2019.07.22版

ほぼ完成です。最終回のレポート課題はありません。「基本的な考え方と解析戦略の変遷（スライド17-29あたり）」は確実に省略しますので講義前に17-29については自分で見ておいてください。スライド134-162についても残り時間次第です。最終回ですので、アンケートのほうもよろしくお願いします

農学生命情報科学特論I 第4回

[1]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プログラム
[2]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/

予定です

講義予定

- 第1回（2019年07月01日）
  - カウント情報取得の続き
  - データの正規化（RPK, RPM, RPKM/FPKM）
- 第2回（2019年07月08日）
  - サンプル間クラスタリング、Rのクラスオブジェクト
  - RのReference Manualの読み解き方、データセットの連結
- 第3回（2019年07月22日）
  - 発現変動解析（多重比較問題とFDR）、各種プロット（M-A plot）
  - 発現変動解析（デザイン行列や3群間比較）
- 第4回（2019年07月29日）
  - 機能解析（発現変動遺伝子セット解析）、GSEA、MSigDB
  - GSVAの実行

Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

機能解析

機能解析の実体は、遺伝子セットの発現変動解析。
発現に差のある遺伝子セットを探したい、ということ。

- Gene Ontology (GO)解析(発現に差のあるGO termを探索)
  - 基本3カテゴリ(Cellular Component (CC), Molecular Function (MF), Biological Process (BP))のどれでも可能
    - 例:肝臓の空腹状態 vs. 満腹状態のGO(BP)解析の結果、「脂肪酸β酸化」関連GO term (GO:0006635)が動いていることが分かった
- パスウェイ解析(発現に差のあるパスウェイを探索)
  - KEGG Pathway, BioCarta, Reactome pathway databaseのどれでも可能
    - 例:酸化的リン酸化パスウェイ関連遺伝子セットが糖尿病患者で動いていた
- モチーフ解析(発現に差のあるモチーフを探索)
  - 同じ3'-UTR microRNA結合モチーフをもつ遺伝子セット
  - 同じ転写因子結合領域(TATA-boxなど)をもつ遺伝子セット
    - 例:TATA-boxをもつ遺伝子セットがG1群 vs. G2群比較で動いていた
- ...

機能解析の全体像

①で機能解析（発現変動遺伝子セット解析）に関する全体像を述べています。

(Rで)塩基配列解析

(last modified 2019/07/17, since 2010)

このウェブページの
必要なパッケージを
Macintosh2019.0
しました。(2018/

What's new? (過

- 「解析 | 発現変
  TCC(Sun 2013
  たからです(山本

- 解析 | 発現変動 | 5群間 | 対応なし | 複製あり | TCC(Sun 2013) (last modified 2015/11/05)推奨
- 解析 | 発現変動 | 時系列 | について (last modified 2018/06/27) NEW
- 解析 | 発現変動 | 時系列 | maSigPro(Nueda 2014) (last modified 2015/08/16)
- 解析 | 発現変動 | 時系列 | Bayesian model-based clustering(Nascimento 2012) (last modified 2012/09/10)
- 解析 | 発現変動 | exon/isoform | について (last modified 2018/04/12)
- 解析 | 発現変動 | exon/isofo | DEXseq(Anders 2012) (last modified 2014/06/23)
- 解析 | 機能解析 | について ① modified 2018/06/24) NEW
- 解析 | 機能解析 | GMTファイル取得 | について (last modified 2018/06/27) NEW
- 解析 | 機能解析 | GMTファイル取得 | EGSEAdata(Alhamdoosh 2017) (last modified 2018/06/27) NEW
- 解析 | 機能解析 | GMTファイル取得 | GeneSetDB(Araki 2012) (last modified 2018/06/27) NEW
- 解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian 2005) (last modified 2018/06/25) NEW
- 解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan 2018) (last modified 2018/06/25) NEW
- 解析 | 機能解析 | 遺伝子セット 解析 | GSVA(Hänzelmann 2013)(last modified 2018/06/26) NEW
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について (last modified 2018/06/27) NEW
- 解析 | 機能解析 | 遺伝
- 解析 | 機能解析 | 遺伝
- 解析 | 機能解析 | 遺伝
- 解析 | 機能解析 | パス
- 解析 | 機能解析 | パス
- 解析 | 機能解析 | パス
- 解析 | 機能解析 | パス
- 解析 | 分類 | について

解析 | 機能解析 | について NEW

多少間違えているかもしれませんが、とりあえず2018年6月現在の私の理解に基づいて、全貌をざっくりと書きます。機能解析という項目ですが、実質的にはGene Set Enrichment Analysis (GSEA)を意識した内容です。GSEAは、発現変動解析の枠組みに属するものです。通常、発現変動解析は発現変動遺伝子(DEG)を個別に検出することを目的として行っています。そして、その後の解析の多くは、「**DEGとして検出されたもの(or 発現変動上位遺伝子群)の中に、何か特定の機能と関連したものが多く濃縮(Enrich)されているかどうか**」を調べるというものでした。例えば、「細胞周期(Cell Cycle)に関連する遺伝子群がDEGの中に多く含まれているかどうか」を調べるというものです。

Mootha et al.,2003の論文では、視点を変えた解析を行っています。特定の機能に属する遺伝子群に関する知識(knowledge)はGene Ontology(GO)やKEGG Pathwayなどで整理されつつあったので、「比較するグループ間で、例えば細胞周期に関連する遺伝子群のような**特定の機能を果たす遺伝子群(Gene Set)が全体として発現変動しているかどうか**を調べる」戦略を提唱しています。そして、このような知識を利用した解析法(knowledge-based analysis)の考え方をbrush upさせたのが、Subramanian et al., 2005の論文のタイトルでもある、Gene Set Enrichment Analysis (GSEA)です。この方法は、遺伝子セット解析(Gene Set Analysis; GSA)とも総称されますが、事実上GSEAの考え方そのものを指します。このような発現変動に関連した機能解析を行う際に、**遺伝子セットとして遺伝子オントロジー(GO)の情報を用いる場合はGO解析になり、遺伝子セットとしてKEGG PathwayやReactomeの情報を用いる場合はパスウェイ解析**になります。GSEAが爆発的に流行ったのは、以下に示すような様々な要因が重なったためと考えられます：

機能解析の全体像

①第1世代の機能解析（ORA）。②GSEAが含まれる第2世代の機能解析（FCS）。ORAやFCSの用語を含む詳細については後述。と書いてありますが講義では省きます。

解析 | 機能解析 | について NEW

多少間違えているかもしれませんが、とりあえず2018年6月現在の私の理解に基づいて、全貌をざっくりと書きます。機能解析という項目ですが、実質的にはGene Set Enrichment Analysis (GSEA)を意識した内容です。GSEAは、発現変動解析の枠組みに属するものです。通常、発現変動解析は発現変動遺伝子(DEG)を個別に検出することを目的として行っています。そして、その後の解析の多くは、「**DEGとして検出されたもの(or 発現変動上位遺伝子群)の中に、何か特定の機能と関連したものが多く濃縮(Enrich)されているかどうか**」を調べるというものでした。例えば、「細胞周期(Cell Cycle)に関連する遺伝子群がDEGの中に多く含まれているかどうか」を調べるというものです。

