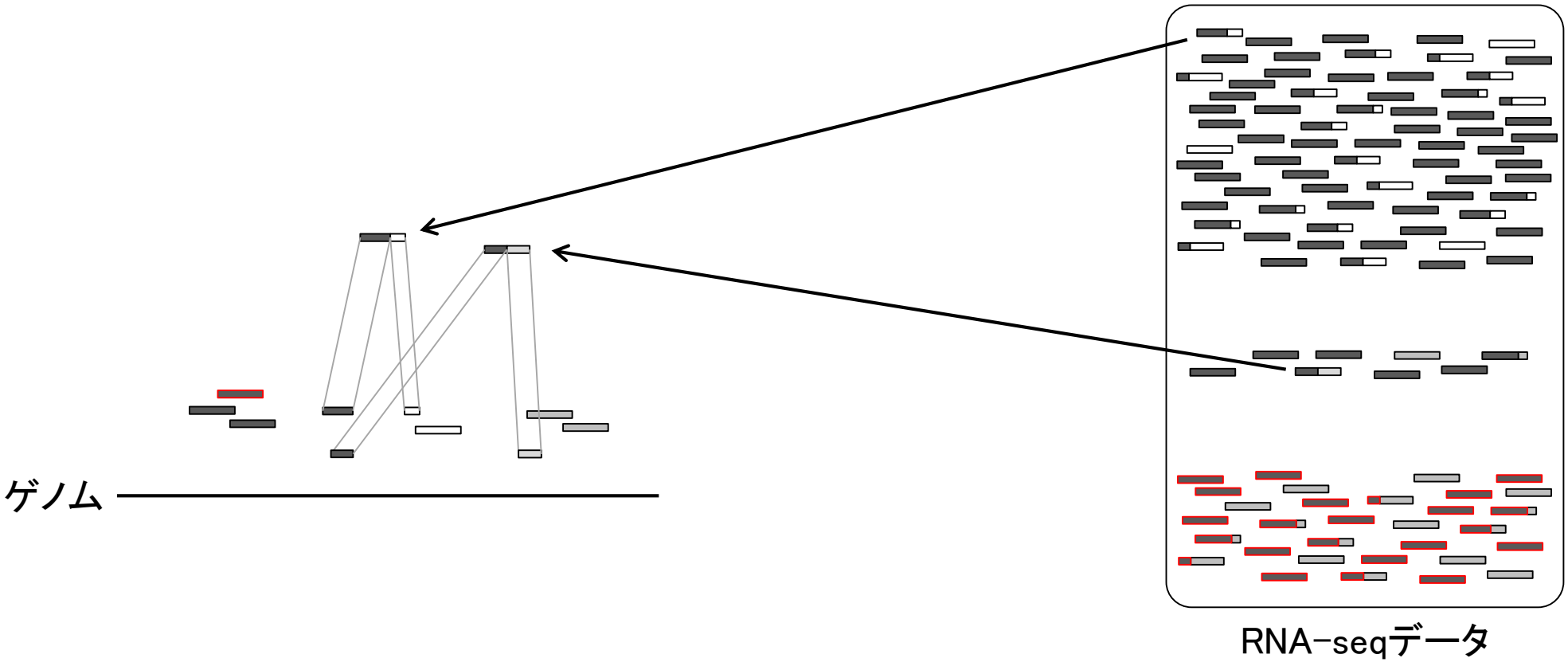
具体的な戦略

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<u>マッピング</u>という作業に相当する



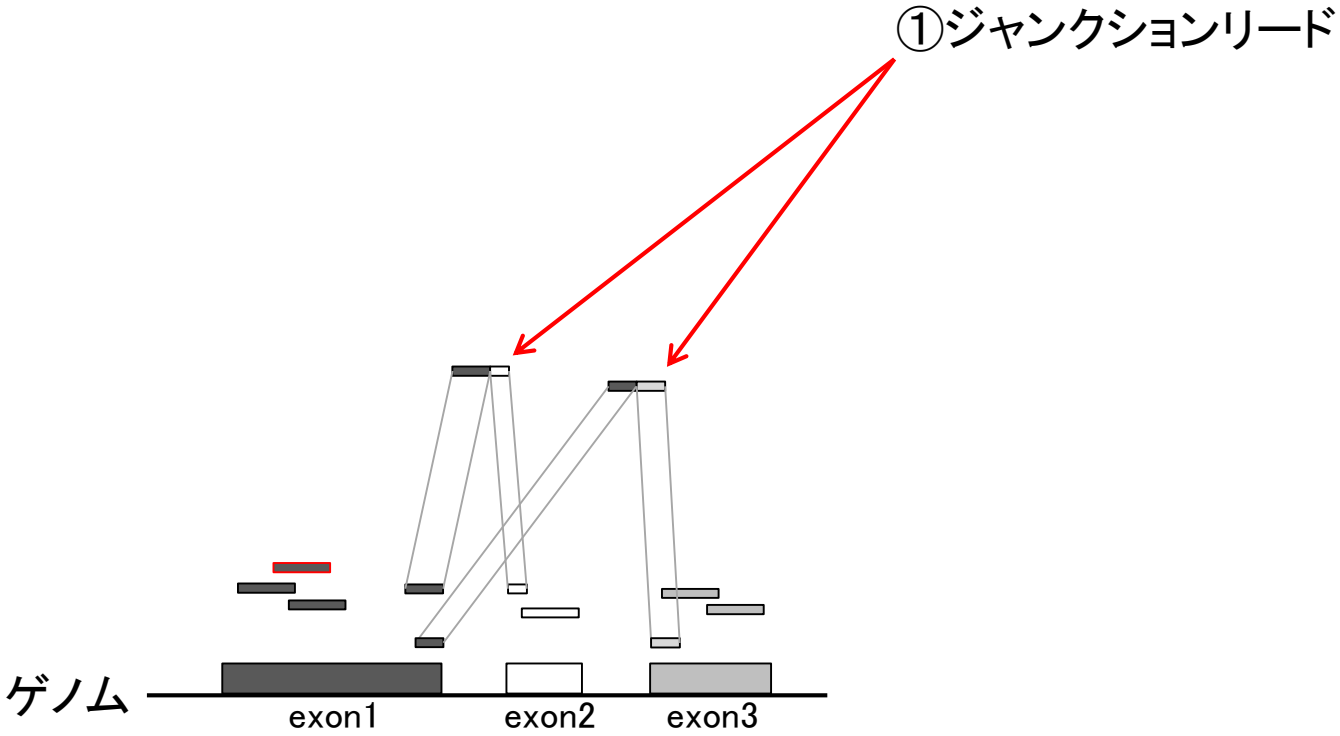
具体的な戦略

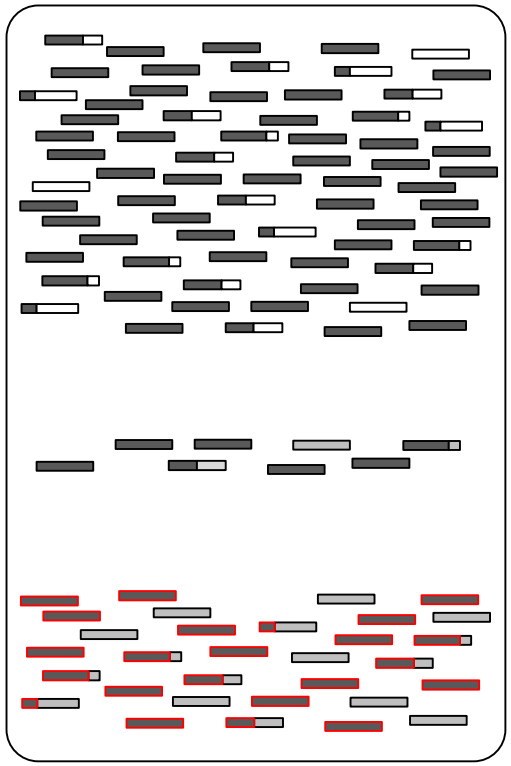
リードの長さが初期は35塩基程度だったが、現在は150塩基程度まで伸びている。そのおかげで、リードを分割してマップすることもできる



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

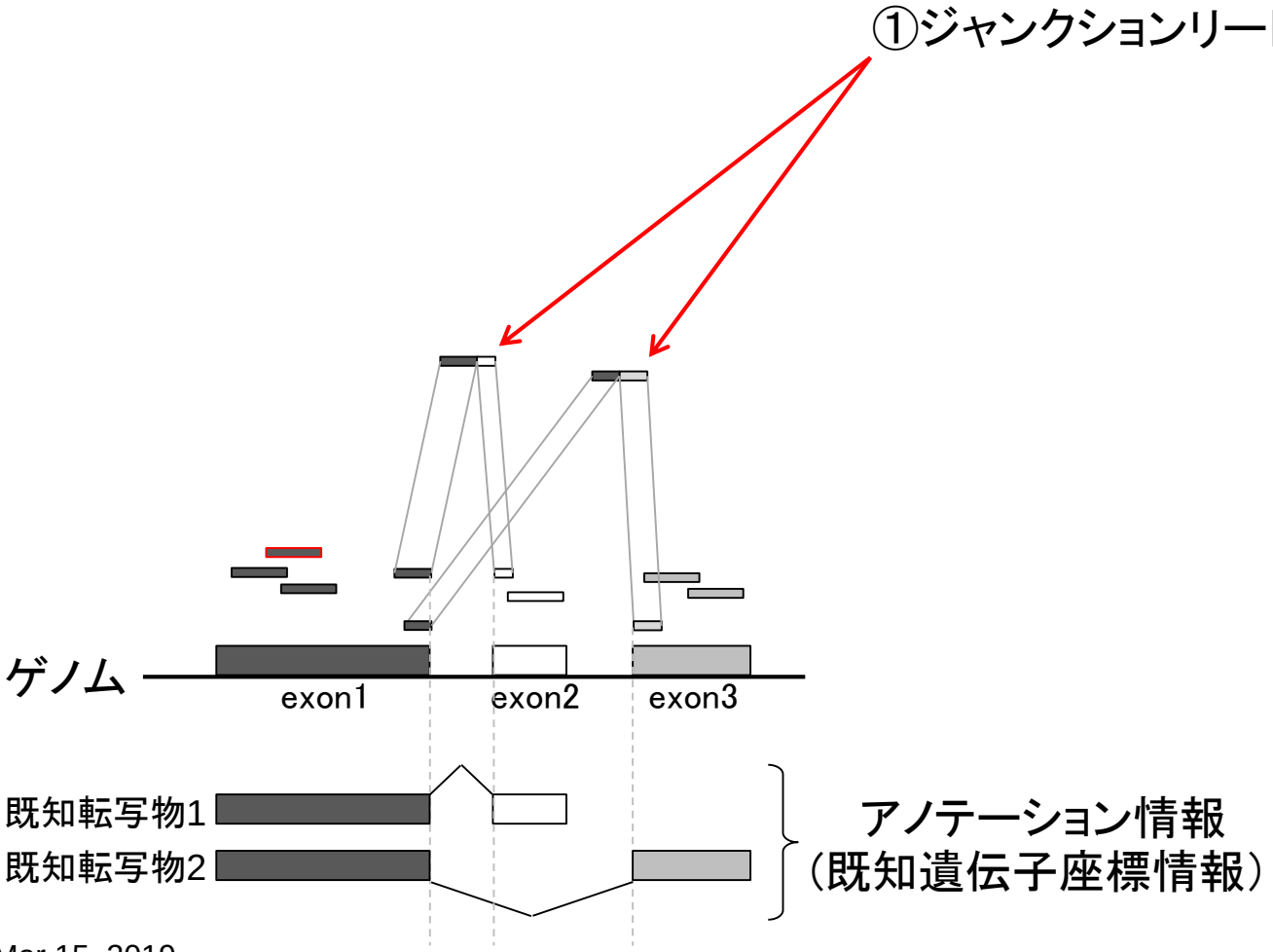


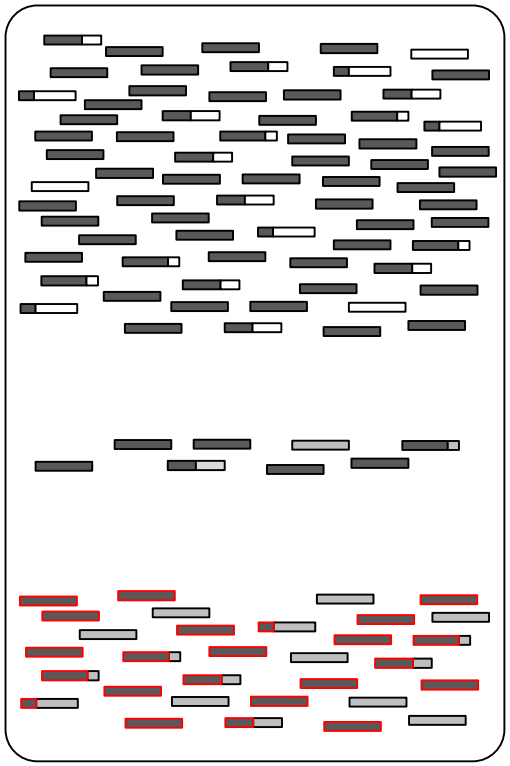


RNA-seqデータ

具体的な戦略

既知遺伝子（転写物）の座標情報と比較することで、答え合わせも可能

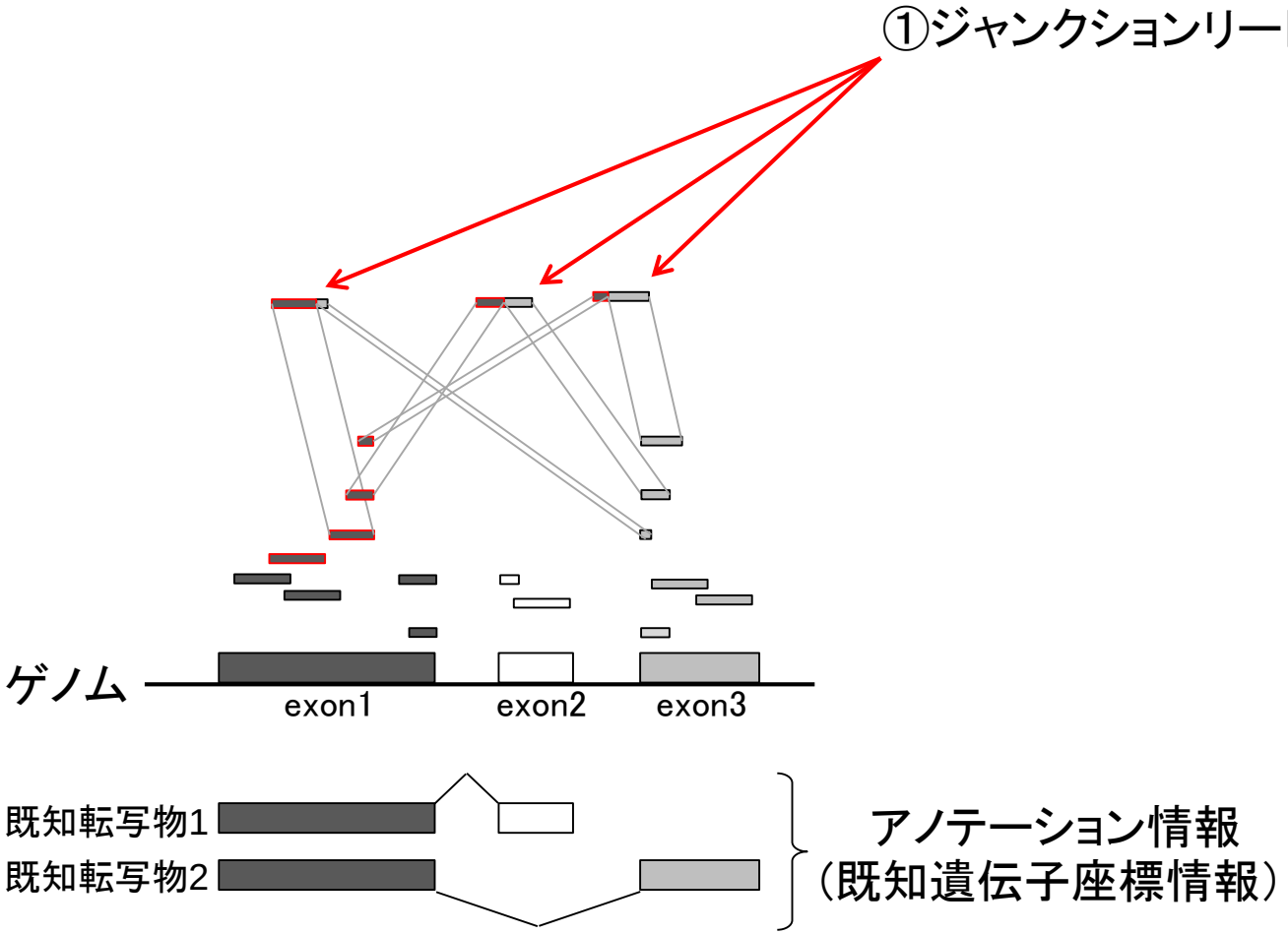


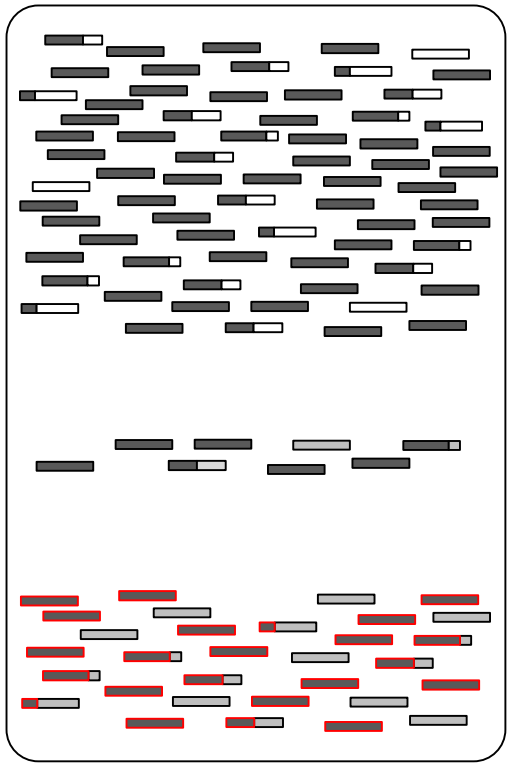


RNA-seqデータ

具体的な戦略

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…

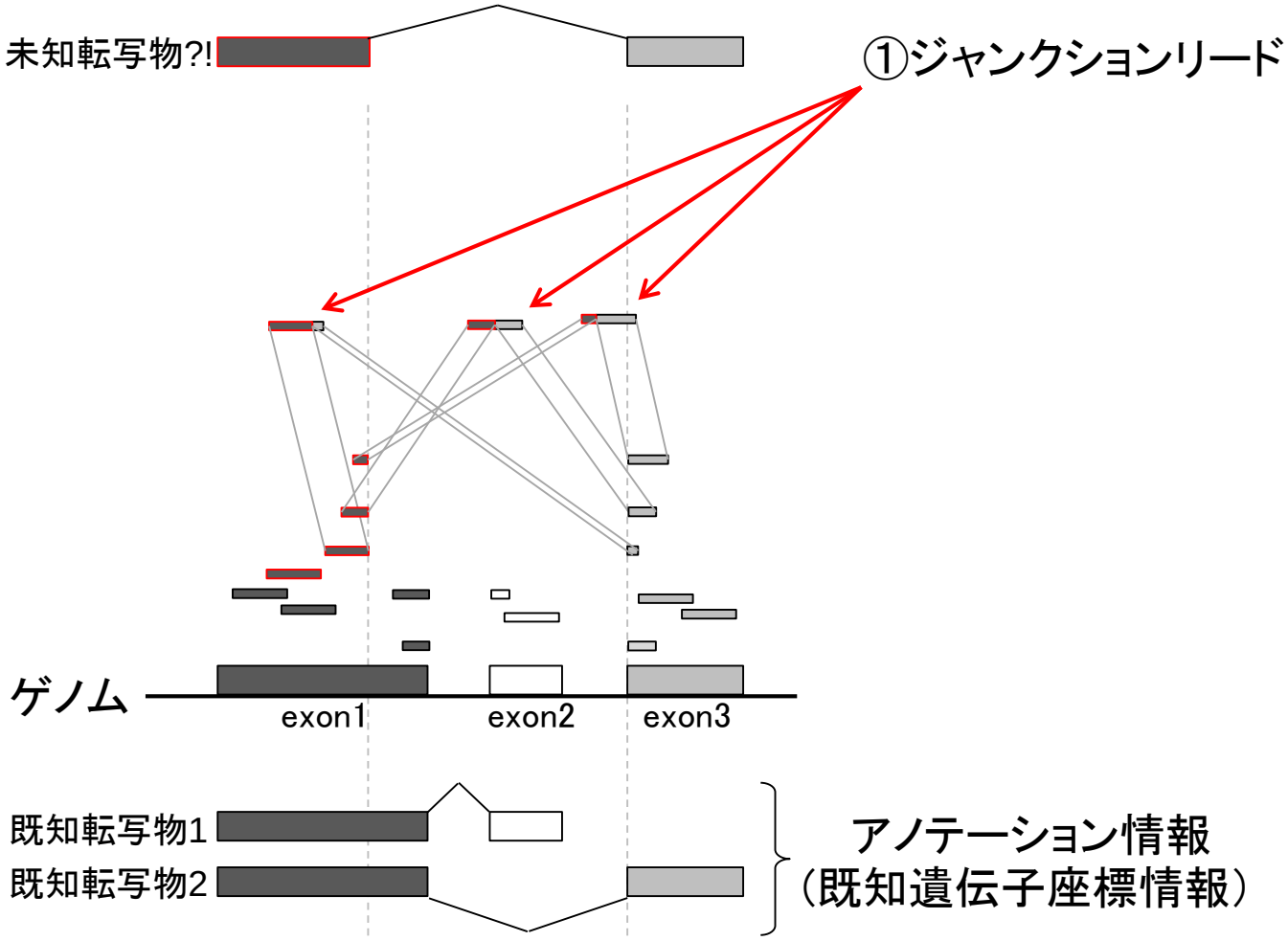


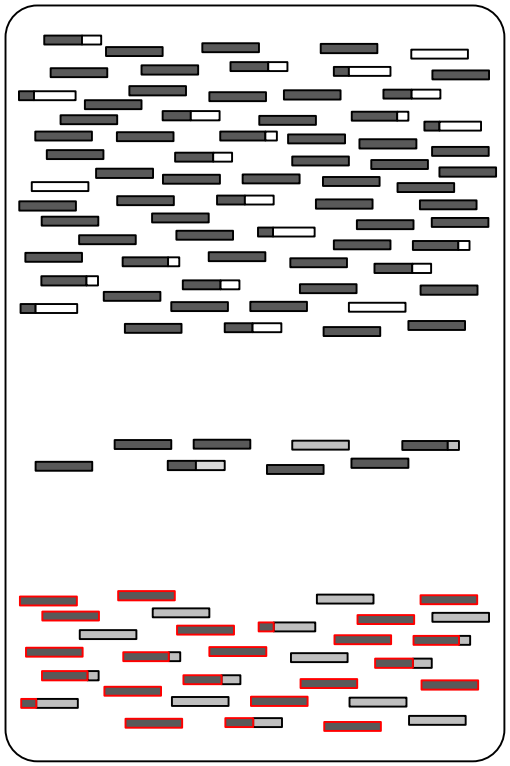


RNA-seqデータ

具体的な戦略

未知転写物（新規isoform）の同定も原理的に可能





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

https://www.ncbi.nlm.nih.gov/pubmed/25650807

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed Advanced

NCBI will be testing https on public web servers from 1:00-4:00 PM EDT (17:00-20:00 UTC) on Monday, October 24. You may experience problems with NCBI services, especially file downloads, during that time. Please plan accordingly. Read more.

Format: Abstract Send to

Genome Biol. 2015 Jan 5;16:7. doi: 10.1186/s13059-014-0558-0.

Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data.

①

Fernandez-Cuesta L, Sun R, Menon R, George J, Lorenz S, Meza-Zepeda LA, Peifer M, Plenker D, Heuckmann JM, Leenders F, Zander T, Dahmen I, Koker M, Schöttle J, Ullrich RT, Altmüller J, Becker C, Nürnberg P, Seidel H, Böhm D, Göke F, Ansén S, Russell PA, Wright GM, Wainer Z, Solomon B, Petersen I, Clement JH, Sänger J, Brustugun OT, Helland Å, Solberg S, Lund-Iversen M, Buettner R, Wolf J, Brambilla E, Vingron M, Perner S, Haas SA, Thomas RK.

Abstract

Genomic translocation events frequently underlie cancer development through generation of gene fusions with oncogenic properties. Identification of such fusion transcripts by transcriptome sequencing might help to discover new potential therapeutic targets. We developed TRUP (Tumor-specimen suited RNA-seq Unified Pipeline) (https://github.com/ruping/TRUP), a computational approach that combines split-read and read-pair analysis with de novo assembly for the identification of chimeric transcripts in cancer specimens. We apply TRUP to RNA-seq data of different tumor types, and find it to be more sensitive than alternative tools in detecting chimeric transcripts, such as secondary rearrangements in EML4-ALK-positive lung tumors, or recurrent inactivating rearrangements affecting RASSF8.

PMID: 25650807 PMCID: PMC4300615 DOI: 10.1186/s13059-014-0558-0

Full text links

Read free full text at BioMed Central

PMC FREE Full text

Save items

Add to Favorites

Similar articles

A transforming KIF5B and RET gene fusion in lung ac [Genome Res. 2012]

Sensitive and specific detection of EML4-ALK rearr [Lung Cancer. 2014]

Exon array profiling detects EML4-ALK fusion ir [Mol Cancer Res. 2009]

Review Sequential combination of

qデータ

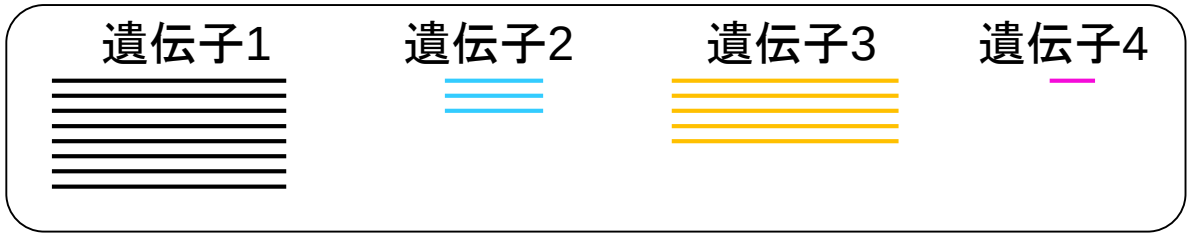
Contents

- 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

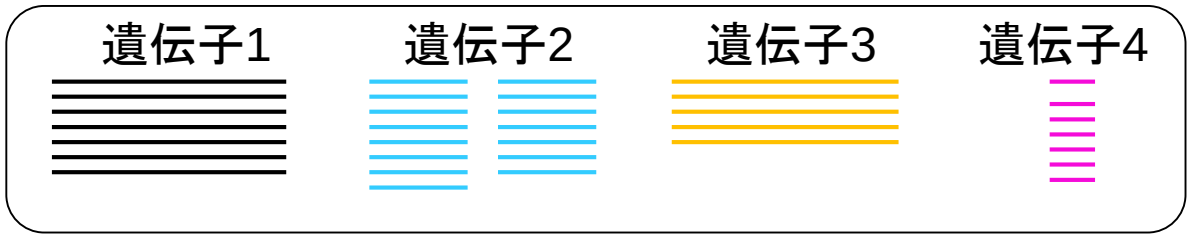
公共カウントデータ

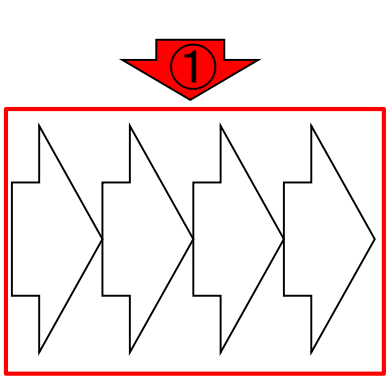
現実問題として、①のところ（データ取得、マッピング、マップされたリード数のカウント）が大変。しかし、公共RNA-seq生データを②の状態にまで行ったものを提供してくれているサイトがあります。それが…

- 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



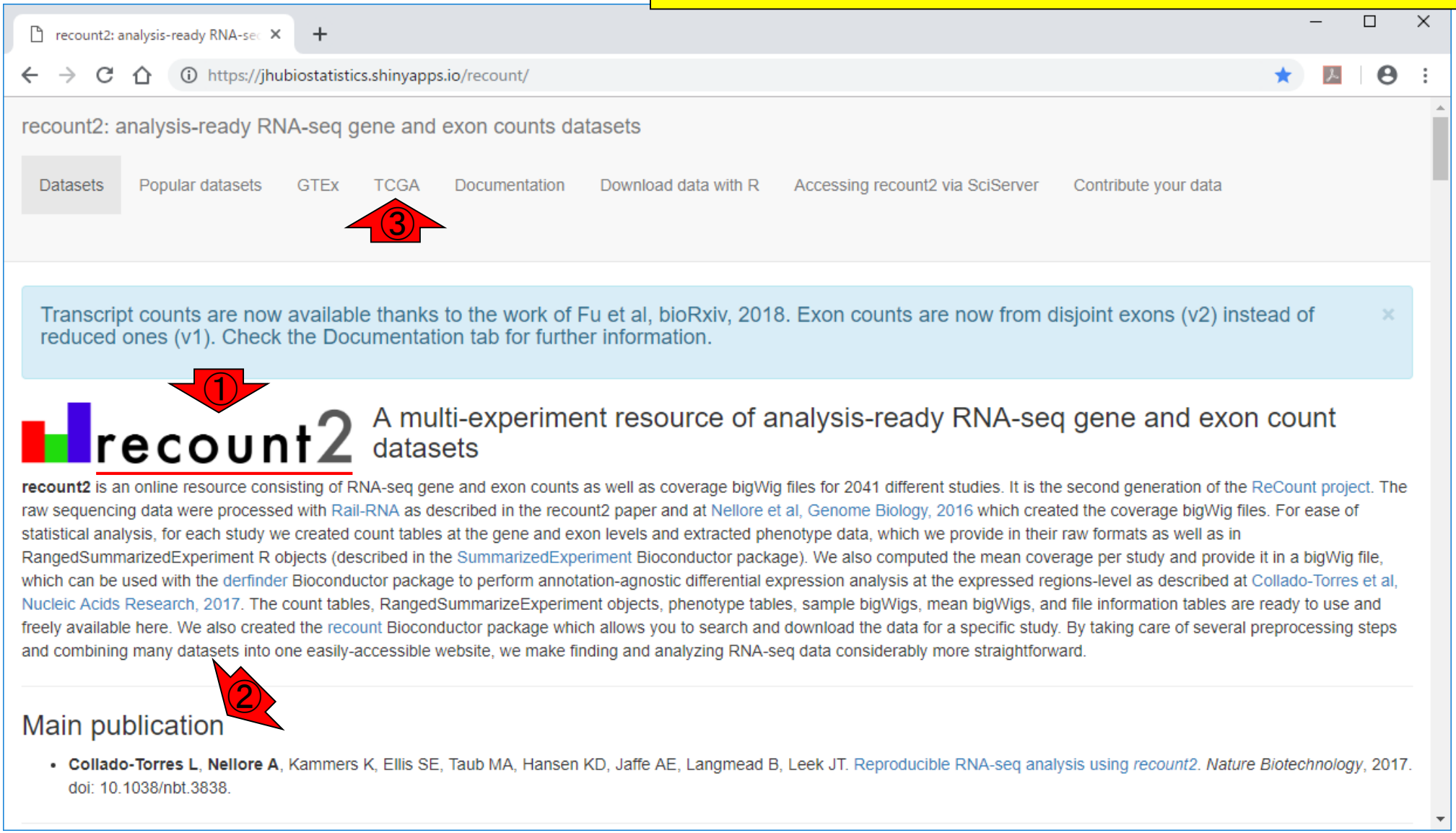


②

	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
…	…	…

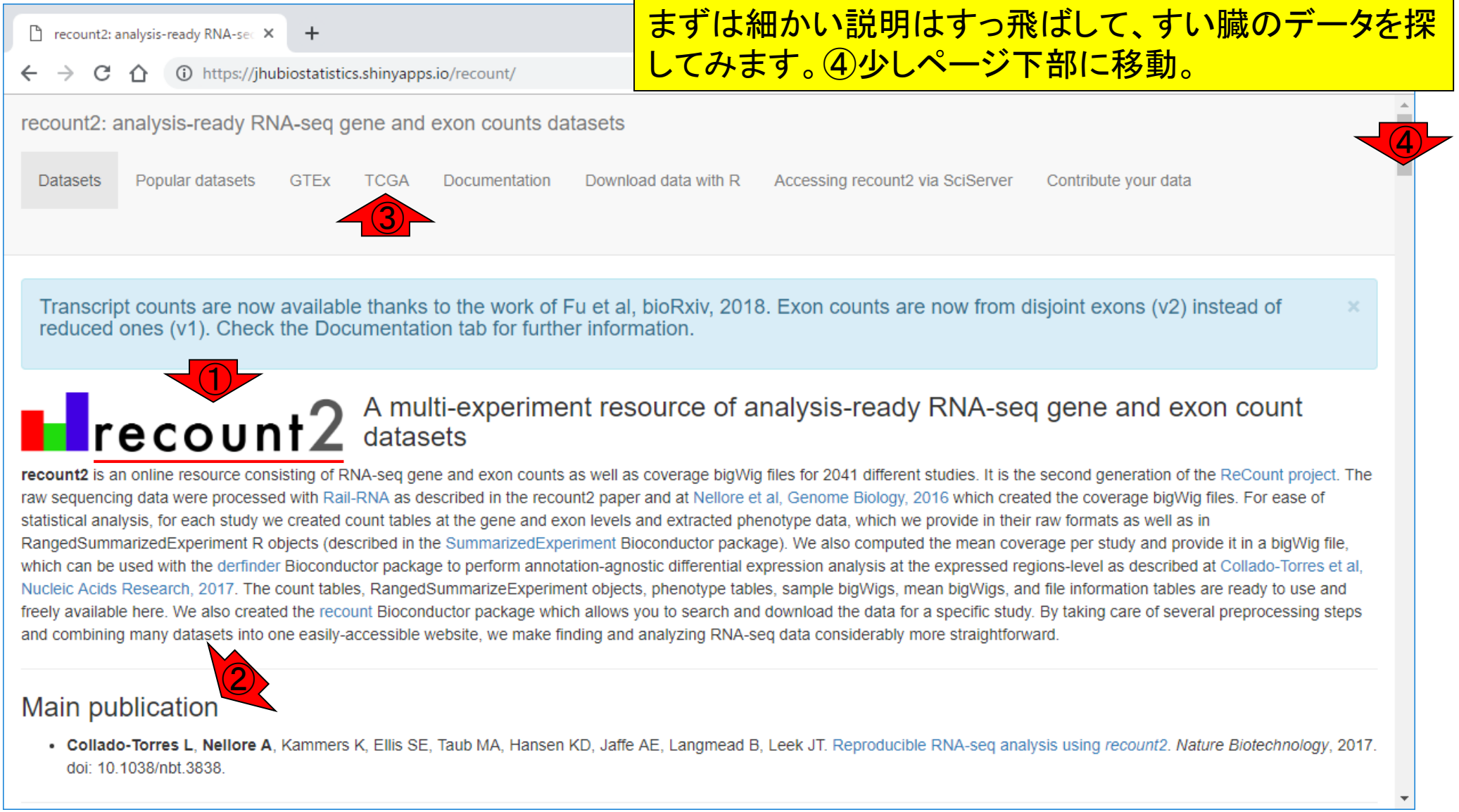
recount2

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。



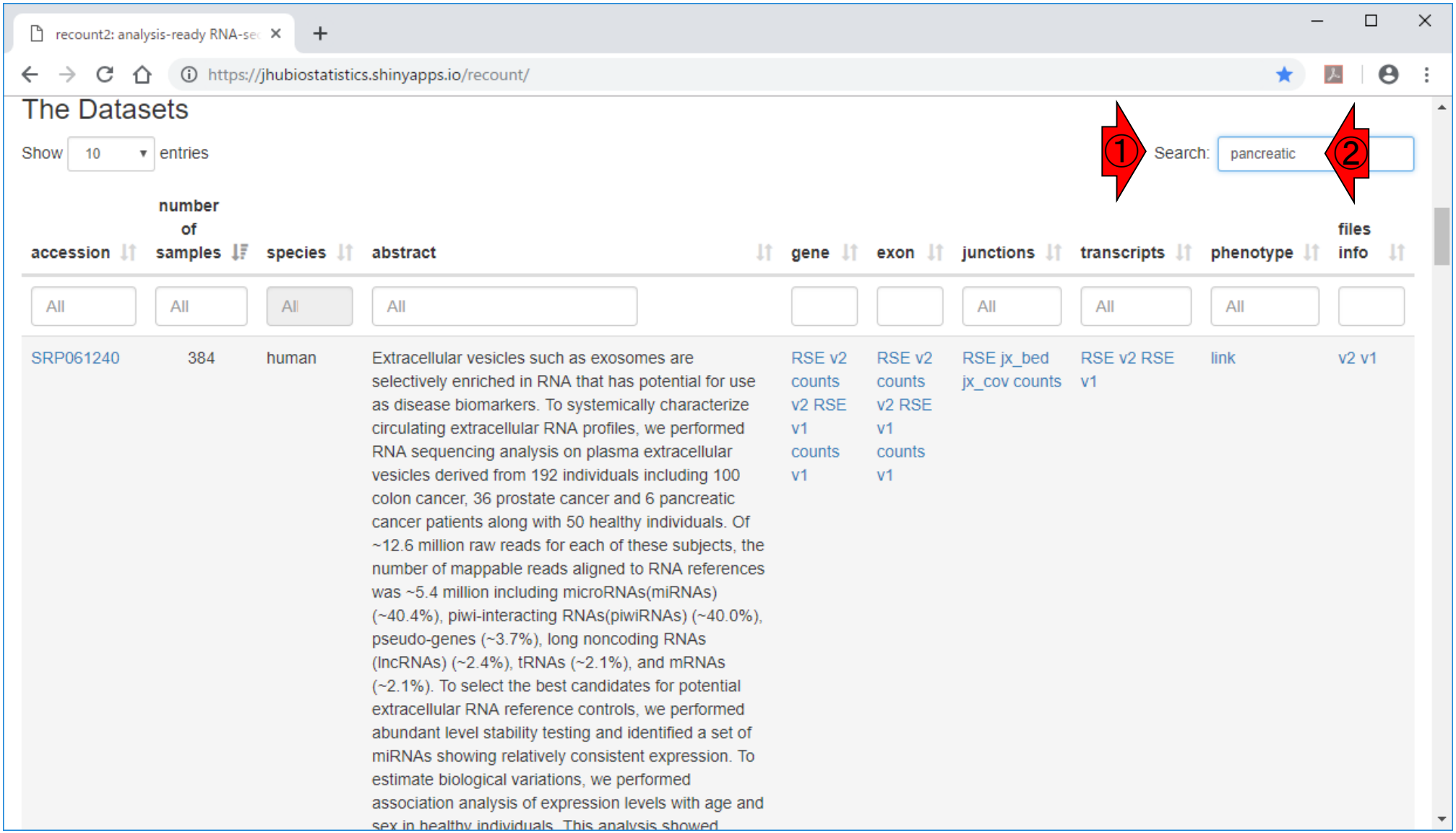
recount2

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。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

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。



recount2

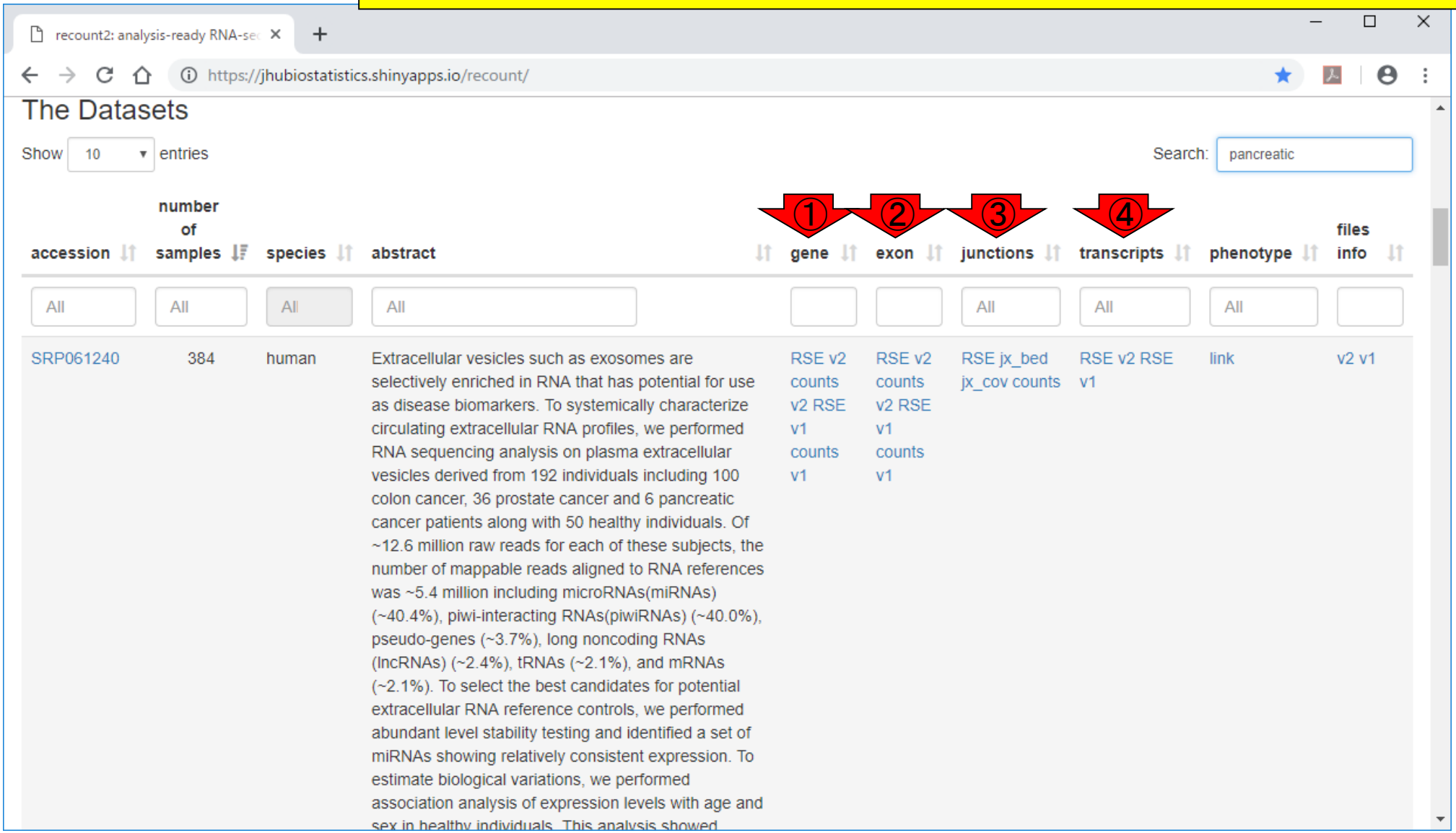
①Searchという検索窓が見えたら、②pancreaticなどそれっぽいキーワードを打ち込んでみる。そうすると、打ち込んでいるそばから赤枠内がごにょごにょ変化する。よく見ると、③打ち込んだキーワードを含むデータセットが表示されていることがわかる。

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。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

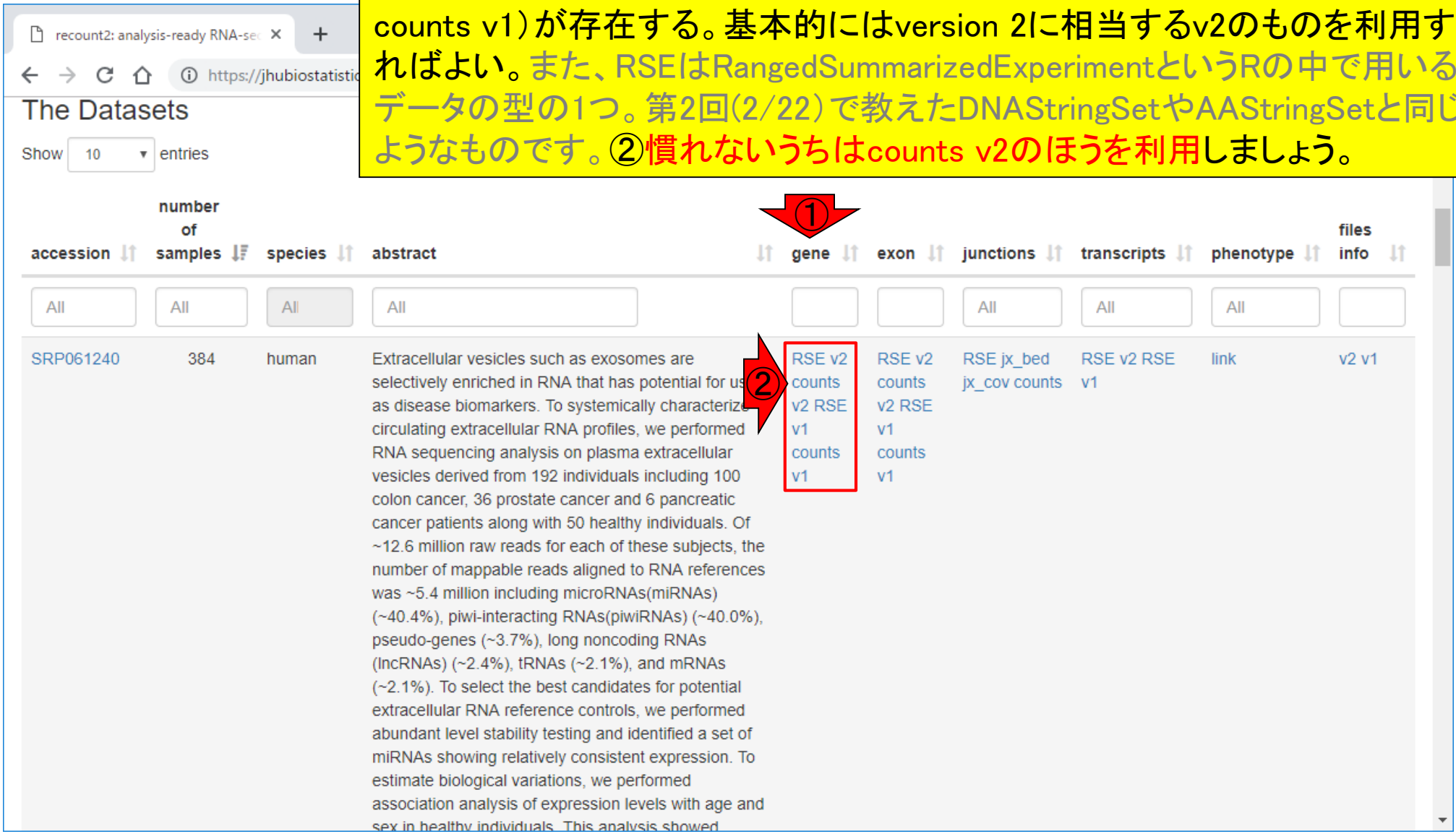
リードのマッピング後にどこの領域上にマップされたリードをカウントするかによって、①～④様々なバリエーションが存在する。通常利用は①geneだと思われます。