Mootha et al.,2003の論文では、視点を変えた解析を行っています。特定の機能に属する遺伝子群に関する知識(knowledge)はGene Ontology(GO)やKEGG Pathwayなどで整理されつつあったので、「比較するグループ間で、例えば細胞周期に関連する遺伝子群のような**特定の機能を果たす遺伝子群(Gene Set)が全体として発現変動しているかどうか**を調べる」戦略を提唱しています。そして、このような知識を利用した解析法(knowledge-based analysis)の考え方をbrush upさせたのが、Subramanian et al., 2005の論文のタイトルでもある、Gene Set Enrichment Analysis (GSEA)です。この方法は、遺伝子セット解析(Gene Set Analysis; GSA)とも総称されますが、事実上GSEAの考え方そのものを指します。このような発現変動に関連した機能解析を行う際に、**遺伝子セットとして遺伝子オントロジー(GO)の情報を用いる場合はGO解析になり、遺伝子セットとしてKEGG PathwayやReactomeの情報を用いる場合はパスウェイ解析**になります。GSEAが爆発的に流行ったのは、以下に示すような様々な要因が重なったためと考えられます：

1. GOやKEGGなど知識の整備が進んでいた(時代背景)
2. マイクロアレイもコストが下がり流行っていた(時代背景)
3. それらをうまく利用して、従来の機能解析を「知識ベースの発現変動解析」に切り替えた(クールな発想の転換)
4. エンドユーザが使いやすいようにMolecular Signatures Database (MSigDB)上で遺伝子セット解析のための基盤情報を提供(丁寧なアフターフォロー)

①

②

①ここがキモ。②遺伝子セット次第でGO解析やPathway解析にもなり得る。

遺伝子セット次第で…

解析 | 機能解析 | について NEW

多少間違えているかもしれませんが、とりあえず2018年6月現在の私の理解に基づいて、全貌をざっくりと書きます。機能解析という項目ですが、実質的にはGene Set Enrichment Analysis (GSEA)を意識した内容です。GSEAは、発現変動解析の枠組みに属するものです。通常、発現変動解析は発現変動遺伝子(DEG)を個別に検出することを目的として行っています。そして、その後の解析の多くは、「**DEGとして検出されたもの(or 発現変動上位遺伝子群)の中に、何か特定の機能と関連したものが多く濃縮(Enrich)されているかどうか**」を調べるというものでした。例えば、「細胞周期(Cell Cycle)に関連する遺伝子群がDEGの中に多く含まれているかどうか」を調べるというものです。

Mootha et al.,2003の論文では、視点を変えた解析を行っています。特定の機能に属する遺伝子群に関する知識(knowledge)はGene Ontology(GO)やKEGG Pathwayなどで整理されつつあったので、「比較するグループ間で、例えば細胞周期に関連する遺伝子群のような**特定の機能を果たす遺伝子群(Gene Set)が全体として発現変動しているかどうか**を調べる」戦略を提唱しています。そして、このような知識を利用した解析法(knowledge-based analysis)の考え方をbrush upさせたのが①Subramanian et al., 2005の論文のタイトルでもある、Gene Set Enrichment Analysis (GSEA)です。この方法は、遺伝子セット解析(Gene Set Analysis; GSA)とも総称されますが、事実上GSEAの考え方そのものを指します。このような発現変動に関連した機能解析を行う際に、**遺伝子セットとして遺伝子オントロジー(GO)の情報を用いる場合はGO解析になり、遺伝子セットとしてKEGG PathwayやReactomeの情報を用いる場合はパスウェイ解析**になります。GSEAが爆発的に流行ったのは、②以下に示すような様々な要因が重なったためと考えられます:

1. GOやKEGGなど知識の整備が進んでいた(時代背景)
2. マイクロアレイもコストが下がり流行っていた(時代背景)
3. それらをうまく利用して、従来の機能解析を「知識ベースの発現変動解析」に切り替えた(クールな発想の転換)
4. エンドユーザが使いやすいようにMolecular Signatures Database (MSigDB)上で遺伝子セット解析のための基盤情報を提供(丁寧なアフターフォロー)

MSigDB

①MSigDBというサイトで、遺伝子セット情報の.gmtという拡張子のついたファイルが提供されています。

解析 | 機能解析 | について NEW

多少間違えているかもしれませんが、とりあえず2018年6月現在の私の理解に基づいて、全貌をざっくりと書きます。機能解析という項目ですが、実質的にはGene Set Enrichment Analysis (GSEA)を意識した内容です。GSEAは、発現変動解析の枠組みに属するものです。通常、発現変動解析は発現変動遺伝子(DEG)を個別に検出することを目的として行っています。そして、その後の解析の多くは、「**DEGとして検出されたもの(or 発現変動上位遺伝子群)の中に、何か特定の機能と関連したものが多く濃縮(Enrich)されているかどうか**」を調べるというものでした。例えば、「細胞周期(Cell Cycle)に関連する遺伝子群がDEGの中に多く含まれているかどうか」を調べるというものです。

Mootha et al.,2003の論文では、視点を変えた解析を行っています。特定の機能に属する遺伝子群に関する知識(knowledge)はGene Ontology(GO)やKEGG Pathwayなどで整理されつつあったので、「比較するグループ間で、例えば細胞周期に関連する遺伝子群のような**特定の機能を果たす遺伝子群(Gene Set)が全体として発現変動しているかどうか**を調べる」戦略を提唱しています。そして、このような知識を利用した解析法(knowledge-based analysis)の考え方をbrush upさせたのが、Subramanian et al., 2005の論文のタイトルでもある、Gene Set Enrichment Analysis (GSEA)です。この方法は、遺伝子セット解析(Gene Set Analysis; GSA)とも総称されますが、事実上GSEAの考え方そのものを指します。このような発現変動に関連した機能解析を行う際に、**遺伝子セットとして遺伝子オントロジー(GO)の情報を用いる場合はGO解析になり、遺伝子セットとしてKEGG PathwayやReactomeの情報を用いる場合はパスウェイ解析**になります。GSEAが爆発的に流行ったのは、以下に示すような様々な要因が重なったためと考えられます:

1. GOやKEGGなど知識の整備が進んでいた(時代背景)
2. マイクロアレイもコストが下がり流行っていた(時代背景)
3. それらをうまく利用して、従来の機能解析を「知識ベースの発現変動解析」に切り替えた(クールな発想の転換)
4. エンドユーザが使いやすいようにMolecular Signatures Database (MSigDB)上で遺伝子セット解析のための基盤情報を提供(丁寧なアフターフォロー)