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

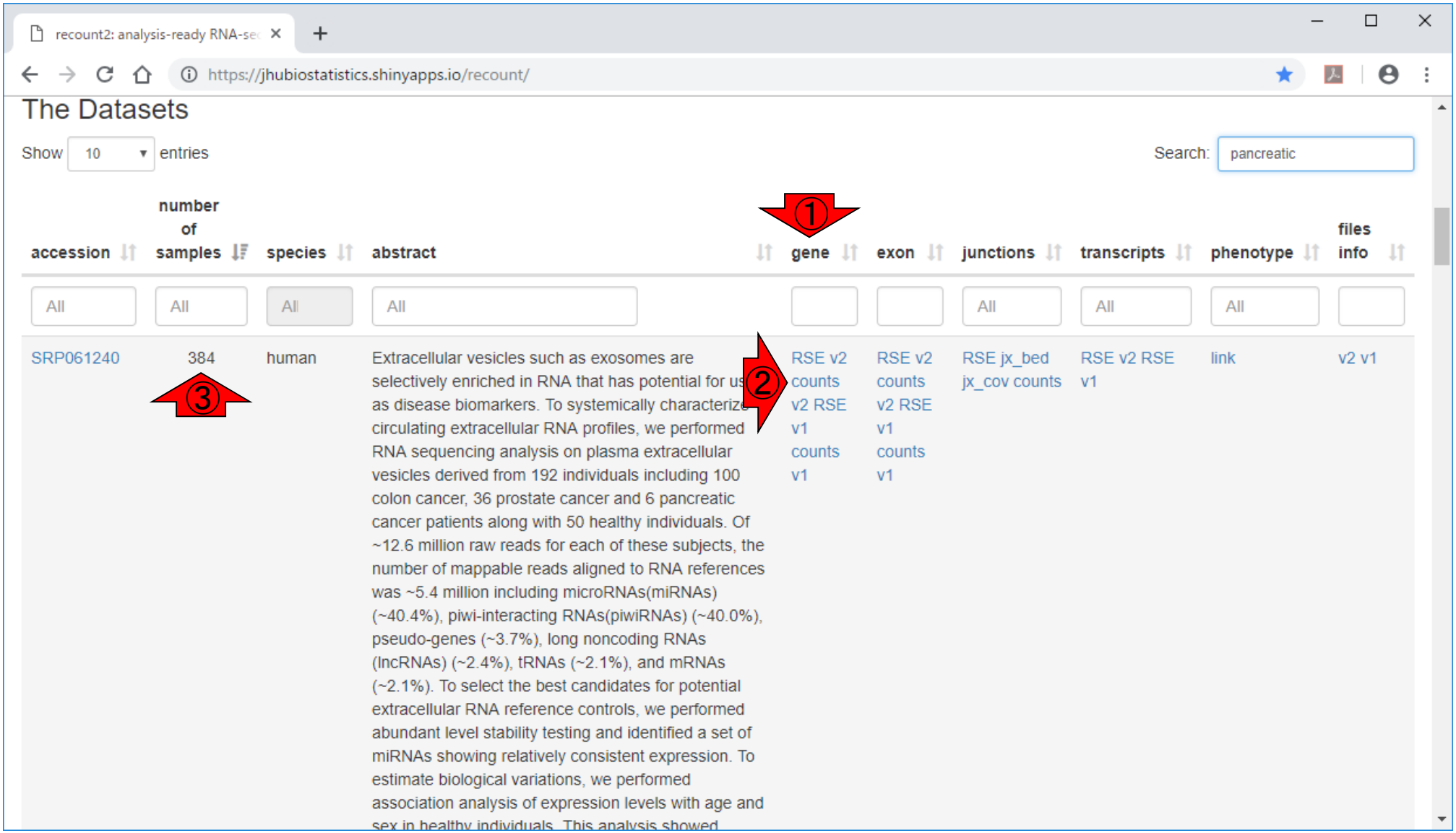
リードのマッピング後にどこの領域上にマップされたリードをカウントするかによって、①～④様々なバリエーションが存在する。通常利用は①geneだと思われます。赤枠内には、4つのリンク先（RSE v2, counts v2, RSE v1, and counts v1）が存在する。基本的にはversion 2に相当するv2のものを利用すればよい。また、RSEはRangedSummarizedExperimentというRの中で用いるデータの型の1つ。第2回(2/22)で教えたDNAStringSetやAAStringSetと同じようなものです。②慣れないうちはcounts v2のほうを利用しましょう。



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 us 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

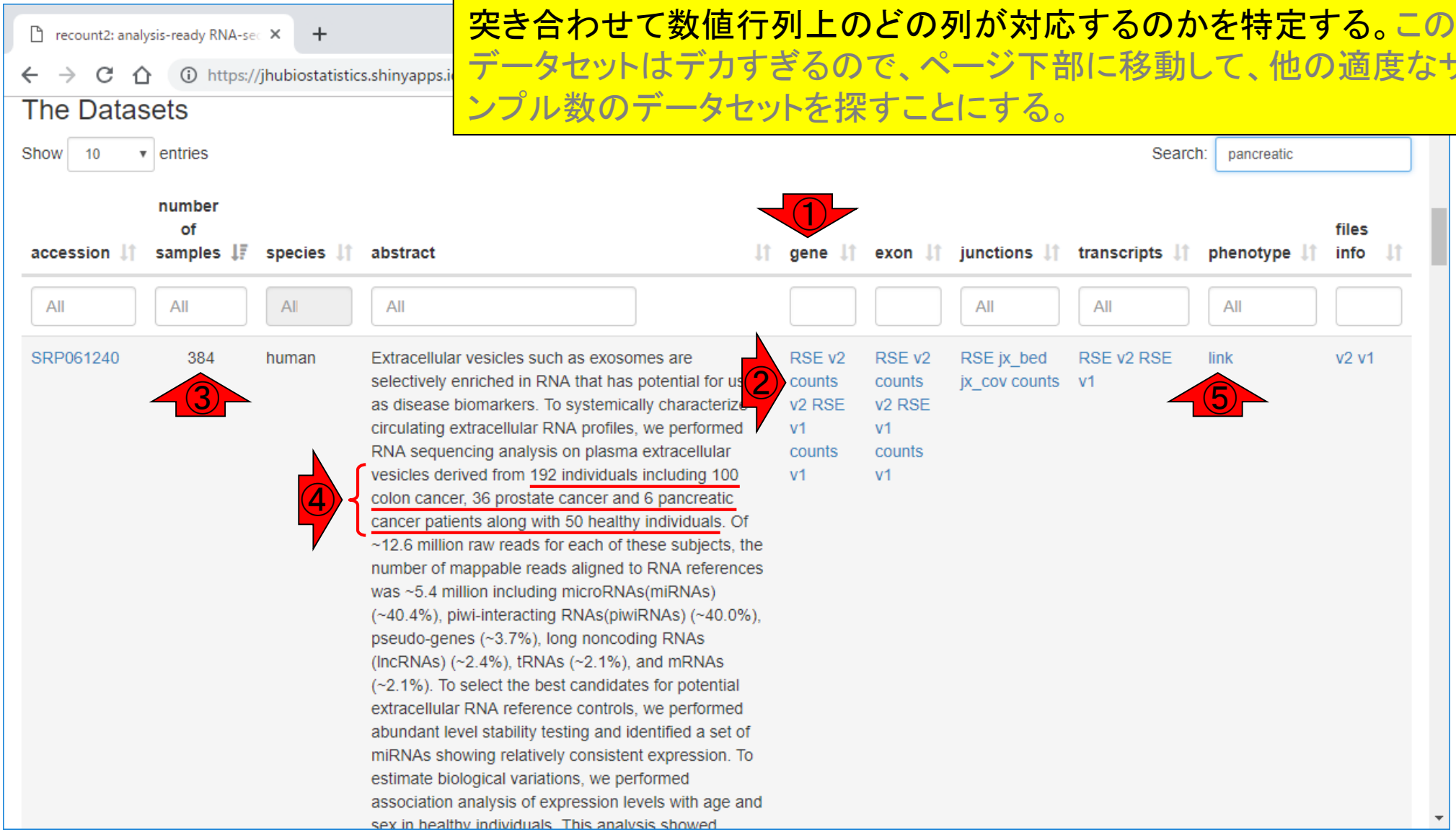
recount2

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。



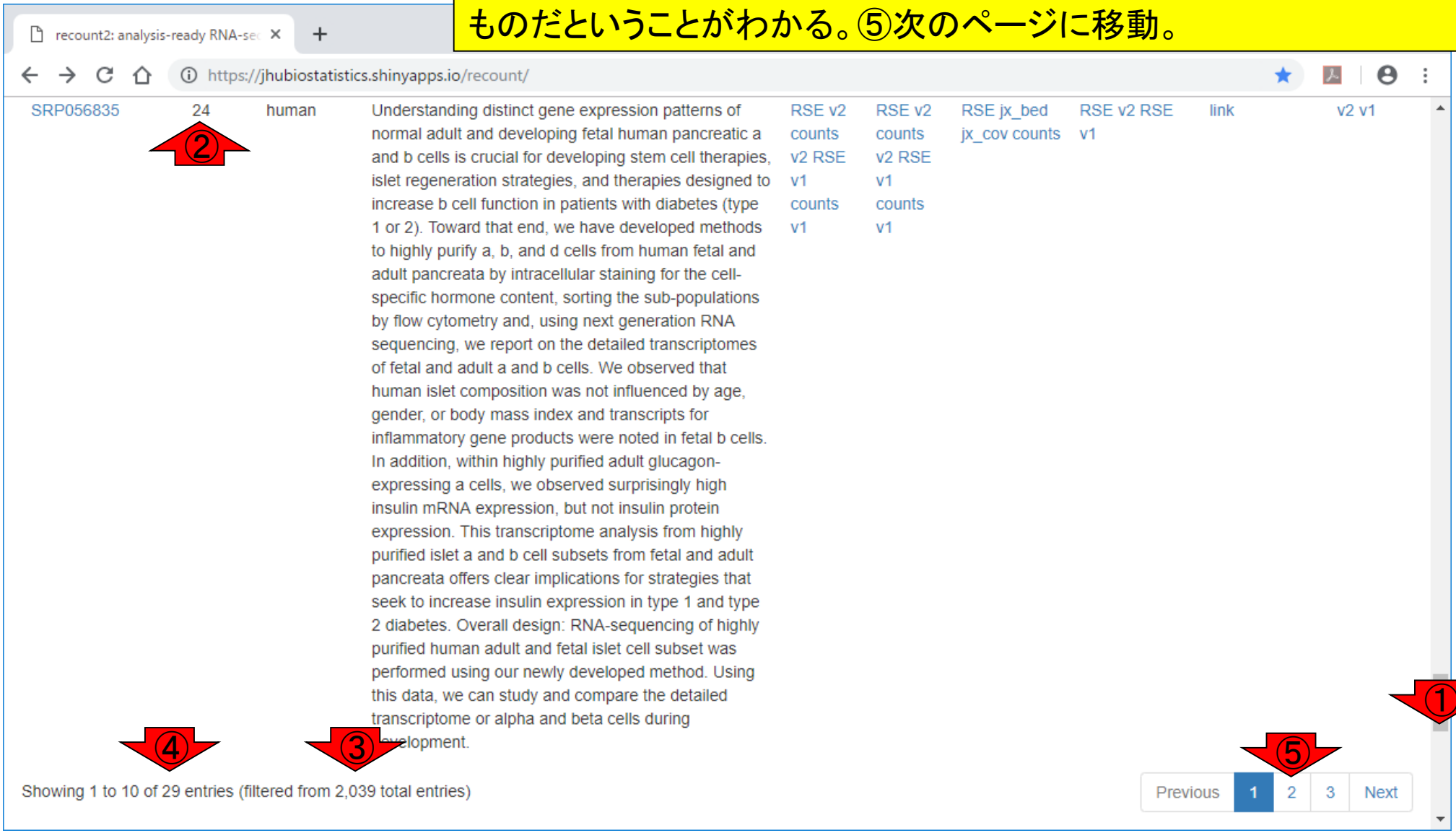
recount2

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。



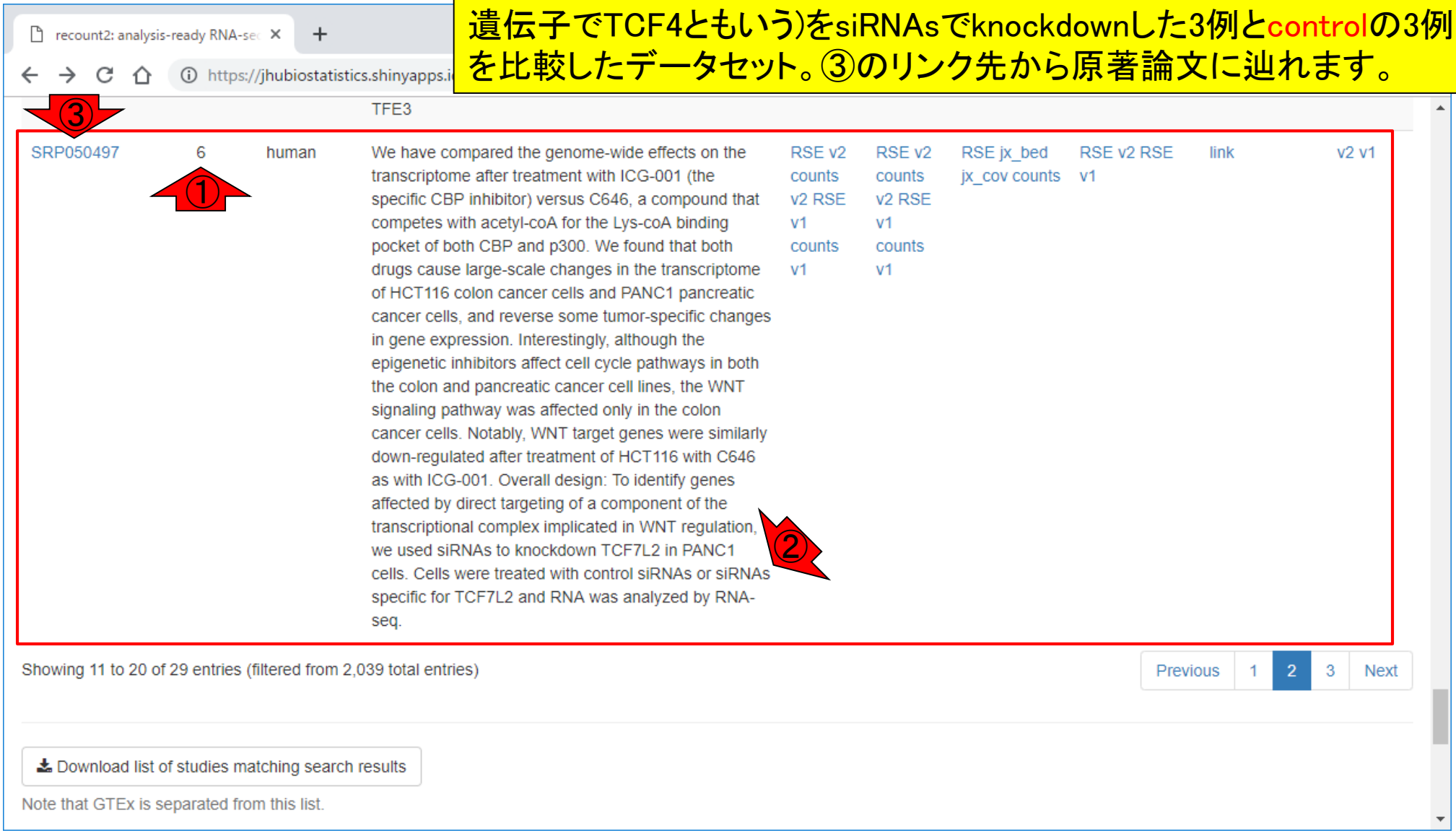
recount2

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

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。



さきほどの③をクリックして辿った、SRP050497の原著論文です。

recount2

参考

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

Altering cancer transcriptomes using epigenomic inhibitors.

Gaddis M[1], Gerrard D[2], Frietze S[2], Farnham PJ[1].

Author information

Abstract

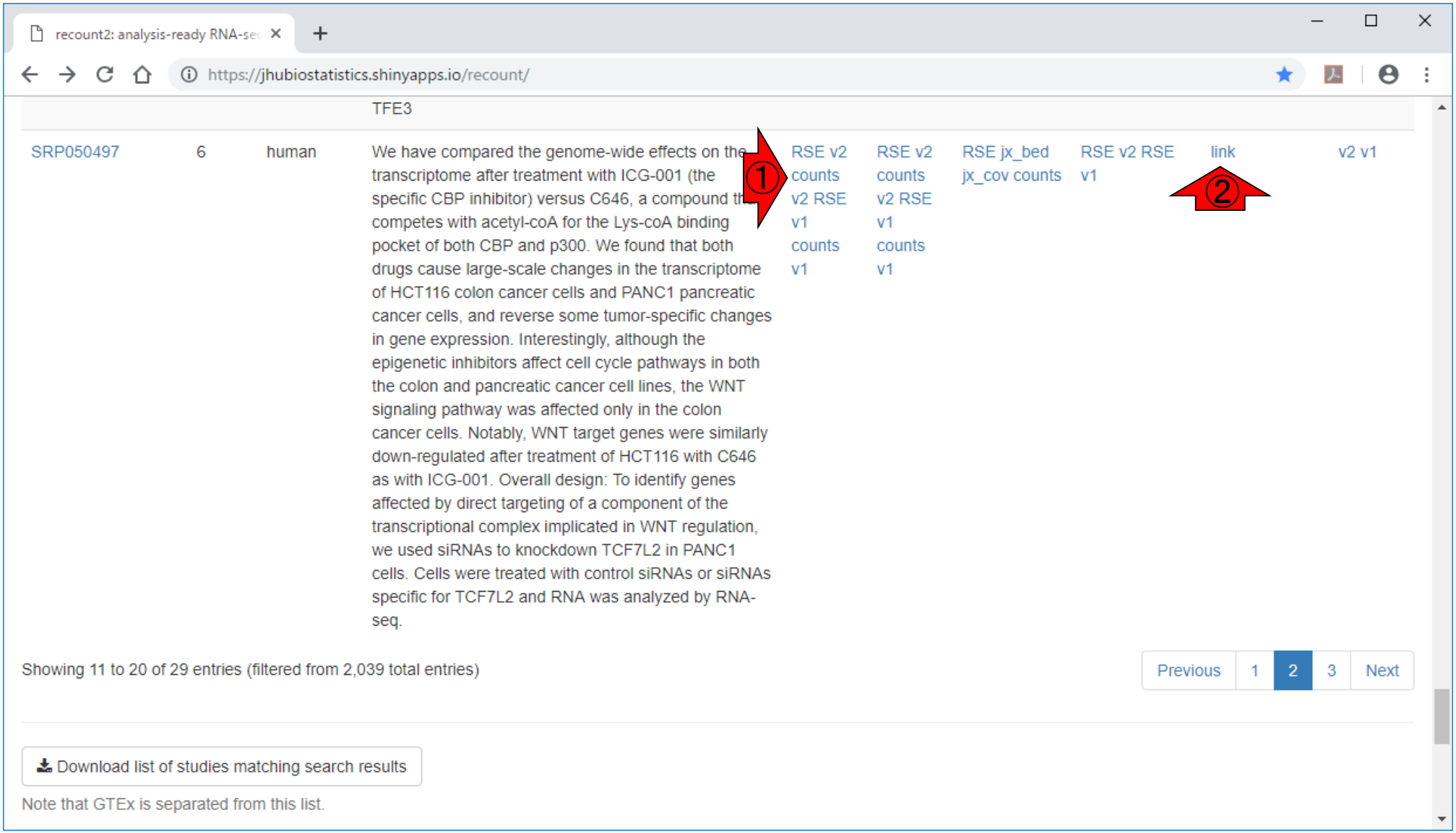
BACKGROUND: Due to the hyper-activation of WNT signaling in a variety of cancer types, there has been a strong drive to develop pathway-specific inhibitors with the eventual goal of providing a chemotherapeutic antagonist of WNT signaling to cancer patients. A new category of drugs, called epigenetic 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。



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。

①

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	ENSG00000000419.12
24666	22830	37528	31189	31249	36804	ENSG00000000457.13
42199	38508	58545	36169	39613	42044	ENSG00000000460.16
0	0	0	0	0	0	ENSG00000000938.12
647	628	794	450	650	446	ENSG00000000971.15
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

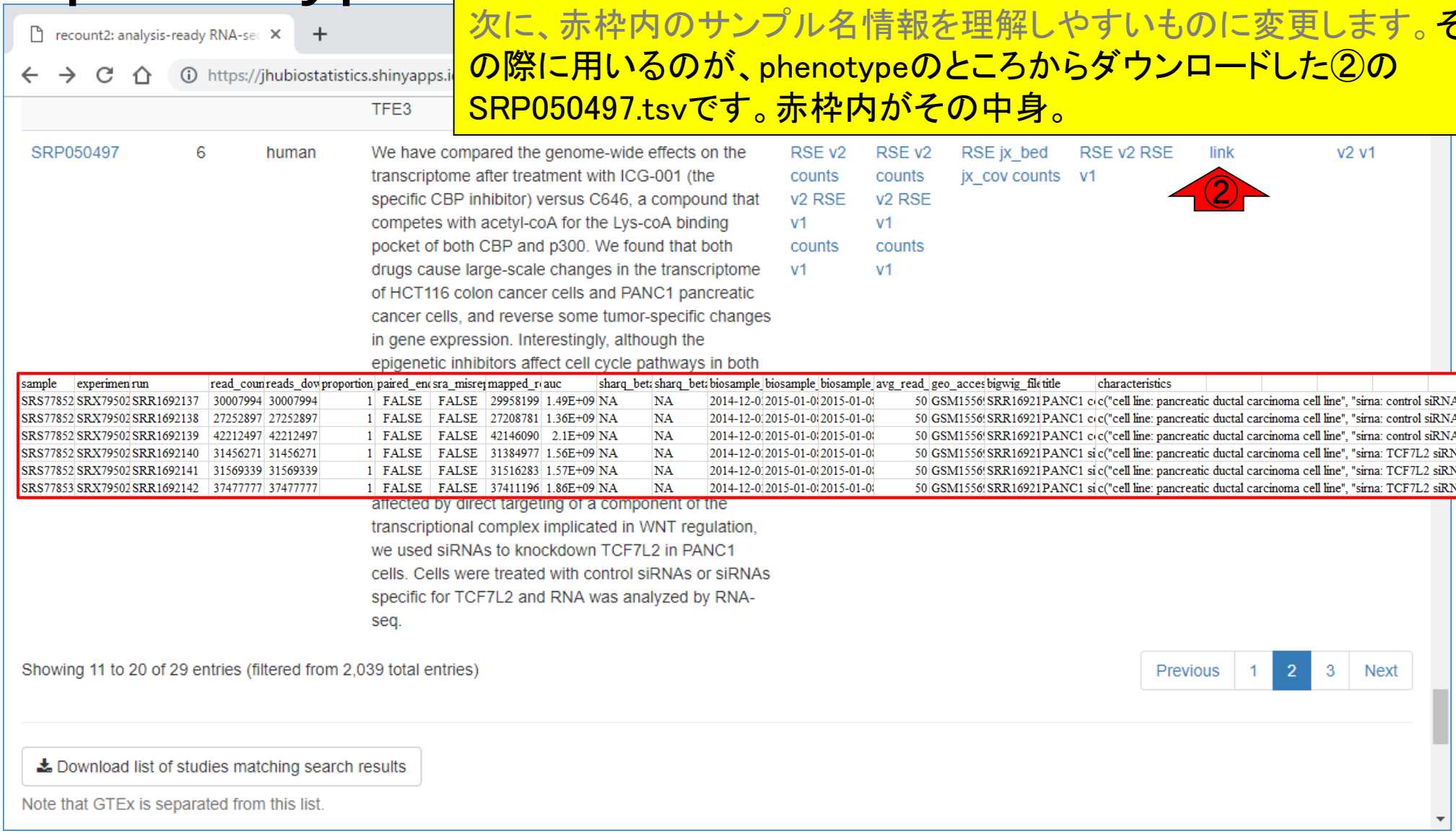
数値行列

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

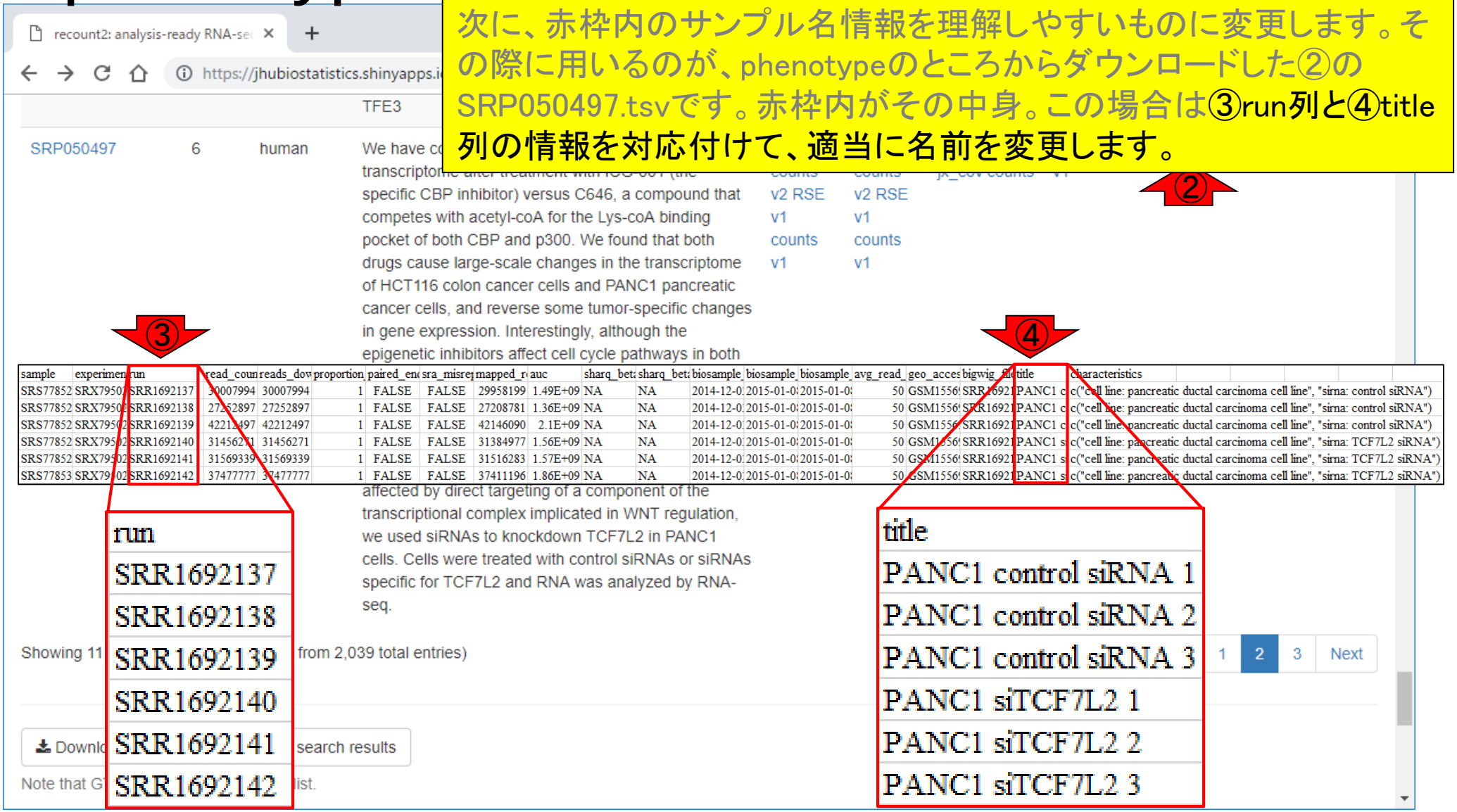
①をダウンロードして解凍すると、counts_gene.tsvが得られます。また、②からSRP050497.tsvが得られます。counts_gene.tsvの中身。①行名情報の列が一番右側になっているので、一番左側に移動させます。次に、赤枠内のサンプル名情報を理解しやすいものに変更します。その際に用いるのが、phenotypeのところからダウンロードした②のSRP050497.tsvです。赤枠内がその中身。



sample	experimen	run	read_coun	reads_dow	proportion	paired_en	sra_misre	mapped_r	auc	sharq_bet	sharq_bet	biosample_	biosample_	biosample_	avg_read_	geo_acces	bigwig_file	title	characteristics
SRS77852	SRX79502	SRR1692137	30007994	30007994	1	FALSE	FALSE	29958199	1.49E+09	NA	NA	2014-12-0	2015-01-0	2015-01-0	50	GSM1556	SRR16921	PANC1 c	c("cell line: pancreatic ductal carcinoma cell line", "sirna: control siRNA")
SRS77852	SRX79502	SRR1692138	27252897	27252897	1	FALSE	FALSE	27208781	1.36E+09	NA	NA	2014-12-0	2015-01-0	2015-01-0	50	GSM1556	SRR16921	PANC1 c	c("cell line: pancreatic ductal carcinoma cell line", "sirna: control siRNA")
SRS77852	SRX79502	SRR1692139	42212497	42212497	1	FALSE	FALSE	42146090	2.1E+09	NA	NA	2014-12-0	2015-01-0	2015-01-0	50	GSM1556	SRR16921	PANC1 c	c("cell line: pancreatic ductal carcinoma cell line", "sirna: control siRNA")
SRS77852	SRX79502	SRR1692140	31456271	31456271	1	FALSE	FALSE	31384977	1.56E+09	NA	NA	2014-12-0	2015-01-0	2015-01-0	50	GSM1556	SRR16921	PANC1 si	c("cell line: pancreatic ductal carcinoma cell line", "sirna: TCF7L2 siRNA")
SRS77852	SRX79502	SRR1692141	31569339	31569339	1	FALSE	FALSE	31516283	1.57E+09	NA	NA	2014-12-0	2015-01-0	2015-01-0	50	GSM1556	SRR16921	PANC1 si	c("cell line: pancreatic ductal carcinoma cell line", "sirna: TCF7L2 siRNA")
SRS77853	SRX79502	SRR1692142	37477777	37477777	1	FALSE	FALSE	37411196	1.86E+09	NA	NA	2014-12-0	2015-01-0	2015-01-0	50	GSM1556	SRR16921	PANC1 si	c("cell line: pancreatic ductal carcinoma cell line", "sirna: TCF7L2 siRNA")

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。



sample	experimen	run	read_coun	reads_dow	proportion	paired_en	sra_misre	mapped_r	auc	sharq_bet	sharq_bet	biosample_	biosample_	biosample_	avg_read_	geo_acces	bigwig_fil	title	characteristics
SRS77852	SRX79502	SRR1692137	30007994	30007994	1	FALSE	FALSE	29958199	1.49E+09	NA	NA	2014-12-0	2015-01-0	2015-01-0	50	GSM1556	SRR16921	PANC1 c	c("cell line: pancreatic ductal carcinoma cell line", "sirna: control siRNA")
SRS77852	SRX79502	SRR1692138	27252897	27252897	1	FALSE	FALSE	27208781	1.36E+09	NA	NA	2014-12-0	2015-01-0	2015-01-0	50	GSM1556	SRR16921	PANC1 c	c("cell line: pancreatic ductal carcinoma cell line", "sirna: control siRNA")
SRS77852	SRX79502	SRR1692139	42212497	42212497	1	FALSE	FALSE	42146090	2.1E+09	NA	NA	2014-12-0	2015-01-0	2015-01-0	50	GSM1556	SRR16921	PANC1 c	c("cell line: pancreatic ductal carcinoma cell line", "sirna: control siRNA")
SRS77852	SRX79502	SRR1692140	31456271	31456271	1	FALSE	FALSE	31384977	1.56E+09	NA	NA	2014-12-0	2015-01-0	2015-01-0	50	GSM1556	SRR16921	PANC1 s	c("cell line: pancreatic ductal carcinoma cell line", "sirna: TCF7L2 siRNA")
SRS77852	SRX79502	SRR1692141	31569339	31569339	1	FALSE	FALSE	31516283	1.57E+09	NA	NA	2014-12-0	2015-01-0	2015-01-0	50	GSM1556	SRR16921	PANC1 s	c("cell line: pancreatic ductal carcinoma cell line", "sirna: TCF7L2 siRNA")
SRS77853	SRX79502	SRR1692142	37477777	37477777	1	FALSE	FALSE	37411196	1.86E+09	NA	NA	2014-12-0	2015-01-0	2015-01-0	50	GSM1556	SRR16921	PANC1 s	c("cell line: pancreatic ductal carcinoma cell line", "sirna: TCF7L2 siRNA")

run
SRR1692137
SRR1692138
SRR1692139
SRR1692140
SRR1692141
SRR1692142

title
PANC1 control siRNA 1
PANC1 control siRNA 2
PANC1 control siRNA 3
PANC1 siTCF7L2 1
PANC1 siTCF7L2 2
PANC1 siTCF7L2 3

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

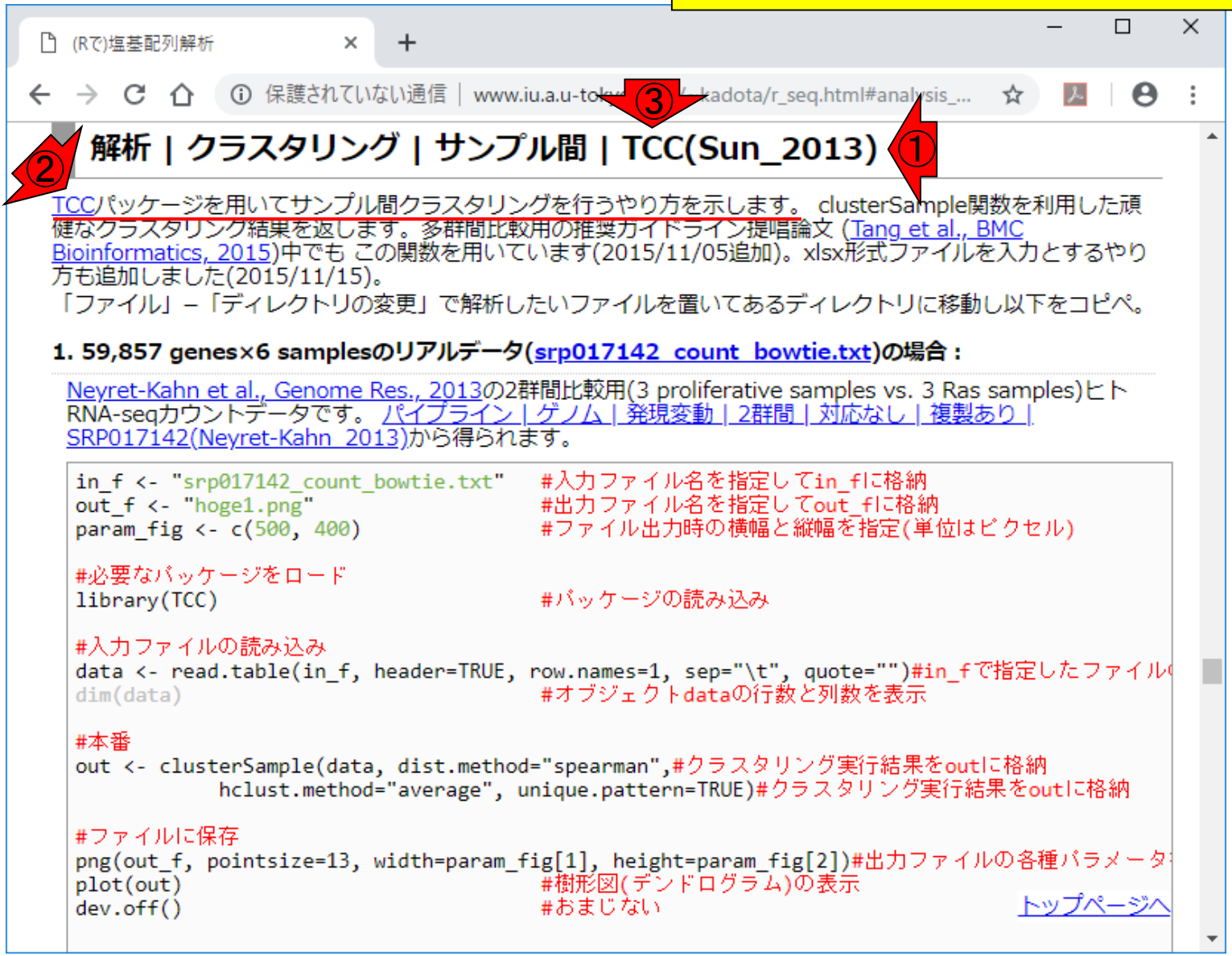
counts_gene.txt

クラスタリング

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。

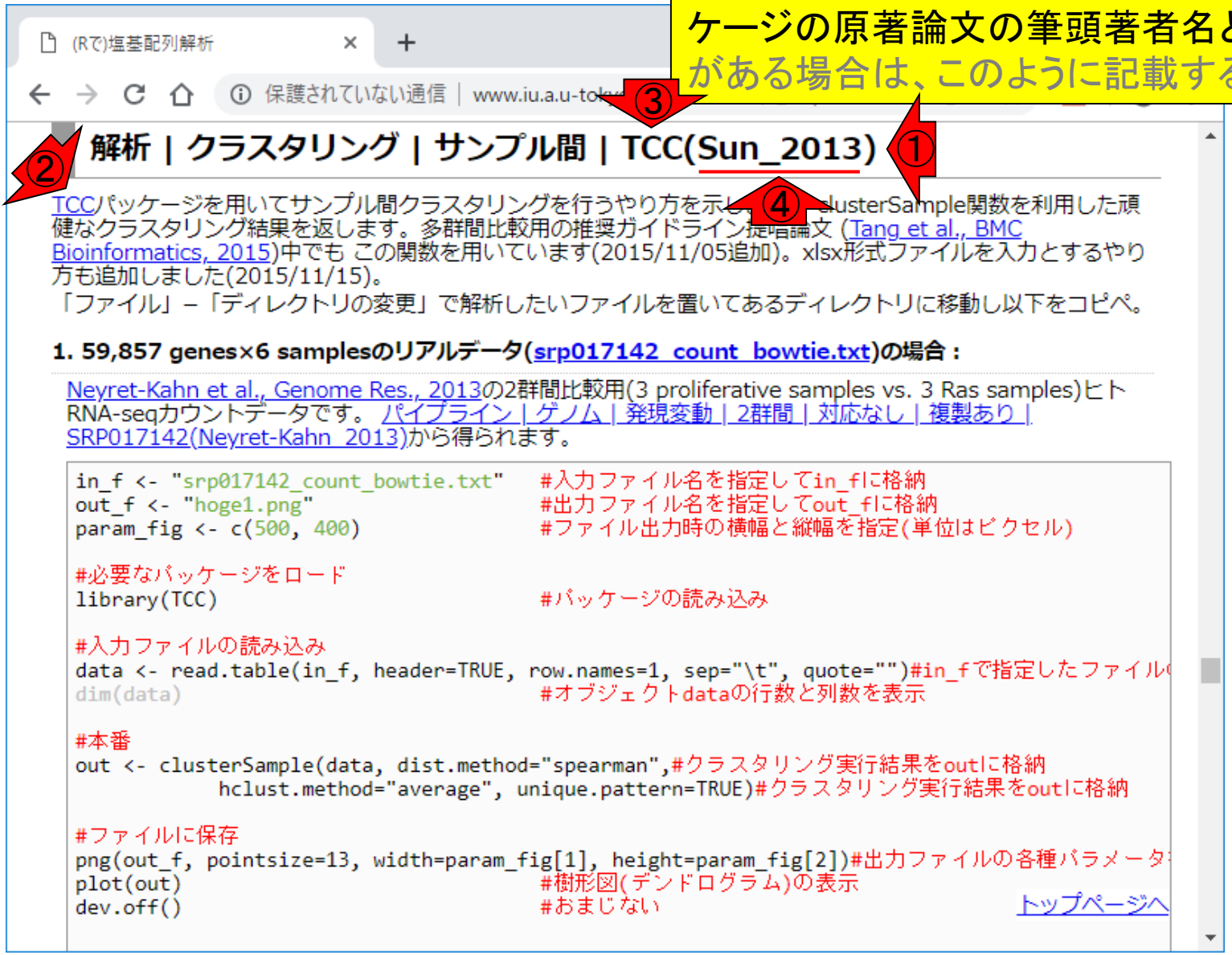
クラスタリング

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。



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多いです。



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ピクセルの大きさで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あるディレクトリに移動し以下をコピペ。

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

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

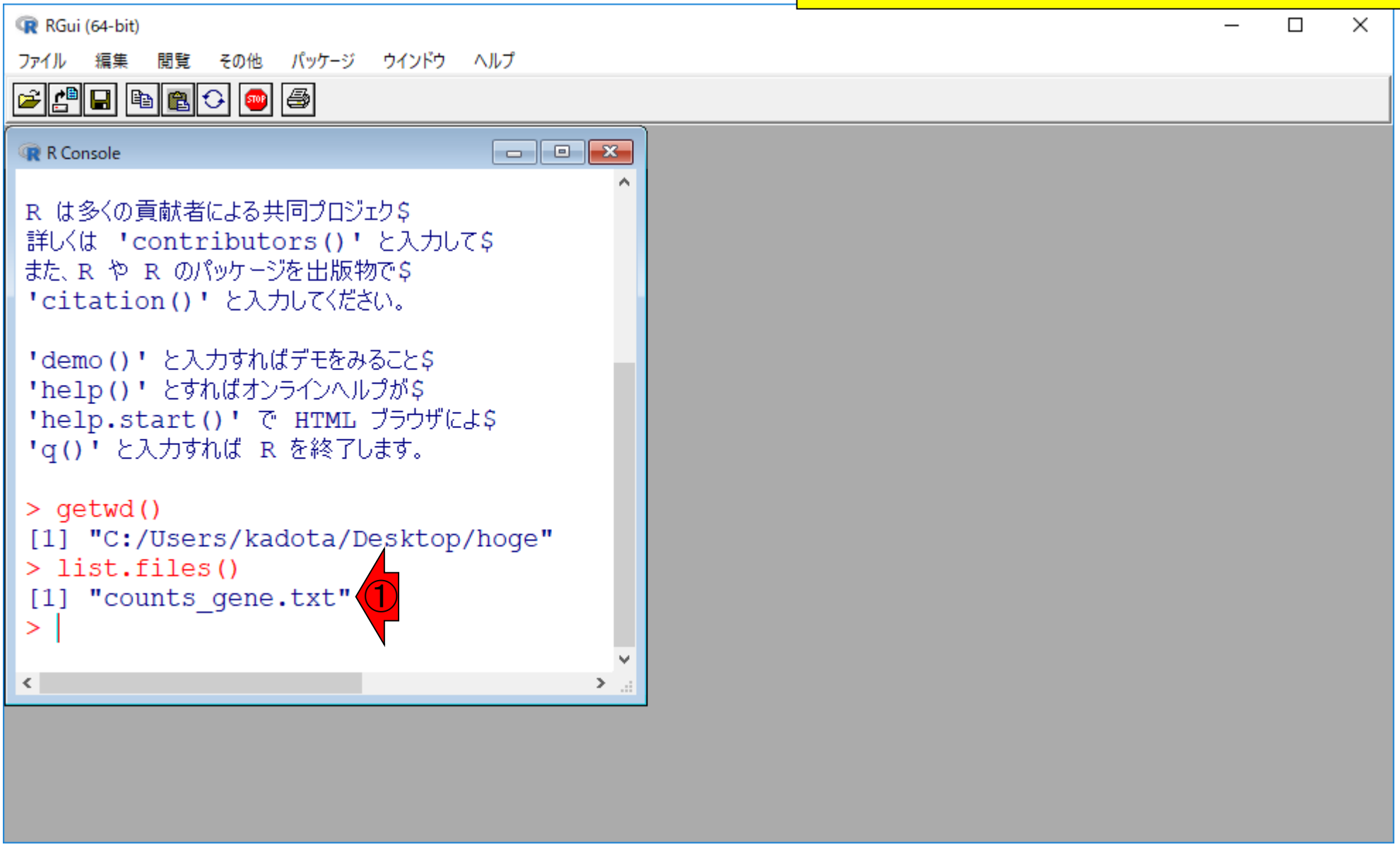
①手元にあるcounts_gene.txtを入力ファイルとして実行するにはどうすればよい?

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

counts_gene.txt

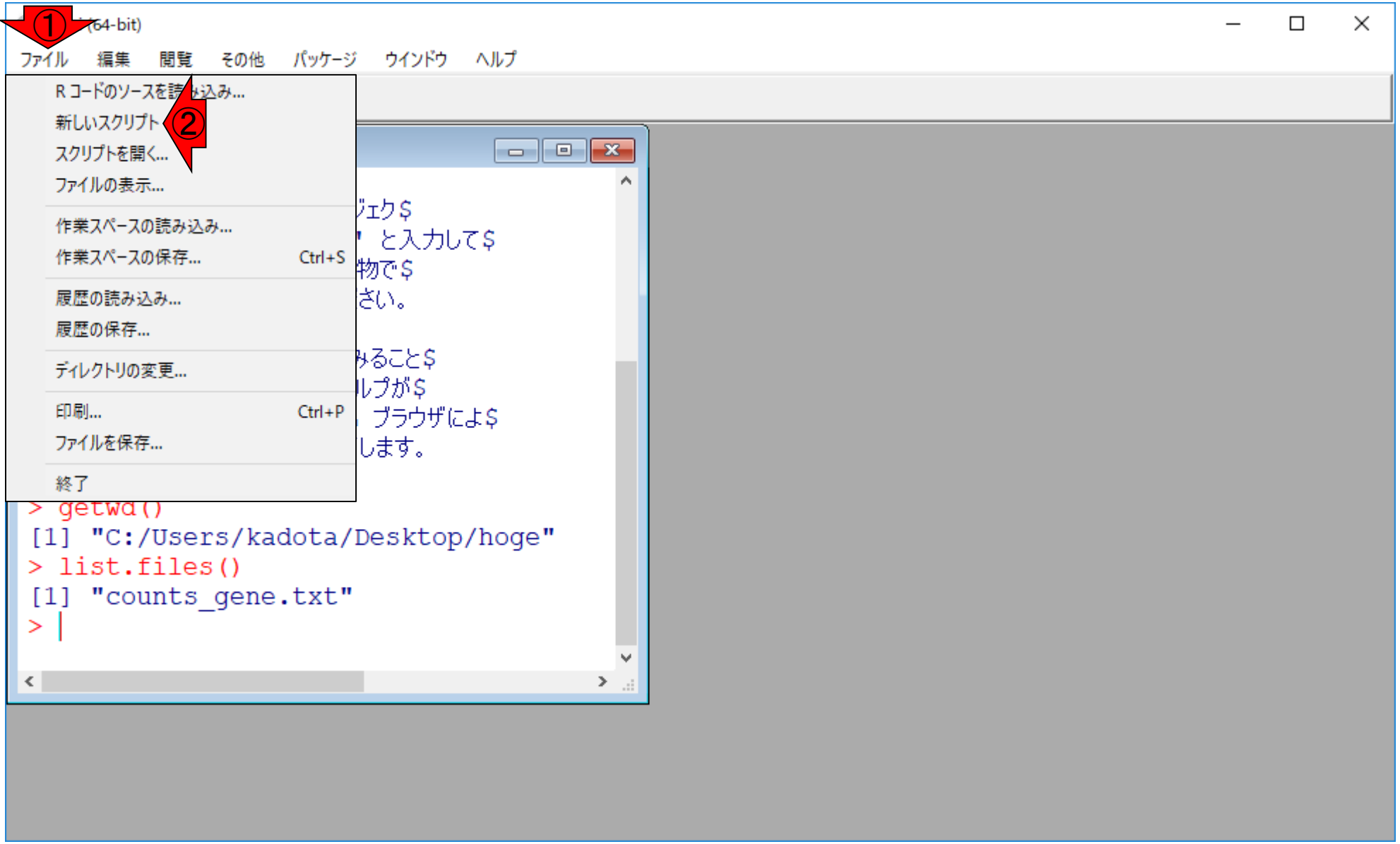
クラスタリング

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。



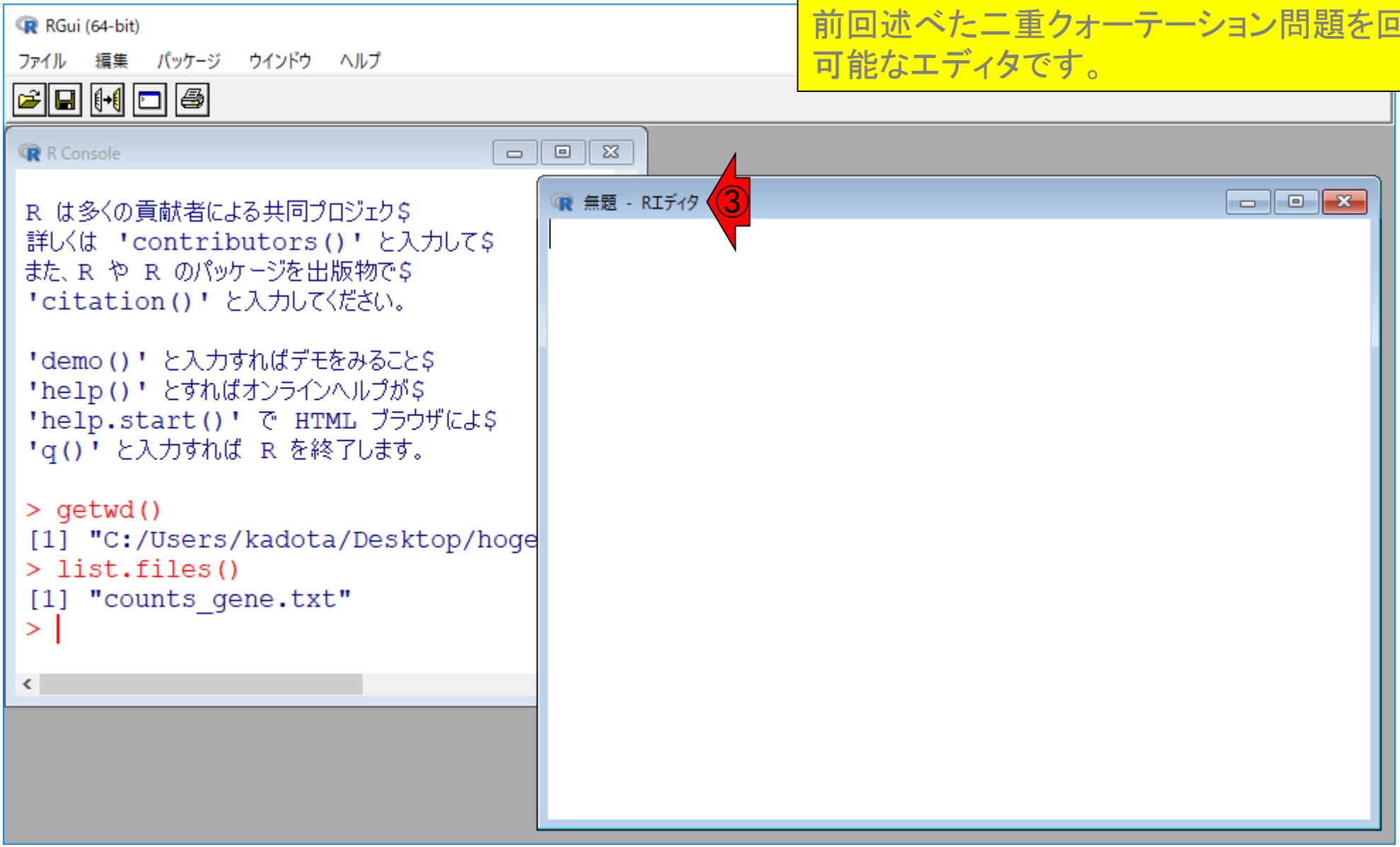
①ファイル、②新しいスクリプト。

クラスタリング



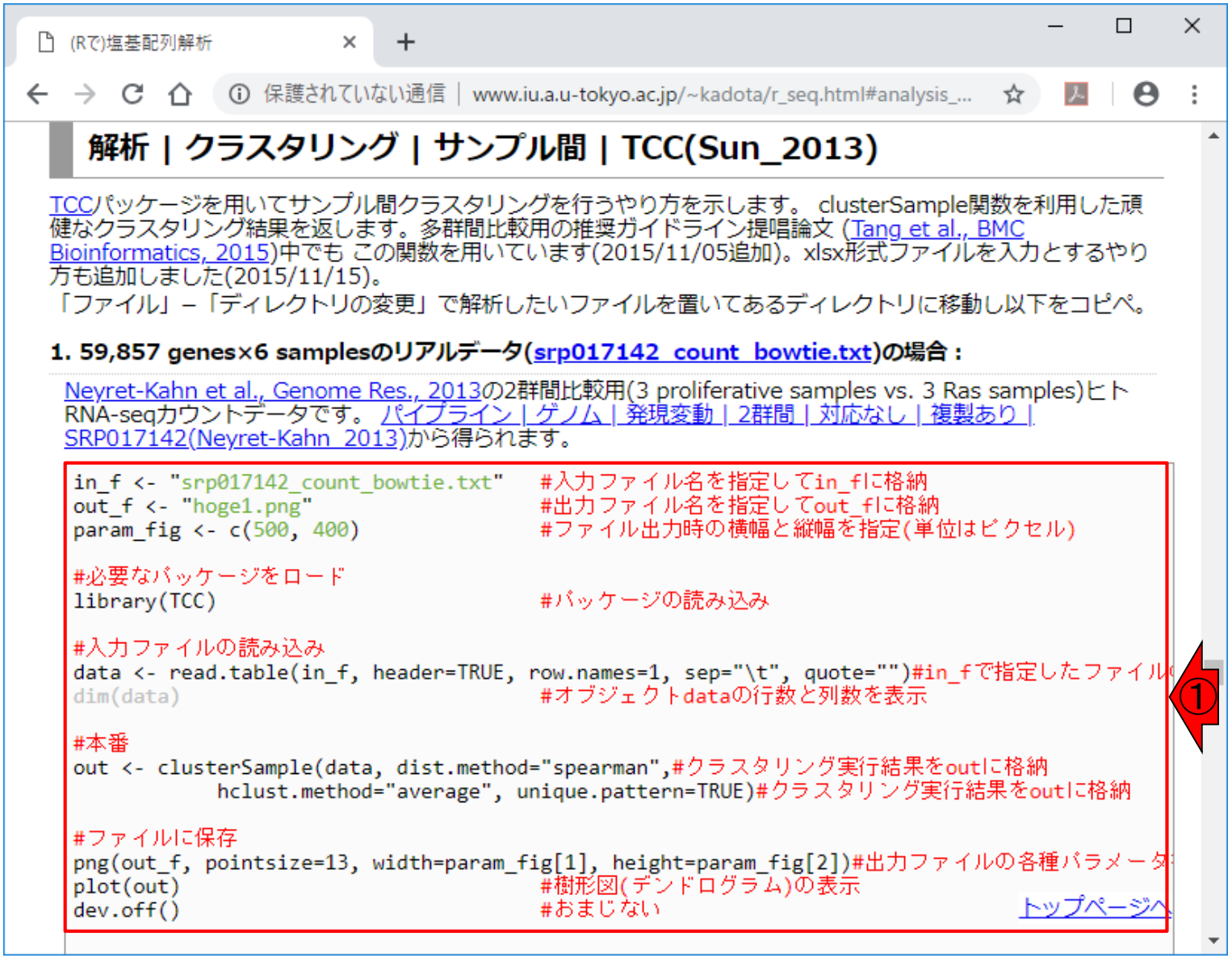
クラスタリング

①ファイル、②新しいスクリプト。こんな感じで③Rエ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。



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…



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
すべてコピーして、②Rエディタ上でペースト。

RGui (64-bit)

ファイル　編集　パッケージ　ウインドウ　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"
>
```

②

無題 - RIディタ

```
in_f <- "srp017142_count_bowtie.txt"    #入力ファイル名を
out_f <- "hoge1.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
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
すべてコピーして、②Rエディタ上でペースト。
必要最小限の変更は、③ここを…

RGui (64-bit)

ファイル　編集　パッケージ　ウインドウ　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"
>
```

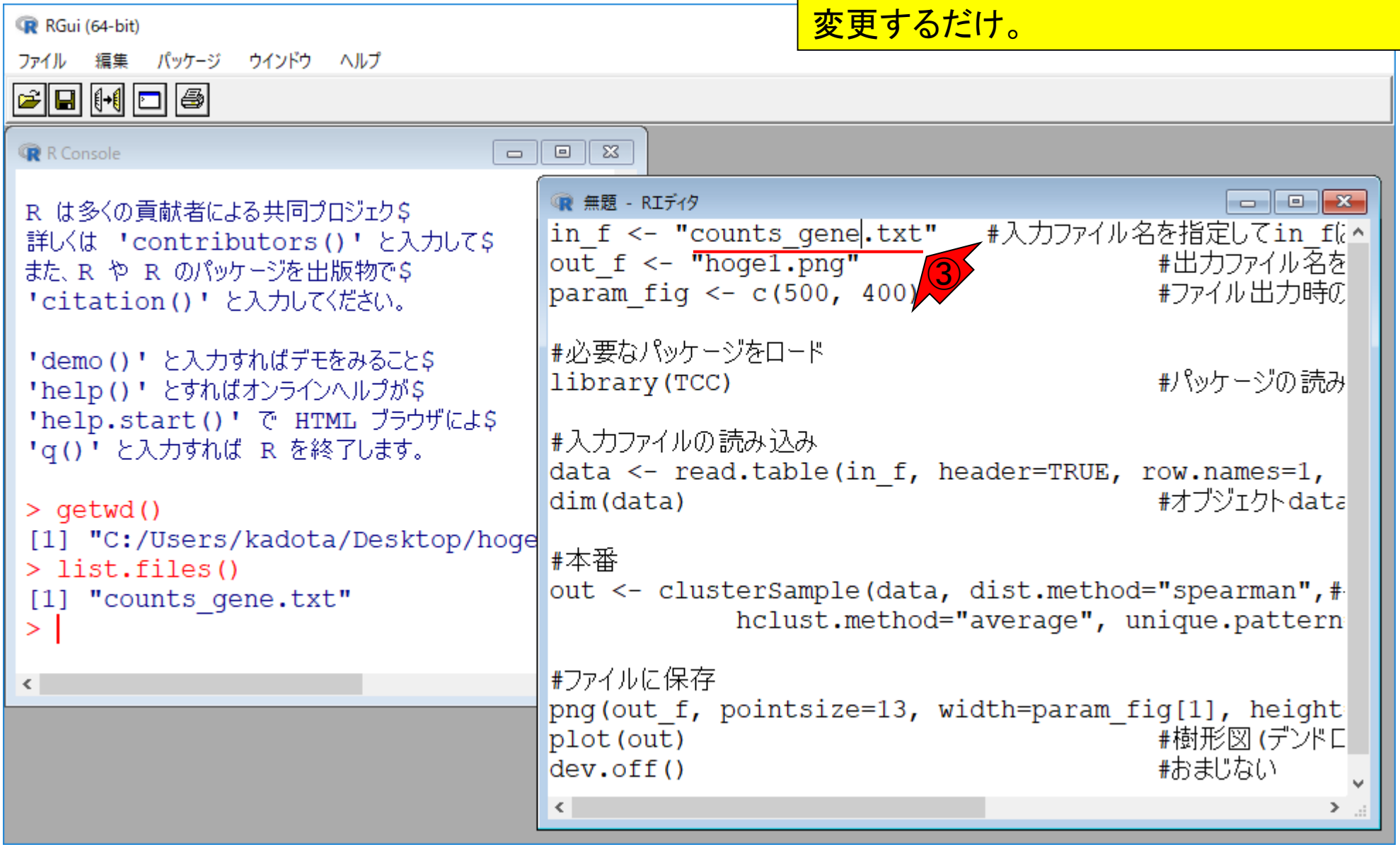
無題 - RIディタ

```
in_f <- "srp017142_count_bowtie.txt"    #入力ファイル名を
out_f <- "hoge1.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
dev.off()                               #おまじない
```

③

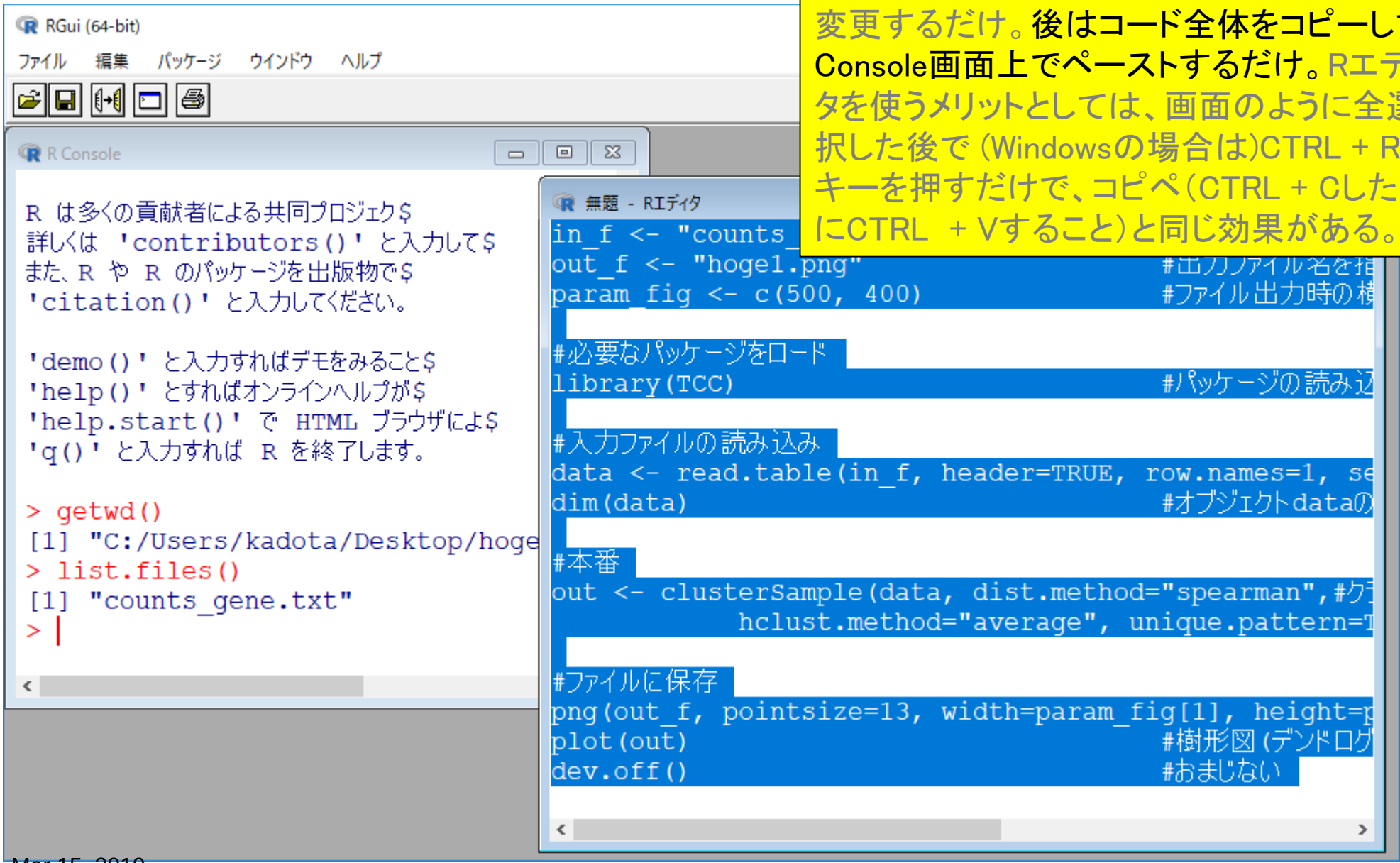
クラスタリング

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。



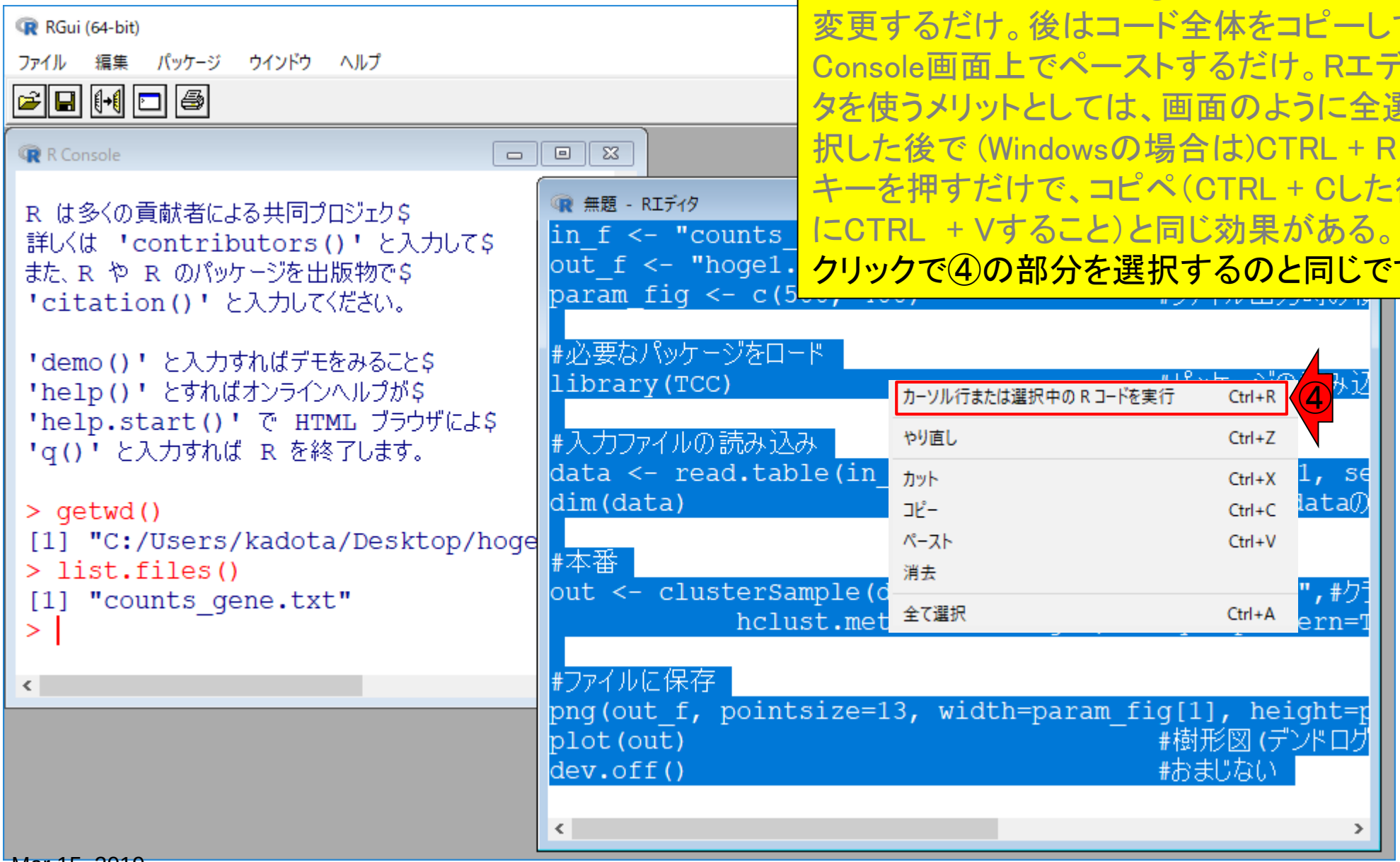
クラスタリング

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RConsole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、コピペ(CTRL + Cした後にCTRL + Vすること)と同じ効果がある。



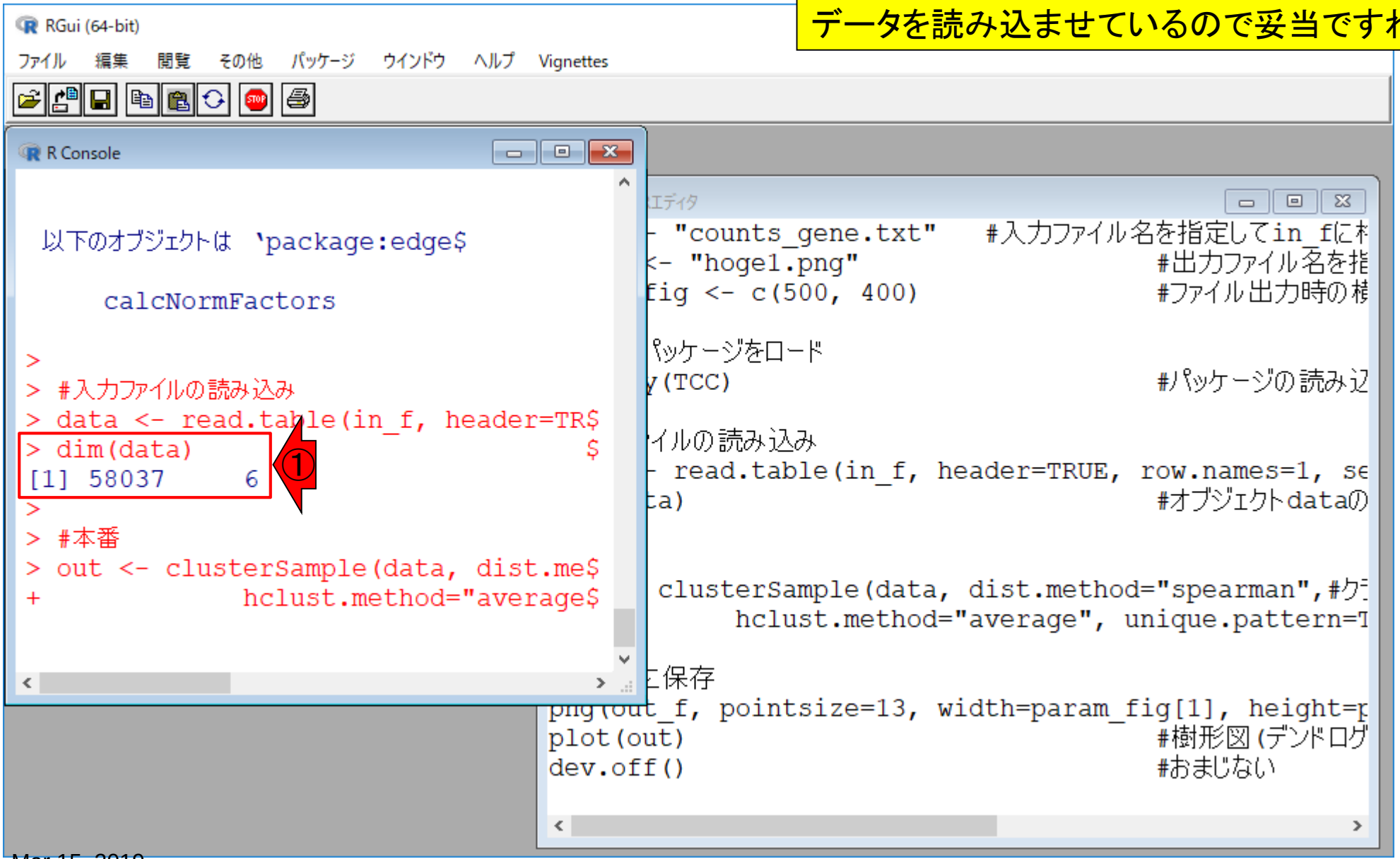
クラスタリング

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、コピペ(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。



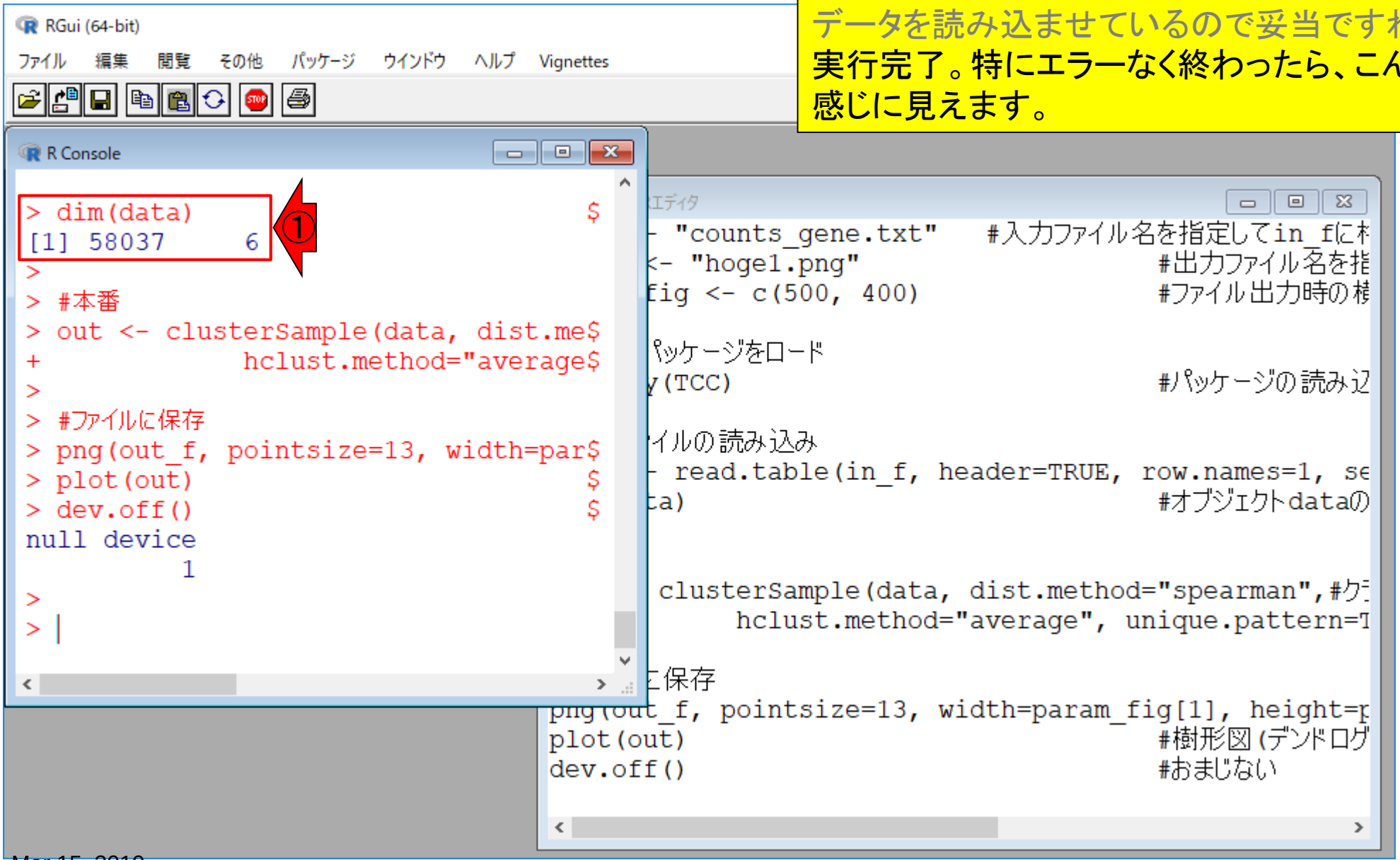
クラスタリング

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。



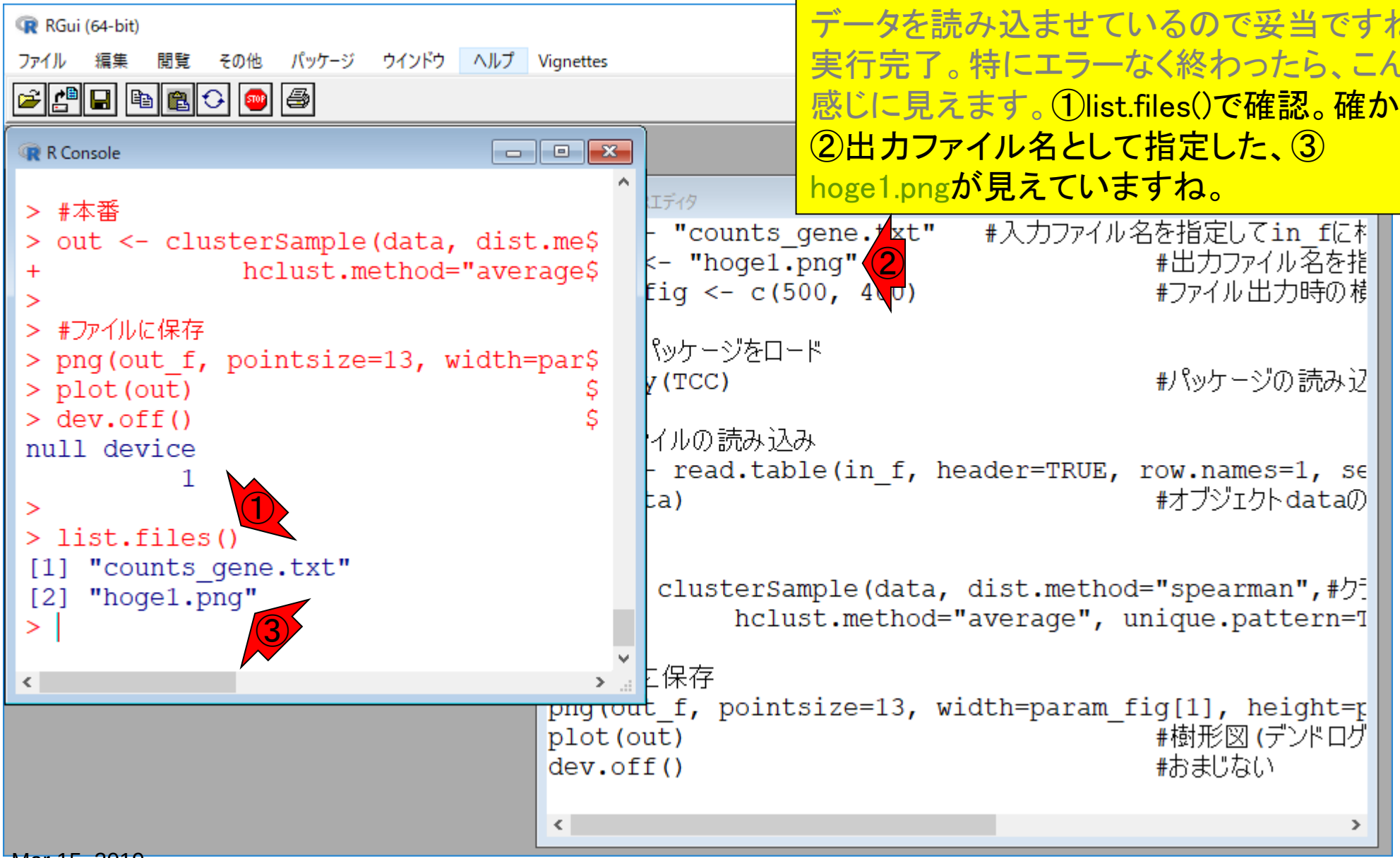
クラスタリング

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。



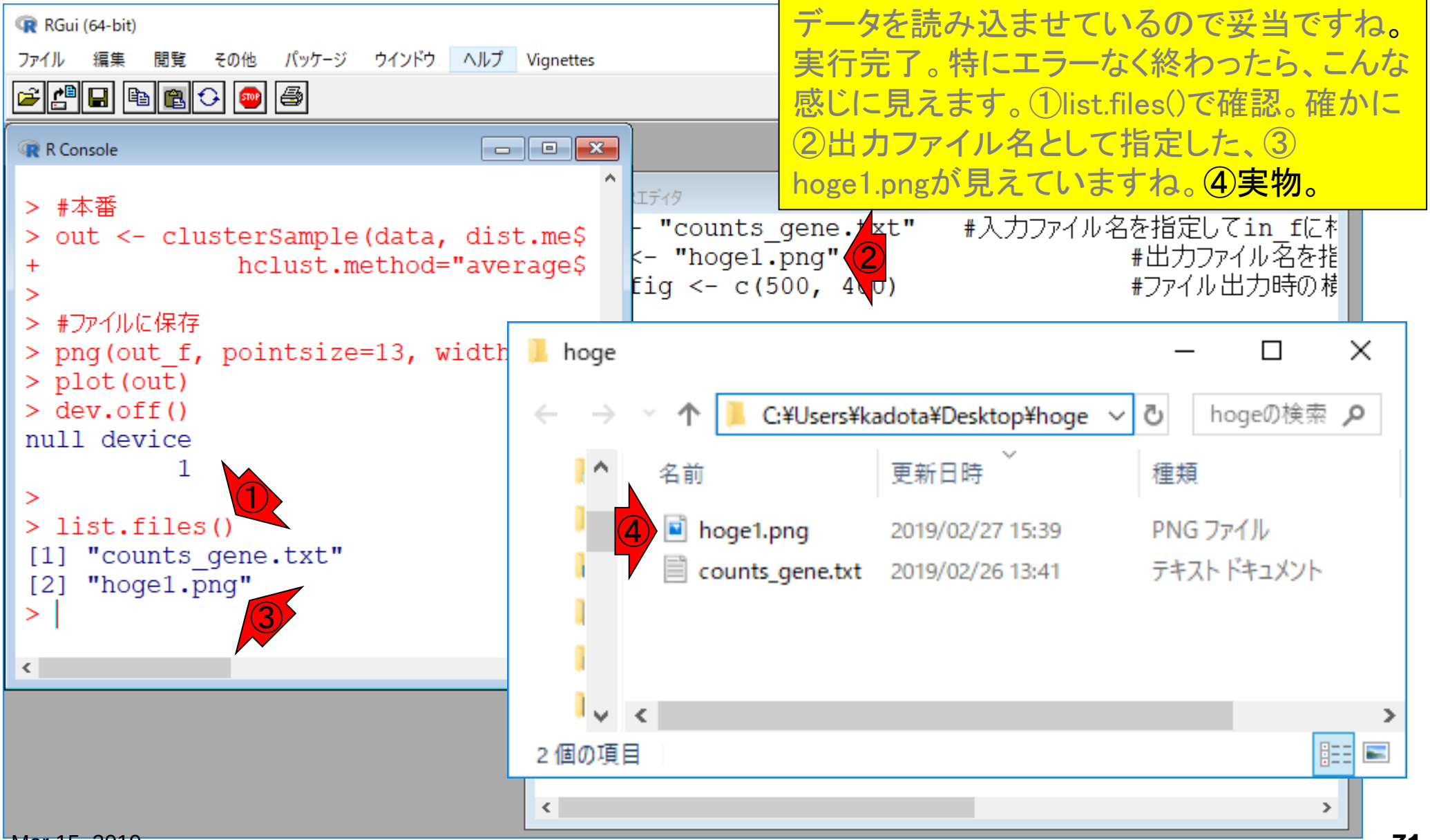
クラスタリング

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。



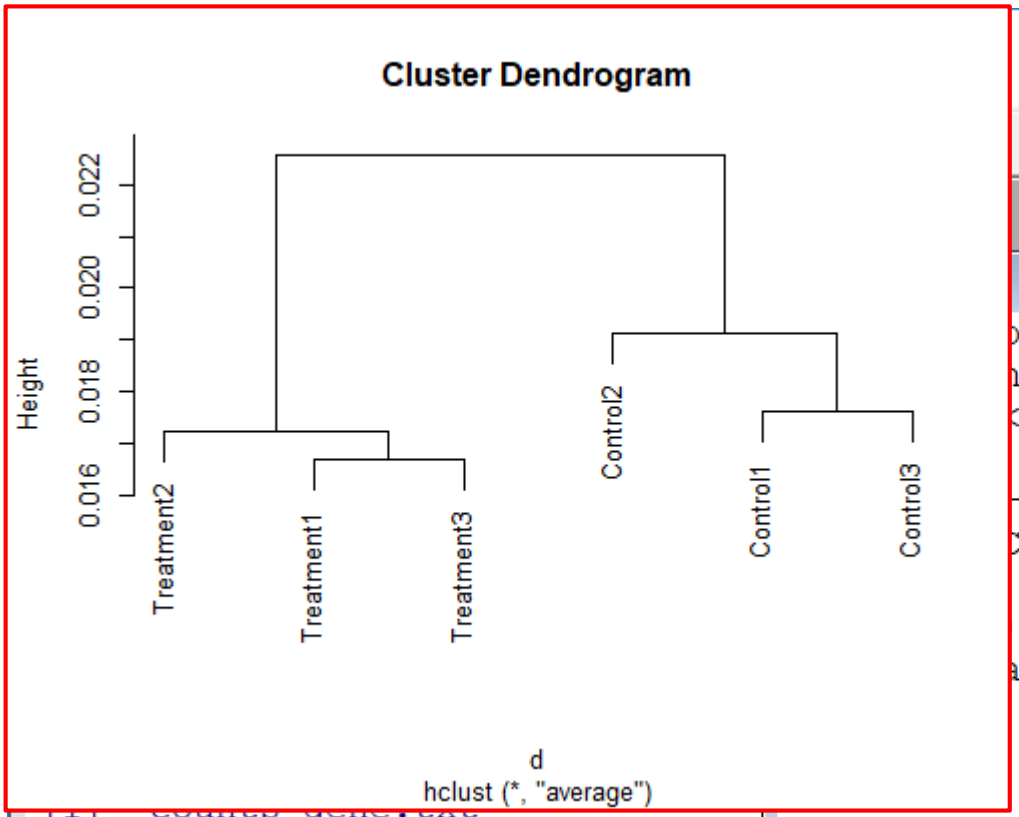
クラスタリング

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。



赤枠が出力ファイル①hoge1.pngの中身。

クラスタリング

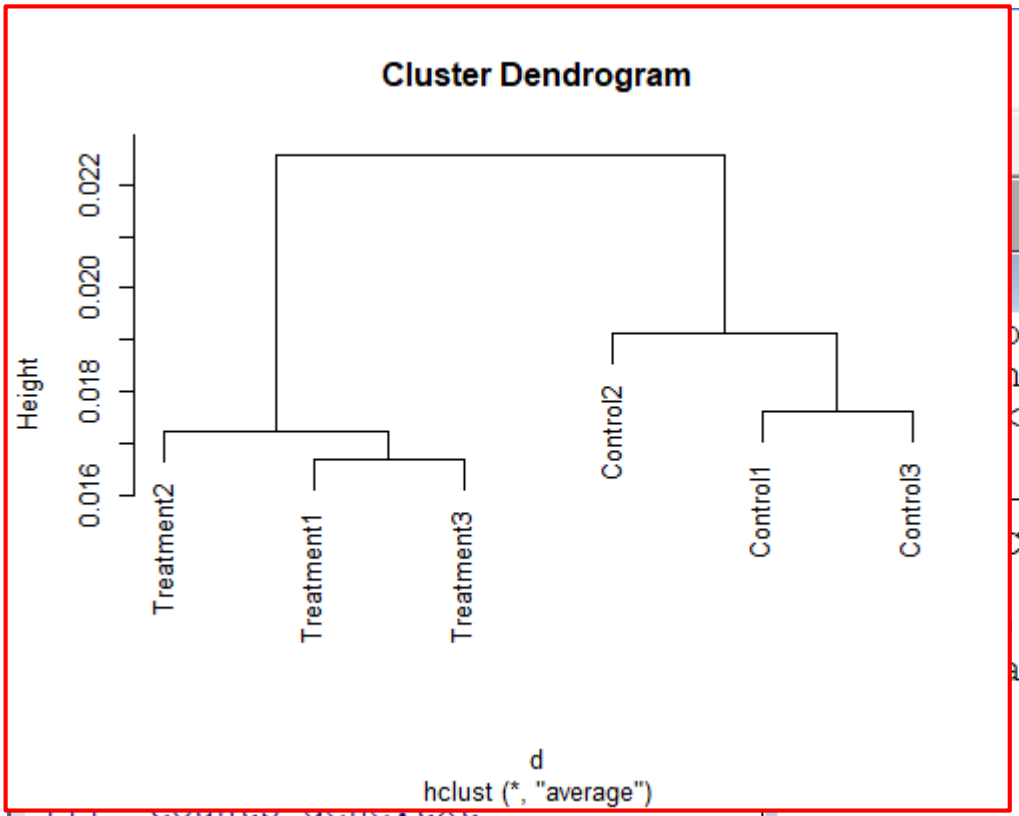



```
[2] "hoge1.png"
>
```

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

ージをロード
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=T

#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。




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

ージをロード
")                                  #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=T

#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
```

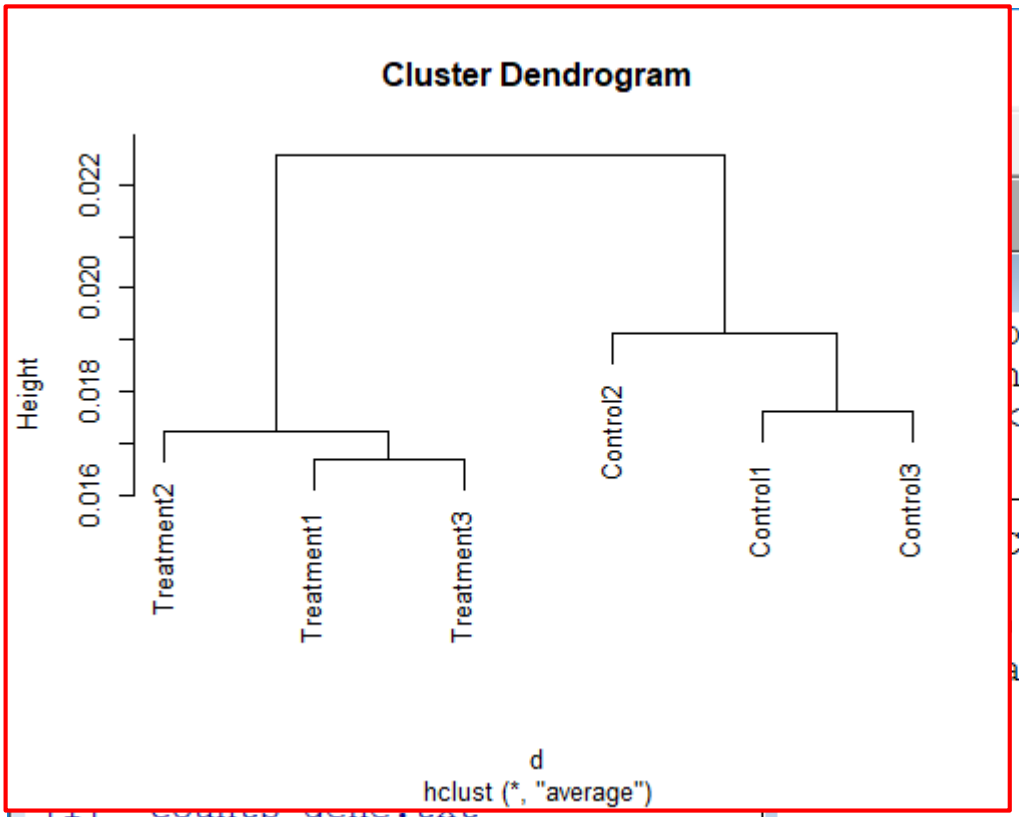
```
[2] "hoge1.png"
>
```

②

③

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。




```
[2] "hoge1.png"
> |
```