①

MSigDB

①MSigDBというサイトで、遺伝子セット情報の.gmtという拡張子のついたファイルが提供されています。

解析 | 機能解析 | について NEW

多少間違えているかもしれませんが、とりあえず2018年6月現在の私の理解に基づいて、全貌をざっくりと書きます。機能解析という項目ですが、実質的にはGene Set Enrichment Analysis (GSEA)を意識した内容です。GSEAは、発現変動解析の枠組みに属するも
して、その後の
たものが多く濃
子群がDEGの
Mootha et al.,20
はGene Ontolog
する遺伝子群の
唱しています。
Subramanian et
解析(Gene Set
した機能解析を
トとしてKEGG
以下に示すよう

1. GOやKE
2. マイクロ
3. それらを
4. エンドユ
   を提供(

MSigDBは、様々な遺伝子セットの情報を含むデータベースです。GSEAのサイトの右上にある図中で、Gene set Databaseと書かれているものに相当します。ユーザは、MSigDBから自分が調べたい遺伝子セット情報を含むGMTファイル(.gmt)を予めダウンロードしておく必要があります。従って、入力ファイルとして必要なものは2種類(マイクロアレイやRNA-seqで得られた発現行列のファイルと.gmtファイル)になります。これらを入力として、GSEAプログラムそのものや、その後提案された様々な遺伝子セット解析用プログラムを実行するのです。発現変動に関連した機能解析用プログラムの中には、パスウェイ解析に特化したもの、GO解析もできるもの、遺伝子セット解析全般ができるものなどいろいろあります。

現実問題として、**エンドユーザが特に手元にあるRNA-seqの発現データを用いてプログラムを実行する障壁は非常に高い**です。理由は、発現情報ファイル中のfeature IDと.gmtファイル中のIDとの対応付けを行う部分が厄介だからです。featureという曖昧な用語を用いているのは、発現行列の各行が仮にgeneを指し示すIDに限定されていたとしても、Ensembl gene ID、Entrez gene ID、gene symbolsなどが現実にあり得ます。また、exonやtranscriptを指し示すIDかもしれませんし、マイクロアレイデータの場合は各メーカーによって異なる独自のID (例えばAffymetrixのID)になります。それゆえ、**featureという曖昧な表現がよく使われる**のです。

現在MSigDBでは、Entrez gene ID(ファイル名の最後のほうが.entrez.gmt)とgene symbols(ファイル名の最後のほうが.symbols.gmt)の2種類が提供されています。それゆえ、手元の発現データファイル中のfeature IDがもしEntrez gene IDなら、.entrez.gmtを利用することになります。feature IDがもしEntrez gene ID以外なら、(多くの場合はgene symbolsとの対応付けは行える状況にあるので)予めfeature IDをgene symbolsにどうにかして変換してから、.symbols.gmtを利用してプログラムを実行することになります。

但し、**発現行列データ側の前処理として、同一feature IDsの重複除去**を行っておく必要もあります。有意な発現変動遺伝子セットを検出する際に、複数個存在する同一feature IDsの情報が過大評価されないようにするのが目的です。これも、実際に重複除去をやろうとすると色々と厄介です。例えば、発現変動遺伝子セット解析用パッケージのGSVAは、**発現行列データの格納形式としてExpressionSetオブジェクトを利用**しています。そして、前処理として重複除去を行う際にgenefilterパッケージ内のnsFilter関数(入力がExpressionSetオブジェクト)を利用しています。ExpressionSetオブジェクトは、特にユーザに意識させることなく(Rで)マイクロアレイデータ解析上でも使っていましたが、このようなデータ形式を取り扱うスキルもきっちり身につけていく必要があります。尚、マイクロアレイデータの頃はExpressionSetオブジェクトがよく使われていましたが、RNA-seqカウントデータの現在はSummarizedExperimentやRangedSummarizedExperimentがよく使われます。

gmtファイルを入手

①MSigDBというサイトで、遺伝子セット情報の.gmtという拡張子のついたファイルを予め入手しておかねばなりません。

解析 | 機能解析 | について NEW

多少間違えているかもしれませんが、とりあえず2018年6月現在の私の理解に基づいて、全貌をざっくりと書きます。機能解析という項目ですが、実質的にはGene Set Enrichment Analysis (GSEA)を意識した内容です。GSEAは、発現変動解析の枠組みに属するもして、その後の**たものが多く濃**子群がDEGの
Mootha et al.,20
はGene Ontolog
する遺伝子群の
唱しています。
Subramanian et
解析(Gene Set
した機能解析を
トとしてKEGG
以下に示すよう

1. GOやKE
2. マイクロ
3. それらを
4. エンドユ
を提供(

MSigDBは、様々な遺伝子セットの情報を含むデータベースです。GSEAのサイトの右上にある図中で、Gene set Databaseと書かれているものに相当します。ユーザは、MSigDBから自分が調べたい遺伝子セット情報を含むGMTファイル(.gmt)を予めダウンロードしておく必要があります。従って、入力ファイルとして必要なものは2種類(マイクロアレイやRNA-seqで得られた発現行列のファイルと.gmtファイル)になります。これらを入力として、GSEAプログラムそのものや、その後提案された様々な遺伝子セット解析用プログラムを実行するのです。発現変動に関連した機能解析用プログラムの中には、パスウェイ解析に特化したもの、GO解析もできるもの、遺伝子セット解析全般ができるものなどいろいろあります。

現実問題として、**エンドユーザが特に手元にあるRNA-seqの発現データを用いてプログラムを実行する障壁は非常に高い**です。理由は、発現情報ファイル中のfeature IDと.gmtファイル中のIDとの対応付けを行う部分が厄介だからです。featureという曖昧な用語を用いているのは、発現行列の各行が仮にgeneを指し示すIDに限定されていたとしても、Ensembl gene ID、Entrez gene ID、gene symbolsなどが現実にあり得ます。また、exonやtranscriptを指し示すIDかもしれませんし、マイクロアレイデータの場合は各メーカーによって異なる独自のID (例えばAffymetrixのID)になります。それゆえ、**featureという曖昧な表現がよく使われる**のです。

現在MSigDBでは、Entrez gene ID(ファイル名の最後のほうが.entrez.gmt)とgene symbols(ファイル名の最後のほうが.symbols.gmt)の2種類が提供されています。それゆえ、手元の発現データファイル中のfeature IDがもしEntrez gene IDなら、.entrez.gmtを利用することになります。feature IDがもしEntrez gene ID以外なら、(多くの場合はgene symbolsとの対応付けは行える状況にあるので)予めfeature IDをgene symbolsにどうにかして変換してから、.symbols.gmtを利用してプログラムを実行することになります。

但し、**発現行列データ側の前処理として、同一feature IDsの重複除去**を行っておく必要もあります。有意な発現変動遺伝子セットを検出する際に、複数個存在する同一feature IDsの情報が過大評価されないようにするのが目的です。これも、実際に重複除去をやろうとすると色々と厄介です。例えば、発現変動遺伝子セット解析用パッケージのGSVAは、**発現行列データの格納形式としてExpressionSetオブジェクトを利用**しています。そして、前処理として重複除去を行う際にgenefilterパッケージ内のnsFilter関数(入力がExpressionSetオブジェクト)を利用しています。ExpressionSetオブジェクトは、特にユーザに意識させることなく(Rで)マイクロアレイデータ解析上でも使っていましたが、このようなデータ形式を取り扱うスキルもきっちり身につけていく必要があります。尚、マイクロアレイデータの頃はExpressionSetオブジェクトがよく使われていましたが、RNA-seqカウントデータの現在はSummarizedExperimentやRangedSummarizedExperimentがよく使われます。