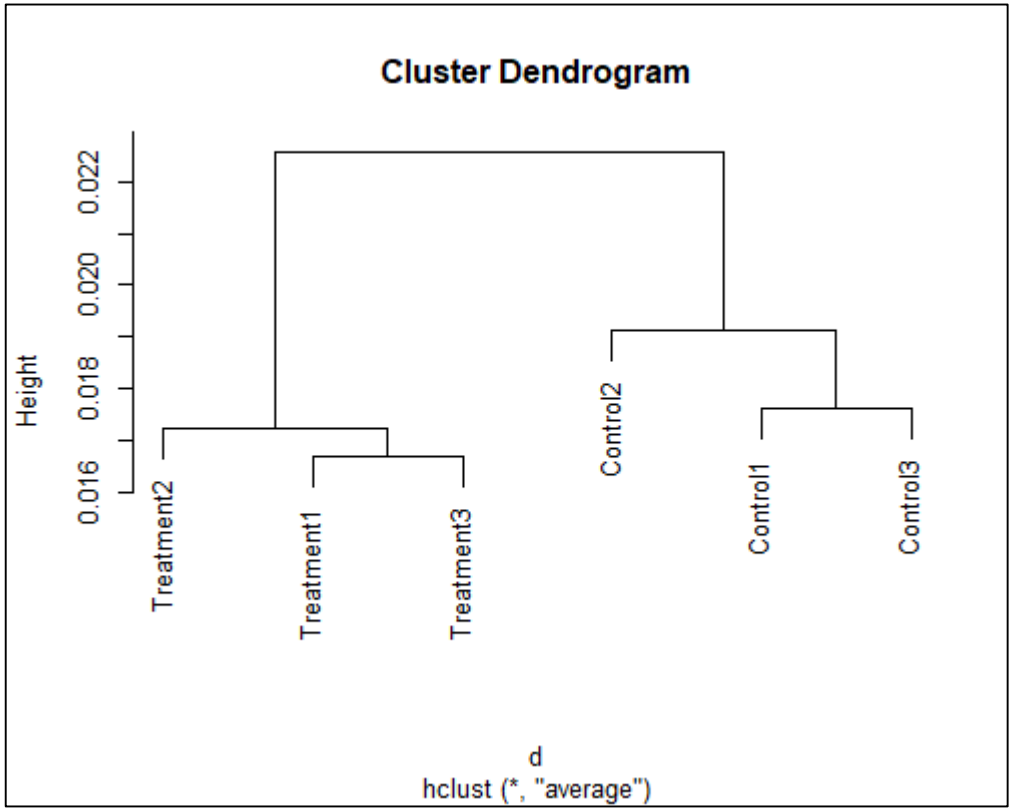
```
ounts_gene.txt"    #入力ファイル名を指定してin_fに
hoge1.png"                          #出力ファイル名を指
<- c(500, 400)                      #ファイル出力時の横

ージをロード
)                                   #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",#クラ
            hclust.method="average", unique.pattern=T

#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

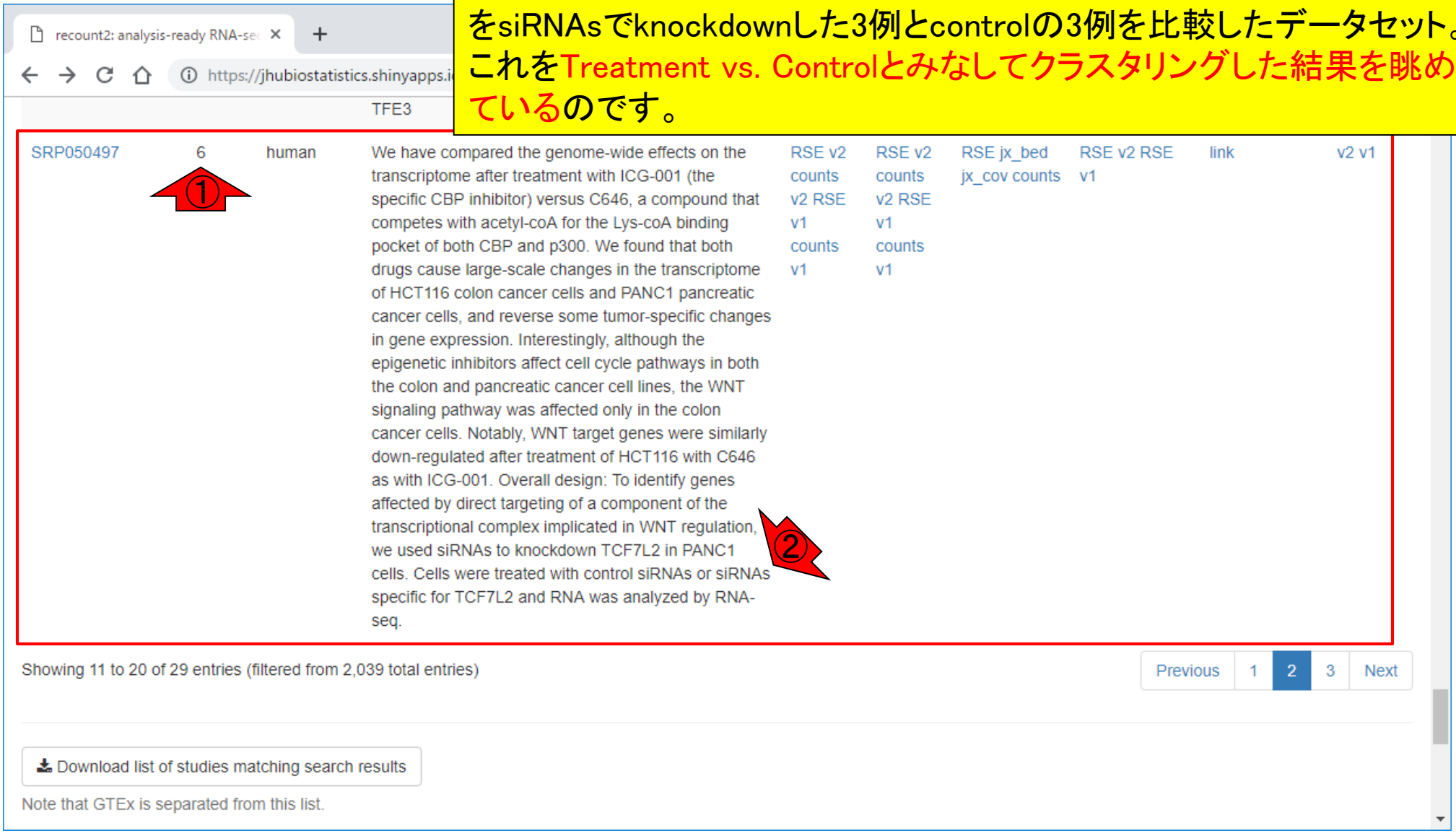
結果の解釈

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でますが、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。



おさらい

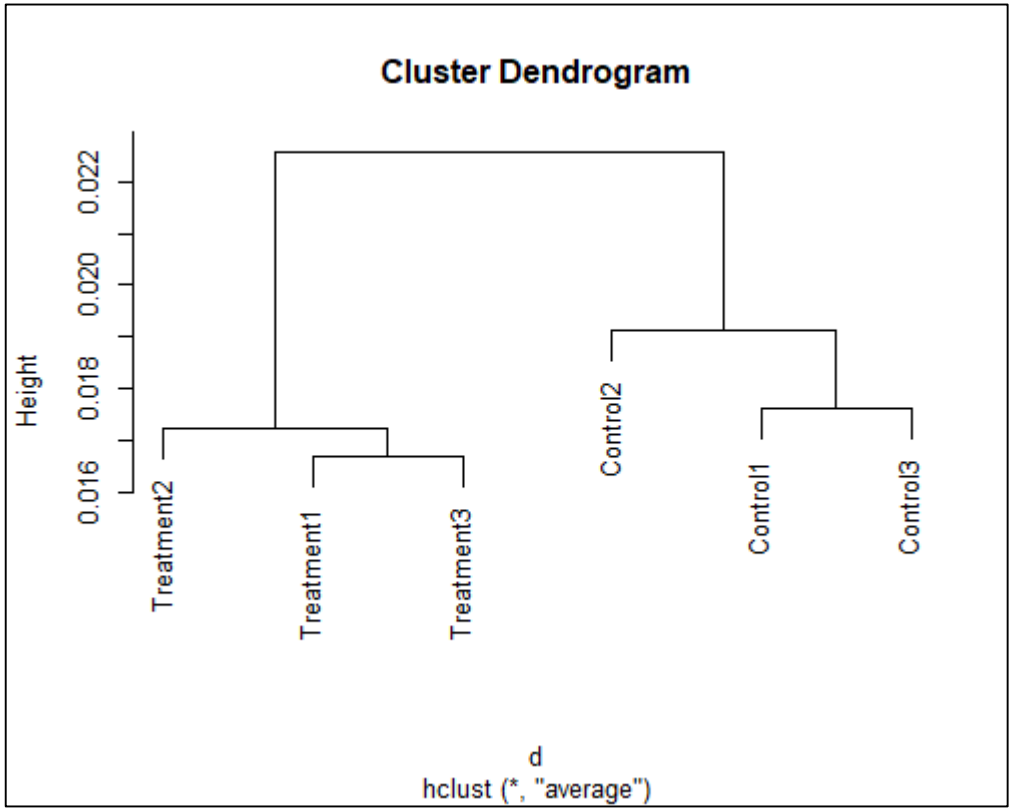
入力データは58,037遺伝子×6サンプル。どのサンプルとどのサンプルが似ているかを眺め、全体としておかしくないか（外れサンプルがあるか）などをチェックする目的で、サンプル間クラスタリングを行いました。

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

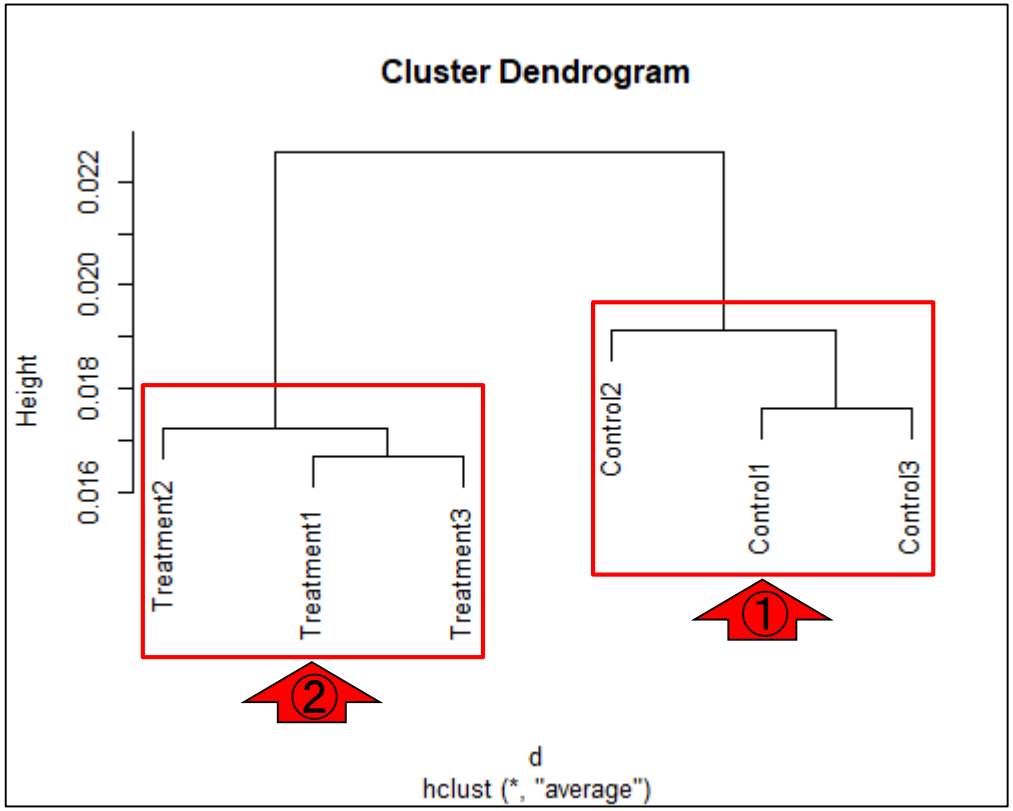
counts_gene.txt

私はこの結果をみて「大丈夫」と判断します。

結果の解釈

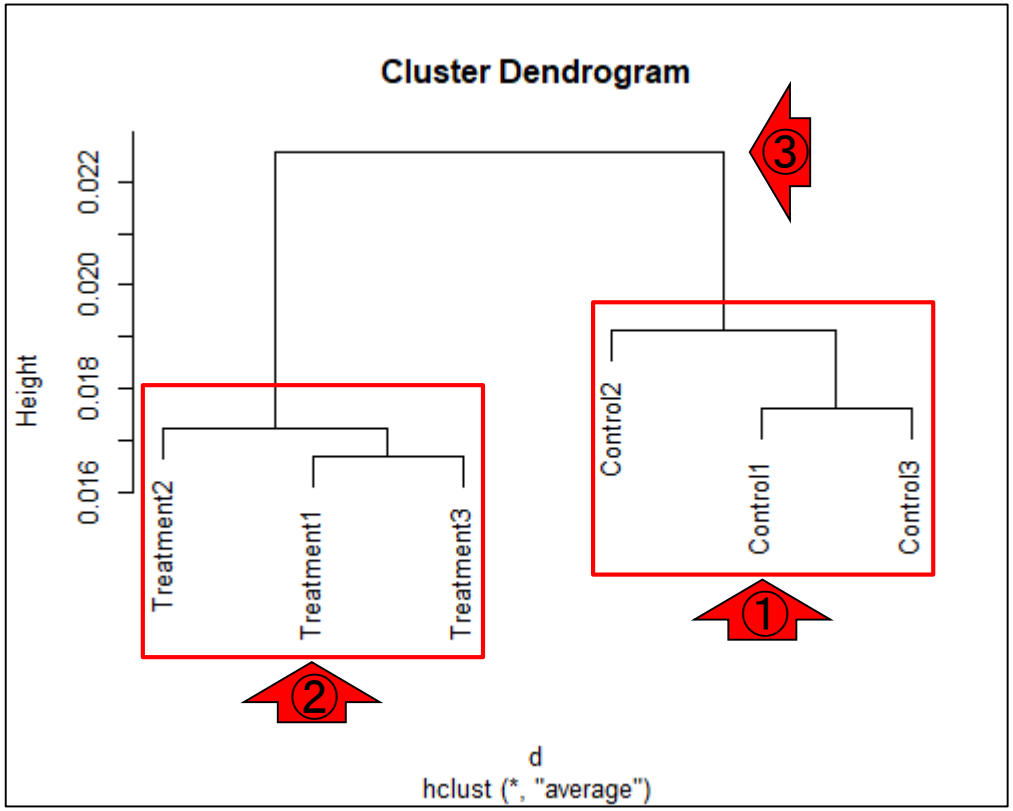


結果の解釈



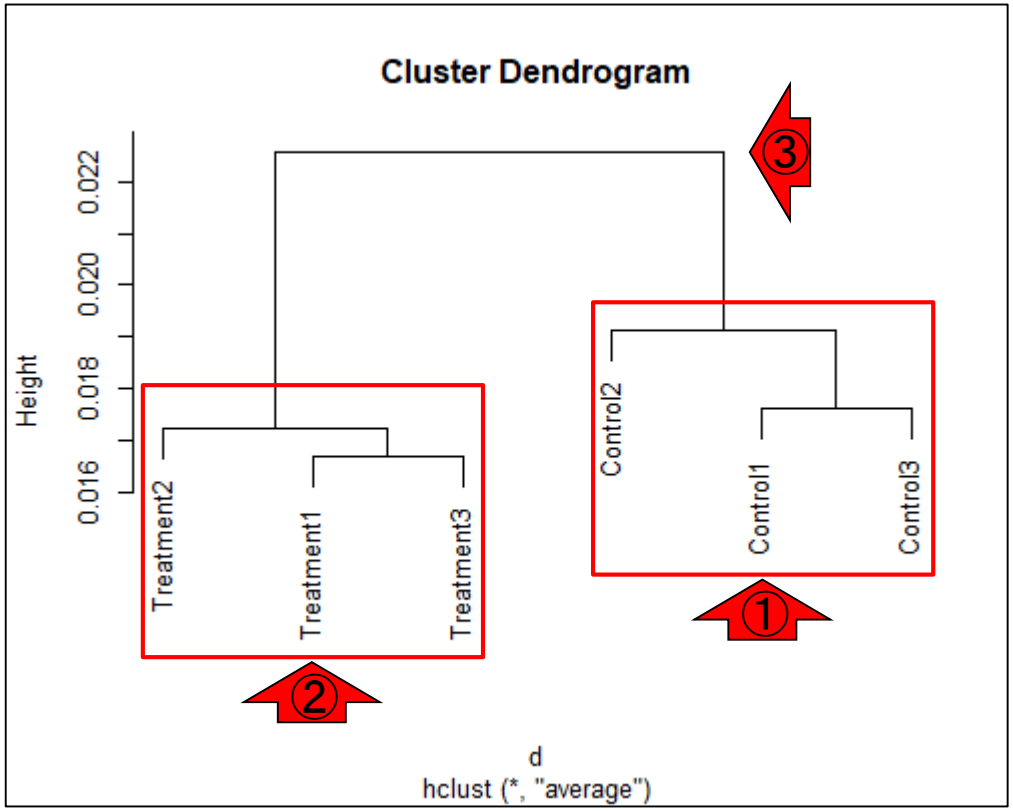
私はこの結果をみて「大丈夫」と判断します。理由は、①Control群、②Treatment群ともに、外れサンプルがないように見えるからです。

結果の解釈



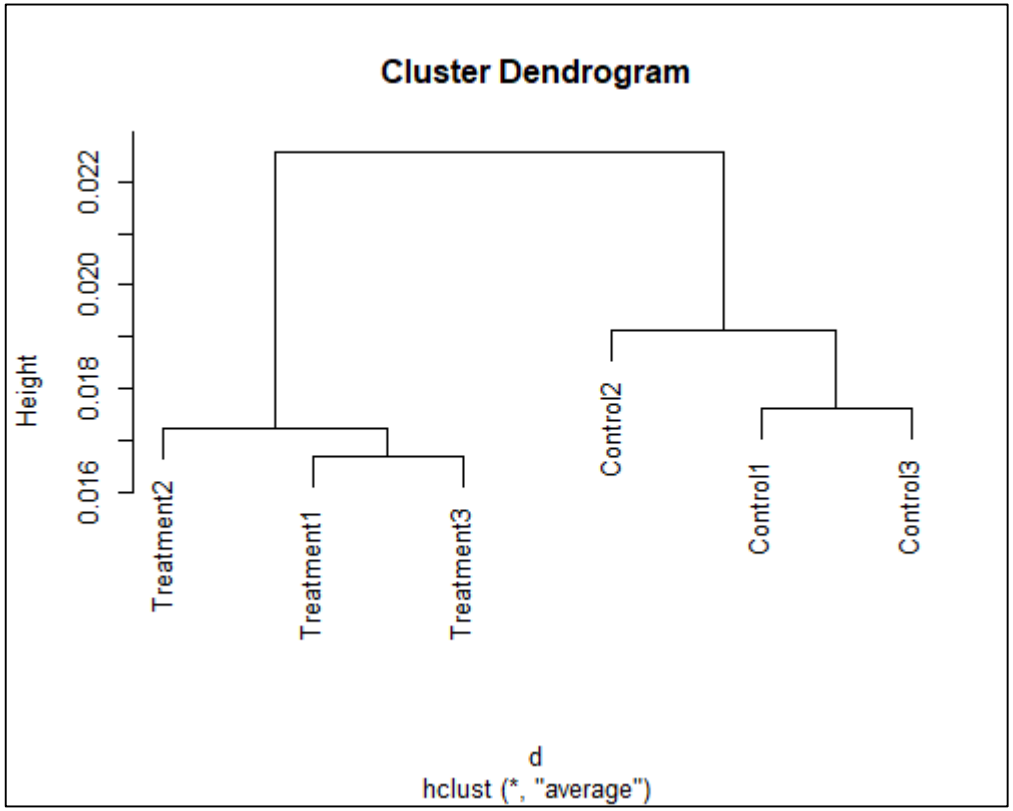
私はこの結果をみて「大丈夫」と判断します。理由は、①Control群、②Treatment群ともに、外れサンプルがないように見えるからです。また、①と②でみられる郡内の平均類似度が、③群間の平均類似度よりも高いからです。

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でみられる郡内の平均類似度が、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でみられる郡内の平均類似度が、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が、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。

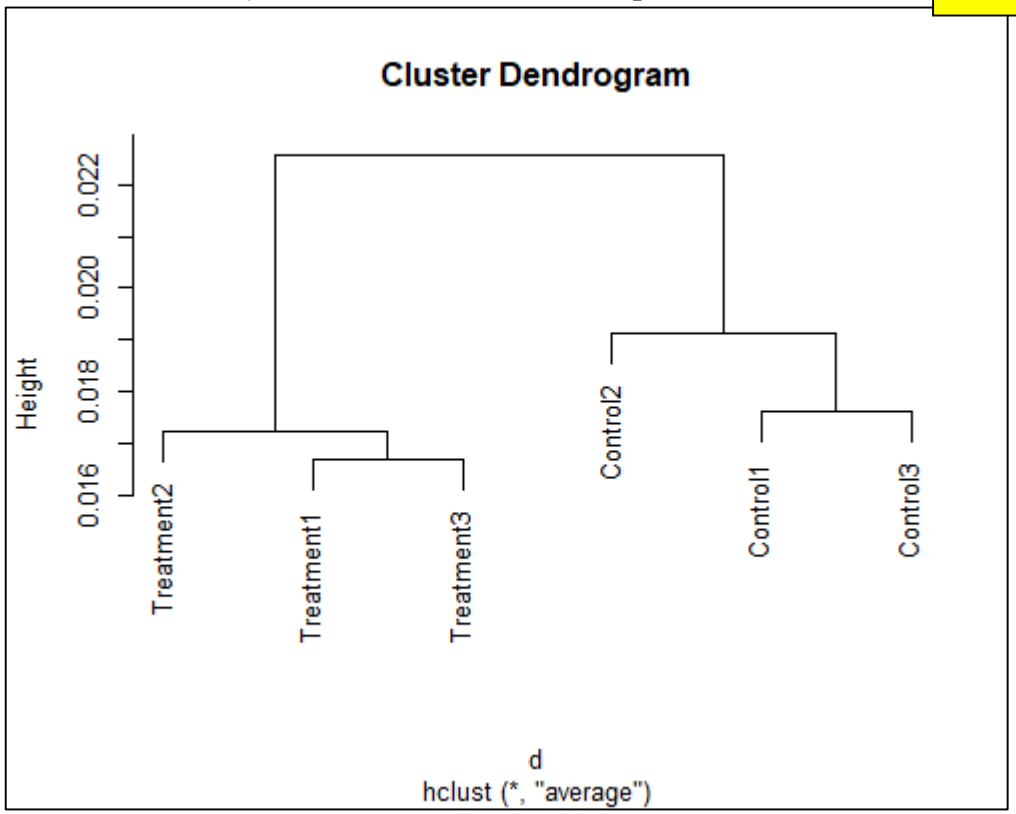
Contents

- 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

参考

客観的な指標

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。



参考

客観的な指標

最適なクラスター数を見積る客観的な指標として有名な平均シルエットスコア（Average Silhouette score; AS）を、任意のグループに置き換えて計算するというシンプルなアイデアの提唱論文。

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 S[1], Sun J[1], Shimizu K[1], Kadota K[1].

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.

参考

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 S[1], Sun J[1], Shimizu K[1], Kadota K[1].

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.

①ここにあります。

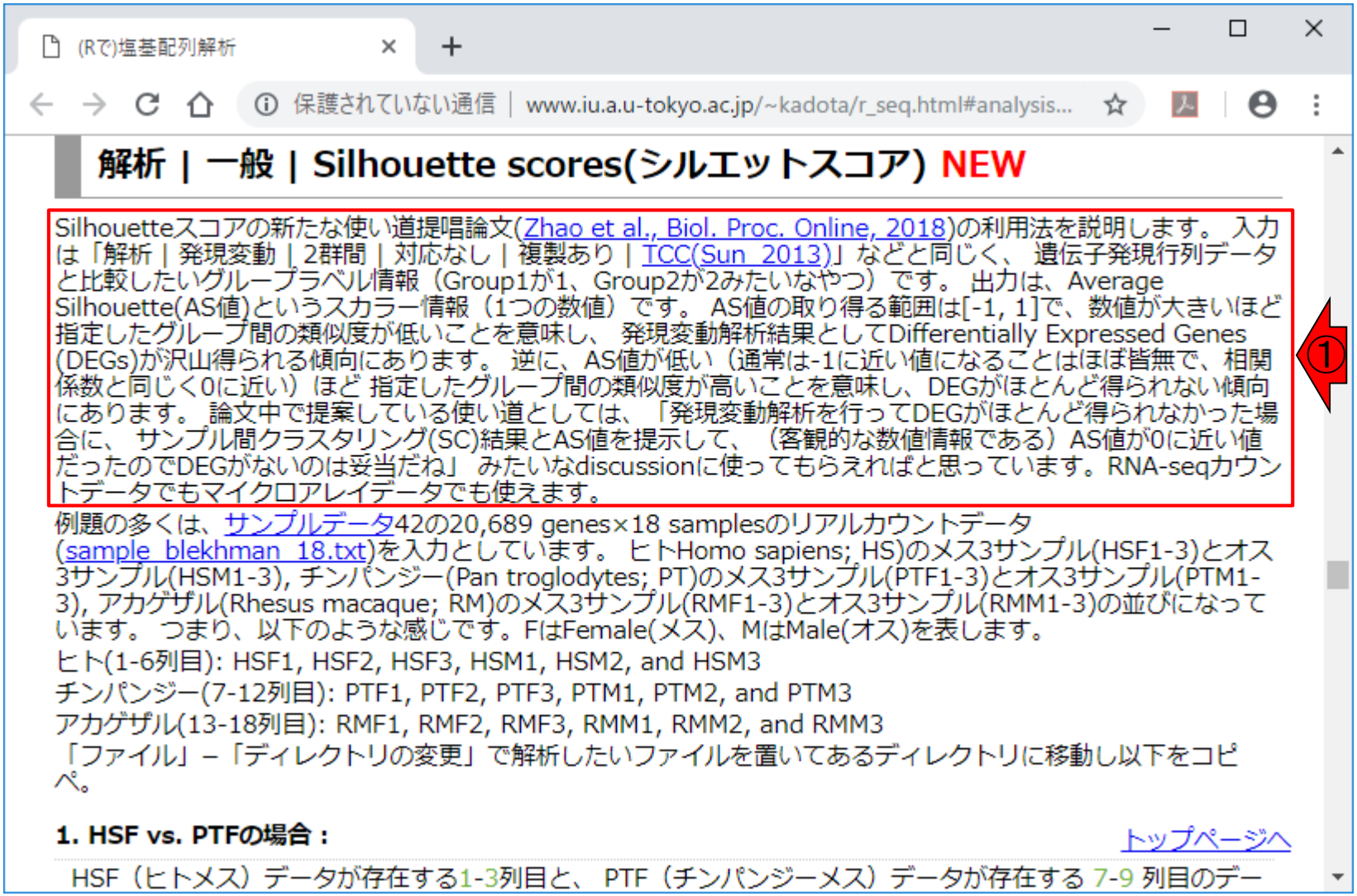
シルエットスコア

- 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 20
- 解析 | 一般 | アラインメント | マルチプル | msa(Bodenhofer_2015) (last modified 201
- 解析 | 一般 | Silhouette scores(シルエットスコア) ①st modified 2018/06/16)
- 解析 | 一般 | パターンマッチング (last modified 2013/06/19)
- 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 modifi
- 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)

参考

シルエットスコア

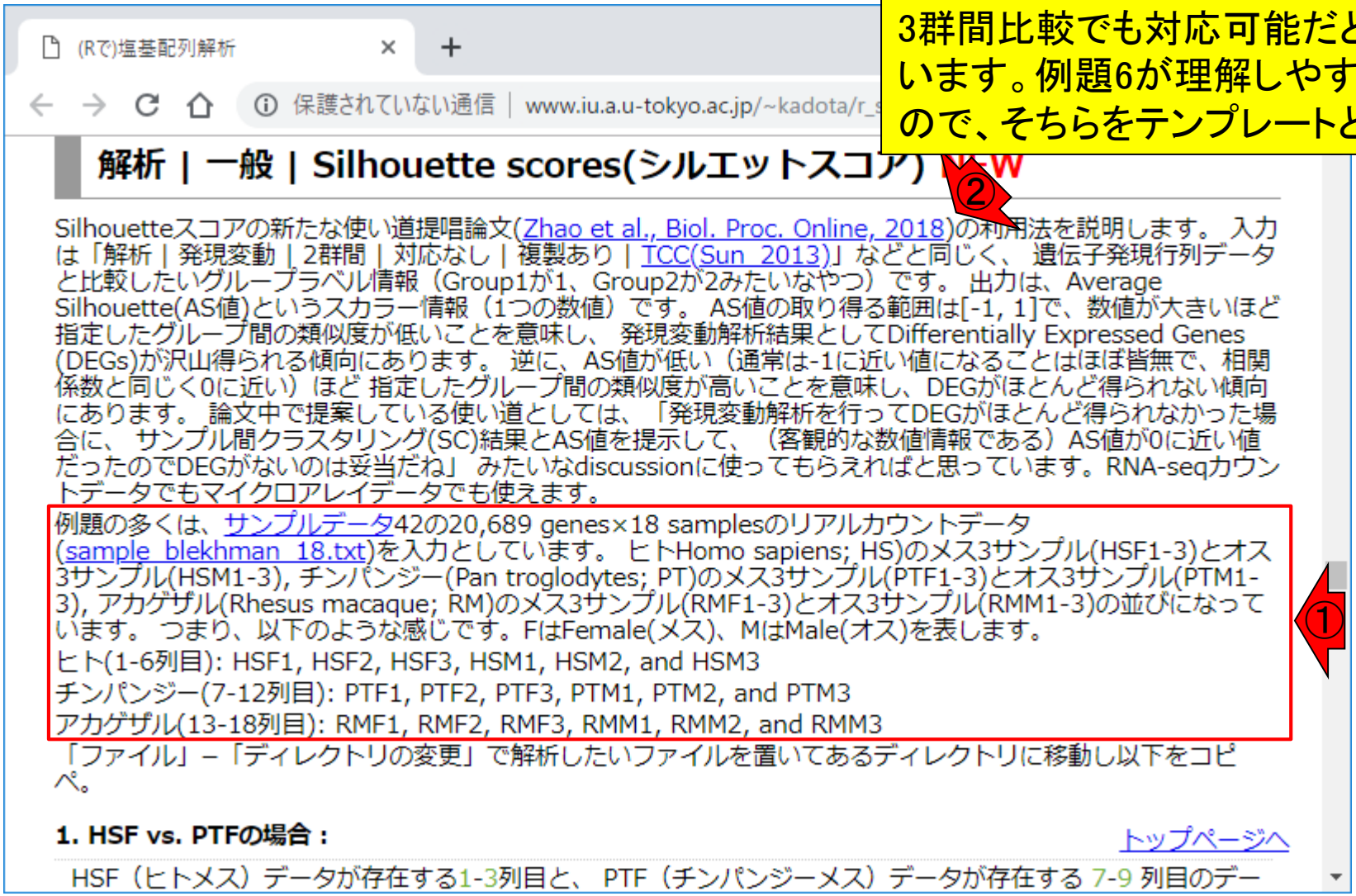
①得られた平均シルエットスコアの解釈について、ごちゃごちゃ書いています。



参考

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。



参考

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

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

③

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

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
gene_10000までがnon-DEGであることが既知です。

④

⑤

```
in_f <- "data_hypdata_3vs3.txt"          #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
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
```

トップページへ

⑥

おさらい

①手元ファイルの入力データは、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…

② ③

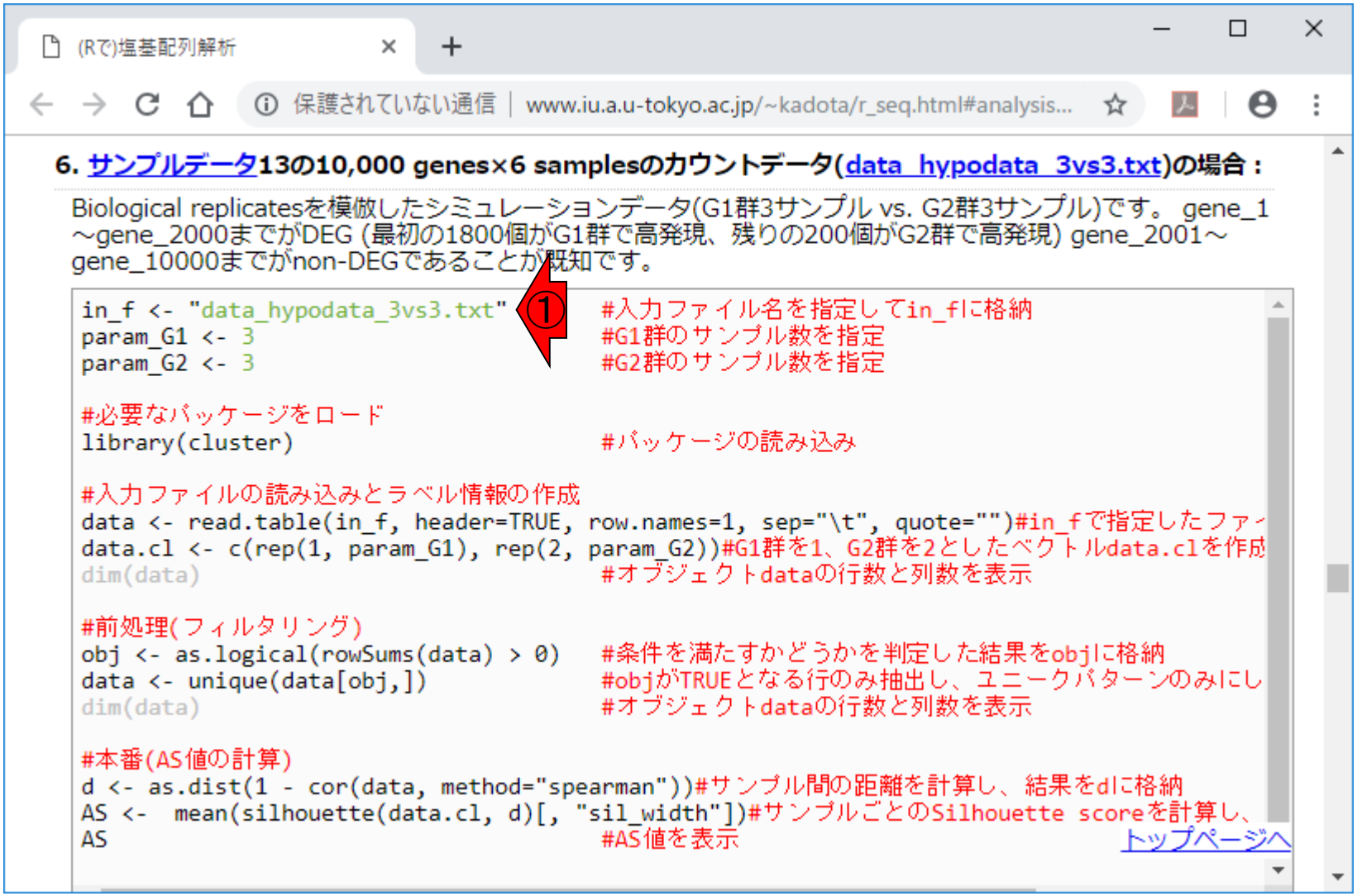
gene_id	Control1	Control2	Control3	Treatment1	Treatment2	Treatment3
ENSG00000000003.14	11156	9957	14914	15968	15138	16747
ENSG00000000005.5	0	0	0	0	0	0
ENSG00000000419.12	64023	55036	88460	62819	67807	70965
ENSG00000000457.13	24666	22830	37528	31189	31249	36804
ENSG00000000460.16	42199	38508	58545	36169	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

参考

①入力ファイル名部分のみ変更すればよい。

シルエットスコア



参考

テンプレートのRコードをコピペして…

シルエットスコア

RGui (64-bit)

ファイル　編集　パッケージ　ウインドウ　ヘルプ

R Console

```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
```

無題 - RIディタ

```
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"))#サンプ
AS <-  mean(silhouette(data.cl, d)[, "sil_width"])#
```

参考

シルエットスコア

テンプレートのRコードをコピペして、①入力ファイル名部分のみ変更して…

RGui (64-bit)

ファイル　編集　パッケージ　ウインドウ　ヘルプ

R Console

```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
```

無題 - RIディタ

```
in_f <- "counts_gene.txt"  ①      #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"))#サンプ
AS <-  mean(silhouette(data.cl, d)[, "sil_width"])#
```

参考

シルエットスコア

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

RGui (64-bit)

ファイル　編集　パッケージ　ウインドウ　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
>
>
```

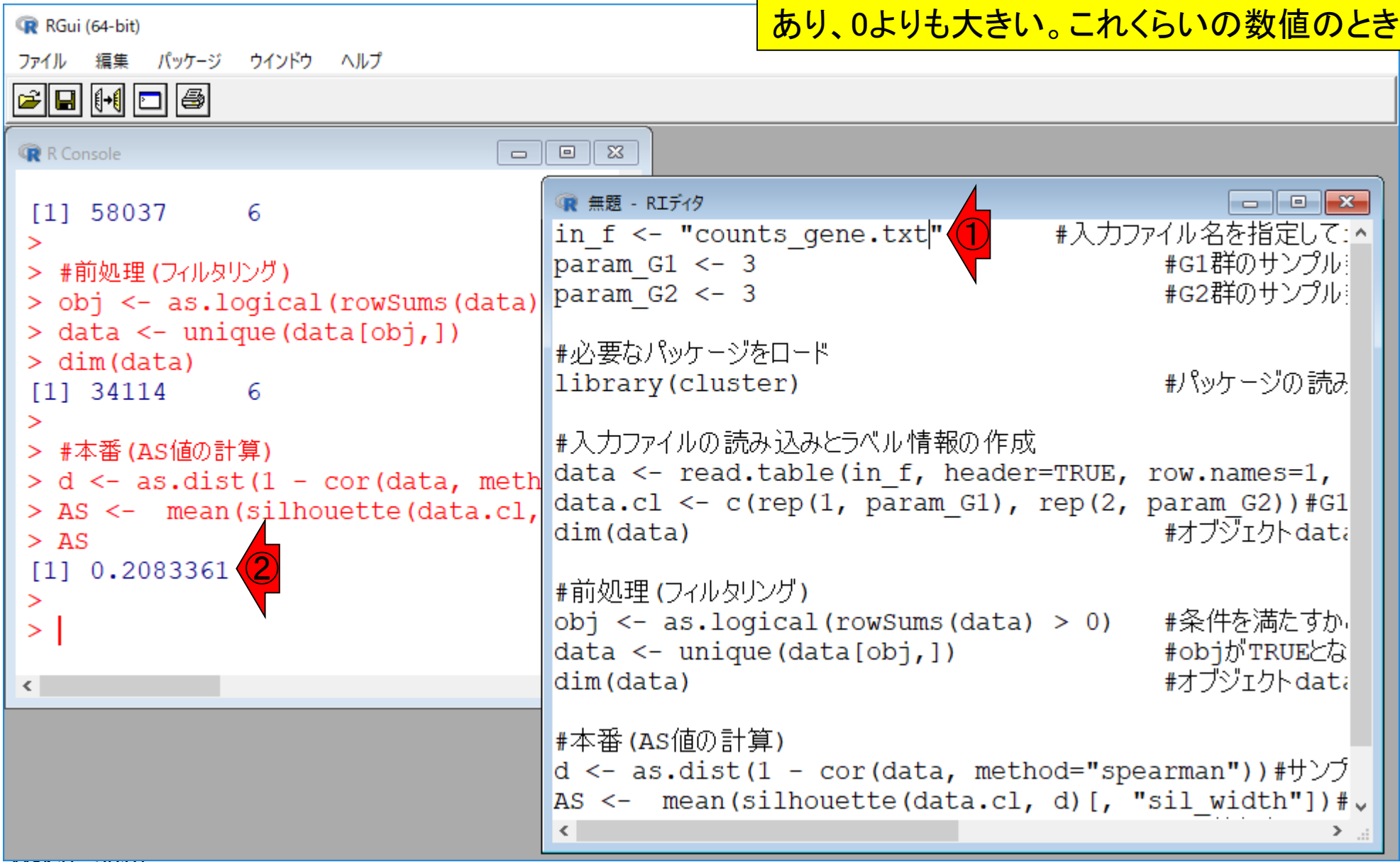
無題 - RIディタ

```
in_f <- "counts_gene.txt"  ①      #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"))#サンプ
AS <-  mean(silhouette(data.cl, d)[, "sil_width"])#
```

参考

シルエットスコア

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



シルエットスコア

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

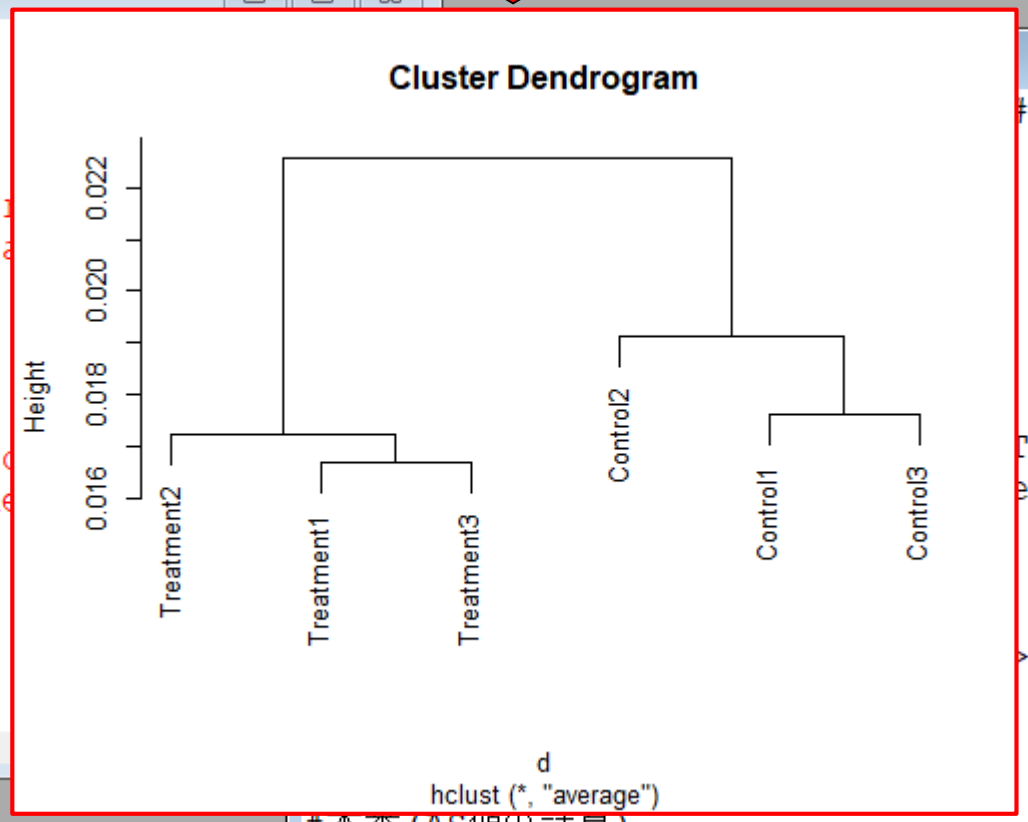
ファイル　編集　パッケージ　ウインドウ　ヘルプ

R Console

```
[1] 58037     6
>
> #前処理(フィルタリング)
> obj <- as.logical(r
> data <- unique(data
> dim(data)
[1] 34114     6
>
> #本番(AS値の計算)
> d <- as.dist(1 - co
> AS <-  mean(silhoue
> AS
[1] 0.2083361
>
>
```

②

③




```
#入力ファイル名を指定して:
                #G1群のサンプル
                #G2群のサンプル


                #パッケージの読み


TRUE, row.names=1,
ep(2, param_G2))#G1
                #オブジェクトdata


> 0)     #条件を満たすか
                #objがTRUEとな
                #オブジェクトdata
#本番(AS値の計算)
d <- as.dist(1 - cor(data, method="spearman"))#サンプ
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

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ペ。

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。これのGUI（グラフィカルユーザインターフェース）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 norm strategies.

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

Author information

①

Abstract

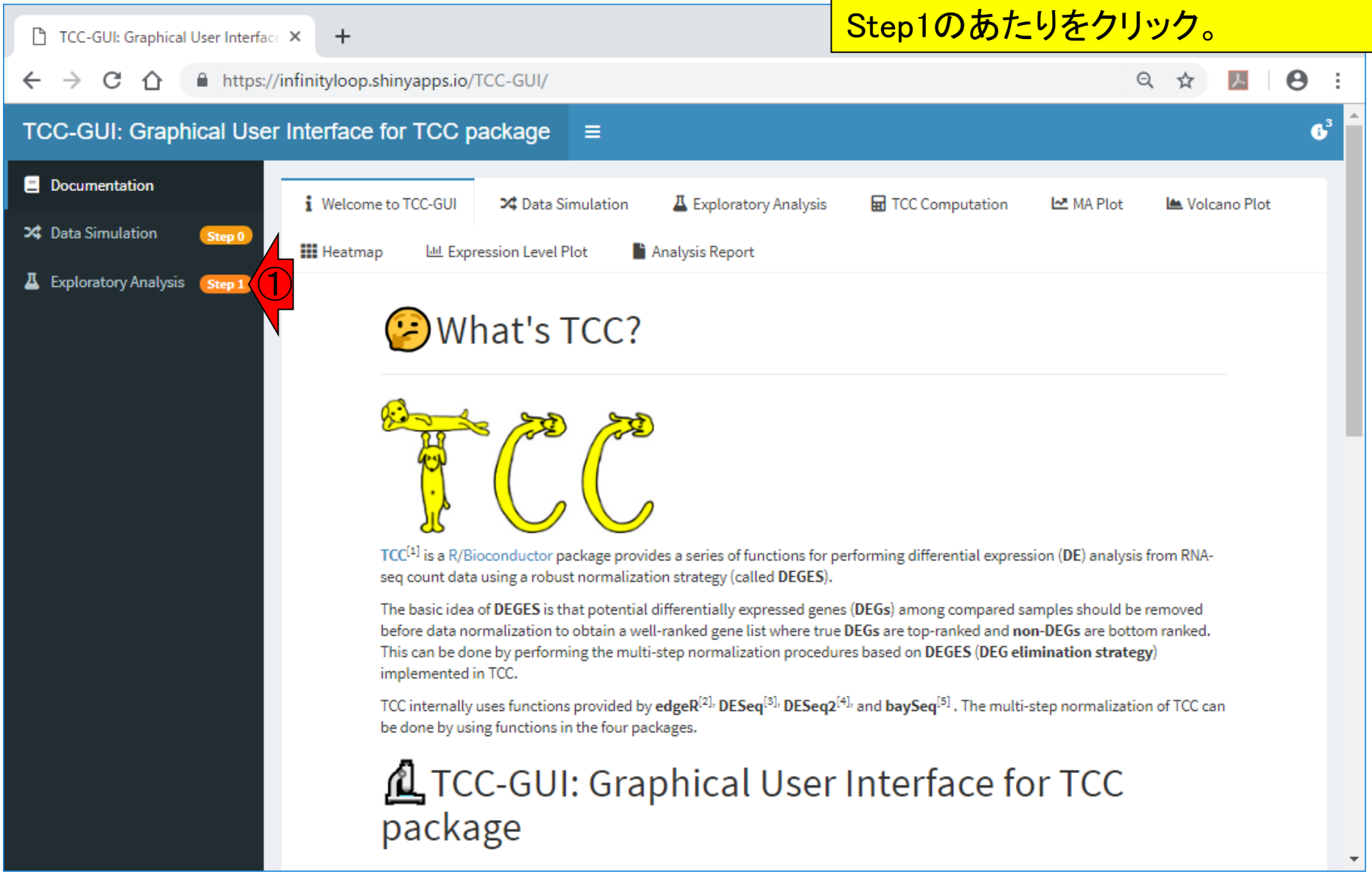
BACKGROUND: Differential expression analysis based on "next-generation" sequencing technologies is a fundamental means of studying RNA expression. We recently developed a multi-step normalization method (called TbT) for two-group RNA-seq data with replicates and demonstrated that the statistical methods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

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

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

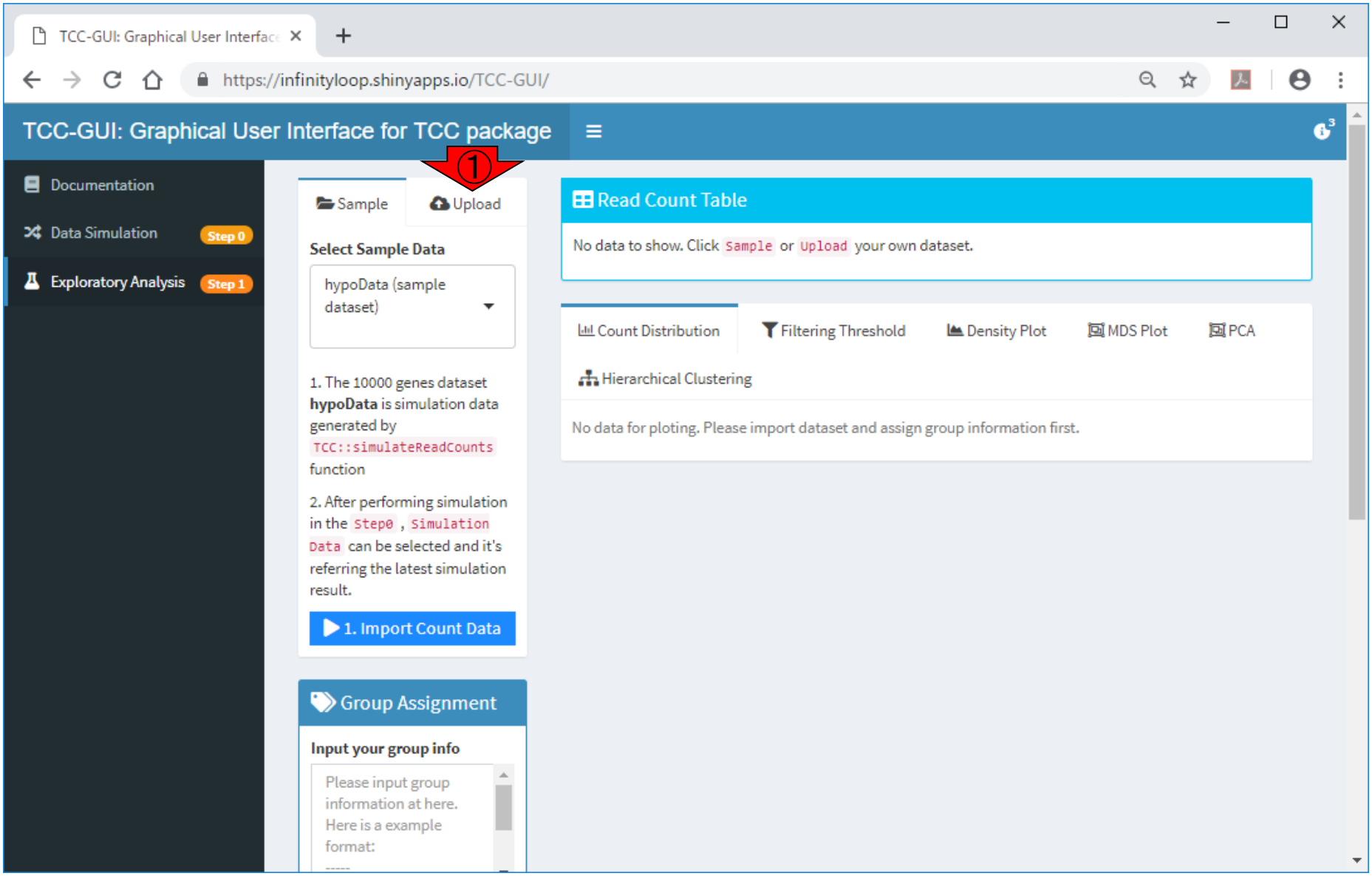
TCC-GUI

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。①Step1のあたりをクリック。



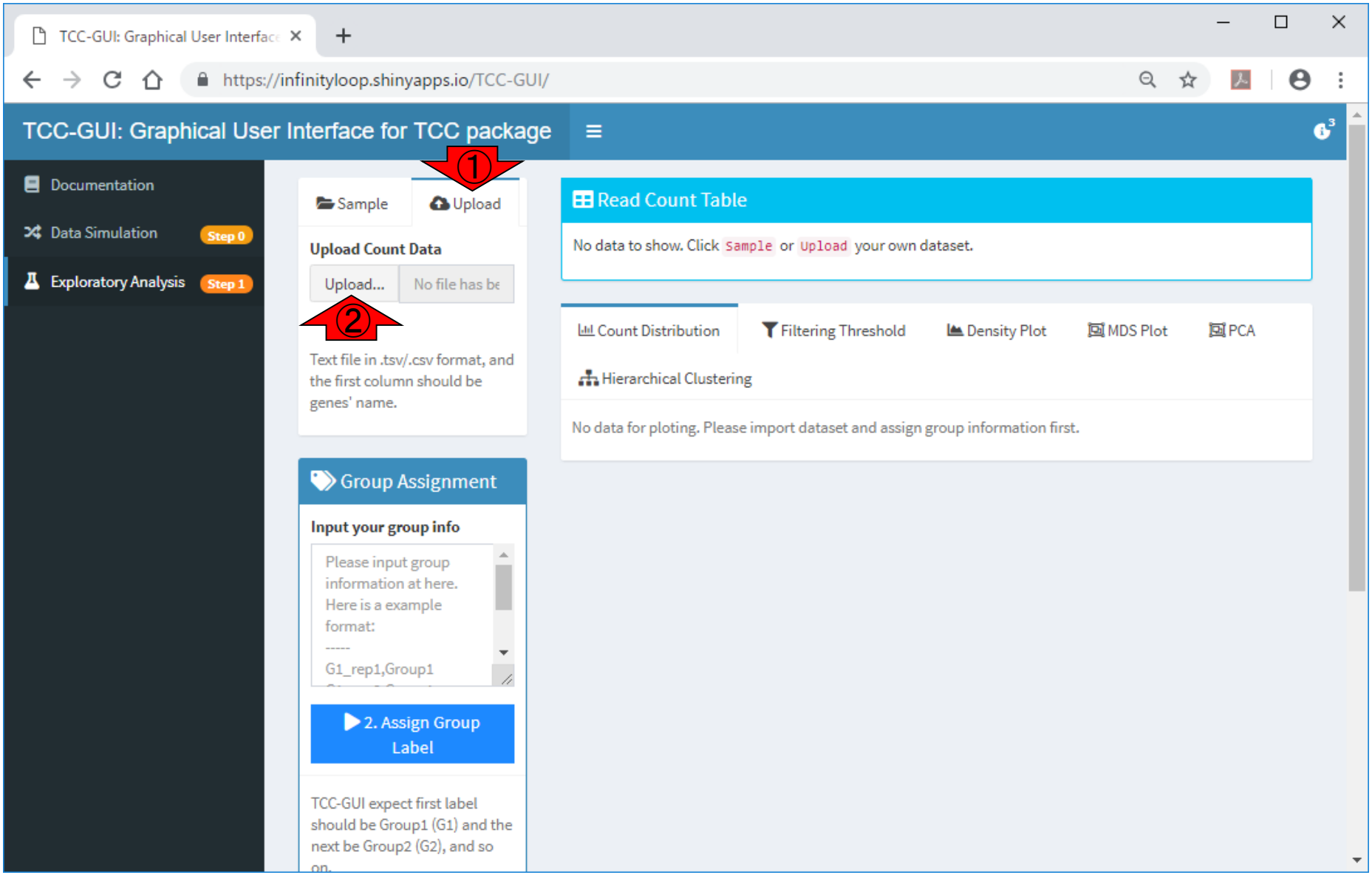
こんな感じの画面に飛びます。①Upload

TCC-GUI



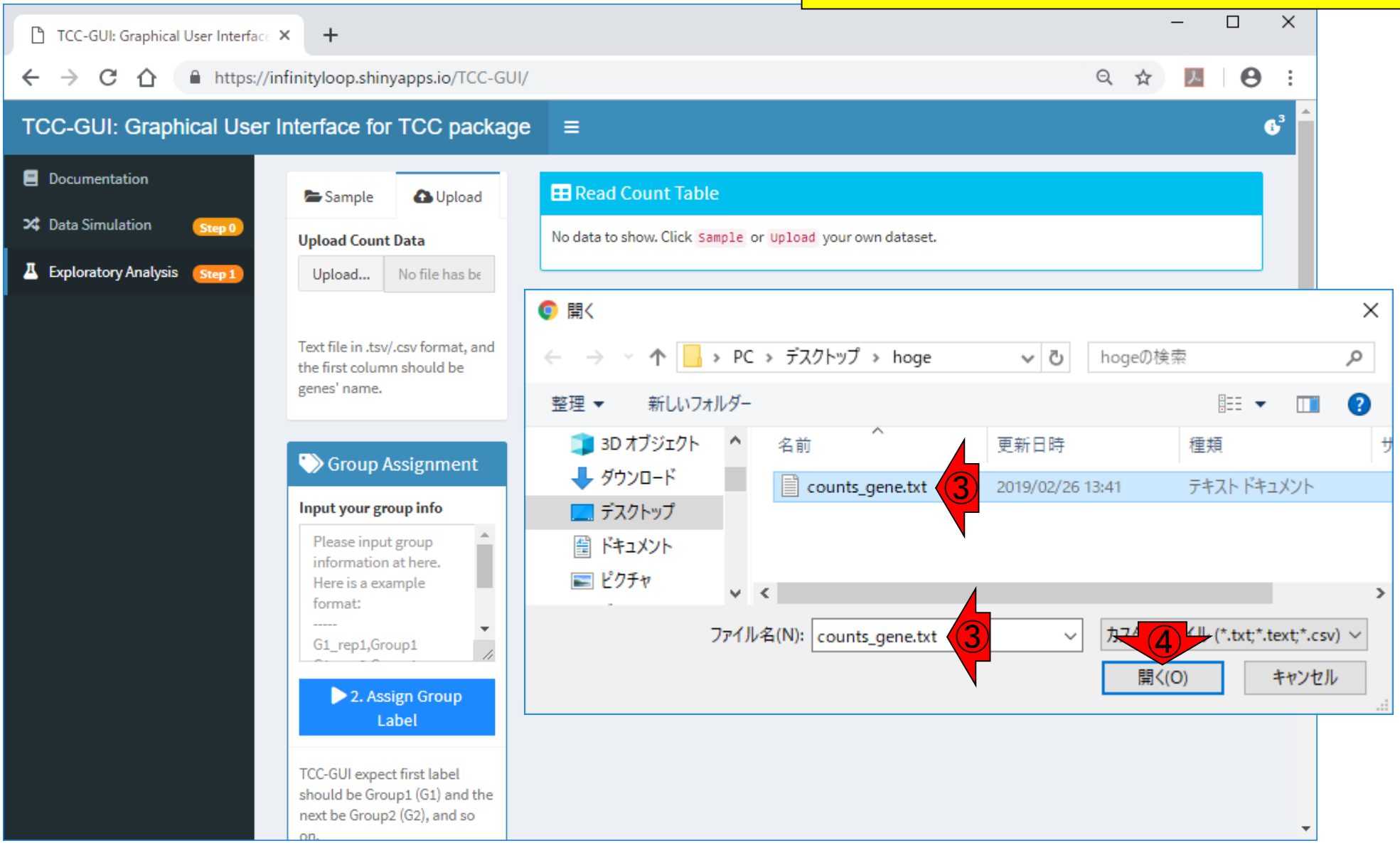
TCC-GUI

こんな感じの画面に飛びます。①Upload。ここでさらに②Upload。



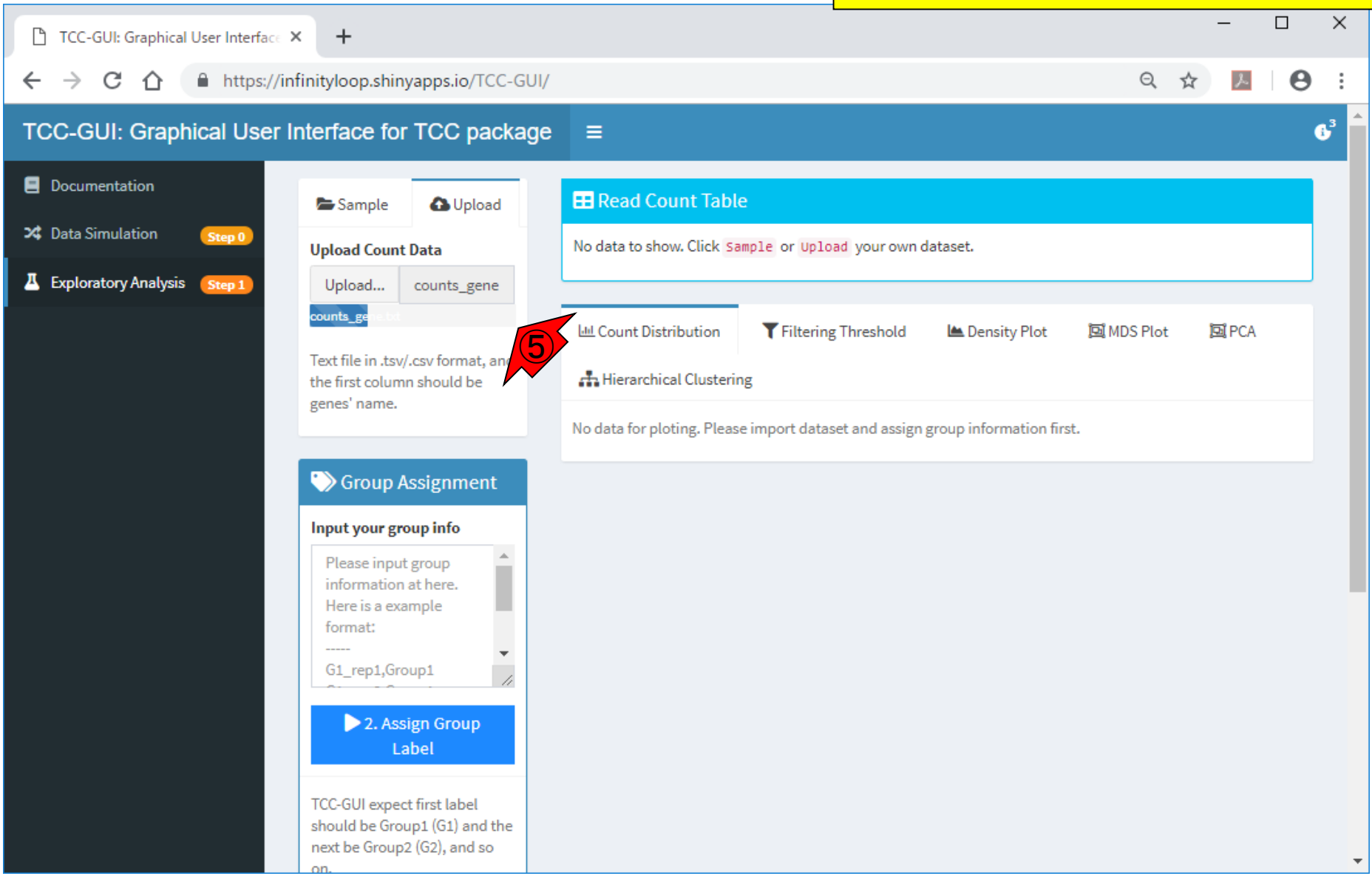
TCC-GUI

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。



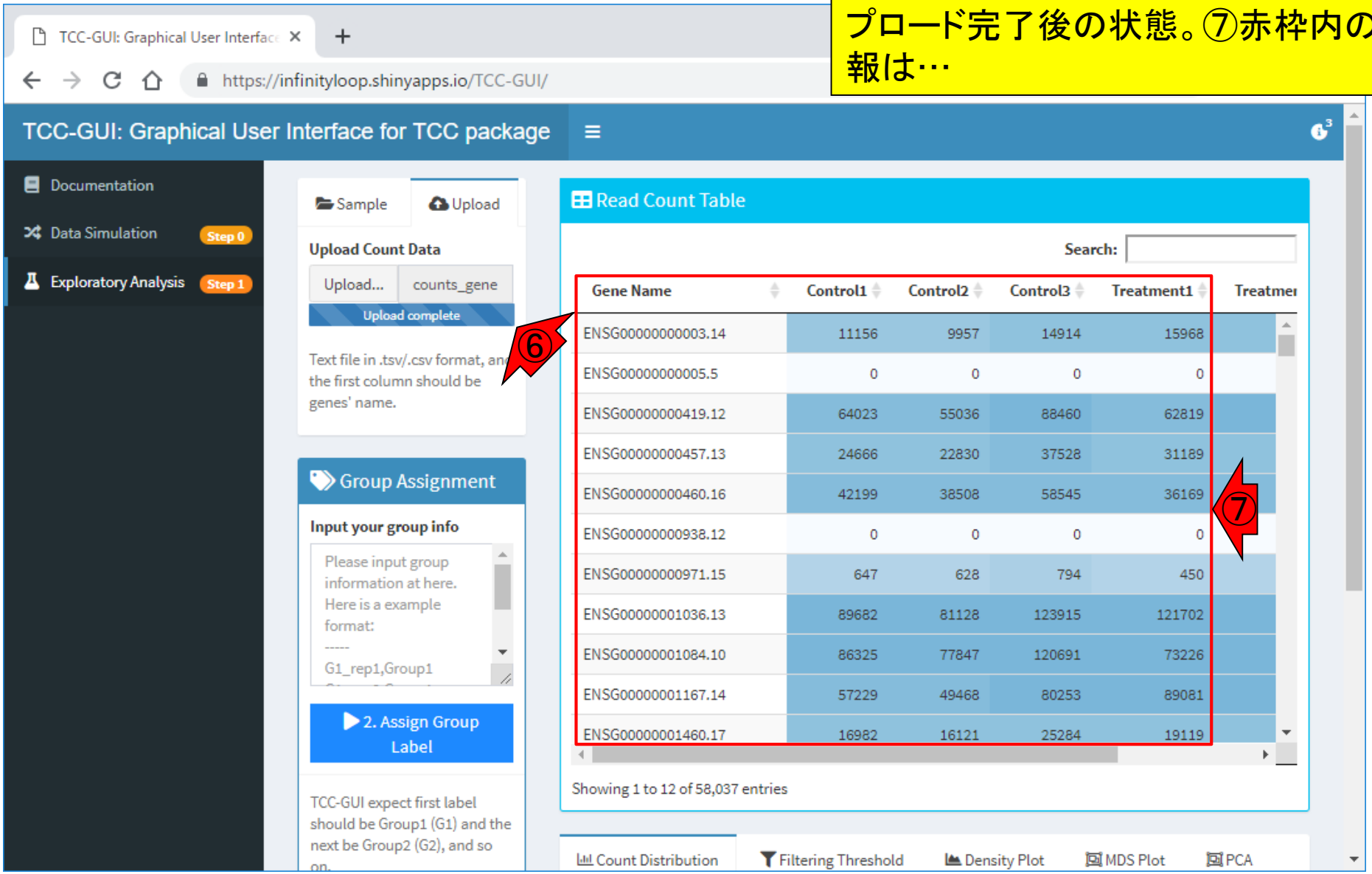
TCC-GUI

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…



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

⑦

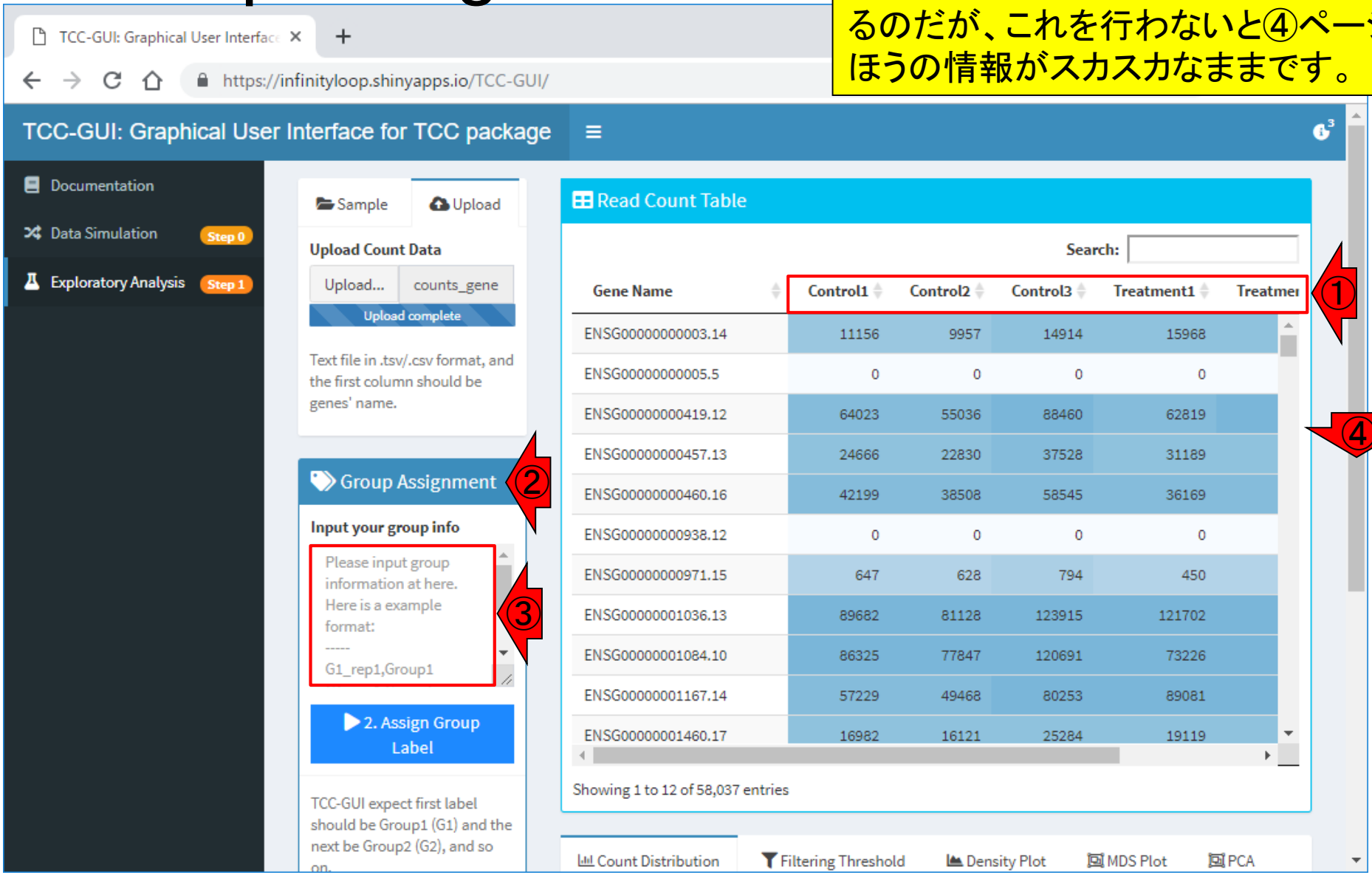
counts_gene.txt

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

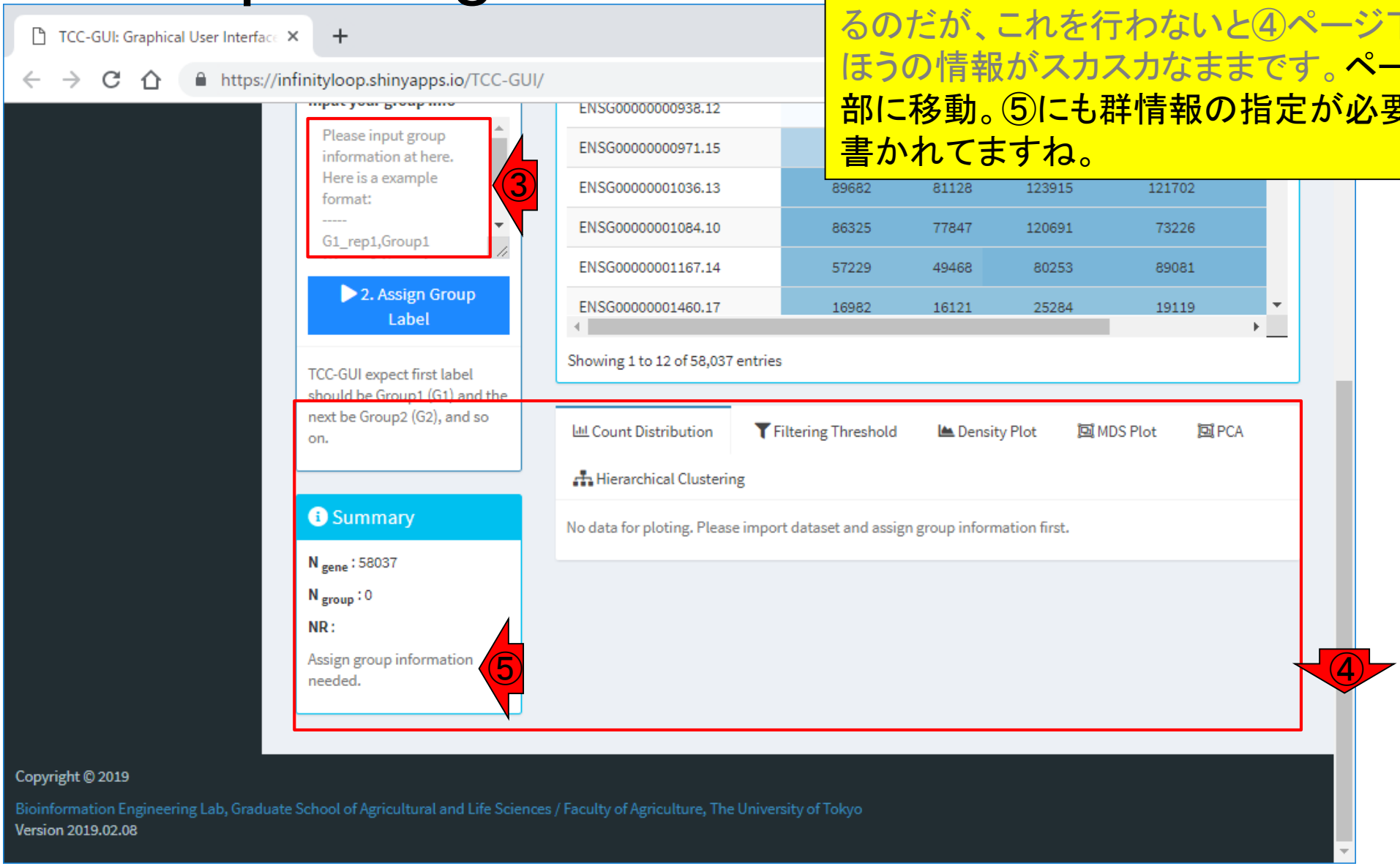
Group Assignment

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。



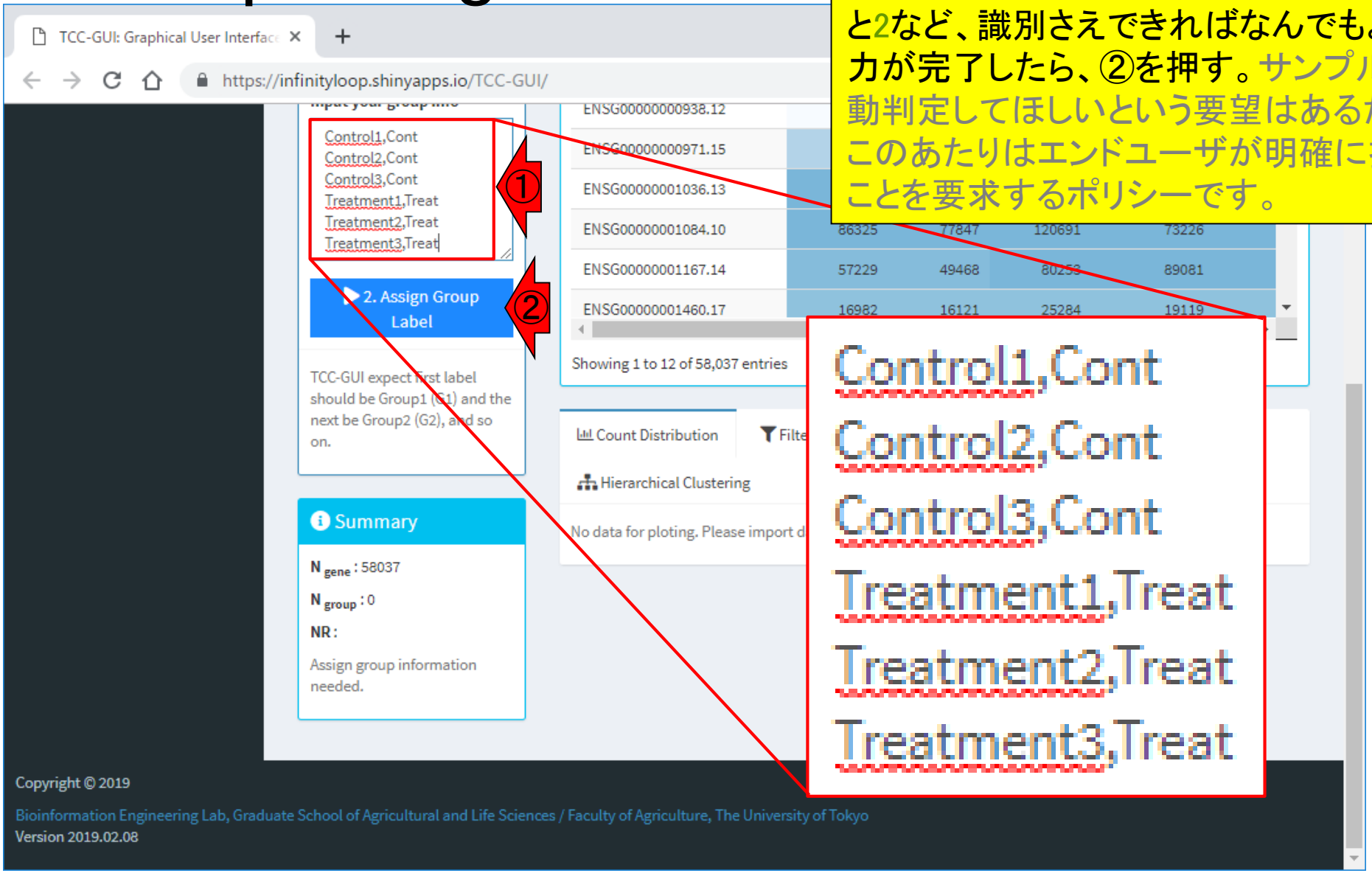
Group Assignment

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。



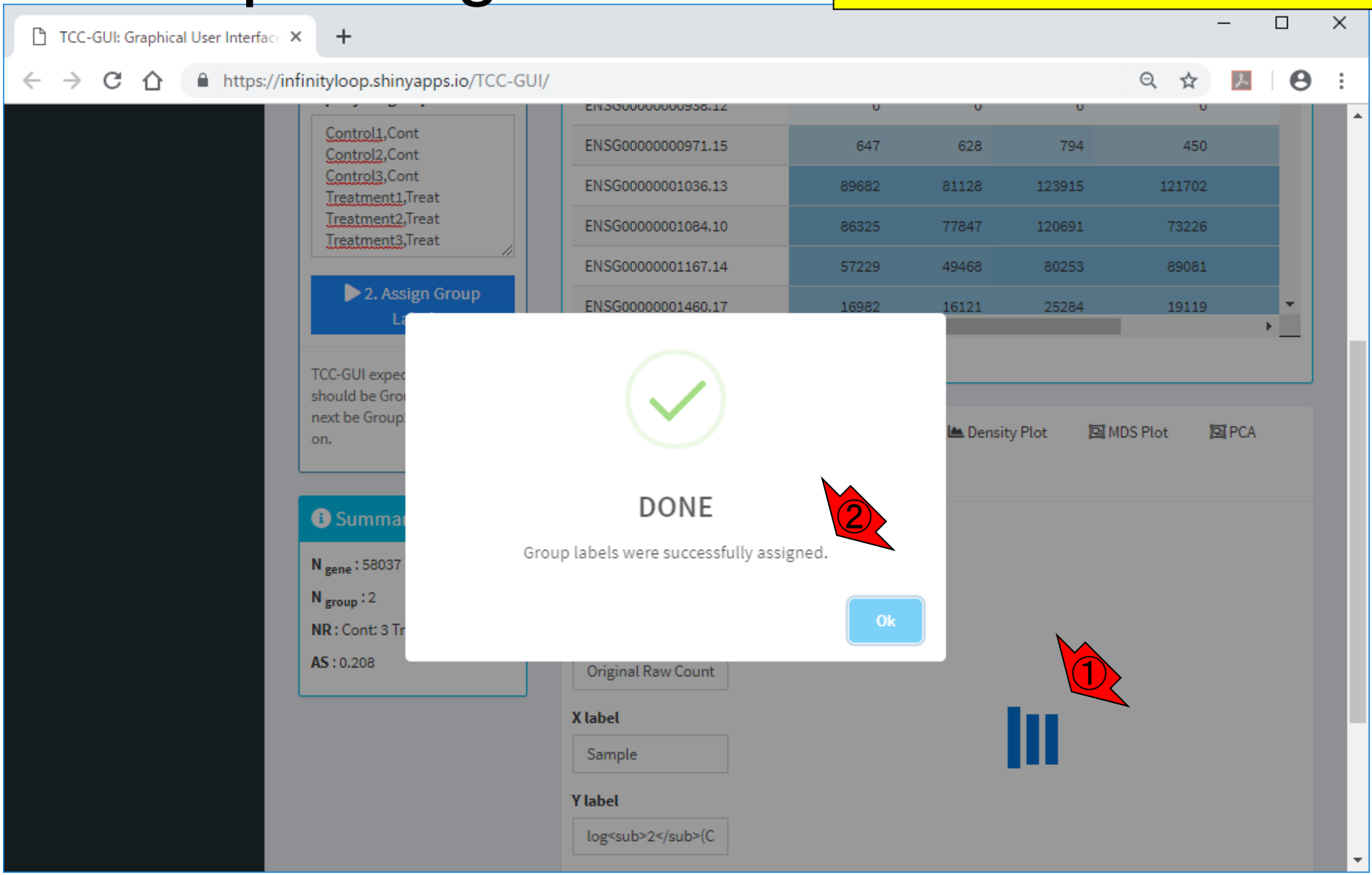
Group Assignment

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。



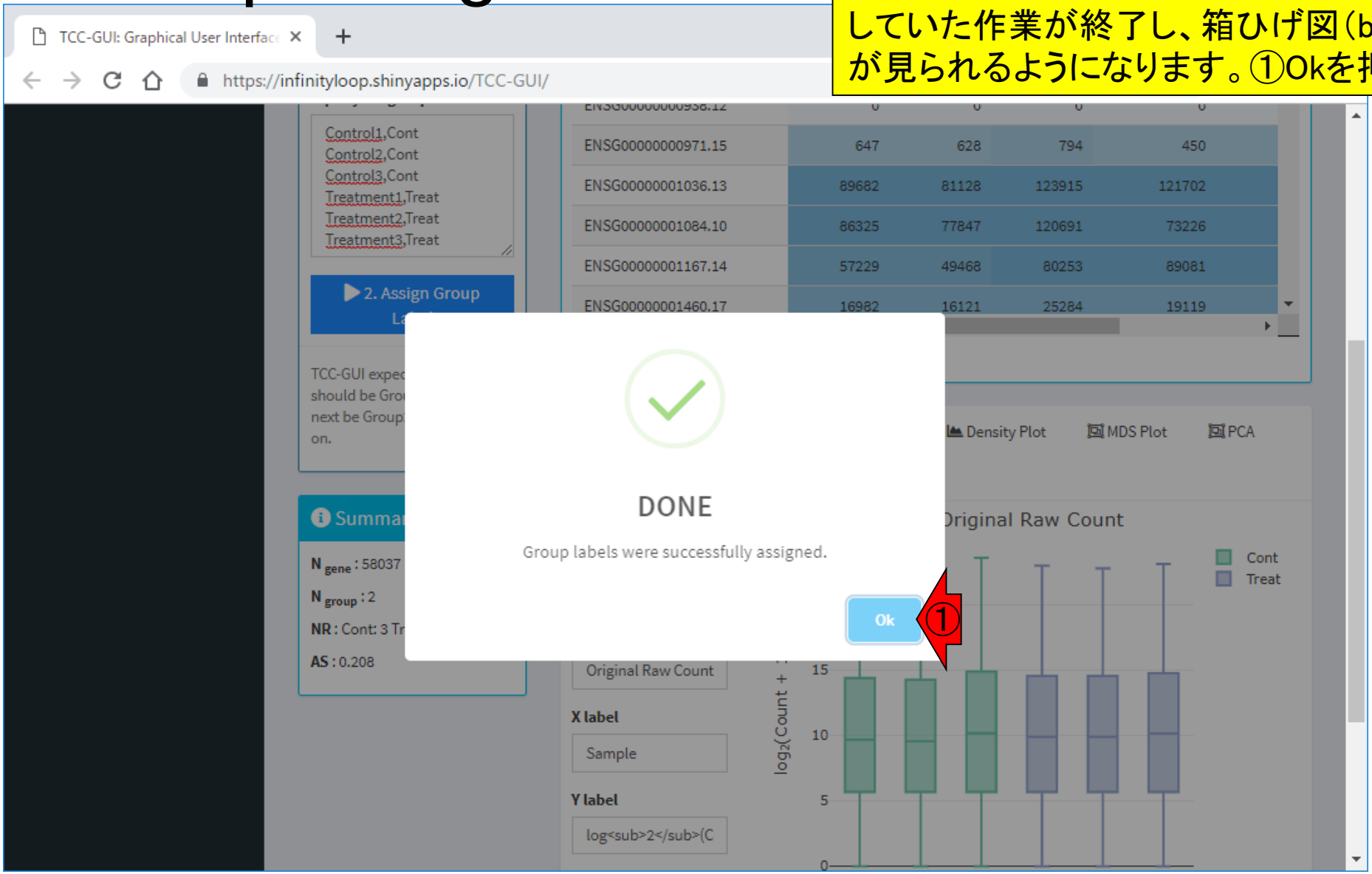
Group Assignment後

①このあたりがごにょごにょなって、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。



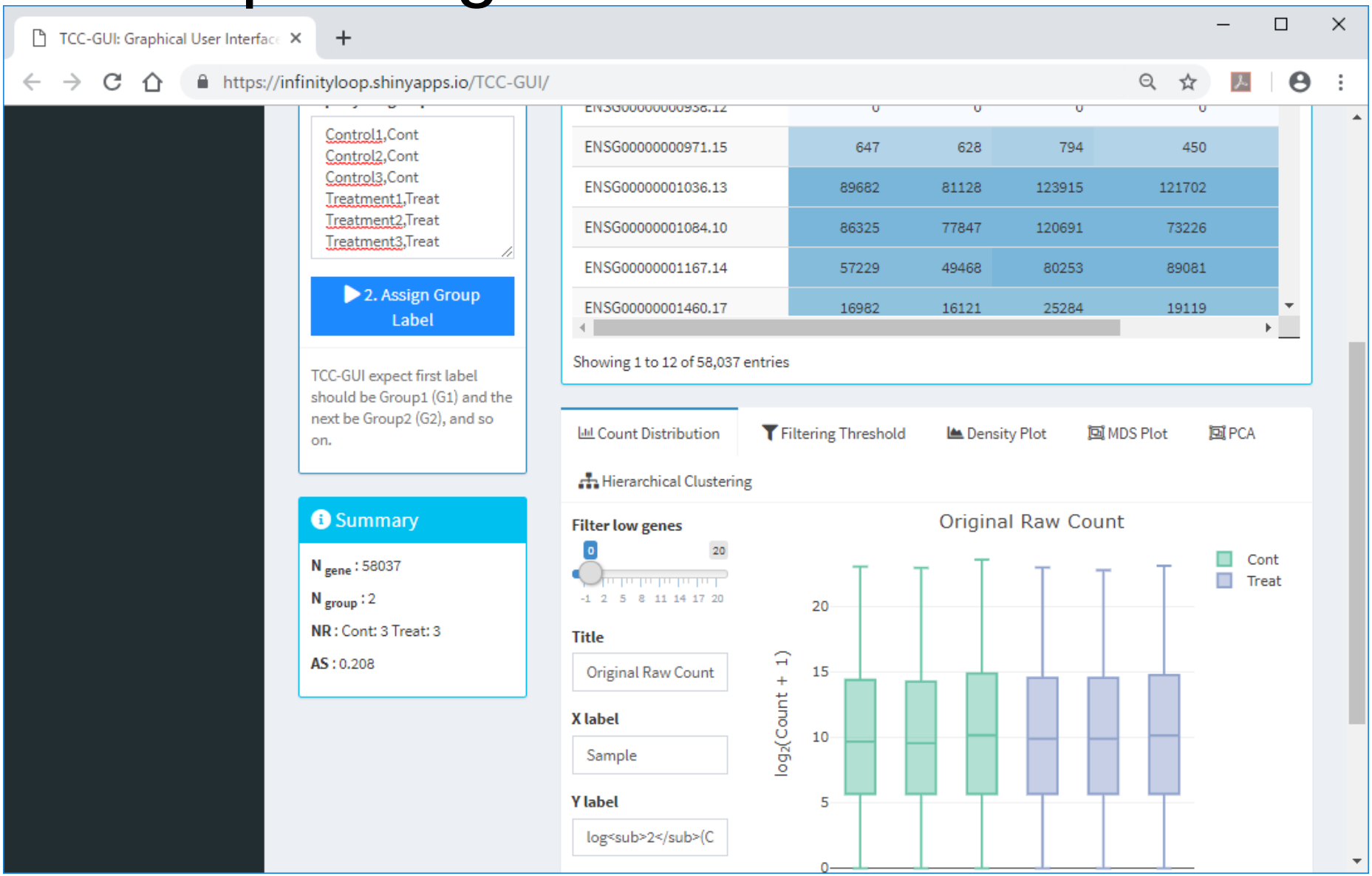
Group Assignment後

①このあたりがごにょごにょなって、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ょごにょ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…



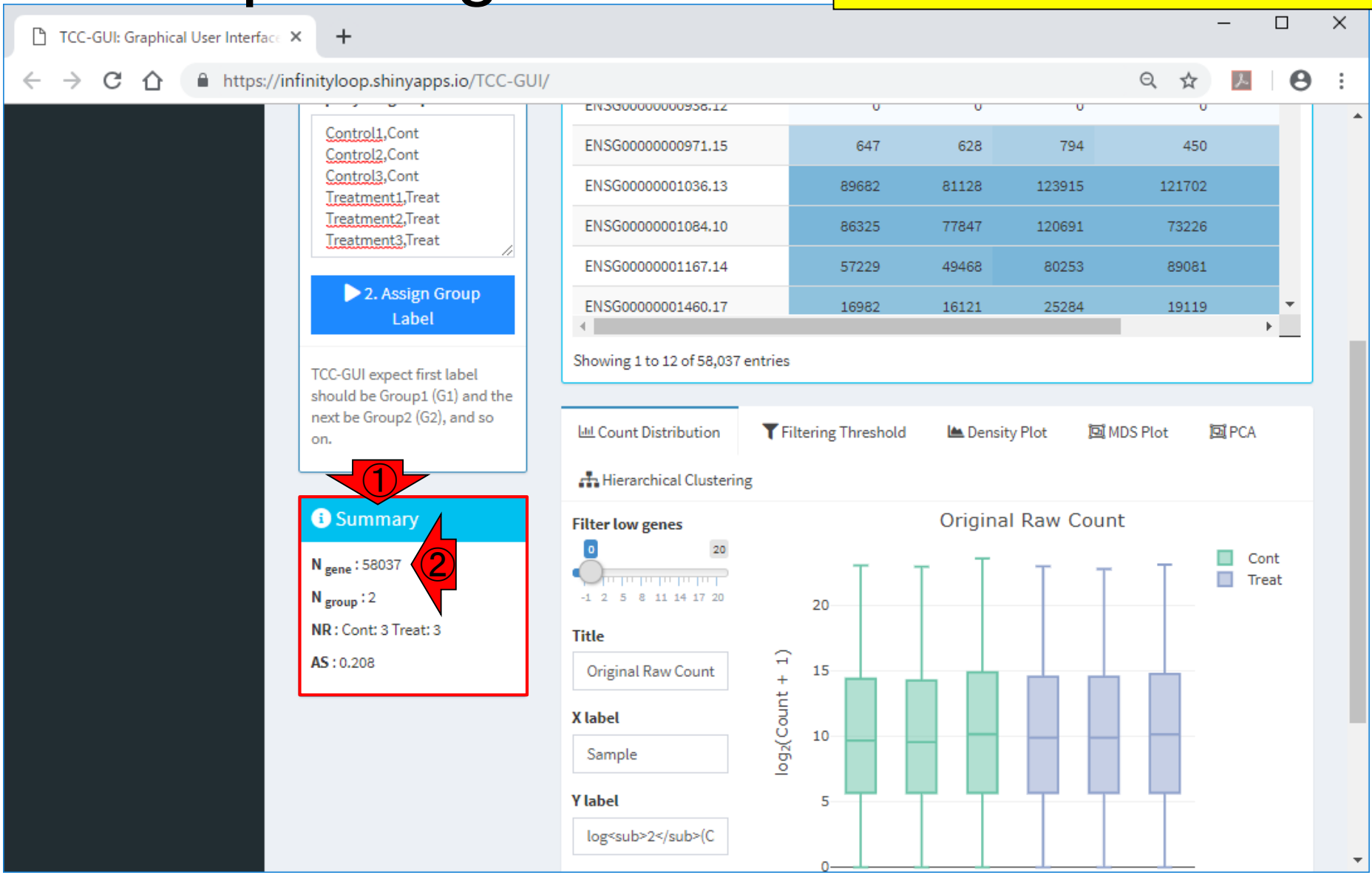
こんな感じになります。

Group Assignment後



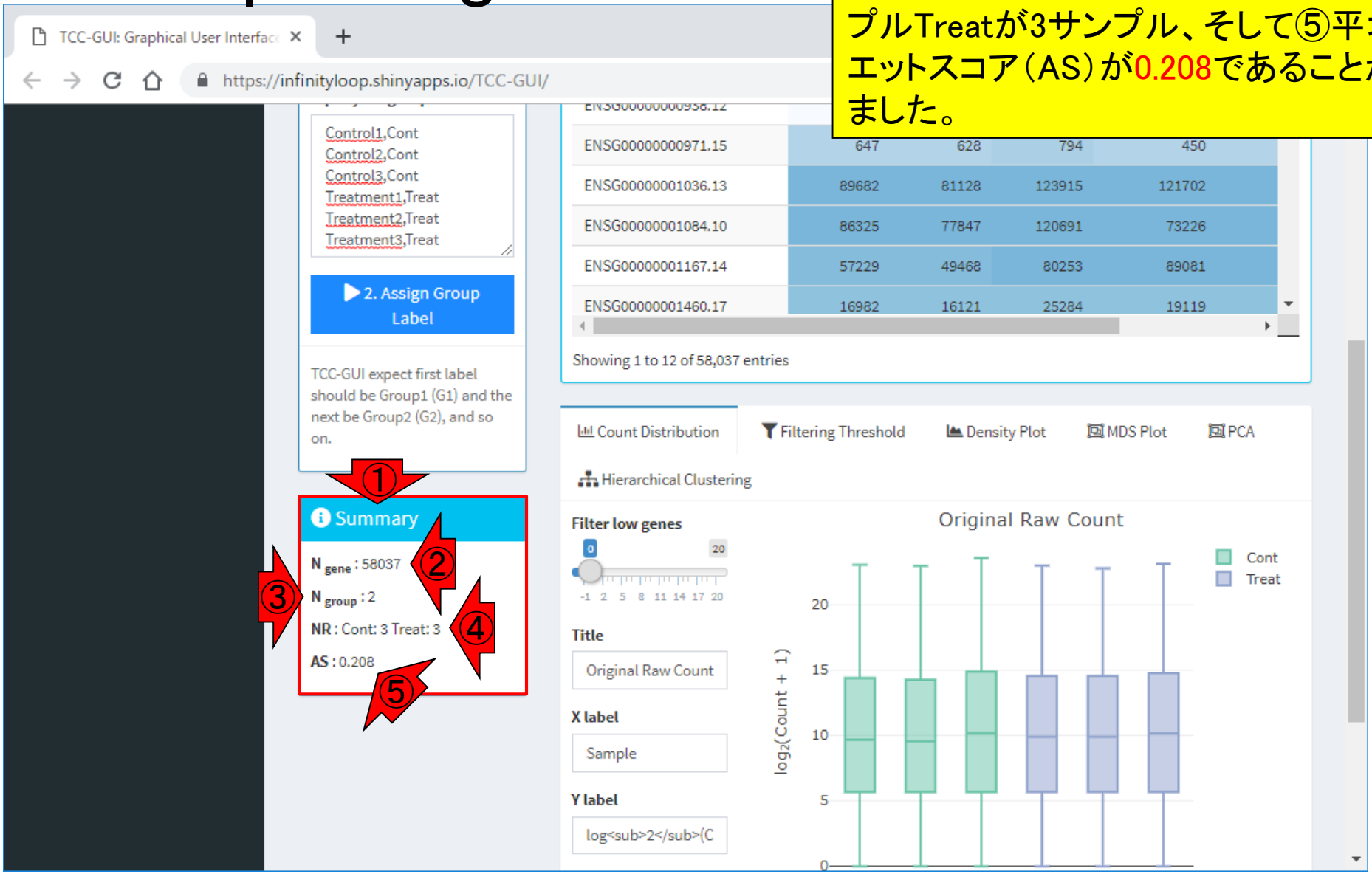
Group Assignment後

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、



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。



参考

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　ウインドウ　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
>
>
```

②

無題 - RIディタ

```
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"))#サンプ
AS <-  mean(silhouette(data.cl, d)[, "sil_width"])#
```