入力ファイルは2つ

発現変動遺伝子検出の場合は、入力が1つ（発現行列データ）であった。①発現変動遺伝子セット解析の場合は、発現データファイルに加えて、どの遺伝子がどの遺伝子セットに属するかという情報を含むgmtファイルも必要です。

解析 | 機能解析 | について NEW

多少間違えているかもしれませんが、とりあえず2018年6月現在の
析という項目ですが、実質的にはGene Set Enrichment Analysis (GSEA)を意識した内容です。GSEAは、発現変動解析の枠
組みに属するも
して、その後の
たものが多く濃
子群がDEGの
Mootha et al.,20
はGene Ontolog
する遺伝子群の
唱しています。
Subramanian et
解析(Gene Set
した機能解析を
トとしてKEGG
以下に示すよう

1. GOやKE
2. マイクロ
3. それらを
4. エンドユ
   を提供(

MSigDBは、様々な遺伝子セットの情報を含むデータベースです。GSEAのサイトの右上にある図中で、Gene set Databaseと書かれているものに相当します。ユーザは、MSigDBから自分が調べたい遺伝子セット情報を含むGMTファイル(.gmt)を予めダウンロードしておく必要があります。従って、入力ファイルとして必要なものは2種類(マイクロアレイやRNA-seqで得られた発現行列のファイルと.gmtファイル)になります。これらを入力として、GSEAプログラムそのものや、その後提案された様々な遺伝子セット解析用プログラムを実行するのです。発現変動に関連した機能解析用プログラムの中には、パスウェイ解析に特化したもの、GO解析もできるもの、遺伝子セット解析全般ができるものなどいろいろあります。

①

現実問題として、**エンドユーザが特に手元にあるRNA-seqの発現データを用いてプログラムを実行する障壁は非常に高い**です。理由は、発現情報ファイル中のfeature IDと.gmtファイル中のIDとの対応付けを行う部分が厄介だからです。featureという曖昧な用語を用いているのは、発現行列の各行が仮にgeneを指し示すIDに限定されていたとしても、Ensembl gene ID、Entrez gene ID、gene symbolsなどが現実にあり得ます。また、exonやtranscriptを指し示すIDかもしれませんし、マイクロアレイデータの場合は各メーカーによって異なる独自のID (例えばAffymetrixのID)になります。それゆえ、**featureという曖昧な表現がよく使われる**のです。

現在MSigDBでは、Entrez gene ID(ファイル名の最後のほうが.entrez.gmt)とgene symbols(ファイル名の最後のほうが.symbols.gmt)の2種類が提供されています。それゆえ、手元の発現データファイル中のfeature IDがもしEntrez gene IDなら、.entrez.gmtを利用することになります。feature IDがもしEntrez gene ID以外なら、(多くの場合はgene symbolsとの対応付けは行える状況にあるので)予めfeature IDをgene symbolsにどうにかして変換してから、.symbols.gmtを利用してプログラムを実行することになります。

但し、**発現行列データ側の前処理として、同一feature IDsの重複除去**を行っておく必要もあります。有意な発現変動遺伝子セットを検出する際に、複数個存在する同一feature IDsの情報が過大評価されないようにするのが目的です。これも、実際に重複除去をやろうとすると色々と厄介です。例えば、発現変動遺伝子セット解析用パッケージのGSVAは、**発現行列データの格納形式としてExpressionSetオブジェクトを利用**しています。そして、前処理として重複除去を行う際にgenefilterパッケージ内のnsFilter関数(入力がExpressionSetオブジェクト)を利用しています。ExpressionSetオブジェクトは、特にユーザに意識させることなく(Rで)マイクロアレイデータ解析上でも使っていましたが、このようなデータ形式を取り扱うスキルもきっちり身につけていく必要があります。尚、マイクロアレイデータの頃はExpressionSetオブジェクトがよく使われていましたが、RNA-seqカウントデータの現在はSummarizedExperimentやRangedSummarizedExperimentがよく使われます。

様々なプログラムがある

①2005年のGSEA論文発表前後を含めて、様々な方法が提案されています。

解析 | 機能解析 | について NEW

多少間違えているかもしれませんが、とりあえず2018年6月現在の私の理解に基づいて、全貌をざっくりと書きます。機能解析という項目ですが、実質的にはGene Set Enrichment Analysis (GSEA)を意識した内容です。GSEAは、発現変動解析の枠組みに属するもして、その後の**たものが多く濃**子群がDEGの
Mootha et al.,20
はGene Ontolog
する遺伝子群の
唱しています。
Subramanian et
解析(Gene Set
した機能解析を
トとしてKEGG
以下に示すよう

1. GOやKE
2. マイクロ
3. それらを
4. エンドユ
   を提供(

MSigDBは、様々な遺伝子セットの情報を含むデータベースです。GSEAのサイトの右上にある図中で、Gene set Databaseと書かれているものに相当します。ユーザは、MSigDBから自分が調べたい遺伝子セット情報を含むGMTファイル(.gmt)を予めダウンロードしておく必要があります。従って、入力ファイルとして必要なものは2種類(マイクロアレイやRNA-seqで得られた発現行列のファイルと.gmtファイル)になります。これらを入力として、GSEAプログラムそのものや、その後提案された様々な遺伝子セット解析用プログラムを実行するのです。発現変動に関連した機能解析用プログラムの中には、パスウェイ解析に特化したもの、GO解析もできるもの、遺伝子セット解析全般ができるものなどいろいろあります。

現実問題として、**エンドユーザが特に手元にあるRNA-seqの発現データを用いてプログラムを実行する①障壁は非常に高い**です。理由は、発現情報ファイル中のfeature IDと.gmtファイル中のIDとの対応付けを行う部分が厄介だからです。featureという曖昧な用語を用いているのは、発現行列の各行が仮にgeneを指し示すIDに限定されていたとしても、Ensembl gene ID、Entrez gene ID、gene symbolsなどが現実にあり得ます。また、exonやtranscriptを指し示すIDかもしれませんし、マイクロアレイデータの場合は各メーカーによって異なる独自のID (例えばAffymetrixのID)になります。それゆえ、**featureという曖昧な表現がよく使われる**のです。

現在MSigDBでは、Entrez gene ID(ファイル名の最後のほうが.entrez.gmt)とgene symbols(ファイル名の最後のほうが.symbols.gmt)の2種類が提供されています。それゆえ、手元の発現データファイル中のfeature IDがもしEntrez gene IDなら、.entrez.gmtを利用することになります。feature IDがもしEntrez gene ID以外なら、(多くの場合はgene symbolsとの対応付けは行える状況にあるので)予めfeature IDをgene symbolsにどうにかして変換してから、.symbols.gmtを利用してプログラムを実行することになります。