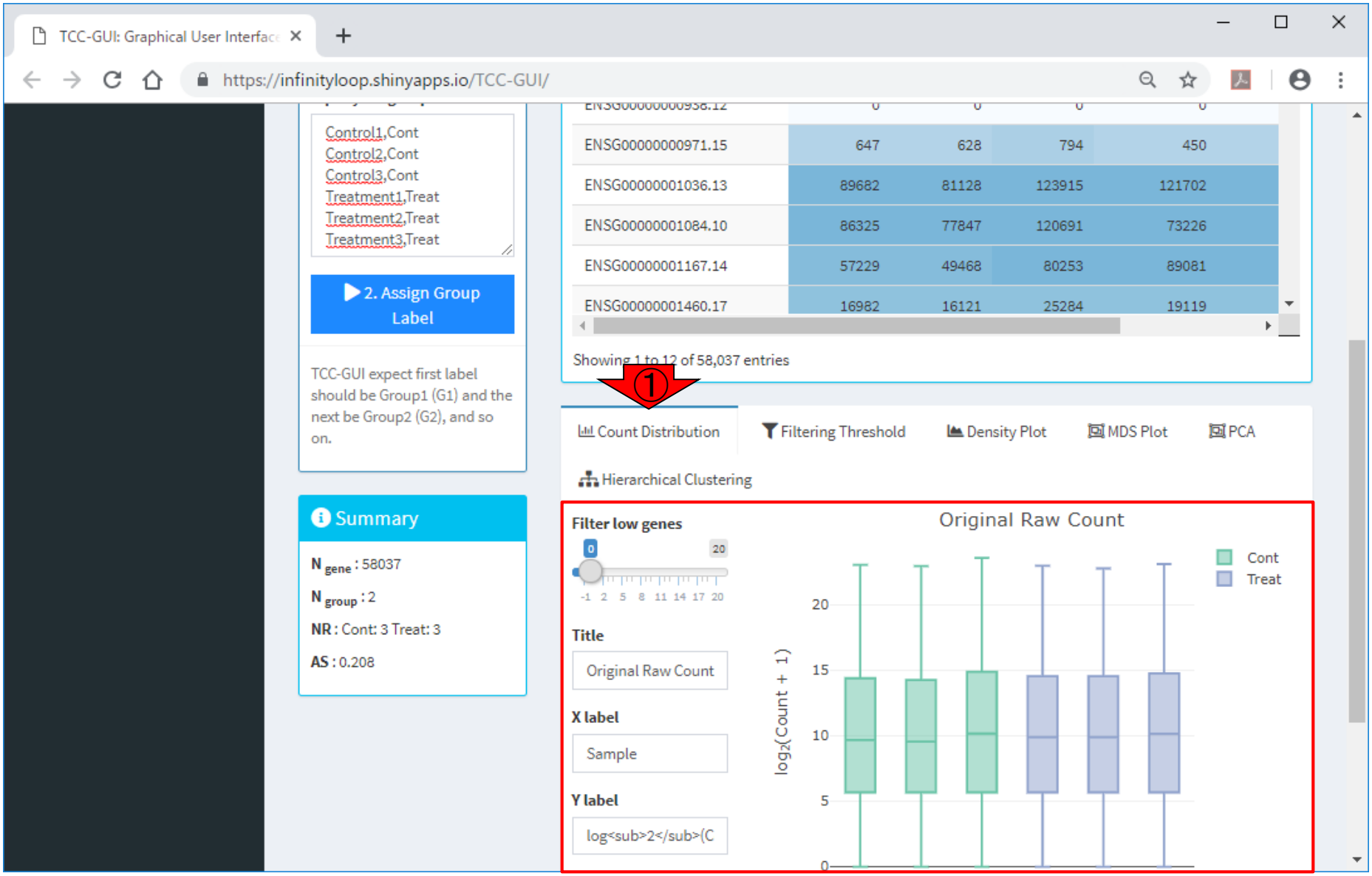
①

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

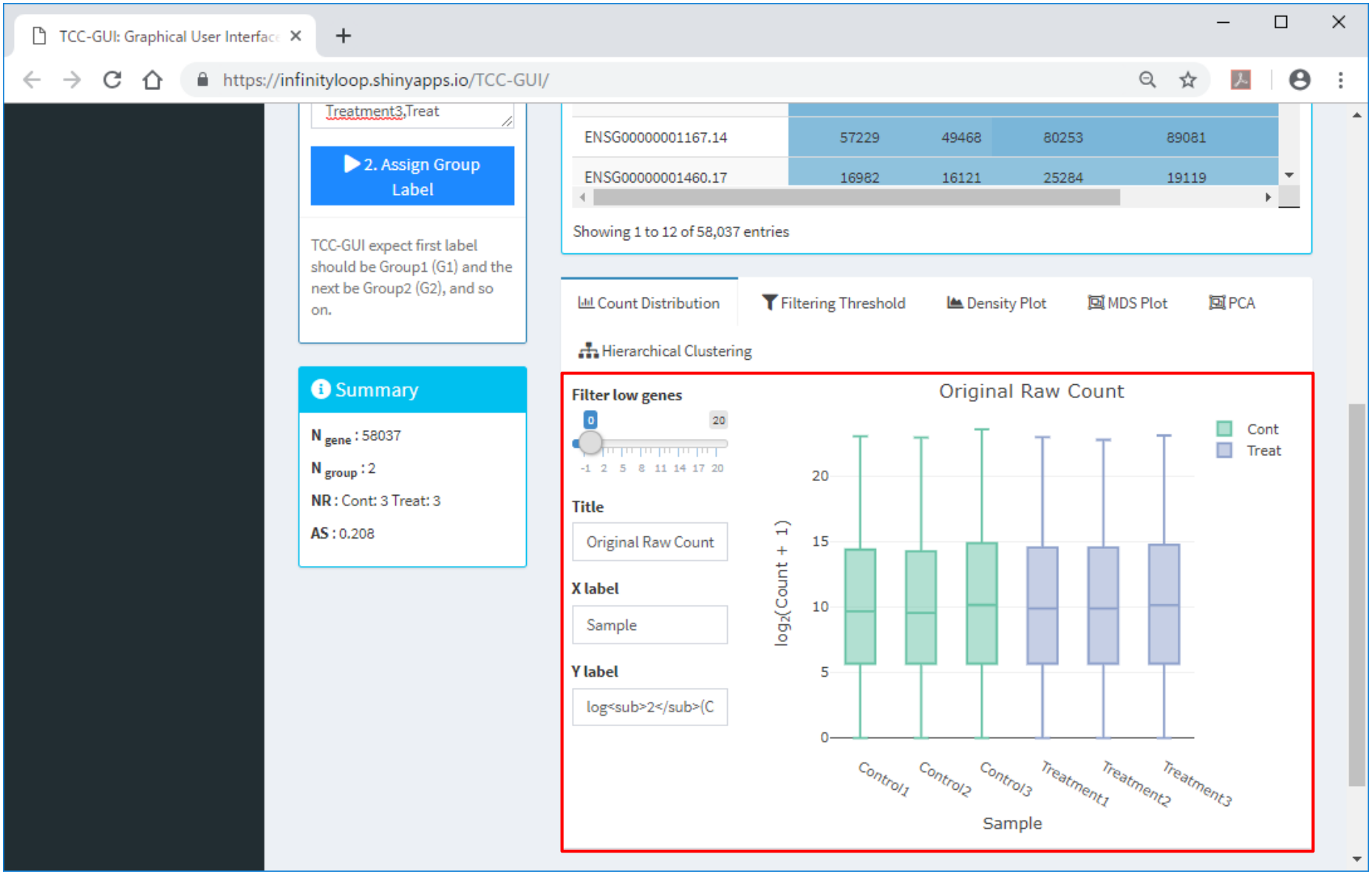
Count Distribution

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。



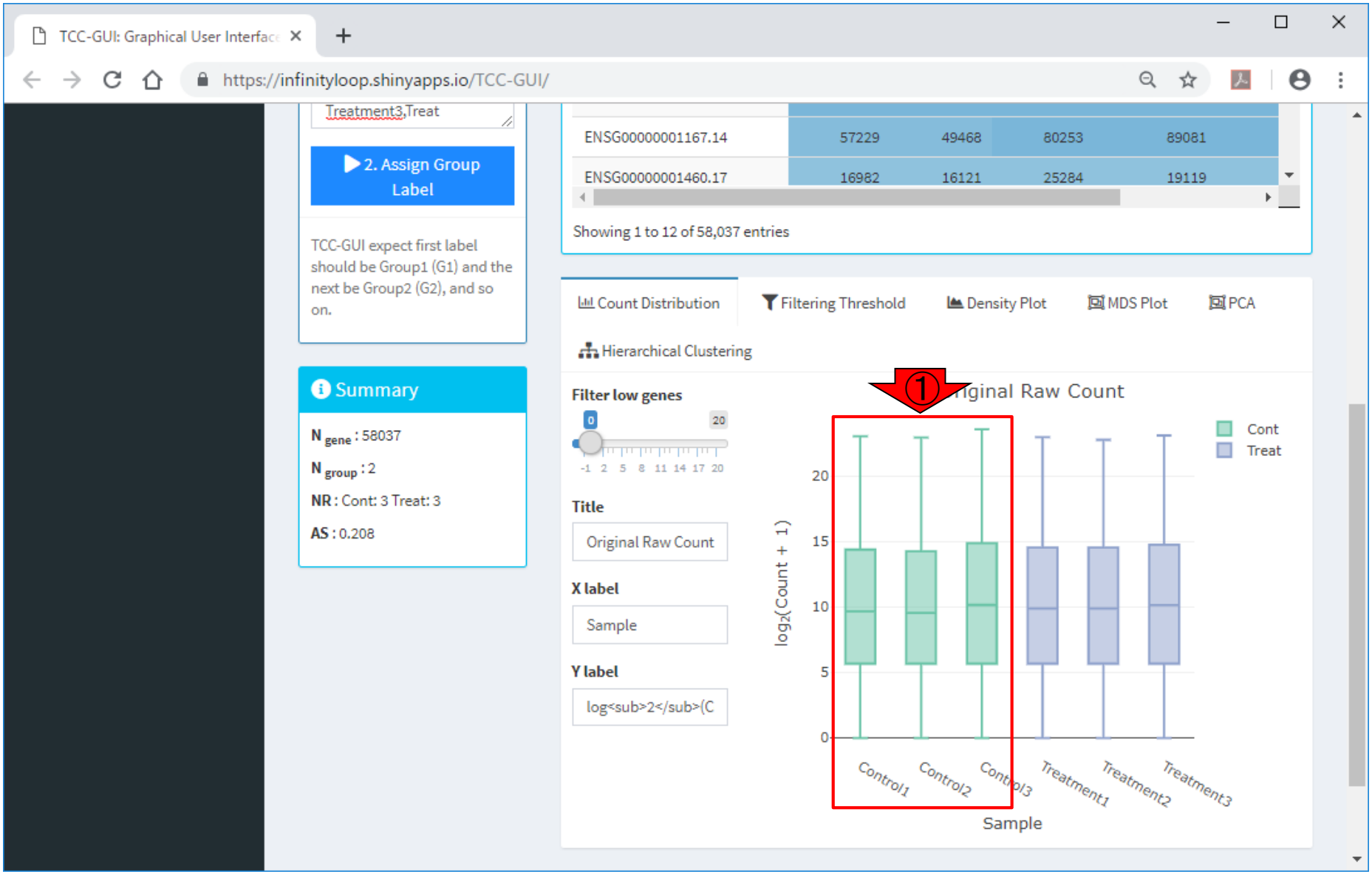
こんな感じになります。

Count Distribution



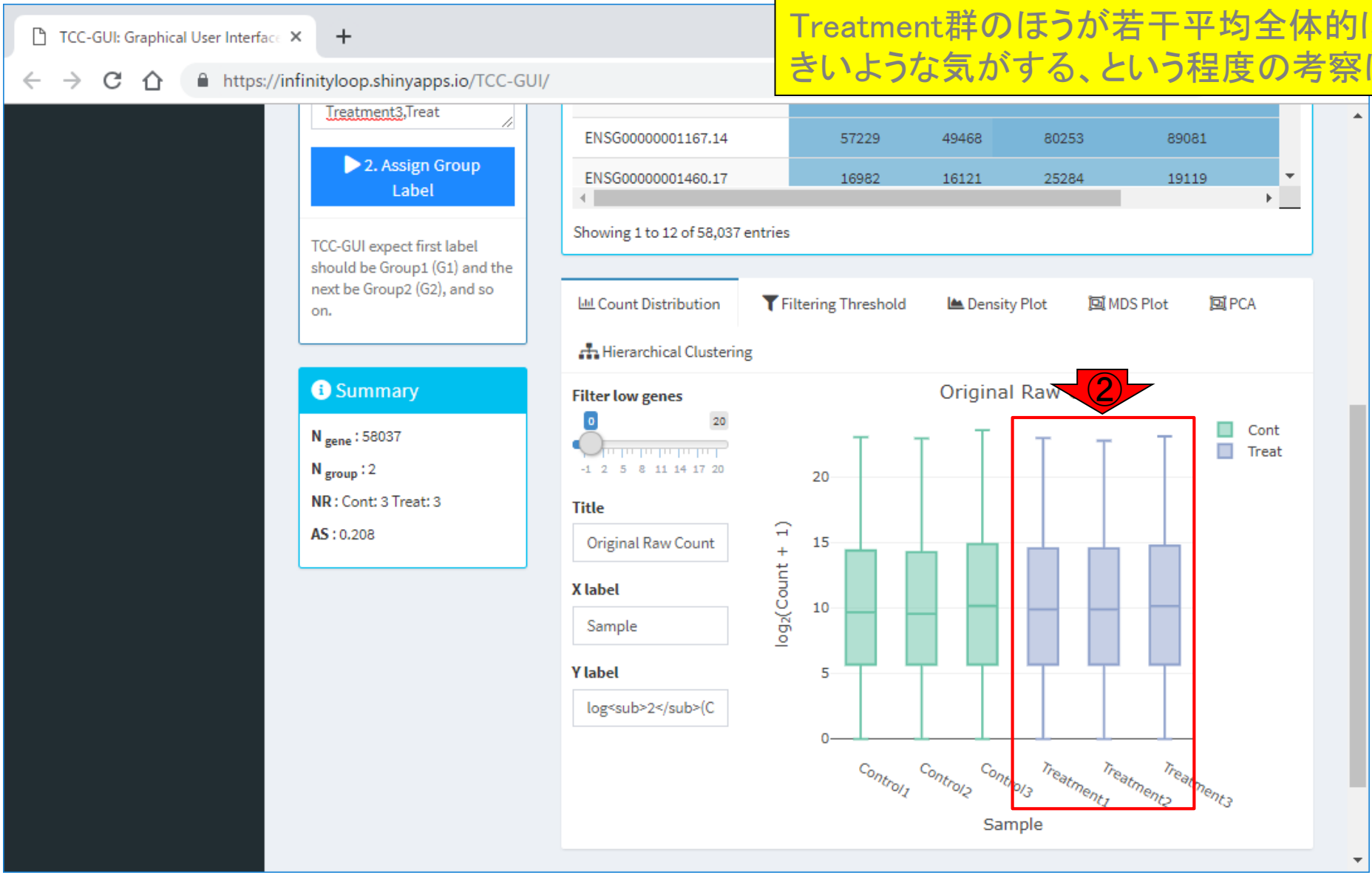
Count Distribution

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。



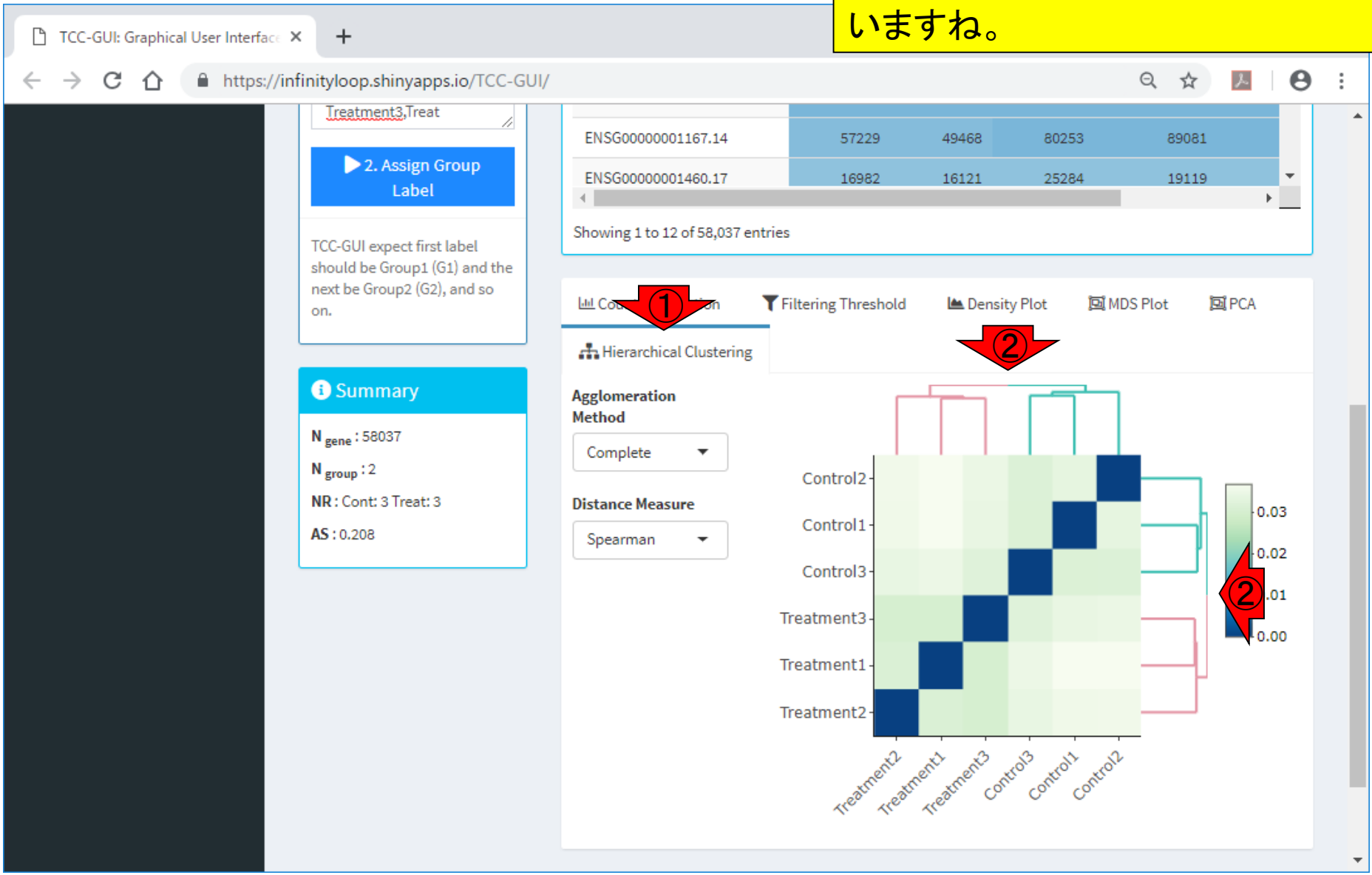
Count Distribution

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。



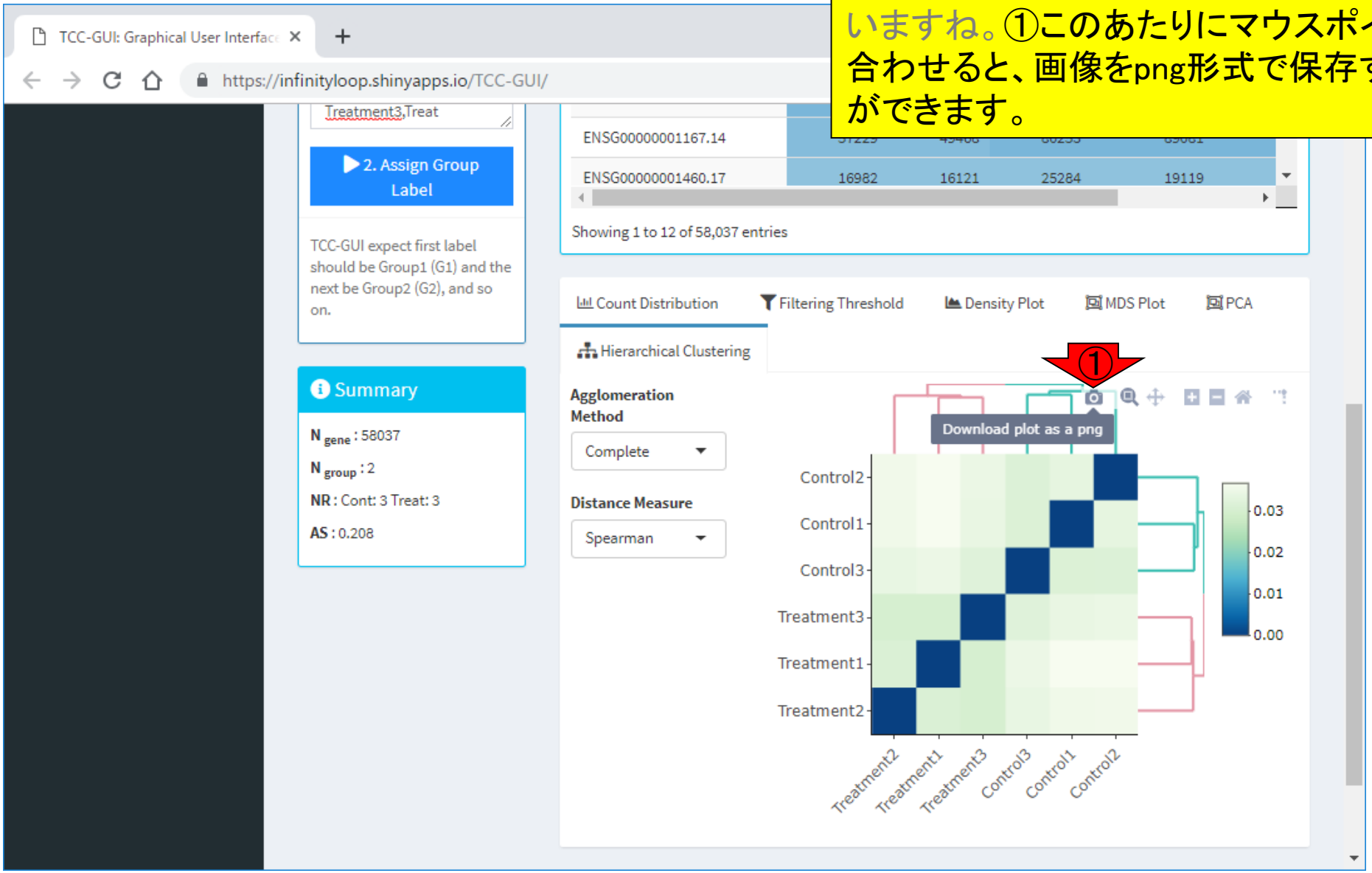
階層的クラスタリング

①Hierarchical 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。



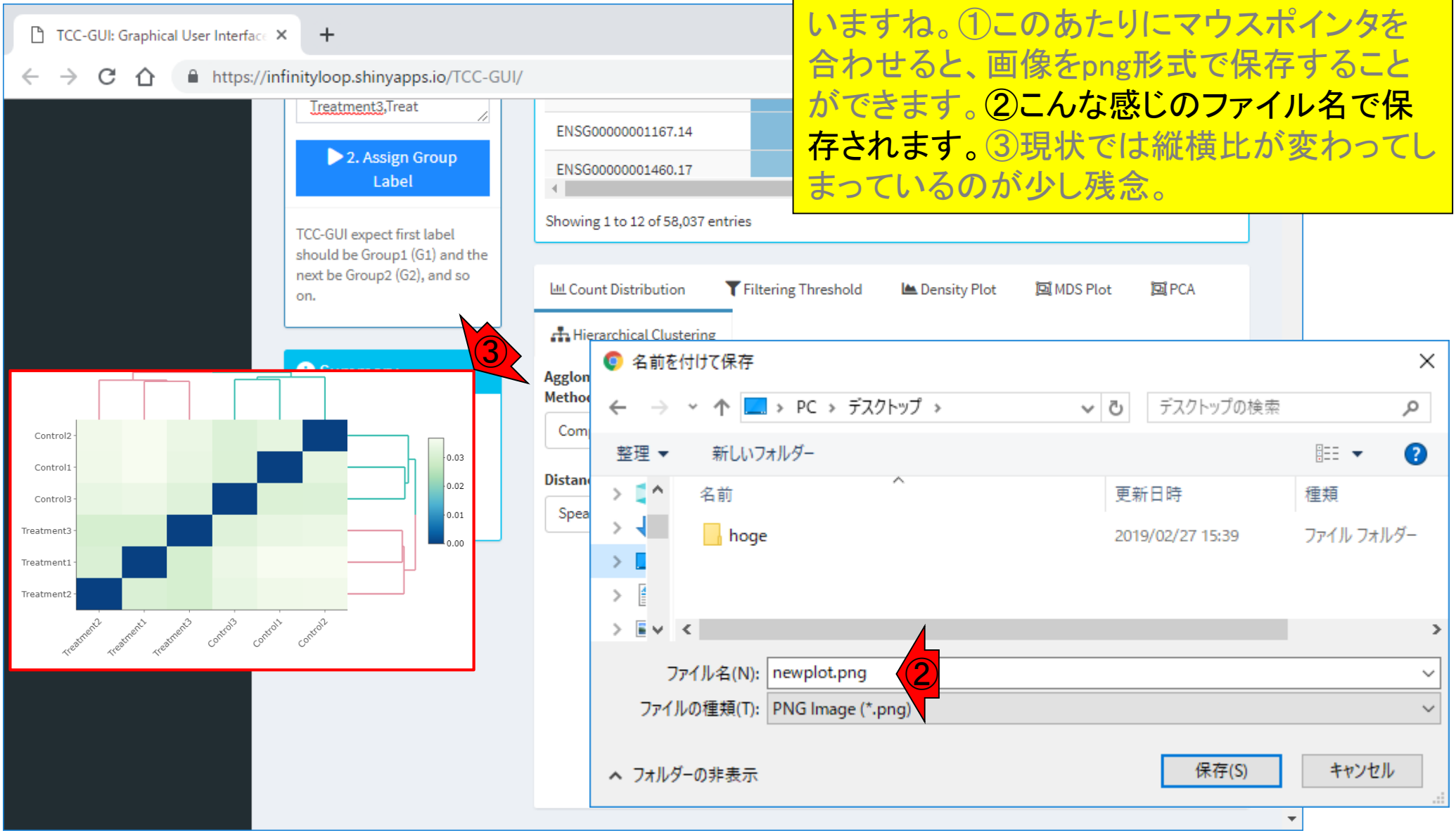
階層的クラスタリング

①Hierarchical 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。①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。



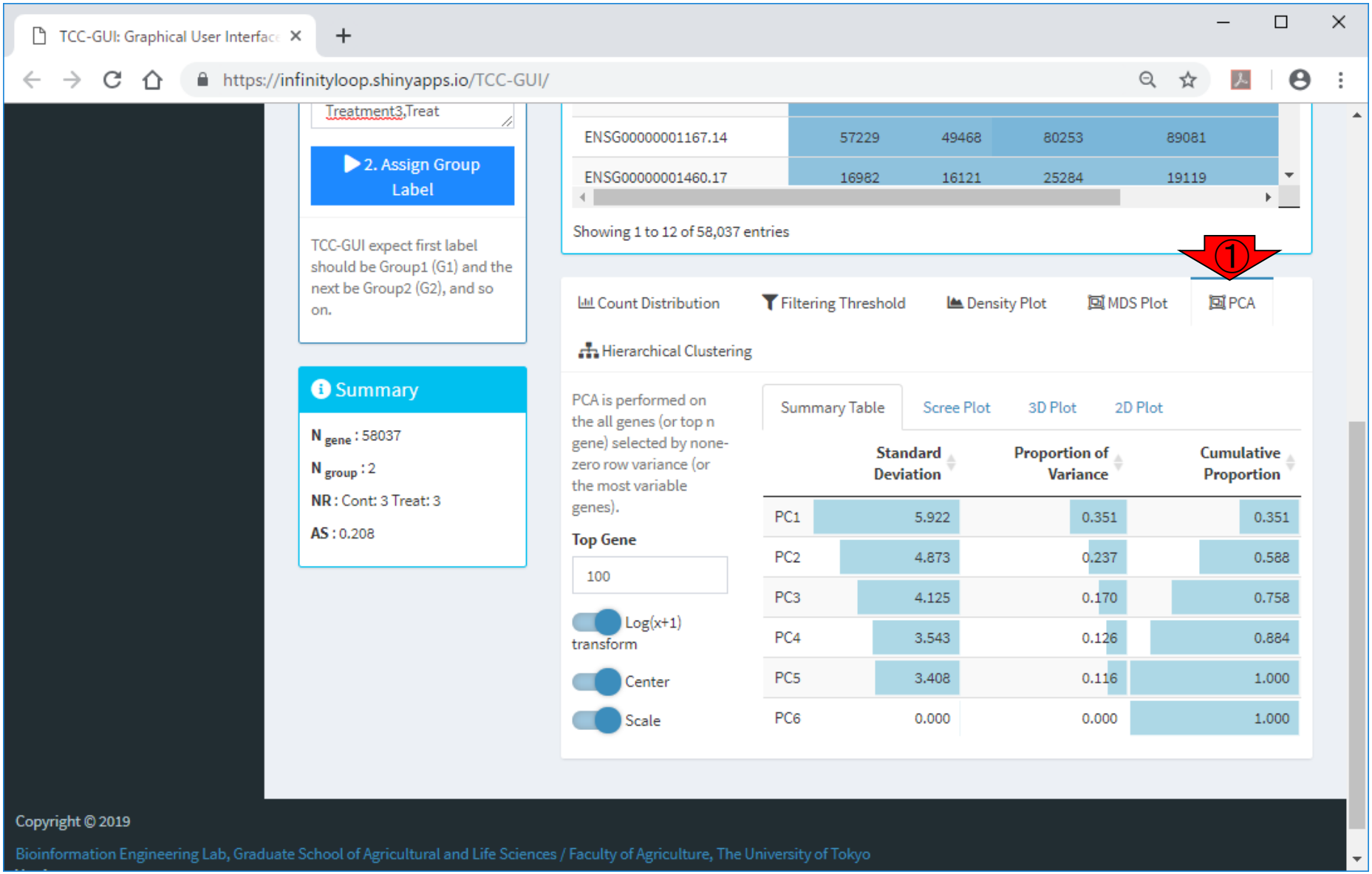
階層的クラスタリング

①Hierarchical 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。①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。②こんな感じのファイル名で保存されます。③現状では縦横比が変わってしまっているのが少し残念。