但し、**発現行列データ側の前処理として、同一feature IDsの重複除去**を行っておく必要もあります。有意な発現変動遺伝子セットを検出する際に、複数個存在する同一feature IDsの情報が過大評価されないようにするのが目的です。これも、実際に重複除去をやろうとすると色々と厄介です。例えば、発現変動遺伝子セット解析用パッケージのGSVAは、**発現行列データの格納形式としてExpressionSetオブジェクトを利用**しています。そして、前処理として重複除去を行う際にgenefilterパッケージ内のnsFilter関数(入力がExpressionSetオブジェクト)を利用しています。ExpressionSetオブジェクトは、特にユーザに意識させることなく(Rで)マイクロアレイデータ解析上でも使っていましたが、このようなデータ形式を取り扱うスキルもきっちり身につけていく必要があります。尚、マイクロアレイデータの頃はExpressionSetオブジェクトがよく使われていましたが、RNA-seqカウントデータの現在はSummarizedExperimentやRangedSummarizedExperimentがよく使われます。

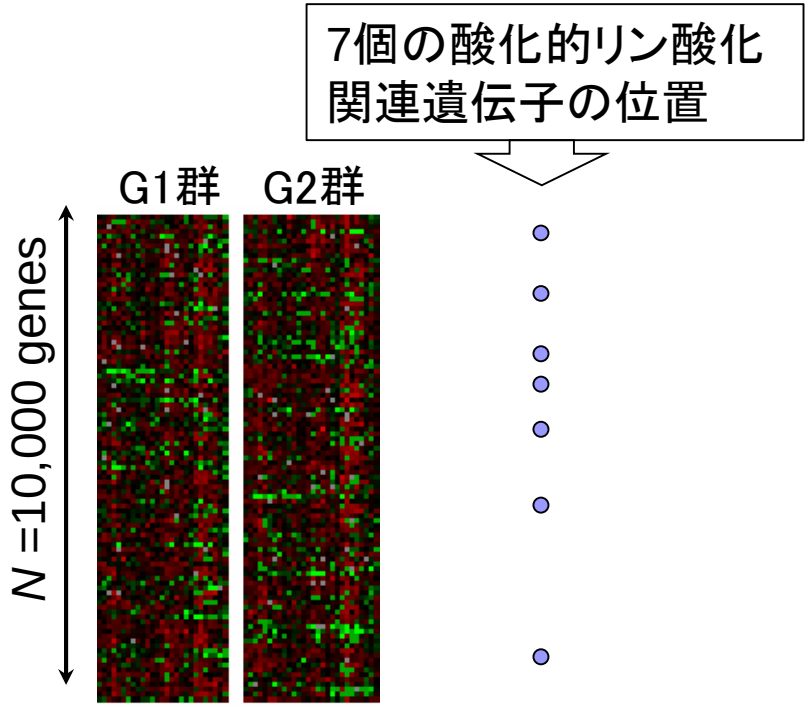
Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

基本的な考え方

酸化的リン酸化関連遺伝子**セット**が変動しているかどうかを調べたい、という問題を考える

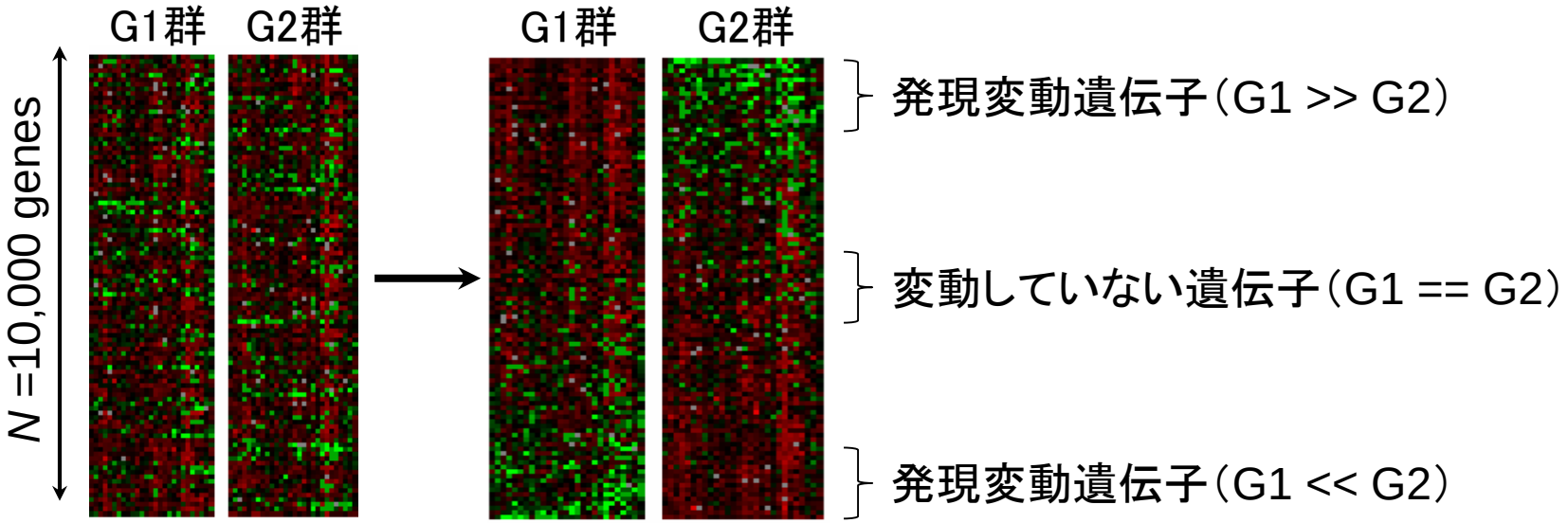
- 発現変動遺伝子セット解析手法（2群間比較用がほとんど）
  - *N* =10,000個の遺伝子からなる2群間比較用データ
  - この中に、XXX関連遺伝子が*n*個含まれている
    - 例：酸化的リン酸化（=XXX）関連遺伝子が7（=*n*）個含まれている



基本はデフォルトでやるようですが、様々な選択肢があります

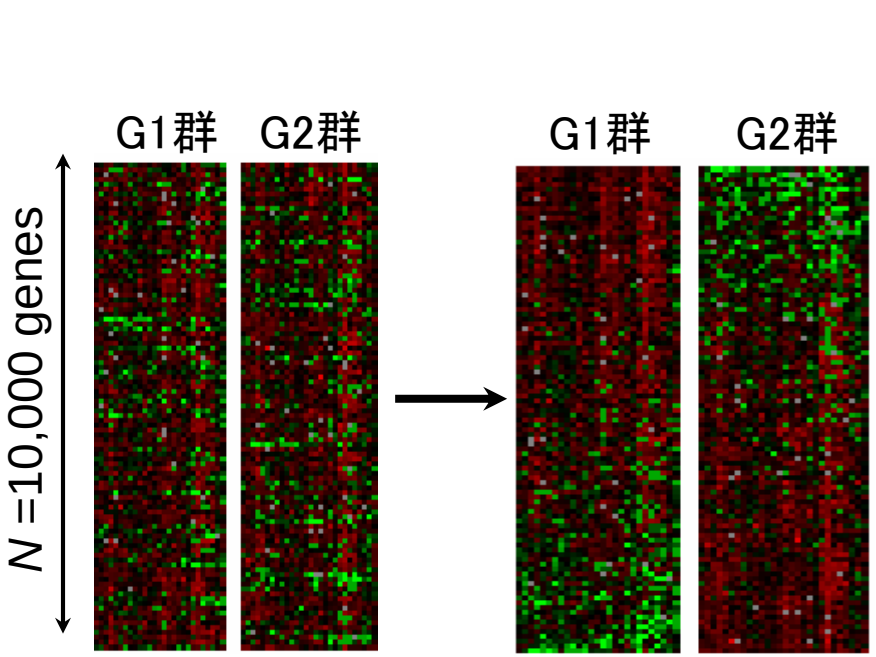
基本的な考え方

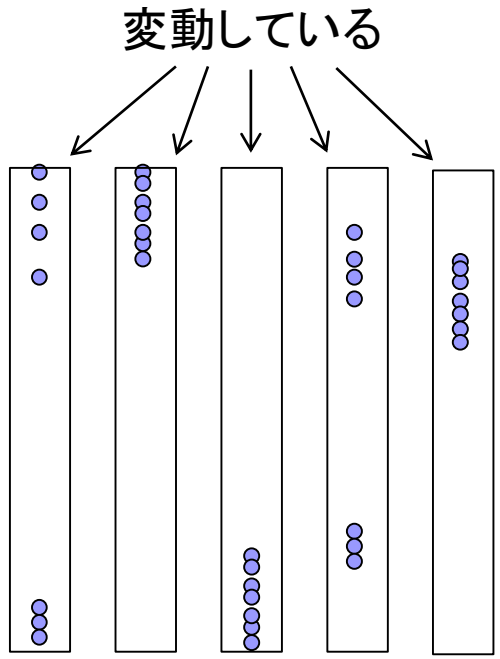
- 遺伝子ごとの発現変動の度合いを数値化
  - 例：t-統計量、log2（G2/G1）、相関係数、…

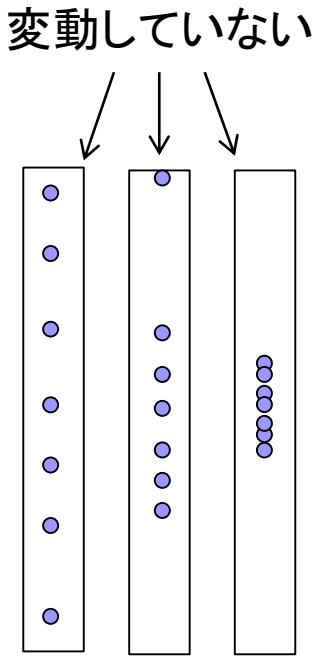


基本的な考え方

- 発現変動順にソート後の酸化的リン酸化関連遺伝子セットのステレオタイプな分布





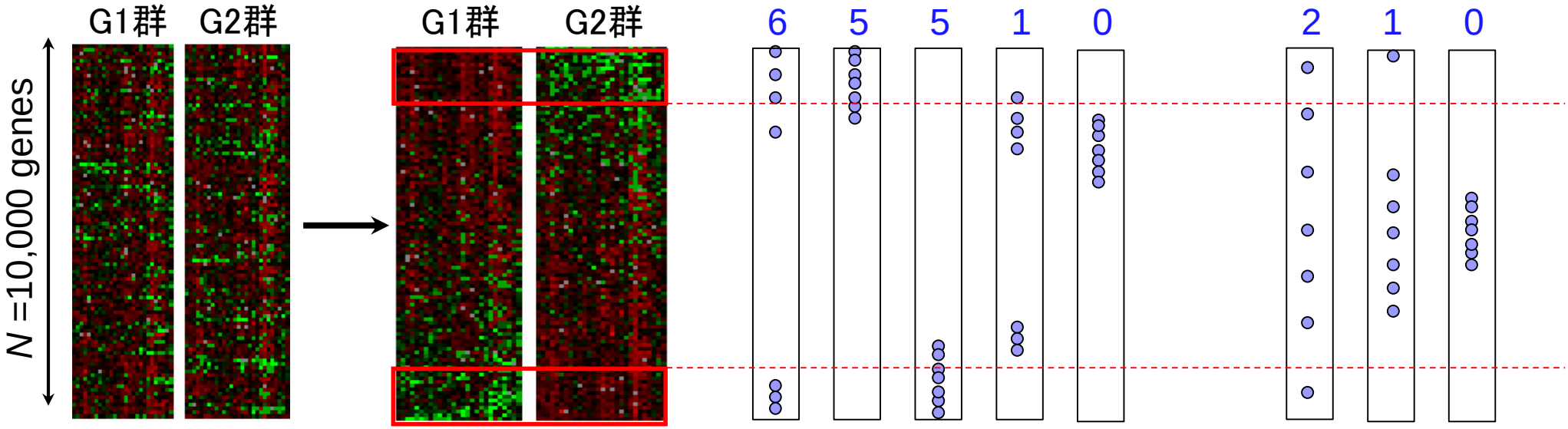


基本的な考え方

基本的な考え方は、「全遺伝子」と「上位のサブセット」のみで、調べたい遺伝子セットの割合が不変という帰無仮説のもとで検定

- Over-Representation Analysis (ORA)
  - 何らかの手段で決めた上位X(=1500)個のうち、x個が酸化的リン酸化関連遺伝子であった

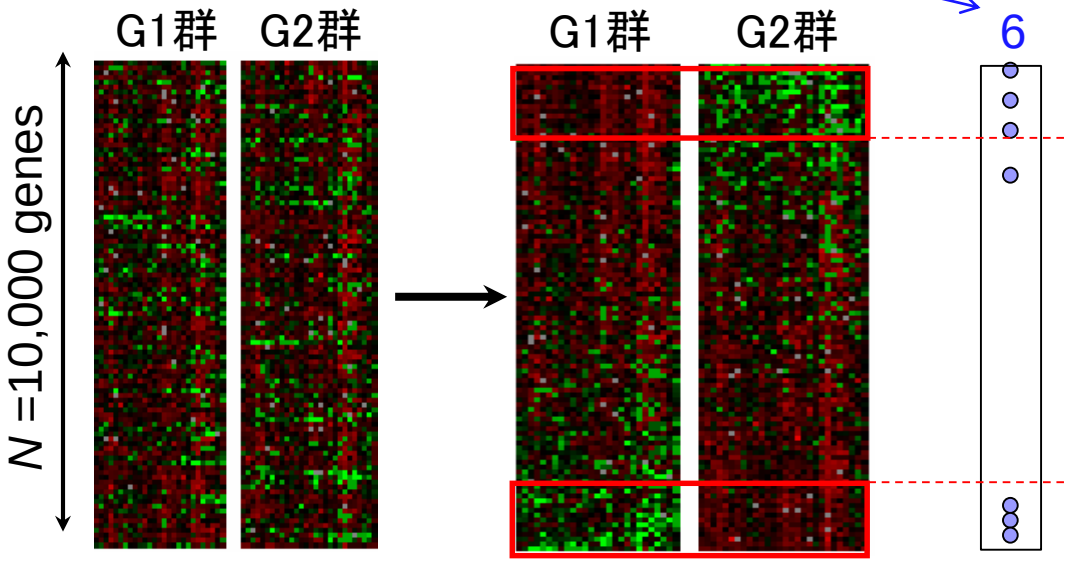
酸化的リン酸化関連遺伝子セット(n=7)が変動していない場合: $x/n \fallingdotseq X/N$ (= 1500/10000)
酸化的リン酸化関連遺伝子セット(n=7)が変動している場合: $x/n >> X/N$ (= 15%)