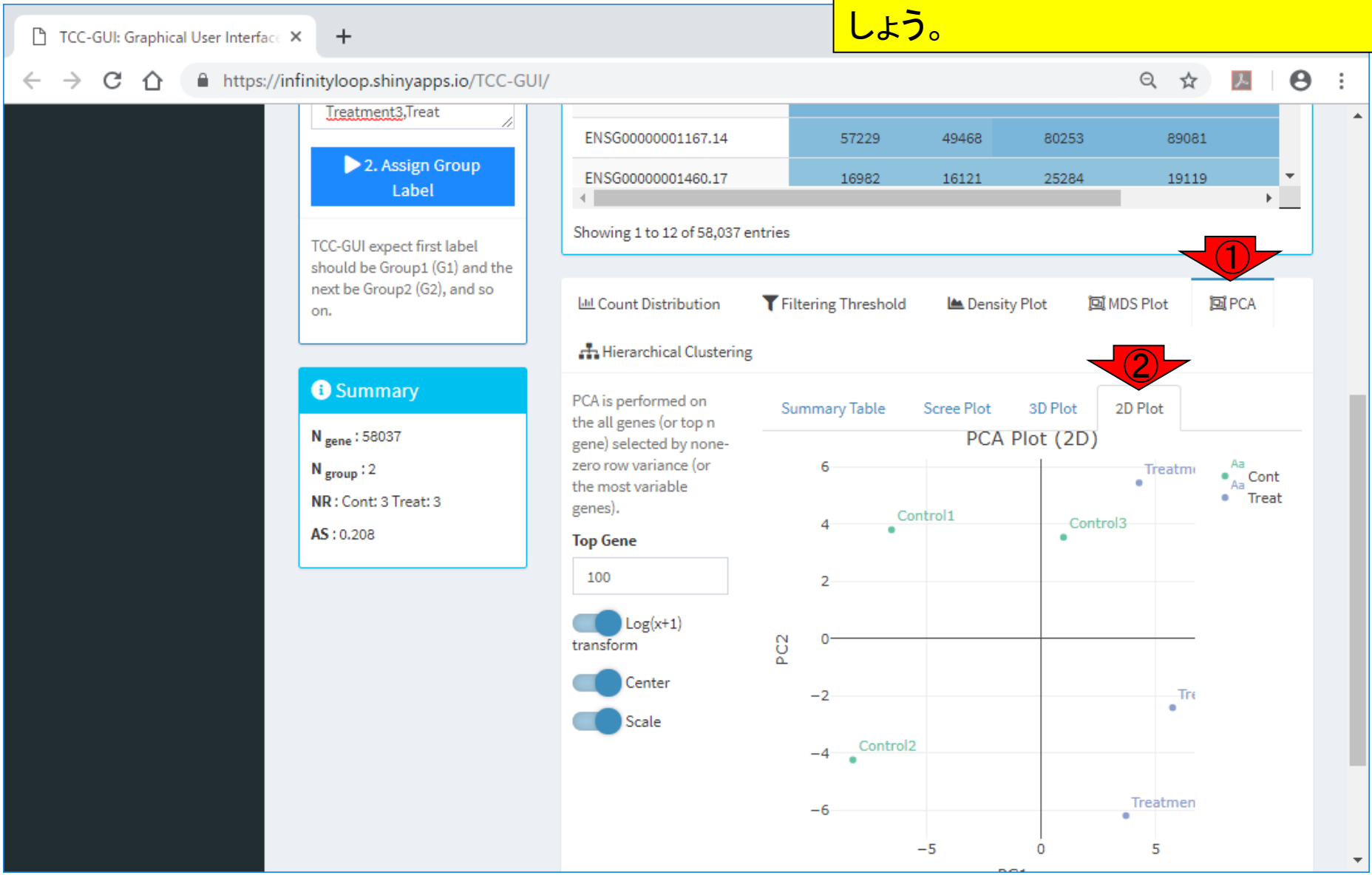
PCA

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



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。

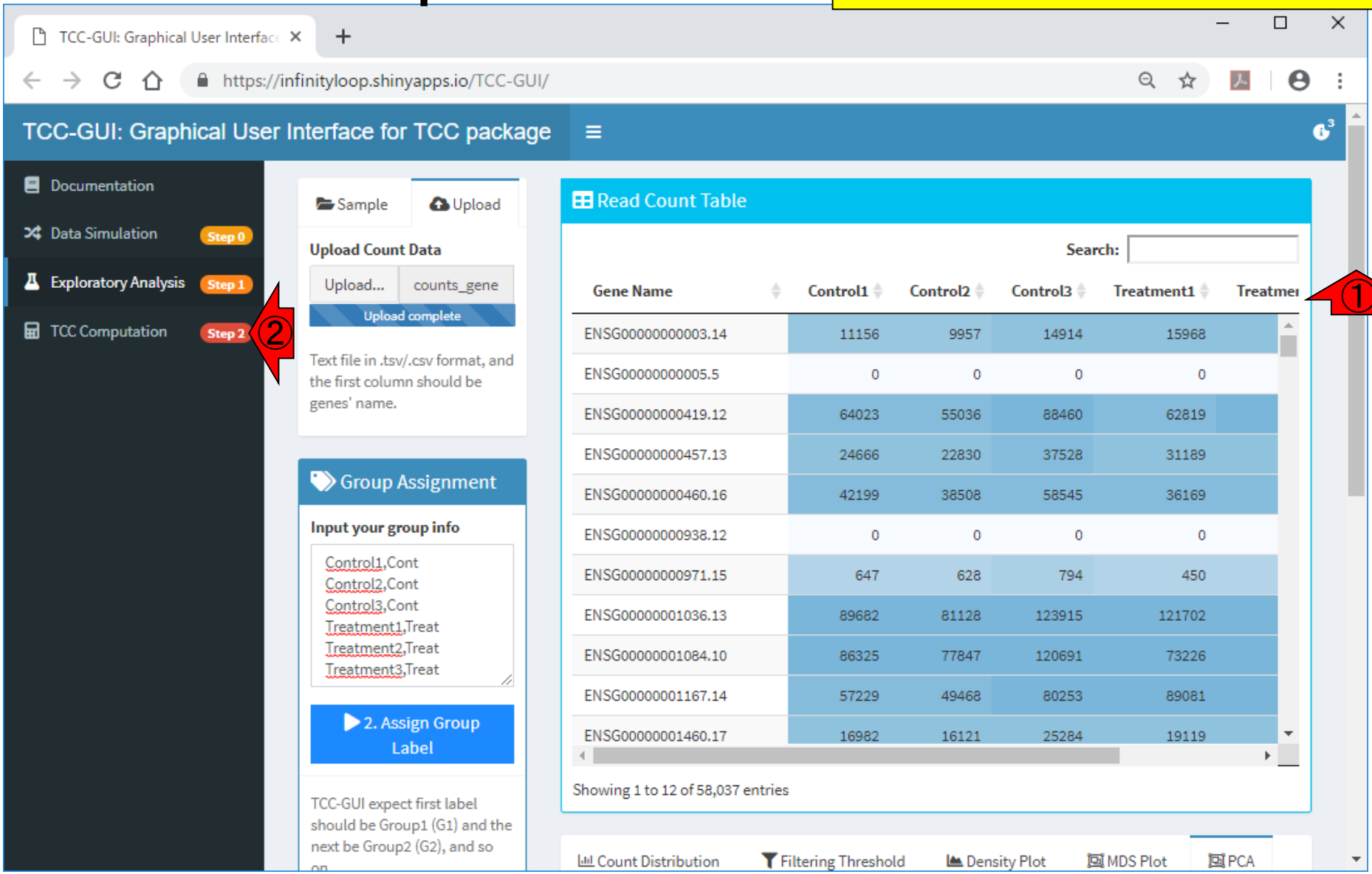


Contents

- 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

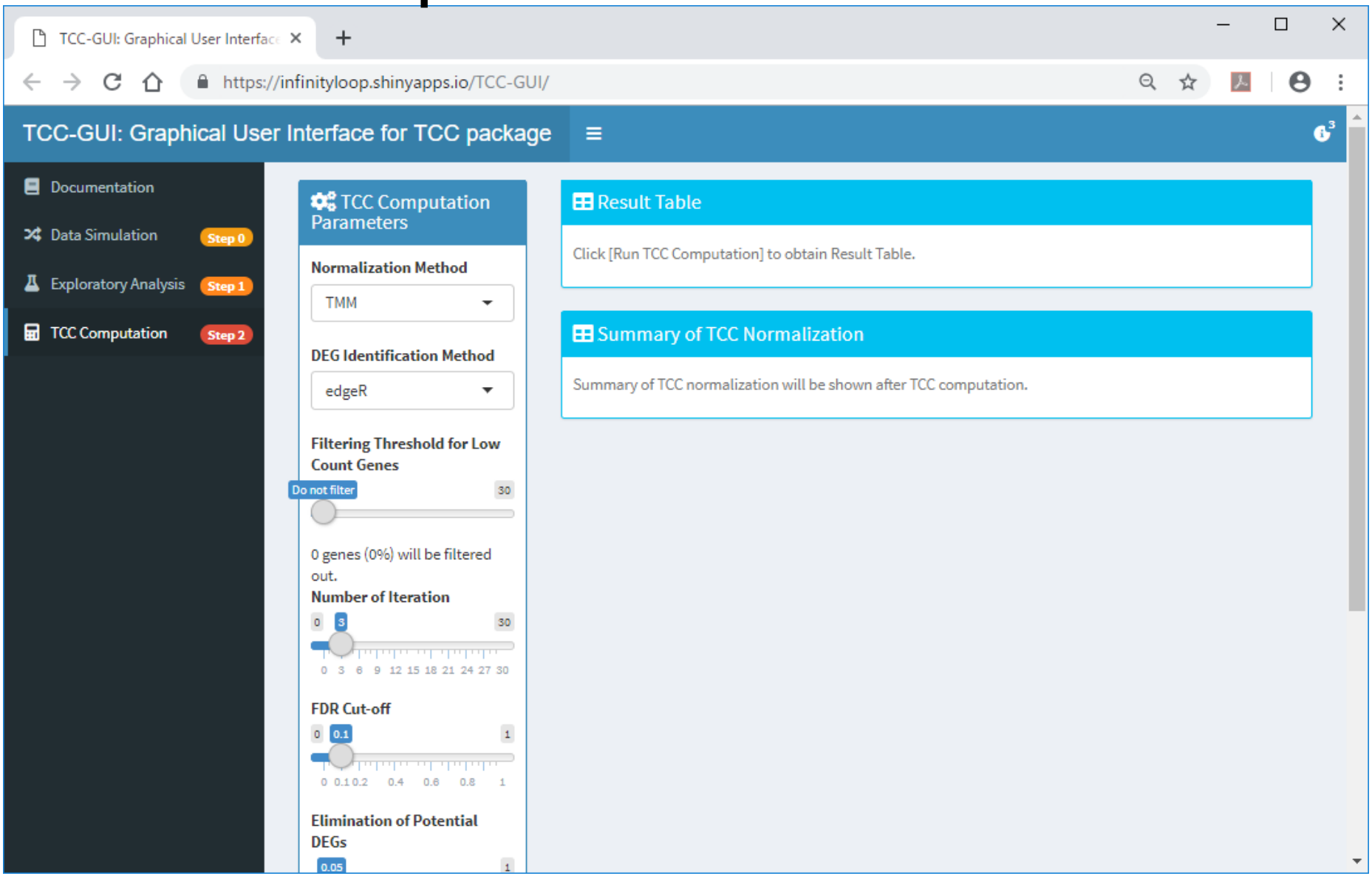
TCC Computation

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



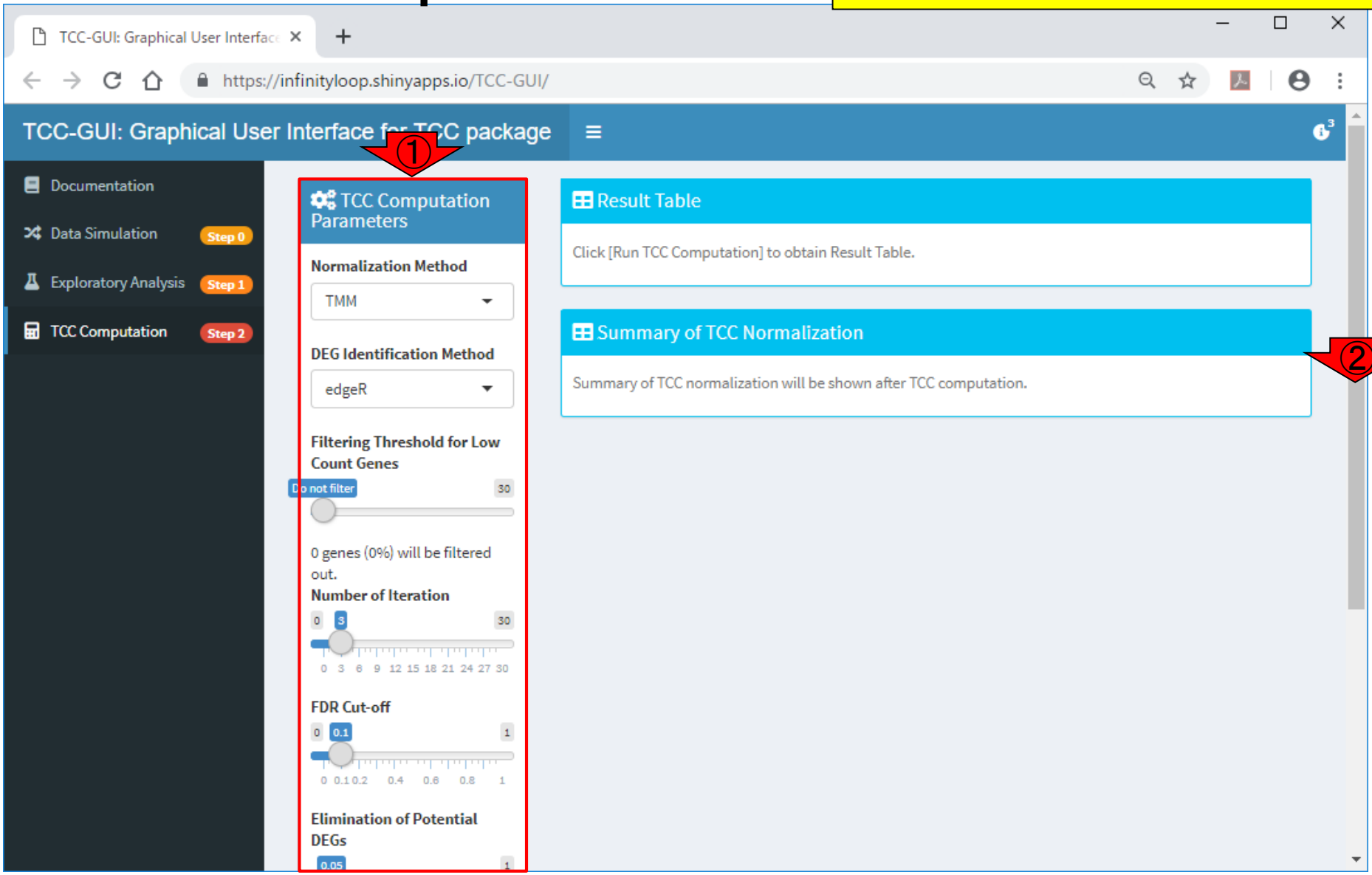
こんな感じになります。

TCC Computation



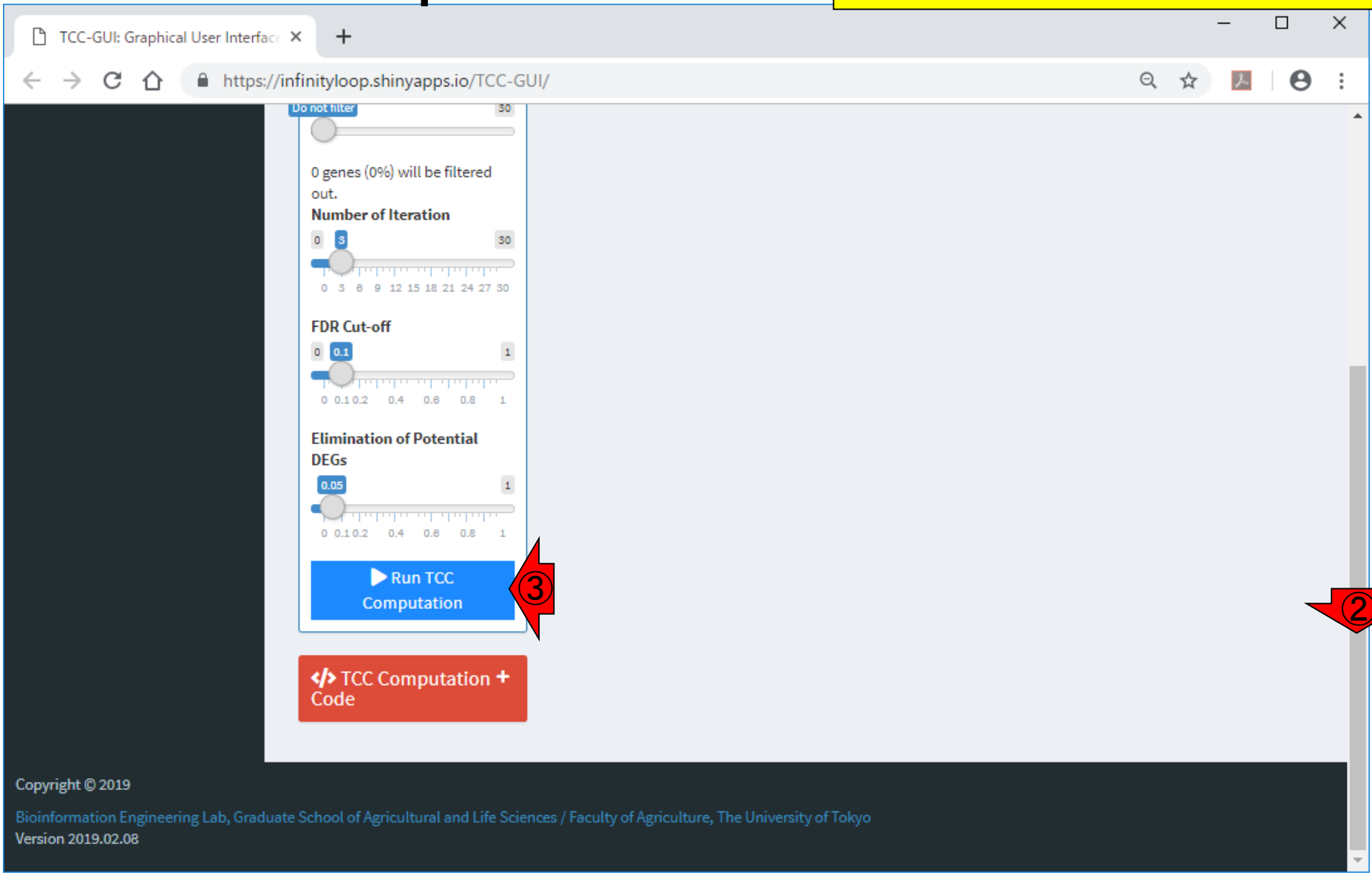
TCC Computation

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。



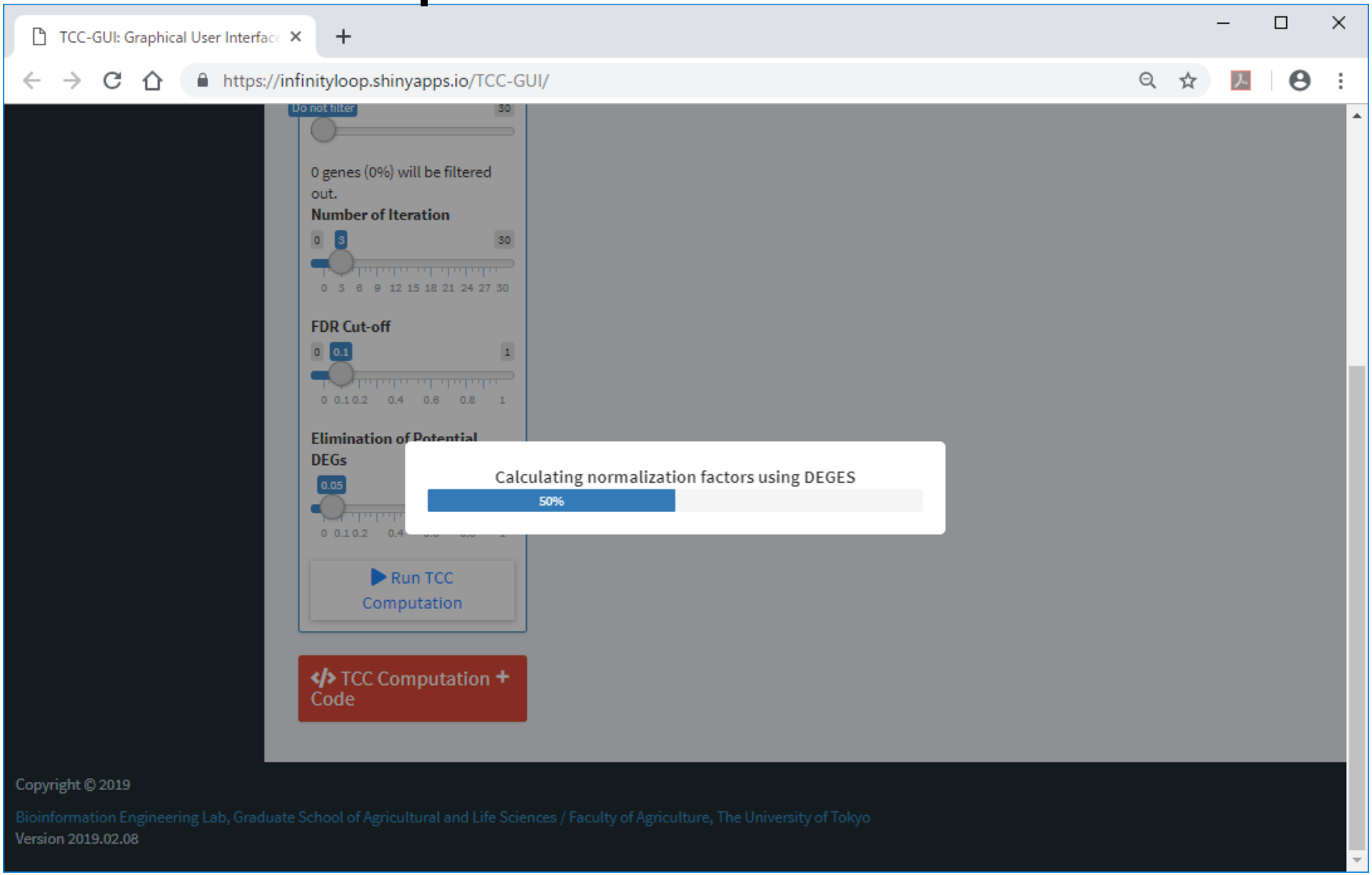
TCC Computation

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。



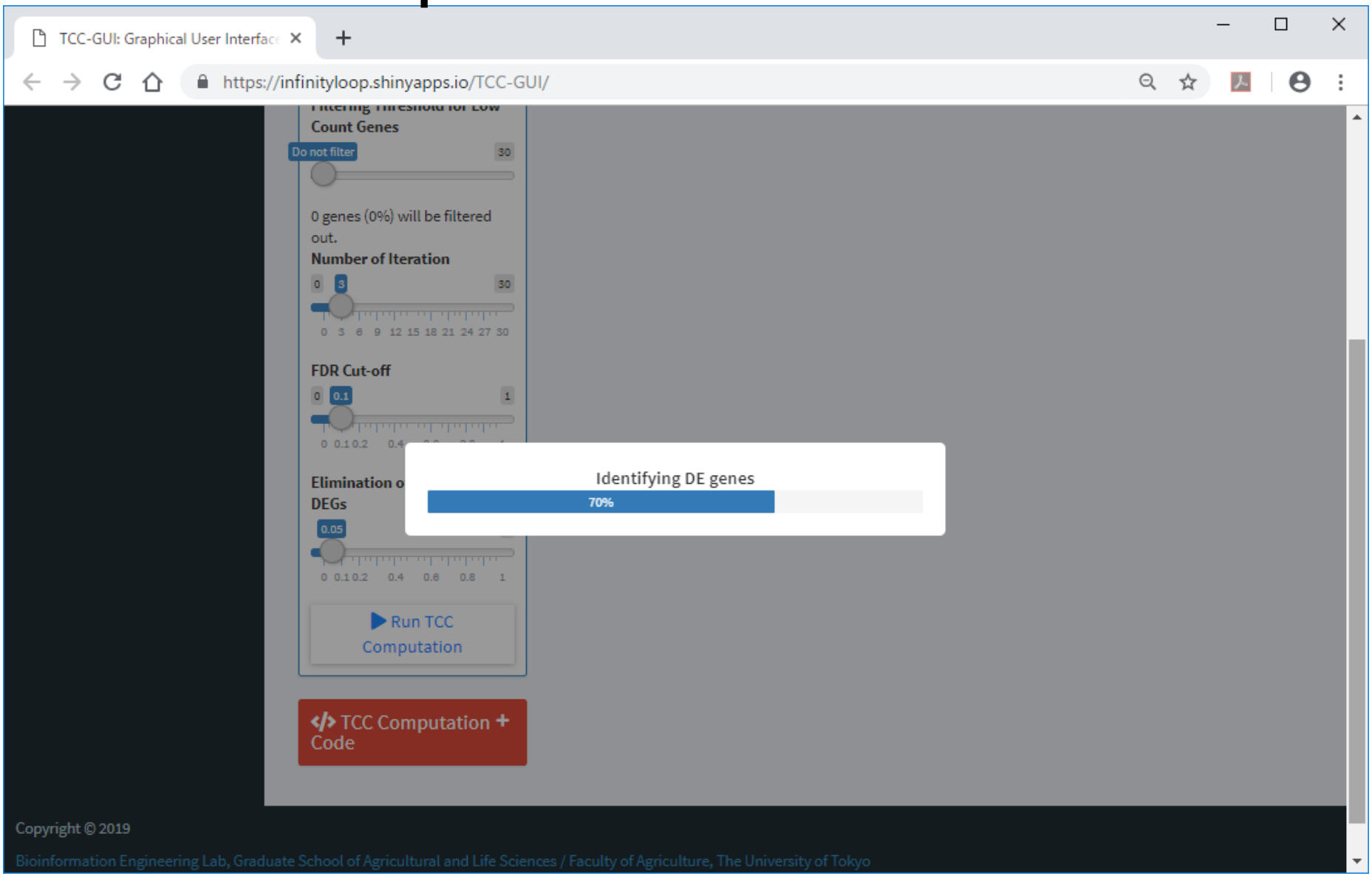
計算中…

TCC Computation



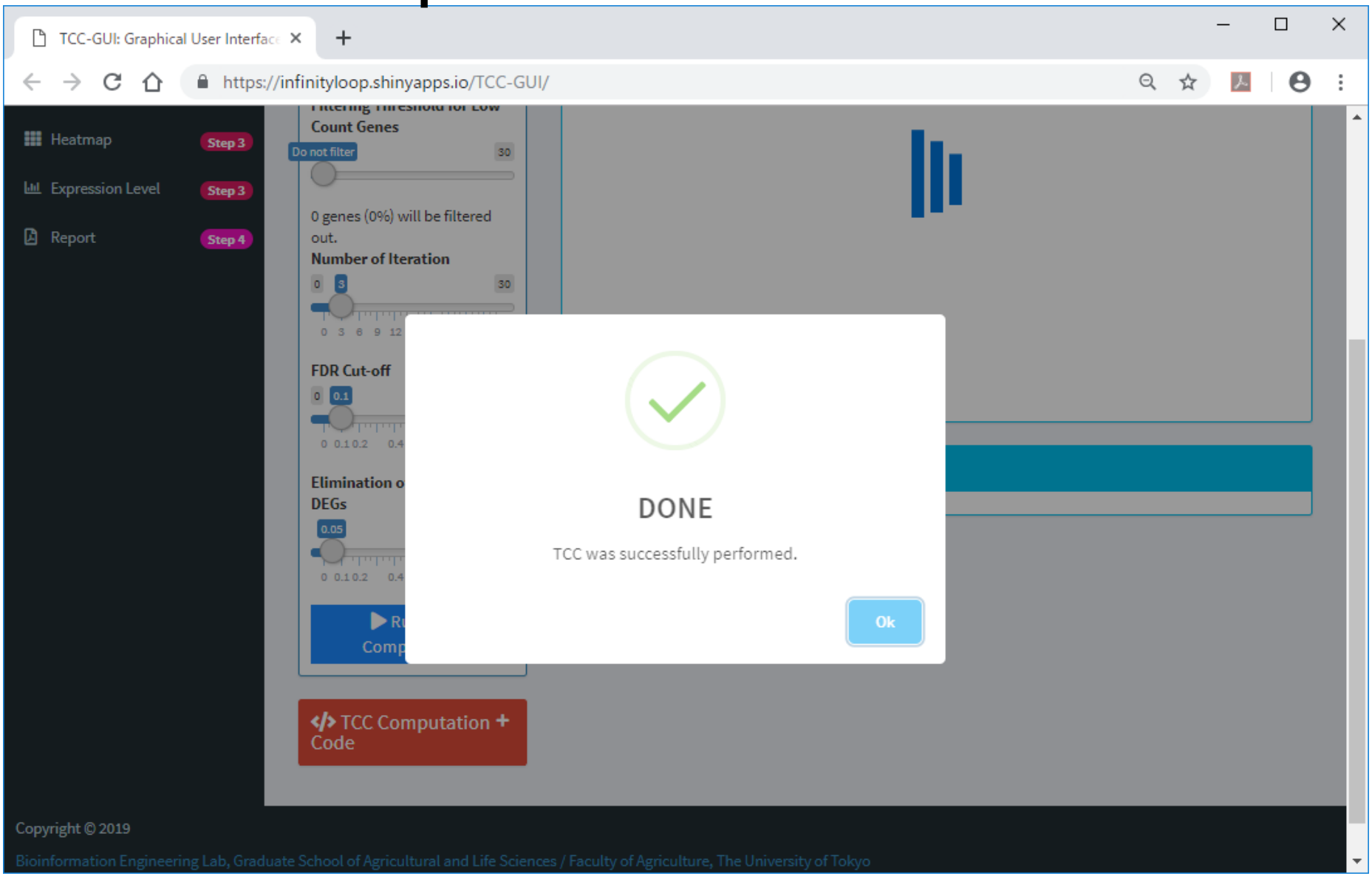
計算中…

TCC Computation



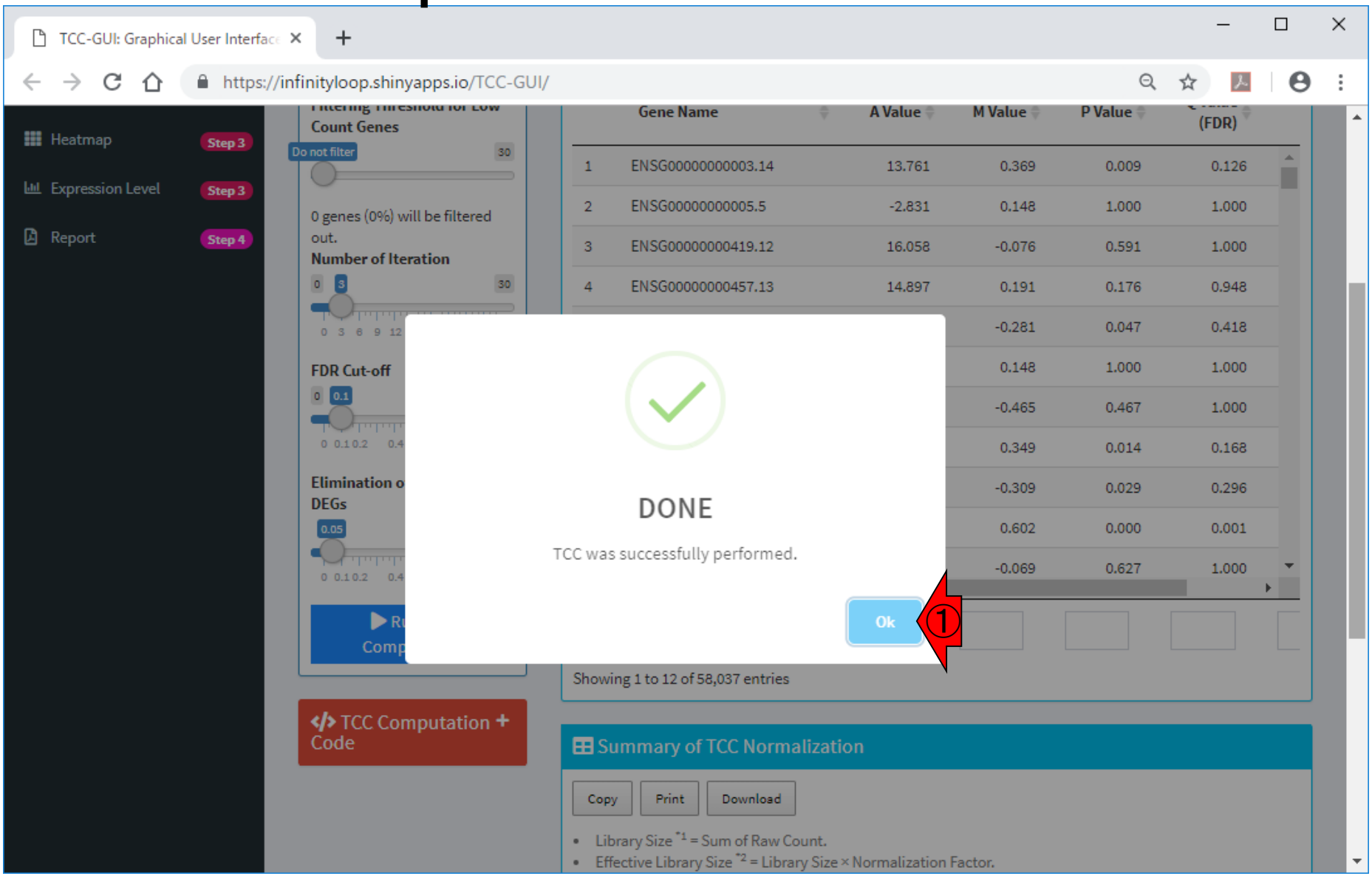
無事計算終了。

TCC Computation



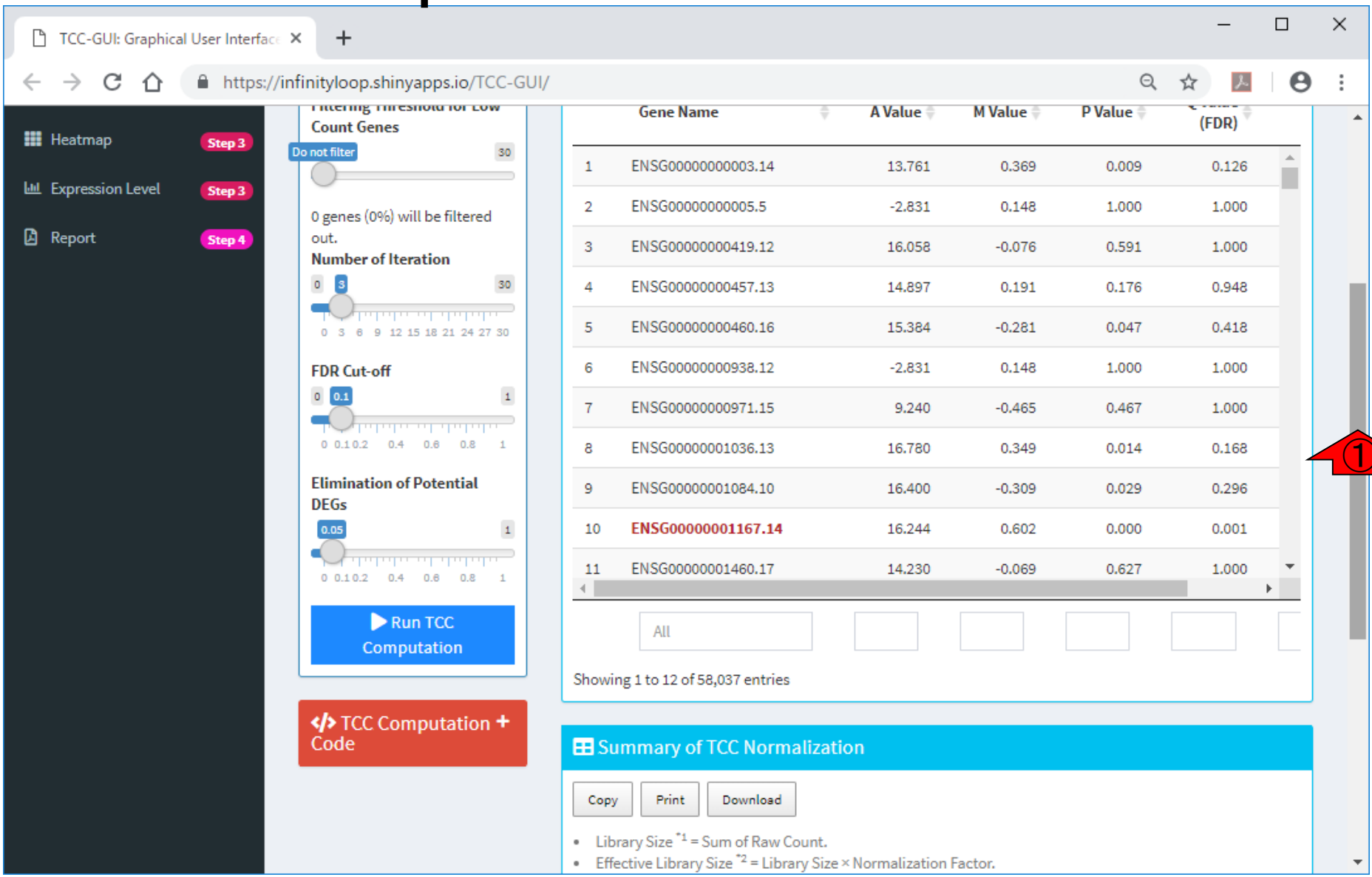
無事計算終了。①Ok

TCC Computation



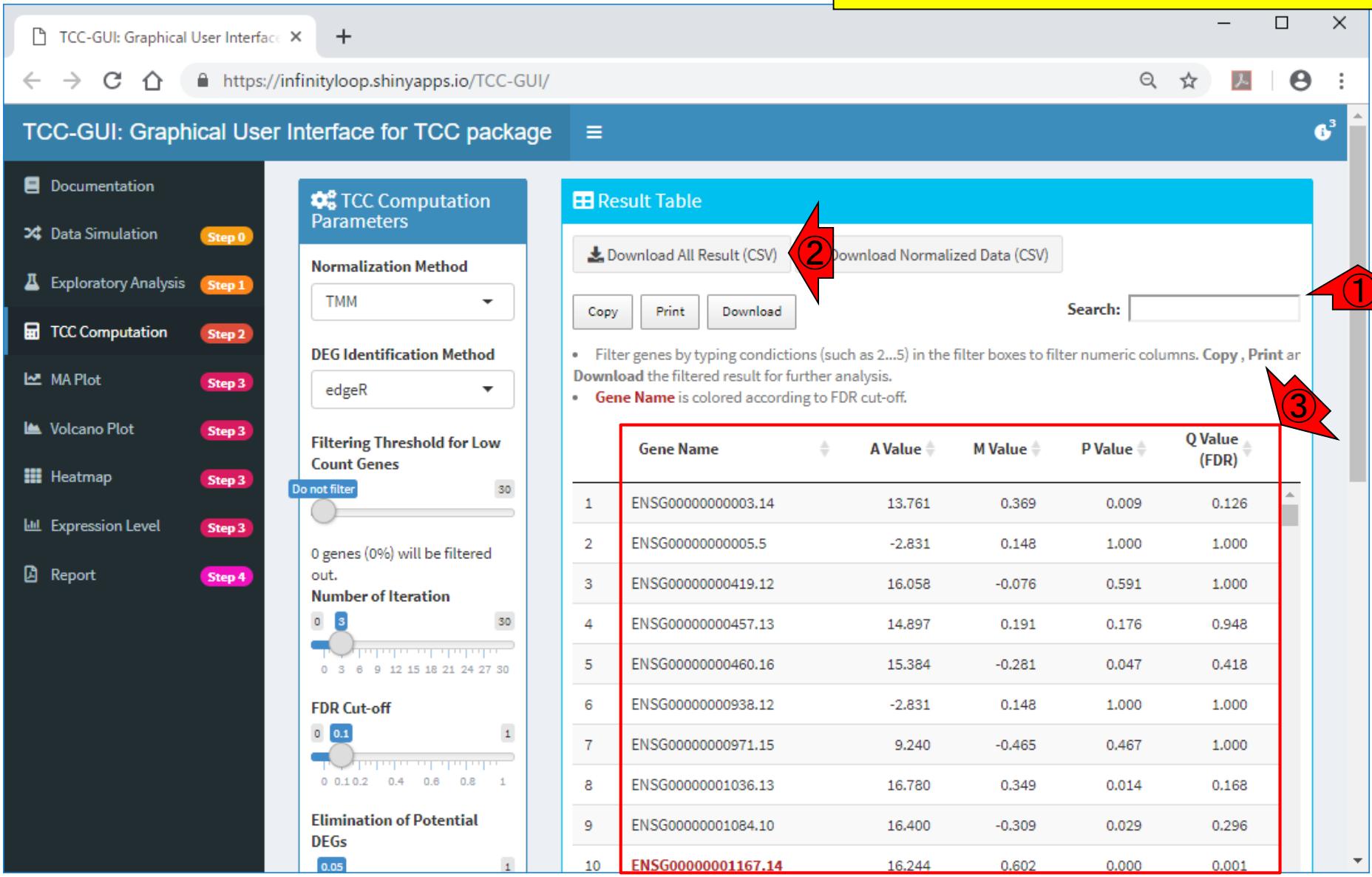
①ページ上部に移動。

TCC Computation



計算結果の保存

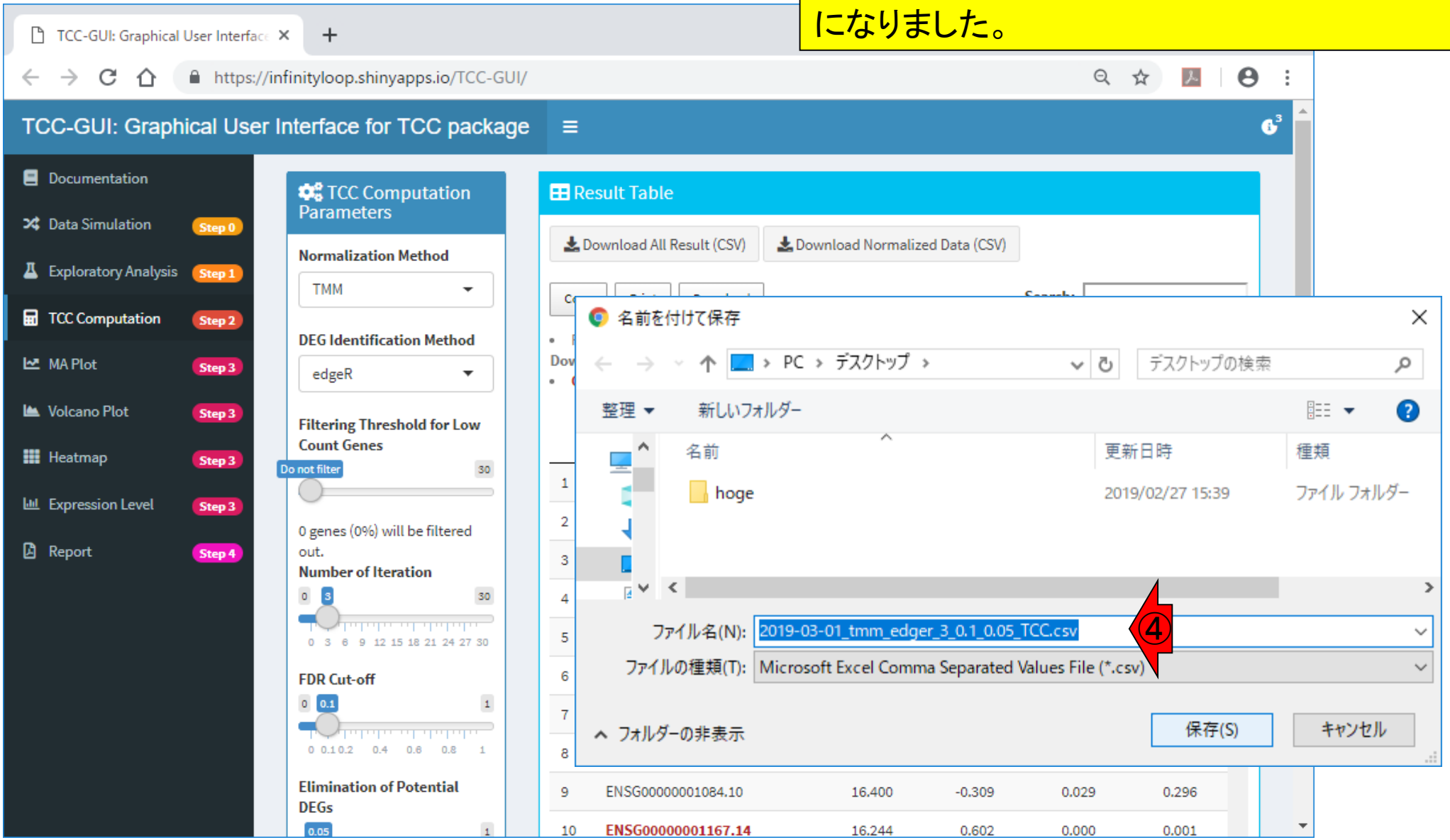
①ページ上部に移動。②を押せば、③赤枠の中身をCSV形式ファイルでダウンロードできます。



	Gene Name	A Value	M Value	P Value	Q 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	ENSG00000000419.12	16.058	-0.076	0.591	1.000
4	ENSG00000000457.13	14.897	0.191	0.176	0.948
5	ENSG00000000460.16	15.384	-0.281	0.047	0.418
6	ENSG00000000938.12	-2.831	0.148	1.000	1.000
7	ENSG00000000971.15	9.240	-0.465	0.467	1.000
8	ENSG00000001036.13	16.780	0.349	0.014	0.168
9	ENSG00000001084.10	16.400	-0.309	0.029	0.296
10	ENSG00000001167.14	16.244	0.602	0.000	0.001

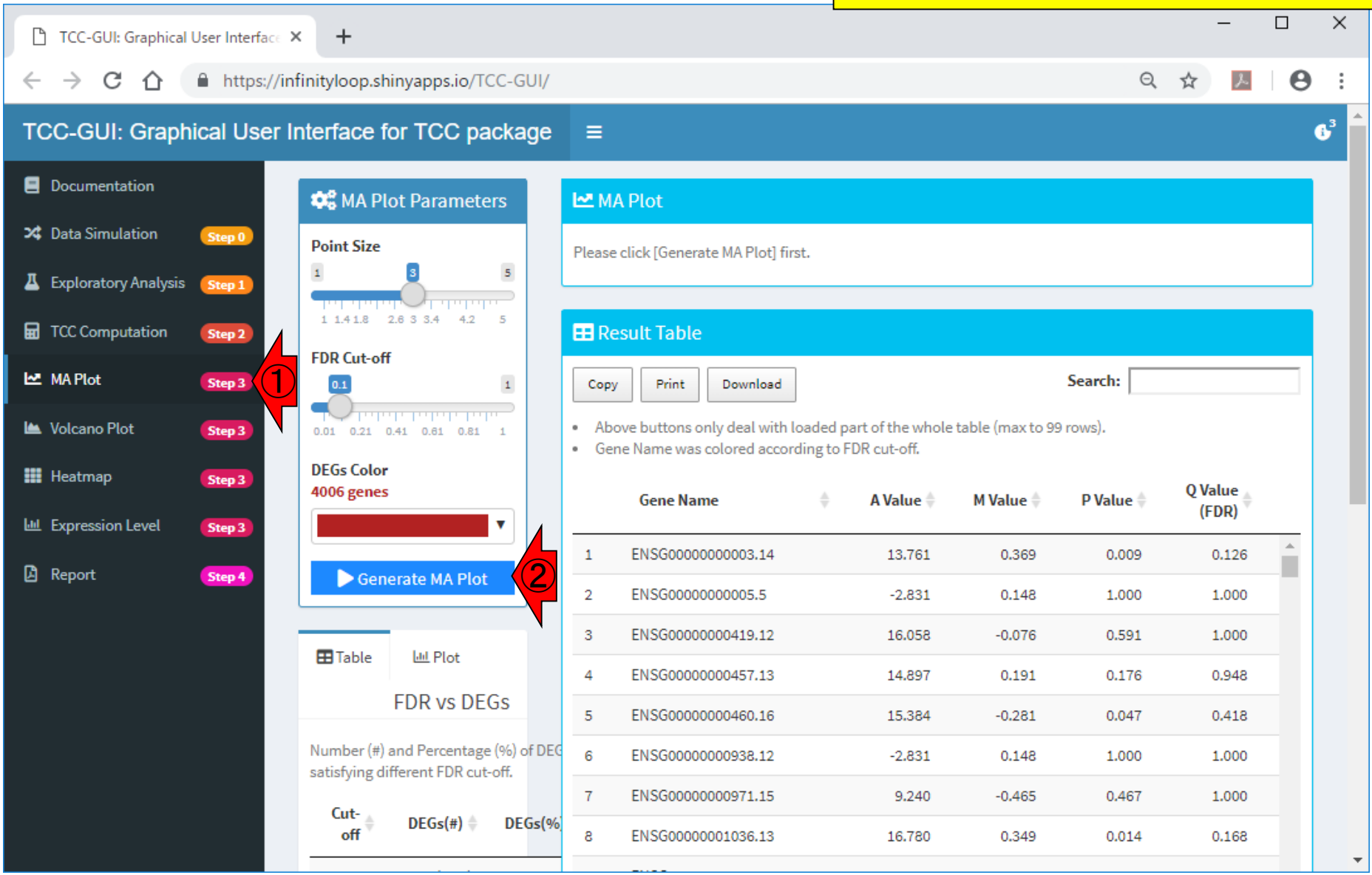
計算結果の保存

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。



MA Plot

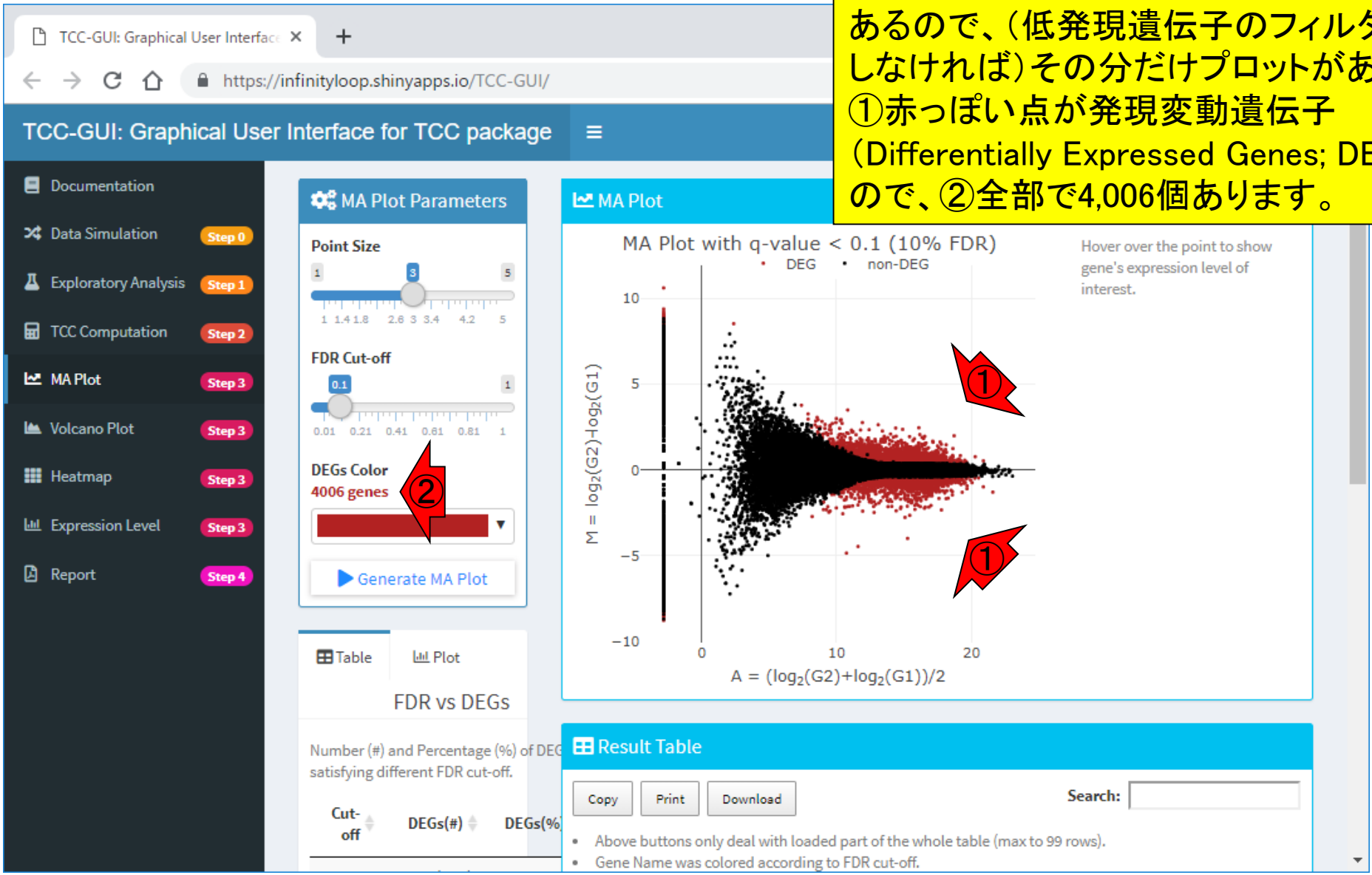
RNA-seq発現変動解析結果を描画する目的でよく利用される①MA Plot。②まずはデフォルトで実行。



	Gene Name	A Value	M Value	P Value	Q 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	ENSG00000000419.12	16.058	-0.076	0.591	1.000
4	ENSG00000000457.13	14.897	0.191	0.176	0.948
5	ENSG00000000460.16	15.384	-0.281	0.047	0.418
6	ENSG00000000938.12	-2.831	0.148	1.000	1.000
7	ENSG00000000971.15	9.240	-0.465	0.467	1.000
8	ENSG00000001036.13	16.780	0.349	0.014	0.168

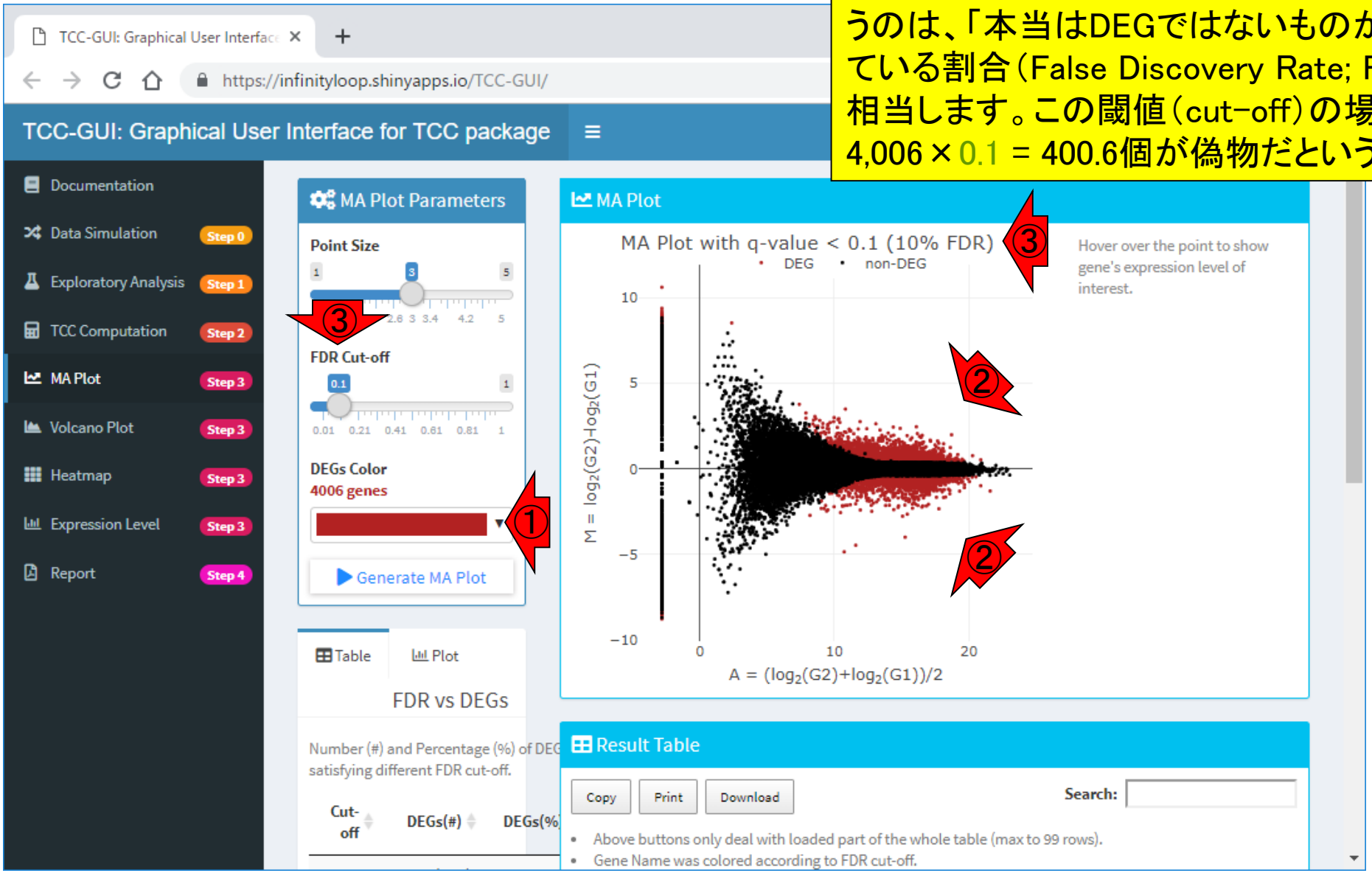
MA Plot

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。



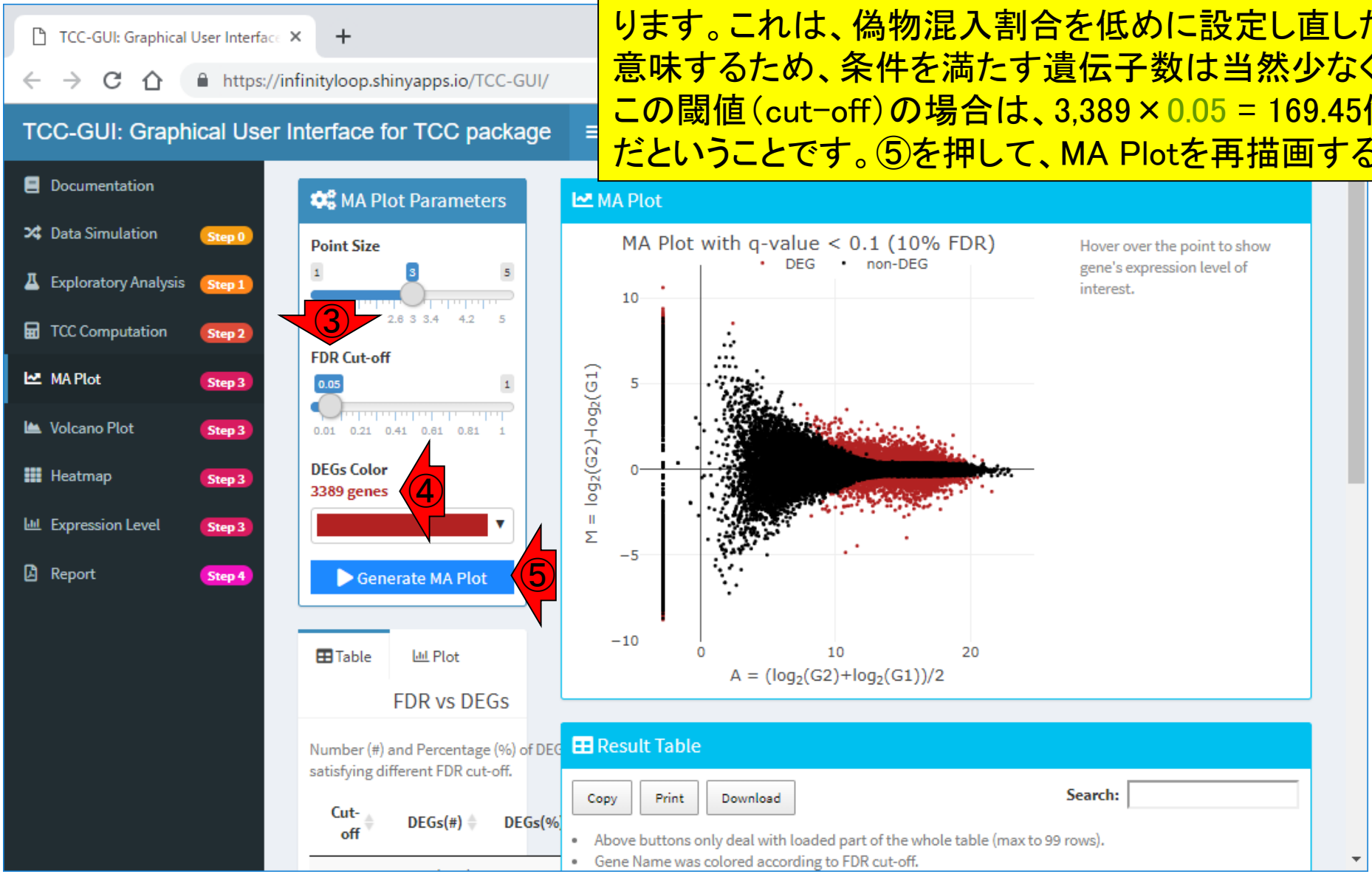
MA Plot

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 × 0.1 = 400.6個が偽物だということです。



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 × 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…



MA Plot

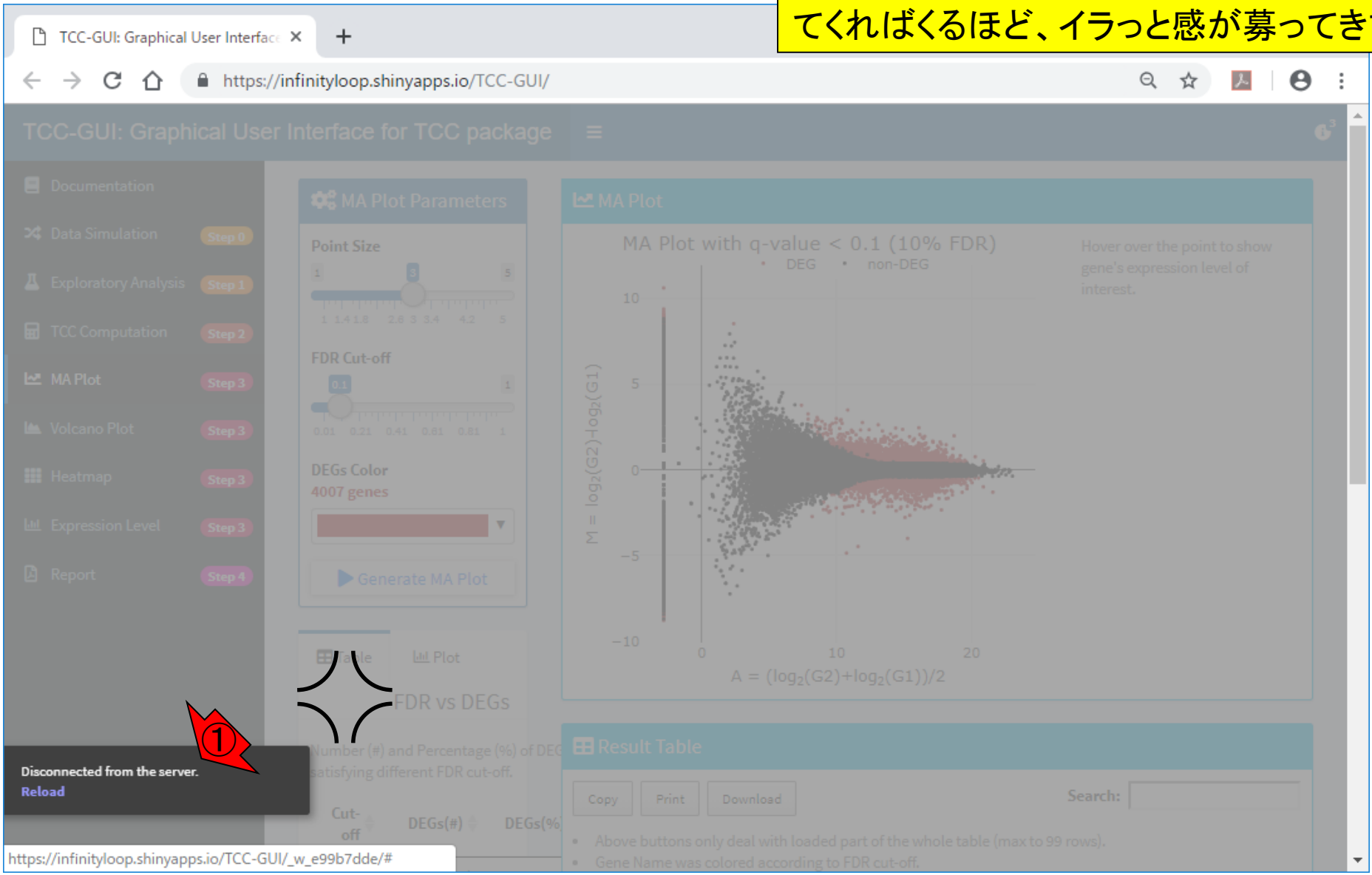
こんな感じになります。①の部分も反映され、②のDEG数も4,006個から3,389個になったのだろうということが黒のプロット領域の増加から想像できます。

Contents

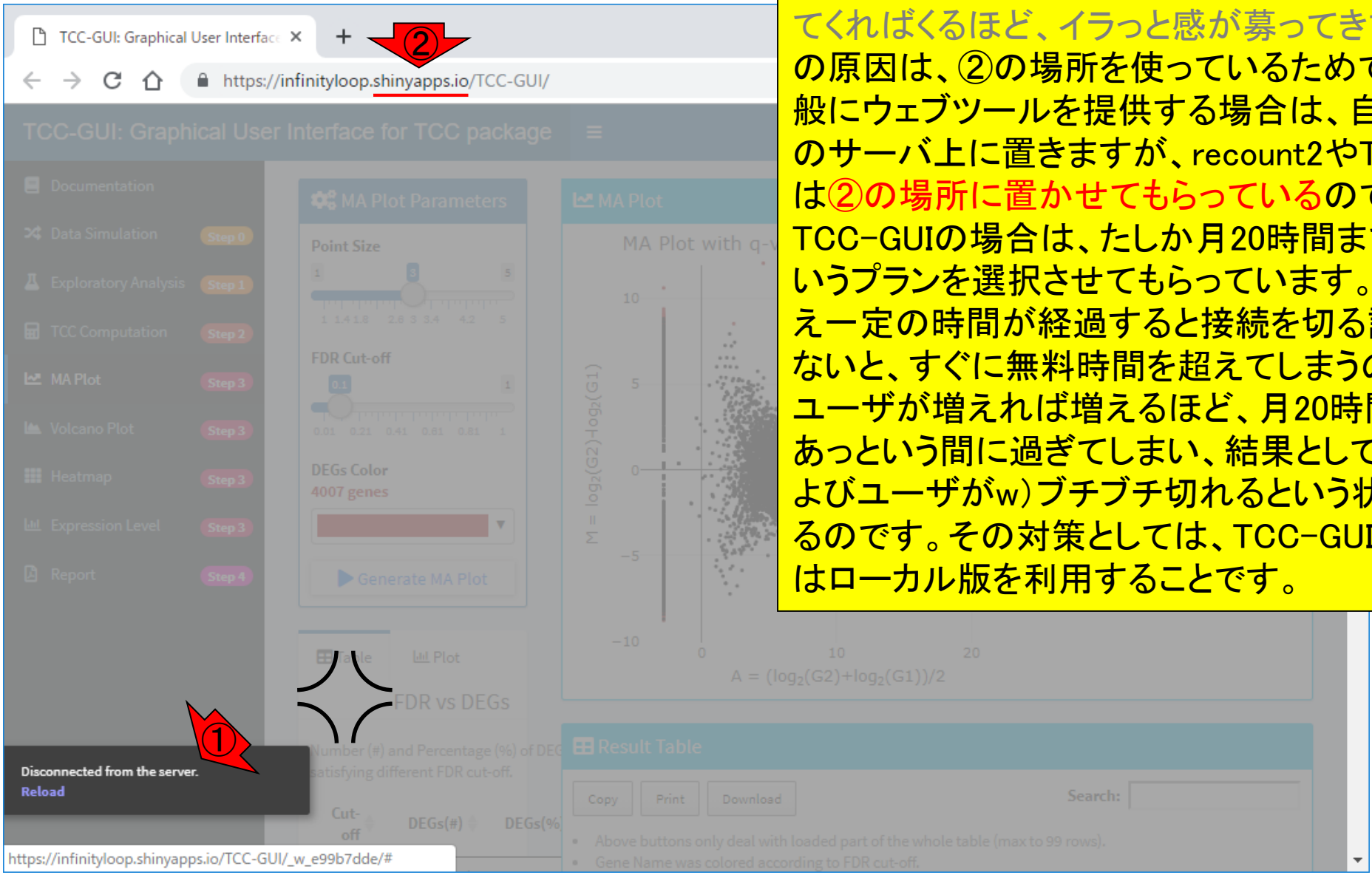
- トランスクリプトーム（RNA-seq）解析の原理、データ解析戦略のイメージ
- 公共カウントデータの取得（recount2）
- サンプル間クラスタリング
  - 手元のデータを実行、結果の解釈
  - グループ間の分離度を客観的に示すスコア
- TCC-GUI
  - ファイルのアップロード、グループラベル情報の付与、平均シルエットスコア（AS値）
  - 探索的解析（Exploratory Analysis）：階層的クラスタリングやPCAなど
  - 発現変動解析（TCC Computation）
- すぐにDisconnectedとなる問題とその対策
  - RStudioのインスト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