ORA

2×2分割表（contingency table)に基づく方法。超幾何検定やカイ二乗検定が利用されます

- Over-Representation Analysis (ORA)
  - 何らかの手段で決めた上位*X*(=1500)個のうち、*x*個が酸化的リン酸化関連遺伝子であった



XXX=酸化的リン酸化関連遺伝子セット

	XXX	XXX以外	計
non-DEG数	1	8500-1	N-X
DEG数	6	1500-6	X
計	n	N-n	

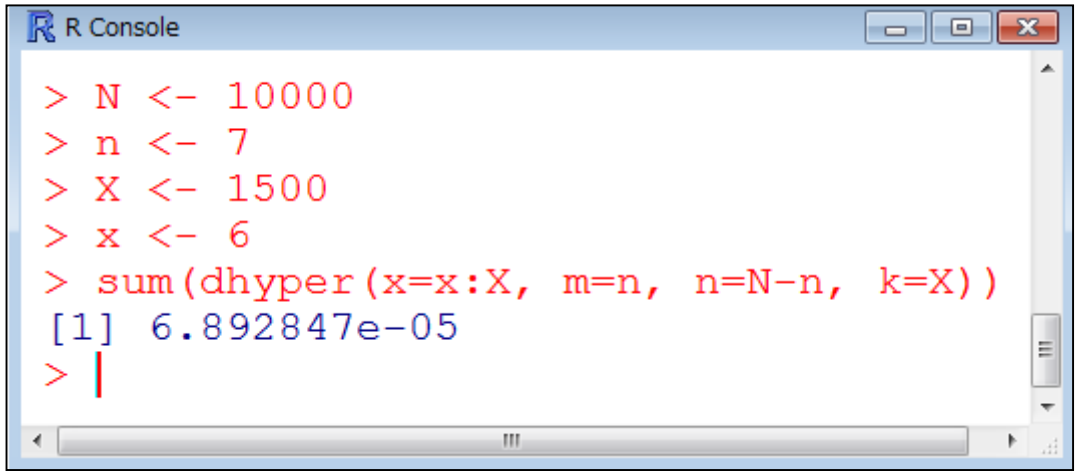
DEGとして1500個抽出したとき、酸化的リン酸化関連遺伝子が6個以上含まれる確率として算出

ORA（超幾何検定）

- *N*=10000個の遺伝子発現データ中にXXX=酸化的リン酸化関連遺伝子は*n*=7個含まれていた。上位*X*=1500個の発現変動遺伝子（DEG）の中に*x*=6個の酸化的リン酸化関連遺伝子が含まれていた
  - 帰無仮説：酸化的リン酸化関連遺伝子の割合はDEGとnon-DEG間で差がない

	XXX	XXX以外	計
non-DEG数	1	8500-1	8500
DEG数	6	1500-6	1500
計	7	9993	10000

	XXX	XXX以外	計
non-DEG数	n-x	$(N$-$n)$-$(X$-$x)$	N-X
DEG数	x	X-x	X
計	n	N-n	N



ORAとしてFisher's hypergeometric test を利用(Tavazoie et al., Nat Genet., **22**: 281–285, 1999)

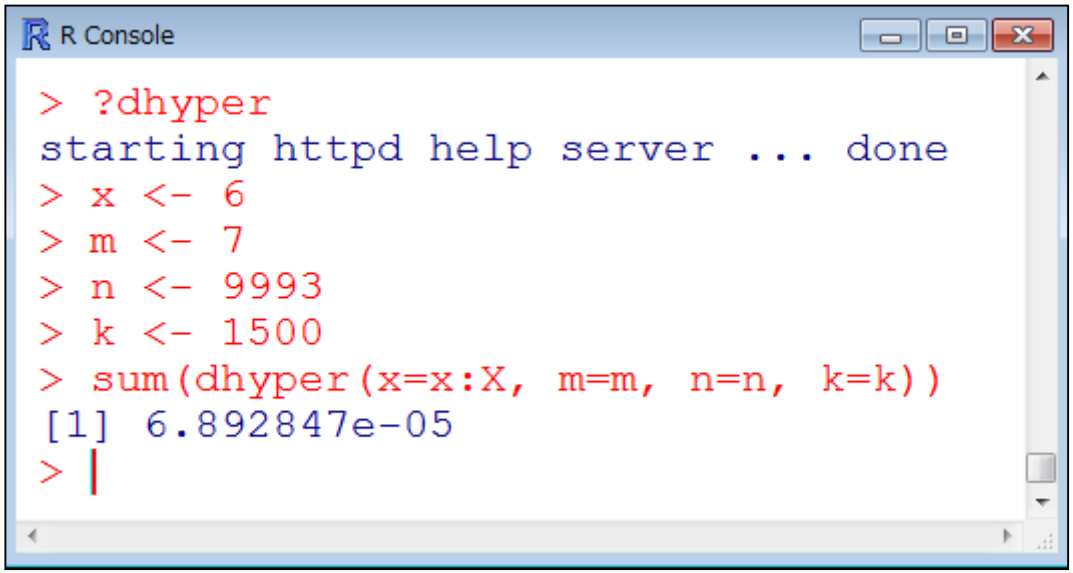
?dhyperマニュアル中の一般的な説明に置き換えるとこんな感じ

ORA（超幾何検定）

- m=7個の白いボールとn=9993個の黒いボールが入った箱があります（トータルで$N=m+n$=10,000個）。この中からk=1500個ランダムに取り出したときにx=6個以上白いボールが含まれる確率を計算しなさい。

	白	黒	計
箱の中	1	9993−(1500−6)	8500
箱の外	6	1500−6	1500
計	7	9993	10000

	白	黒	計
箱の中	m-x	n-(k-x)	m+n-k
箱の外	x	k-x	k
計	m	n	N



ORA(カイ二乗検定)