Disconnected…



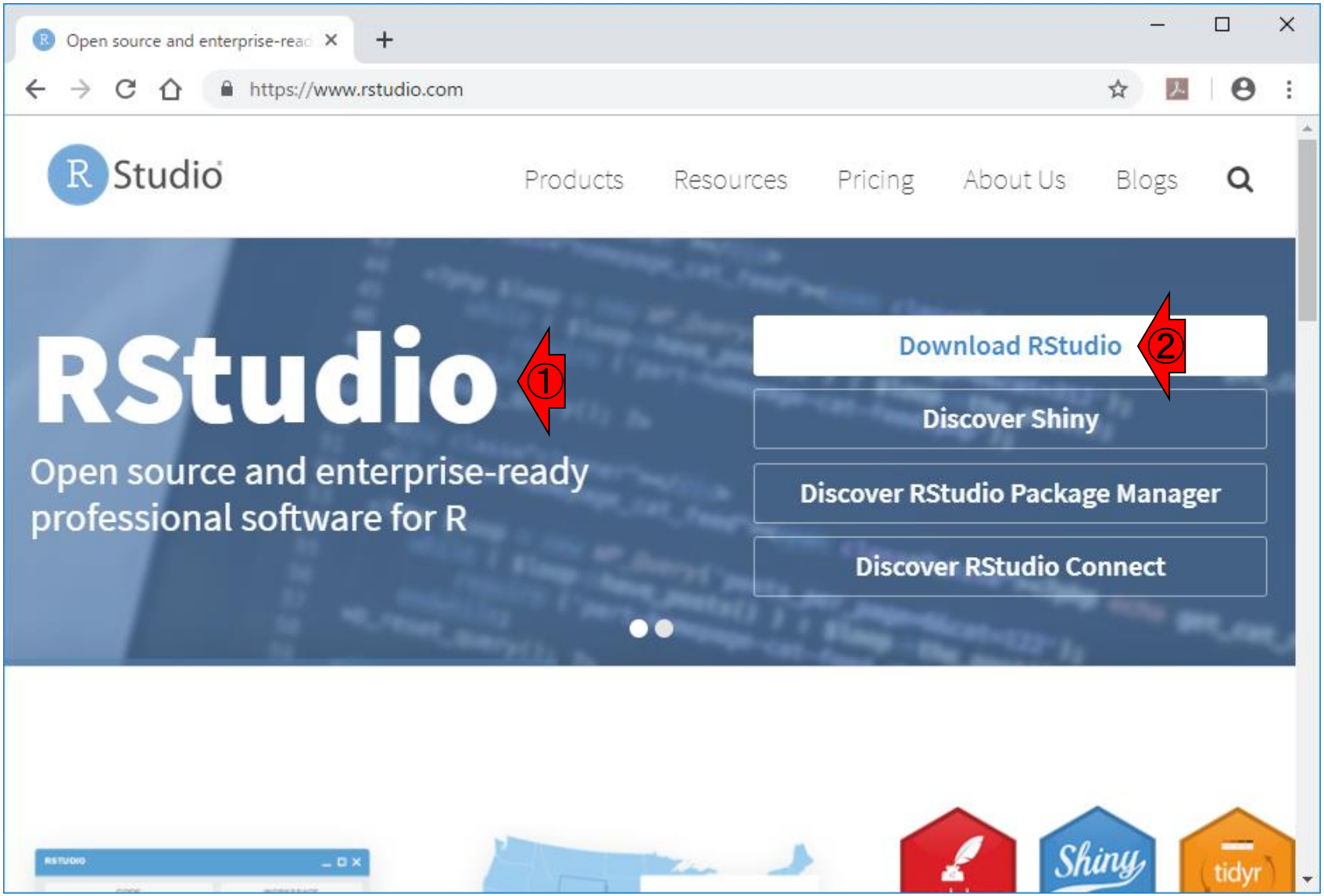
recount2やTCC-GUIのウェブ版は、接続から5分程度で①Disconnectedとなり、最初からやり直す必要が出てきます。これらのウェブツールに慣れてくればくるほど、イラっと感が募ってきます。この原因は、②の場所を使っているためです。一般にウェブツールを提供する場合は、自分たちのサーバ上に置きますが、recount2やTCC-GUIは②の場所に置かせてもらっているのです。TCC-GUIの場合は、たしか月20時間まで無料というプランを選択させてもらっています。それゆえ一定の時間が経過すると接続を切る設定にしないと、すぐに無料時間を超えてしまうのです。ユーザが増えれば増えるほど、月20時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

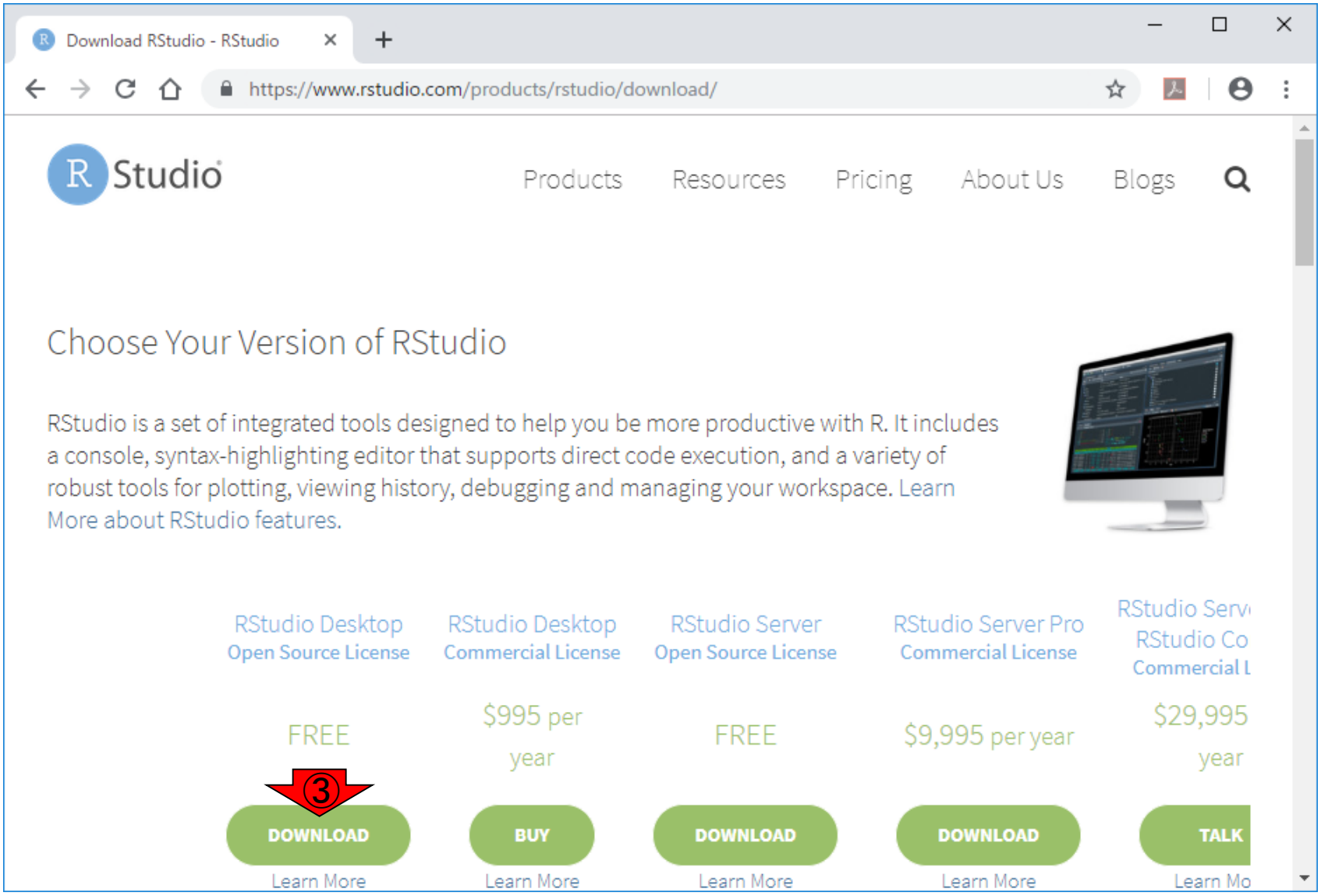
RStudio

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



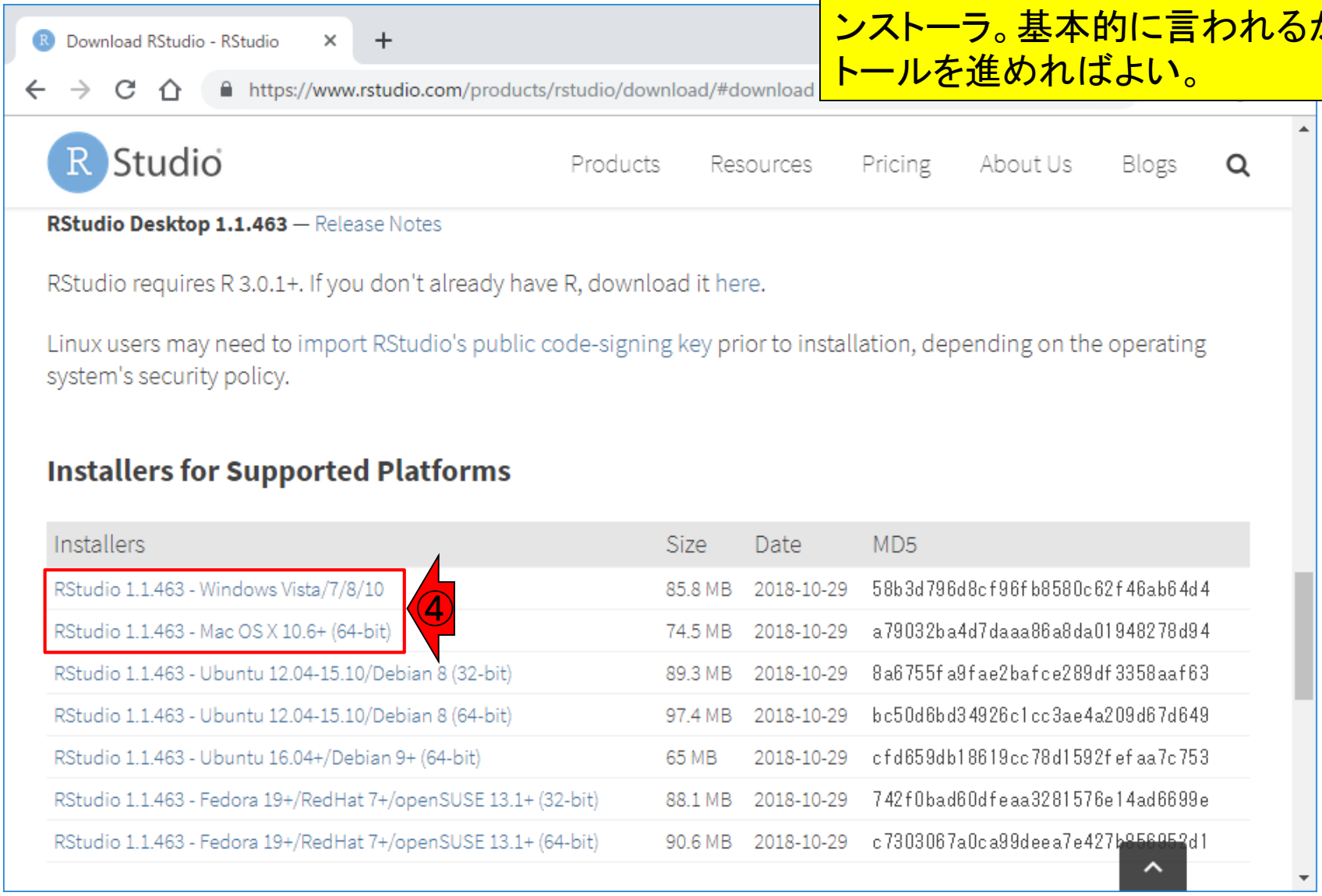
RStudio

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



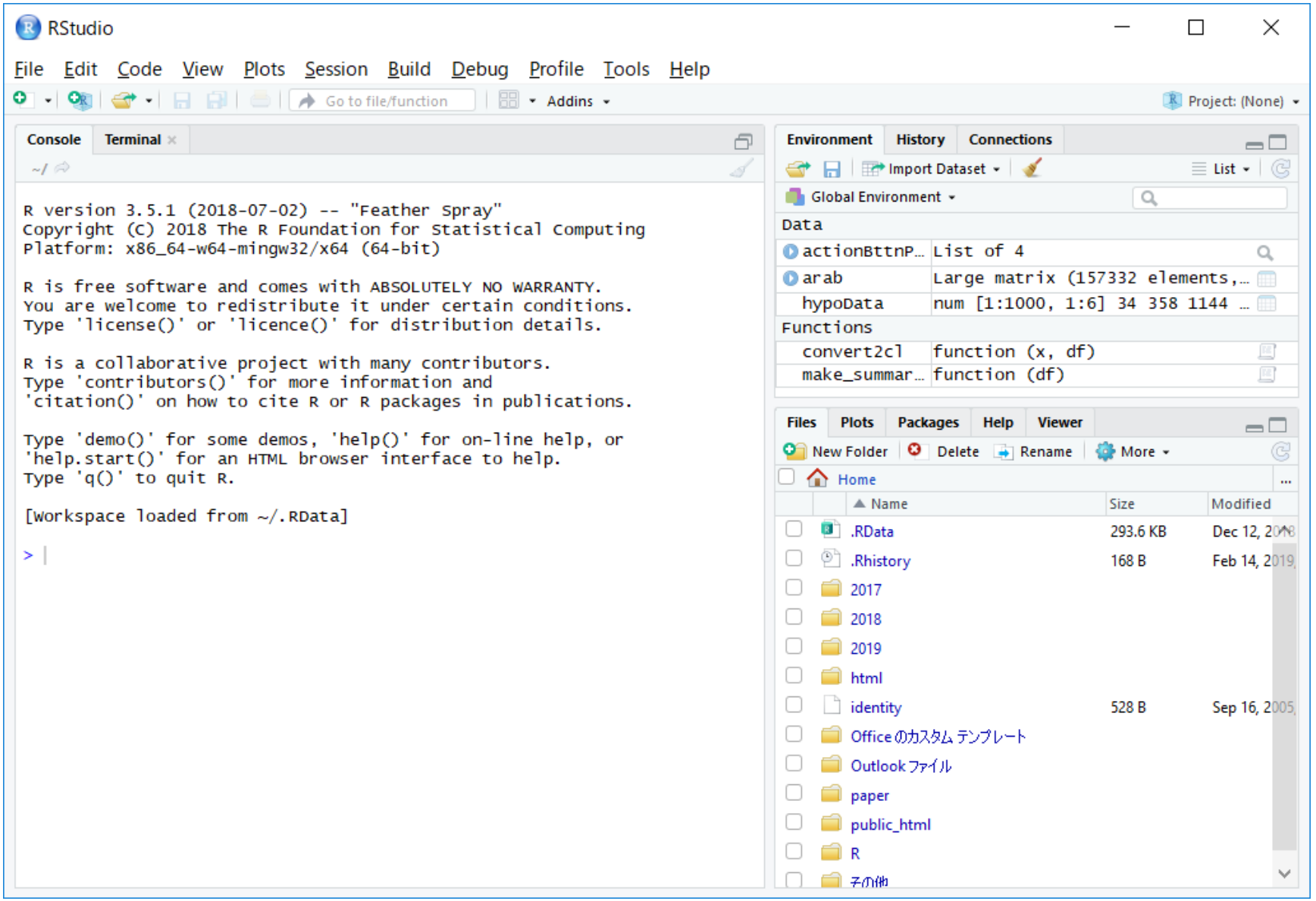
RStudio

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。



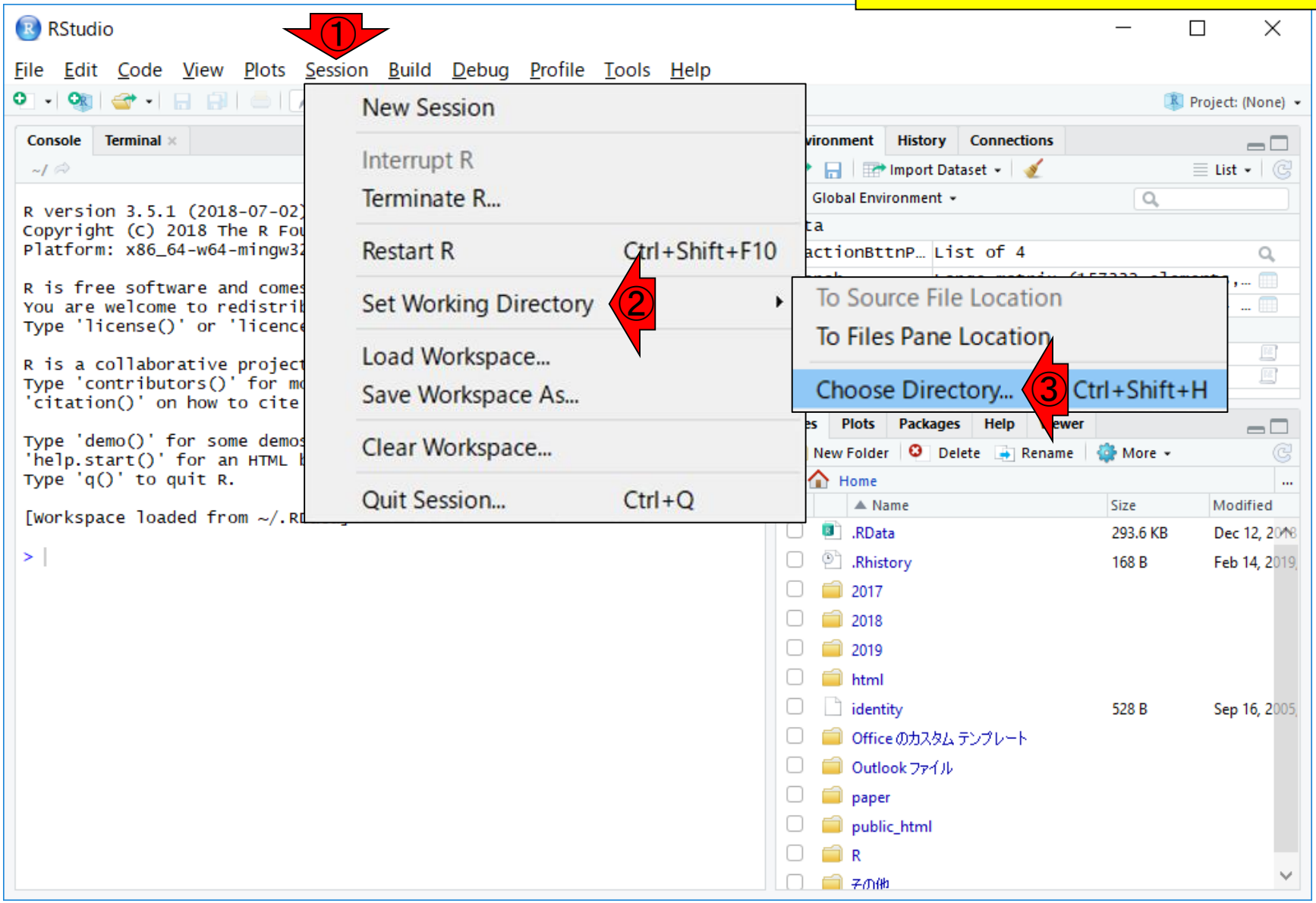
RStudio起動

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。



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、③Choose Directory、のようにすればよいです。

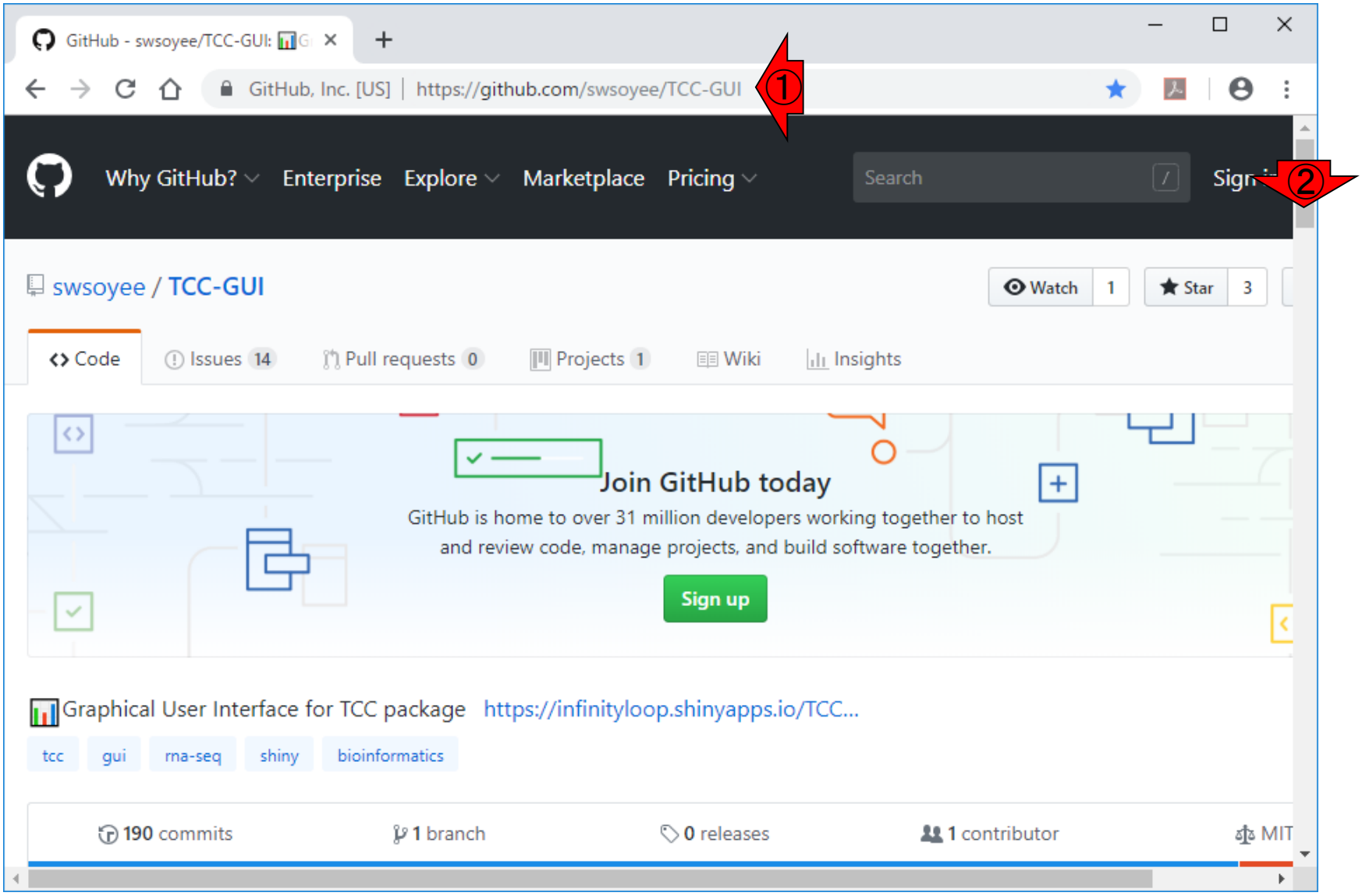


Contents

- 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

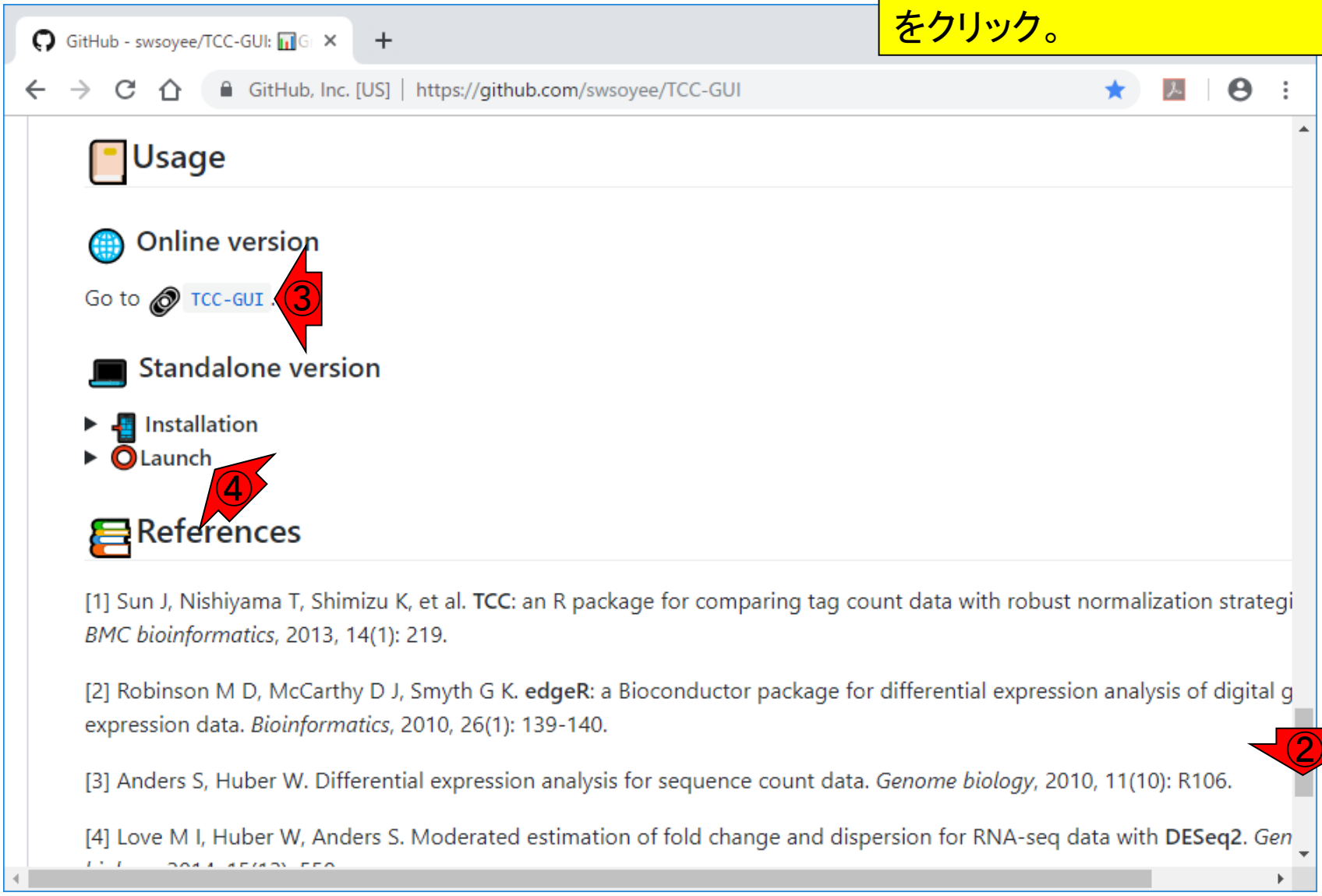
TCC-GUIローカル版

「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。

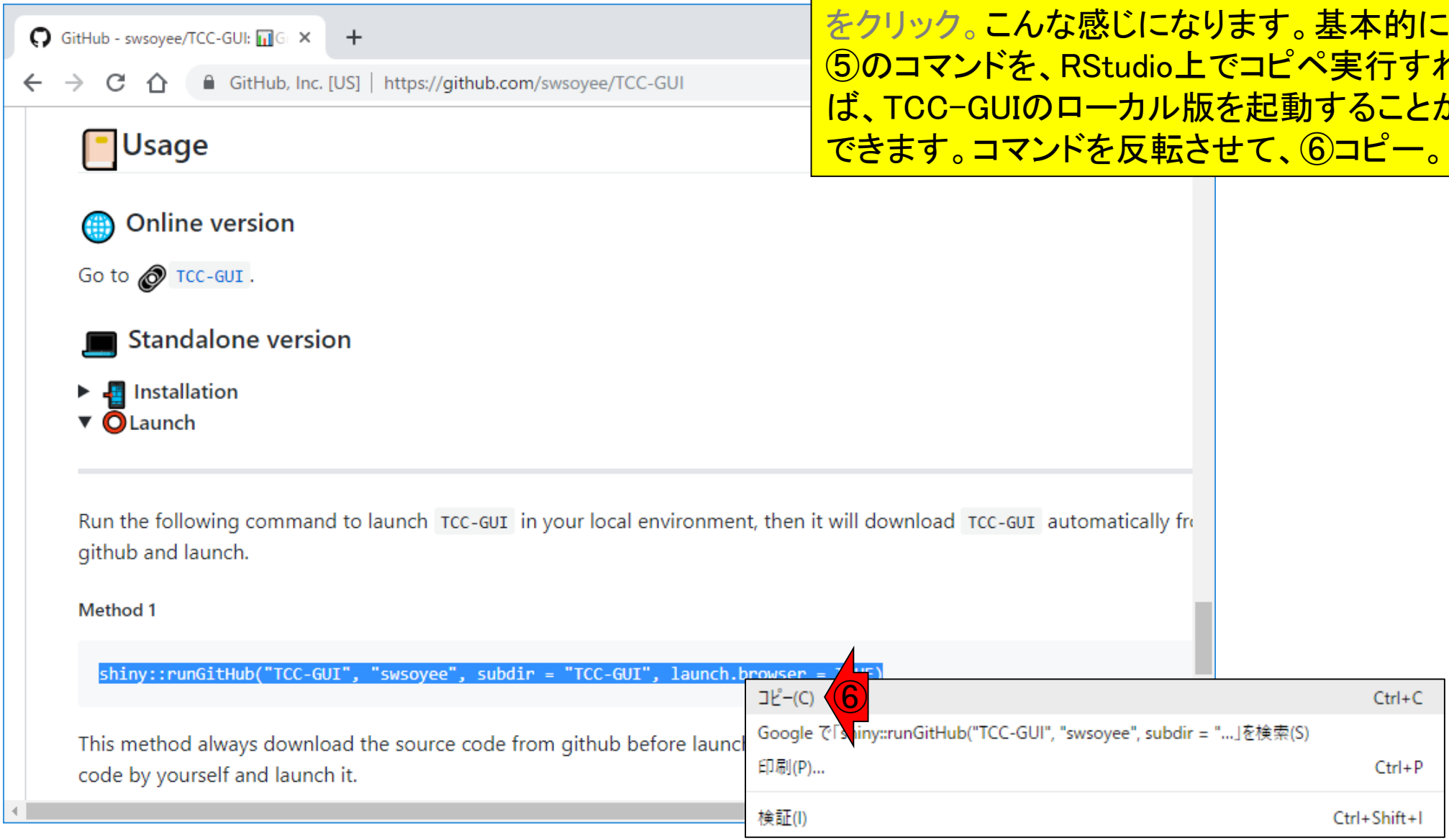


TCC-GUIローカル版

「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上でコピペ実行すれば、TCC-GUIのローカル版を起動することができます。

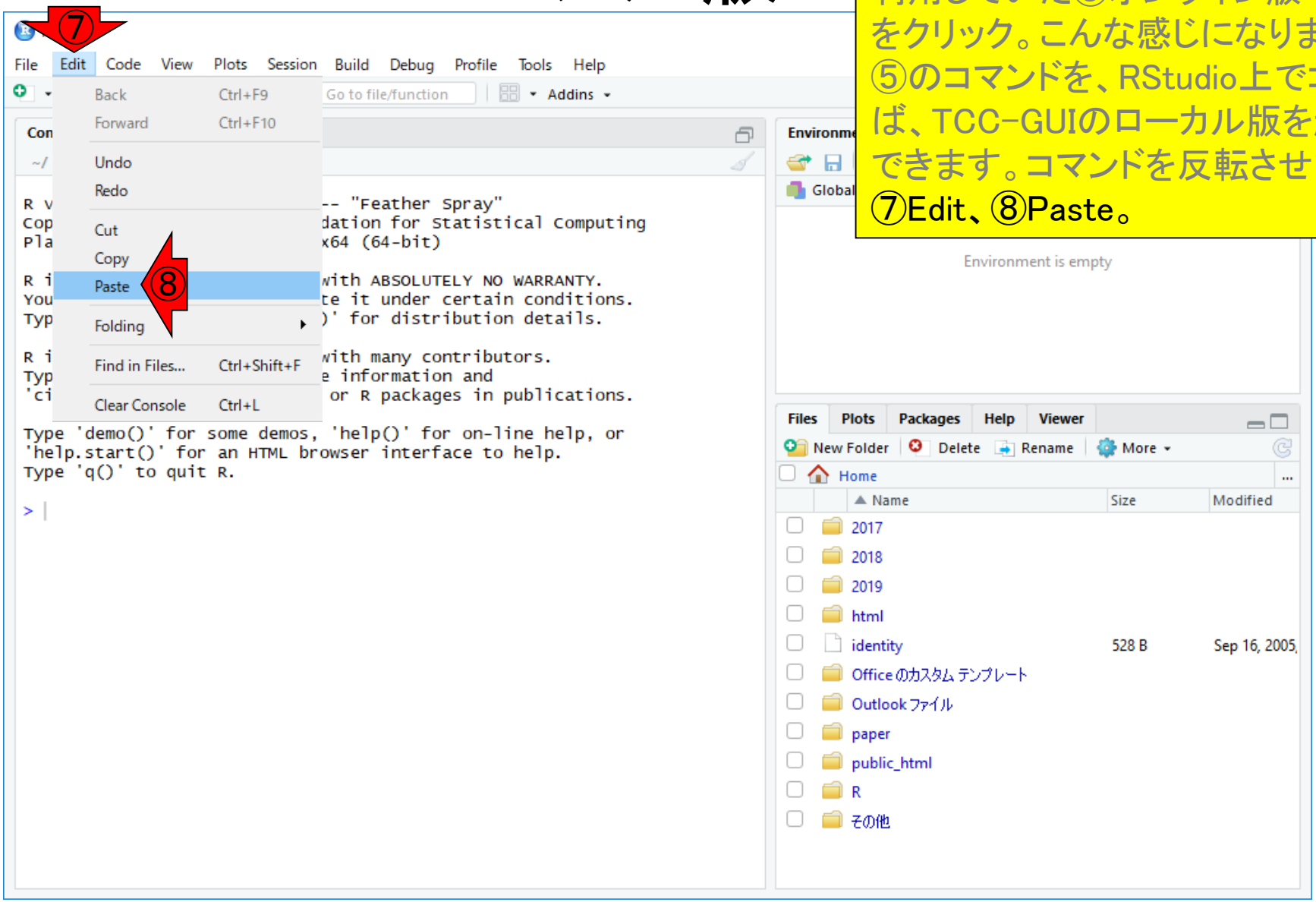
TCC-GUIローカル版

「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上でコピペ実行すれば、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。



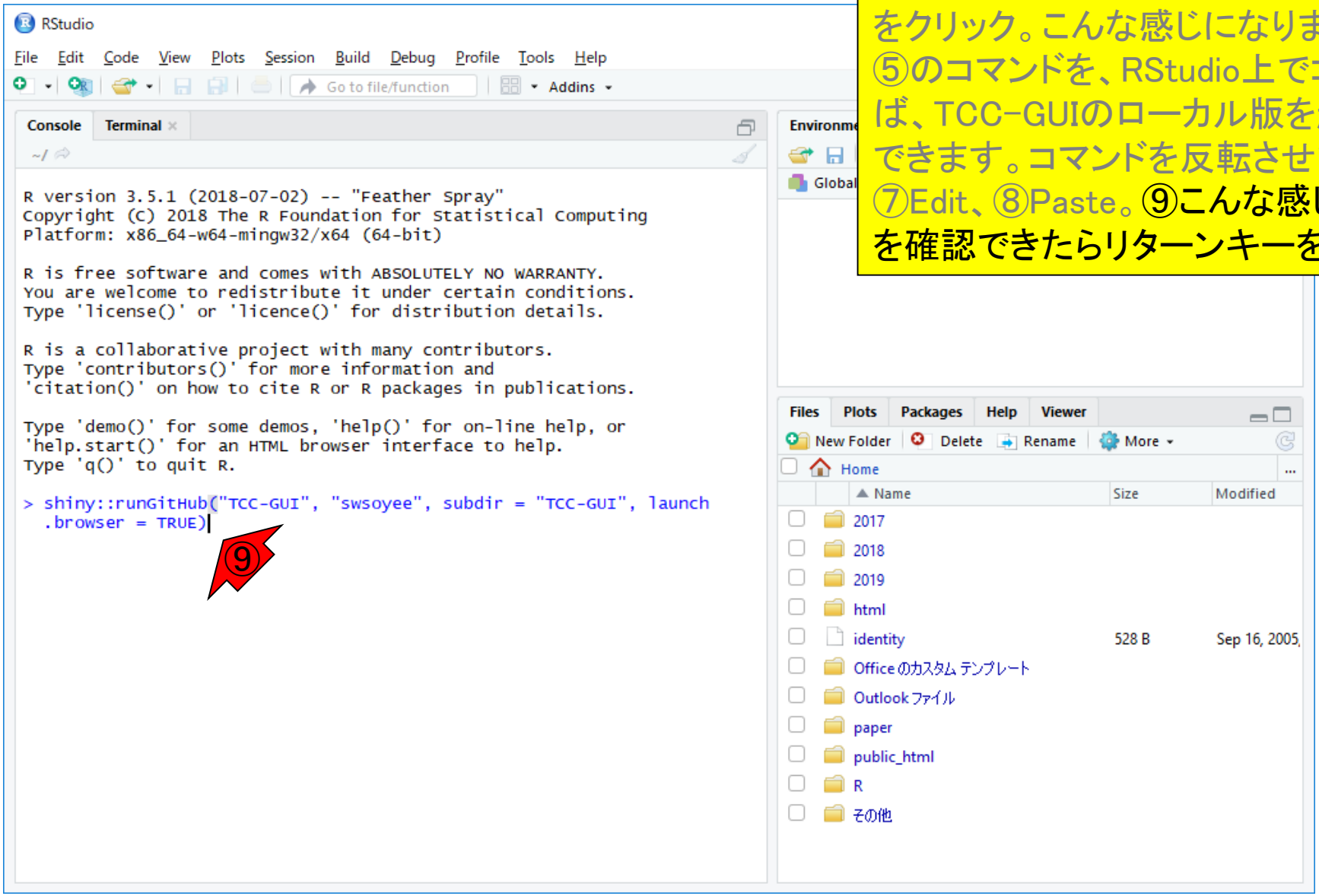
TCC-GUIローカル版

「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上でコピペ実行すれば、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。



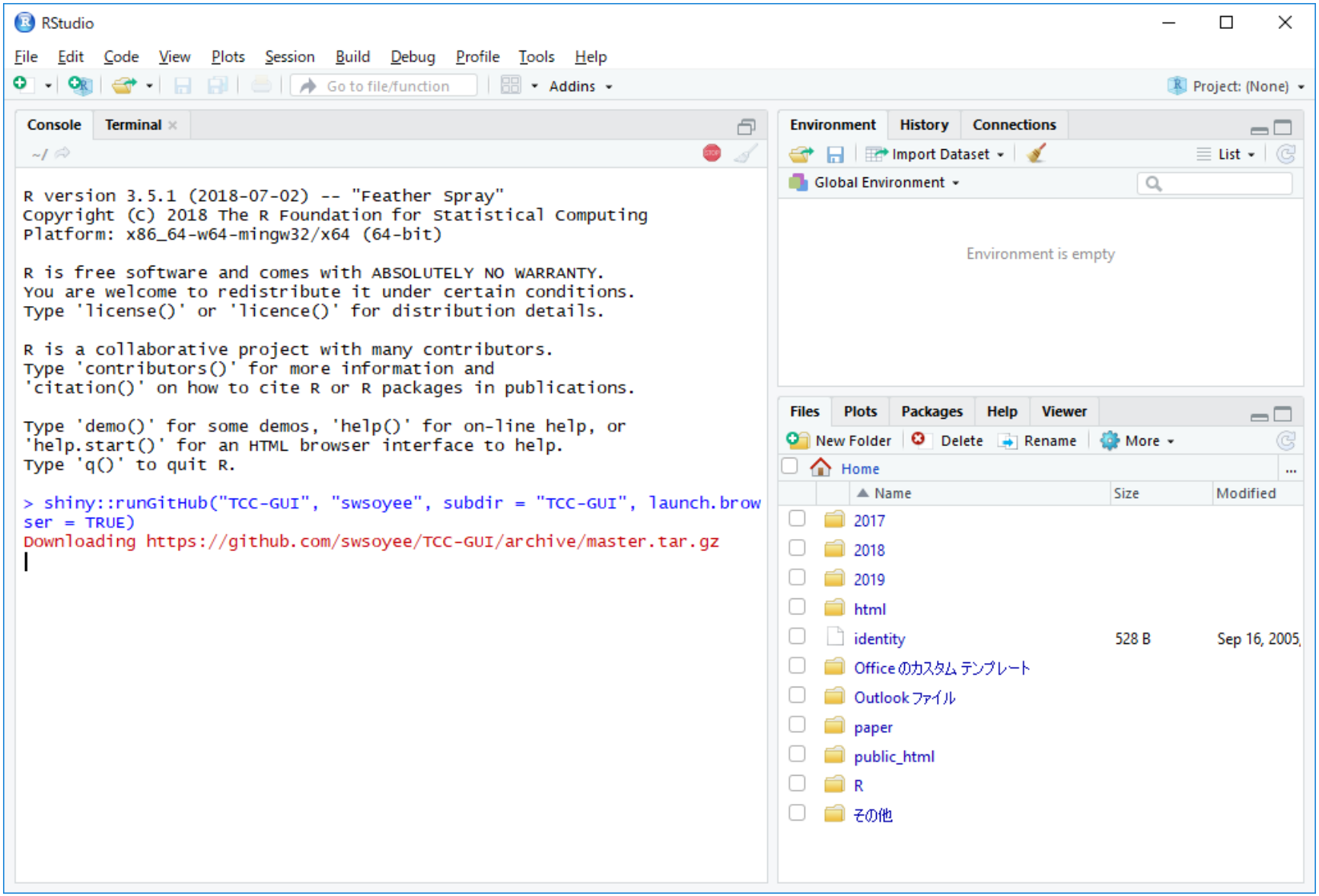
TCC-GUIローカル版

「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上でコピペ実行すれば、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。



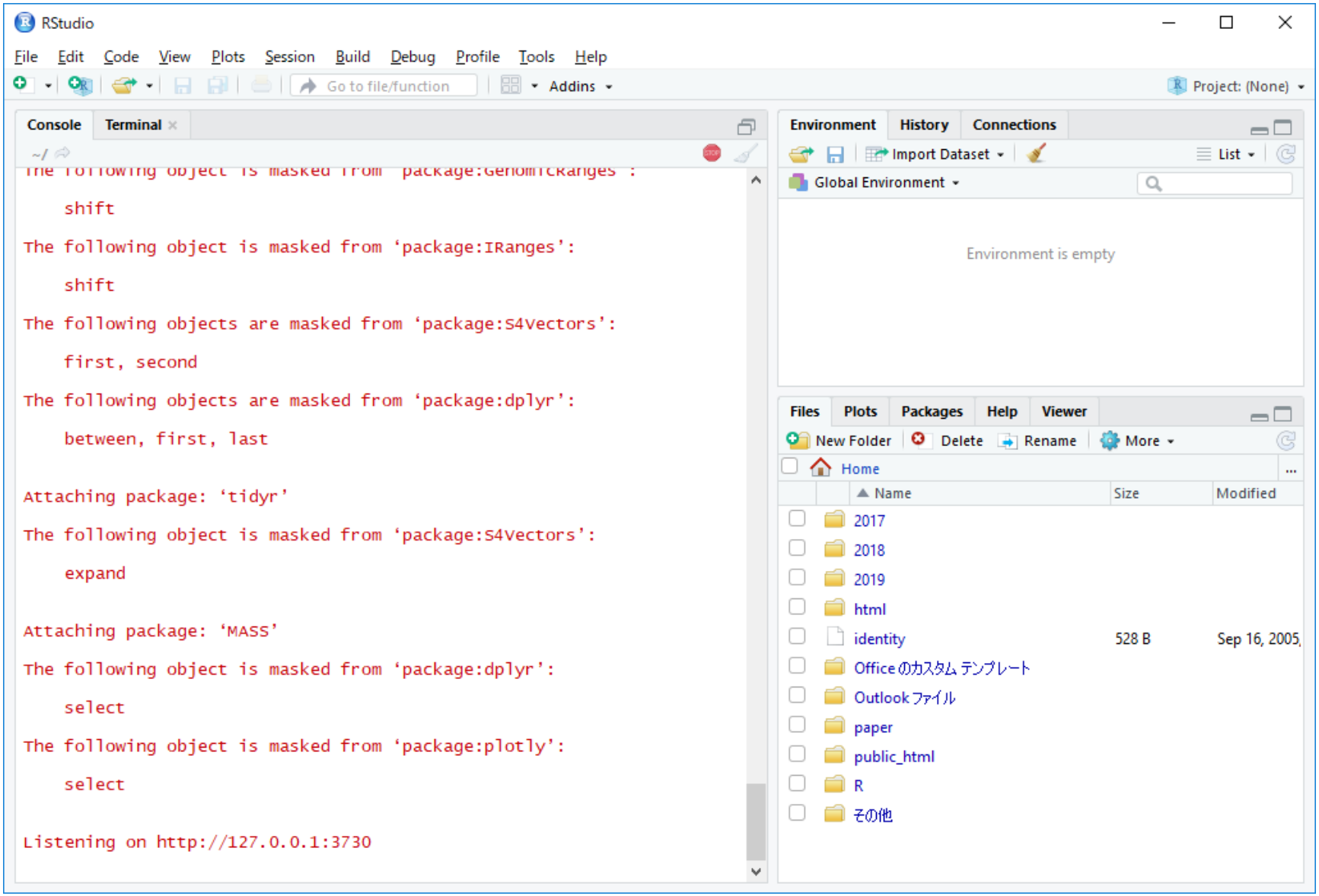
TCC-GUIローカル版

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。



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。



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。