DEGとして1500個抽出したとき、酸化的リン酸化関連遺伝子が6個以上含まれる確率として算出

R Console

```
> N <- 10000
> n <- 7
> X <- 1500
> x <- 6
> data <- matrix(c((n-x), (N-n)-(X-x), x, (X-x)), ncol=2, byrow=T)
> data
     [,1] [,2]
[1,]    1 8499
[2,]    6 1494
> chisq.test(data)

        Pearson's Chi-squared test with Yates' continuity
        correction

data:  data
X-squared = 22.2032, df = 1, p-value = 2.453e-06

 警告ﾒｯｾｰｼﾞ:
In chisq.test(data) :  カイ自乗近似は不正確かもしれません
>
```

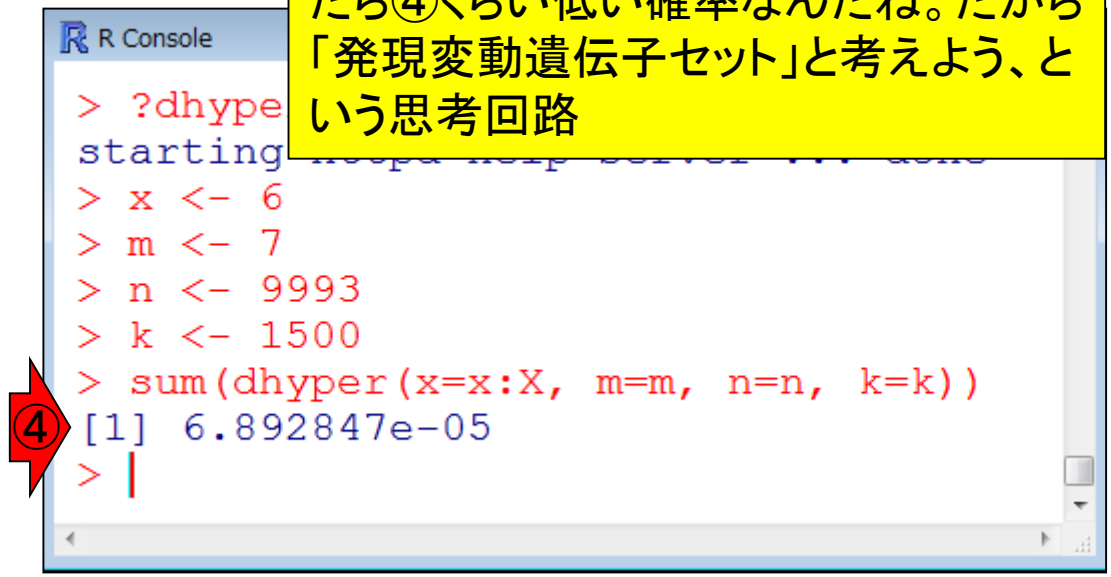
	XXX	XXX以外	計
non-DEG数	1	8500-1	8500
DEG数	6	1500-6	1500
計	7	9993	10000

	XXX	XXX以外	計
non-DEG数	n-x	$(N$-$n)$-$(X$-$x)$	N-X
DEG数	x	X-x	X
計	n	N-n	N

直感は重要

① N=10000個の遺伝子発現データ中にXXX=酸化的リン酸化関連遺伝子は*n*=7個存在する
② 上位*X*=1500個の発現変動遺伝子（DEG）の中に*x*=6個の酸化的リン酸化関連遺伝子が含まれていた
③ 帰無仮説：酸化的リン酸化関連遺伝子の割合はDEGとnon-DEG間で差がない

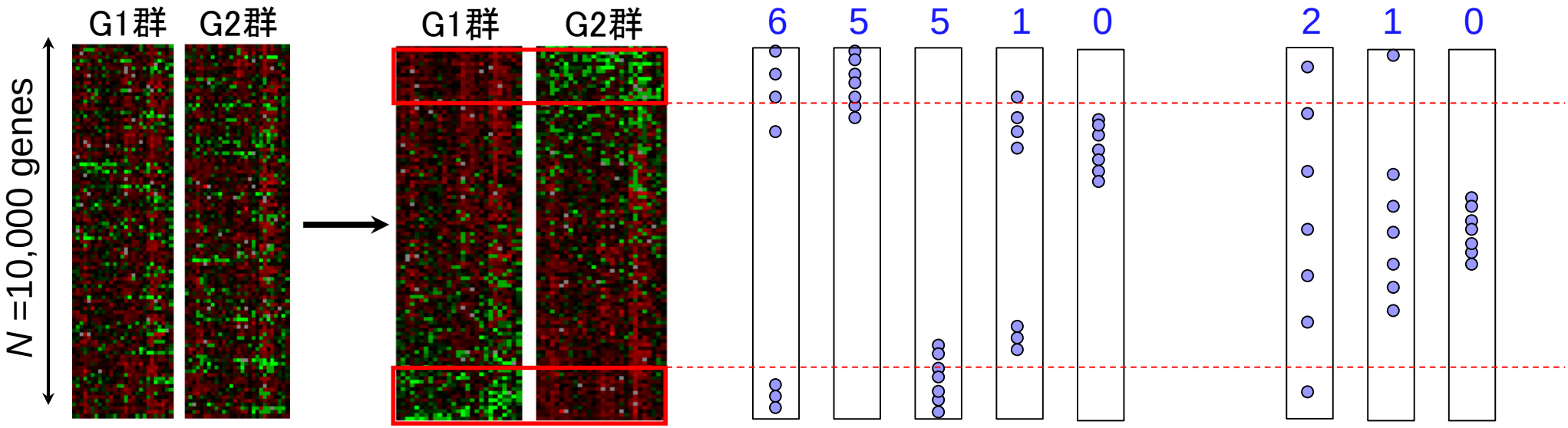
①の段階で、調べたい遺伝子セットは、7/10,000 = 0.07%の割合だと考える。②で6/1,500 = 0.4%の割合に濃縮されていると考える。③今やっているのは、ある2群間比較。もし比較している群間で、この遺伝子セットが全体として発現変動していなかったとしたら…ランダムで1,500個とった時に、この遺伝子セット中の遺伝子が含まれる割合は0.07%なので、個数だと1500×0.07% = 1.05個程度しか含まれないはず。実際に得られたのは6個なので、偶然こんな結果が得られたとは考えにくい。起こるとしたら④くらい低い確率なんだね。だから「発現変動遺伝子セット」と考えよう、という思考回路



ORA

上位1500個のうち、酸化的リン酸化関連遺伝子が7個中4つ以上含まれていれば $p < 0.05$ で検出可能ということを意味する

- Over-Representation Analysis (ORA)



p-value	x=6	x=5	x=4	x=3	x=2	x=1	x=0
超幾何検定	6.89E-05	0.0012	0.0121	0.0737	0.2834	0.6795	1.0000
カイ二乗検定	2.45E-06	0.0003	0.0095	0.1247	0.6337	0.6337	0.5603
Fisher test	6.89E-05	0.0012	0.0121	0.0737	0.2834	1.0000	0.6039

$p < 0.05$を灰色で示した

①GenMAPPは比較的有名

ORA

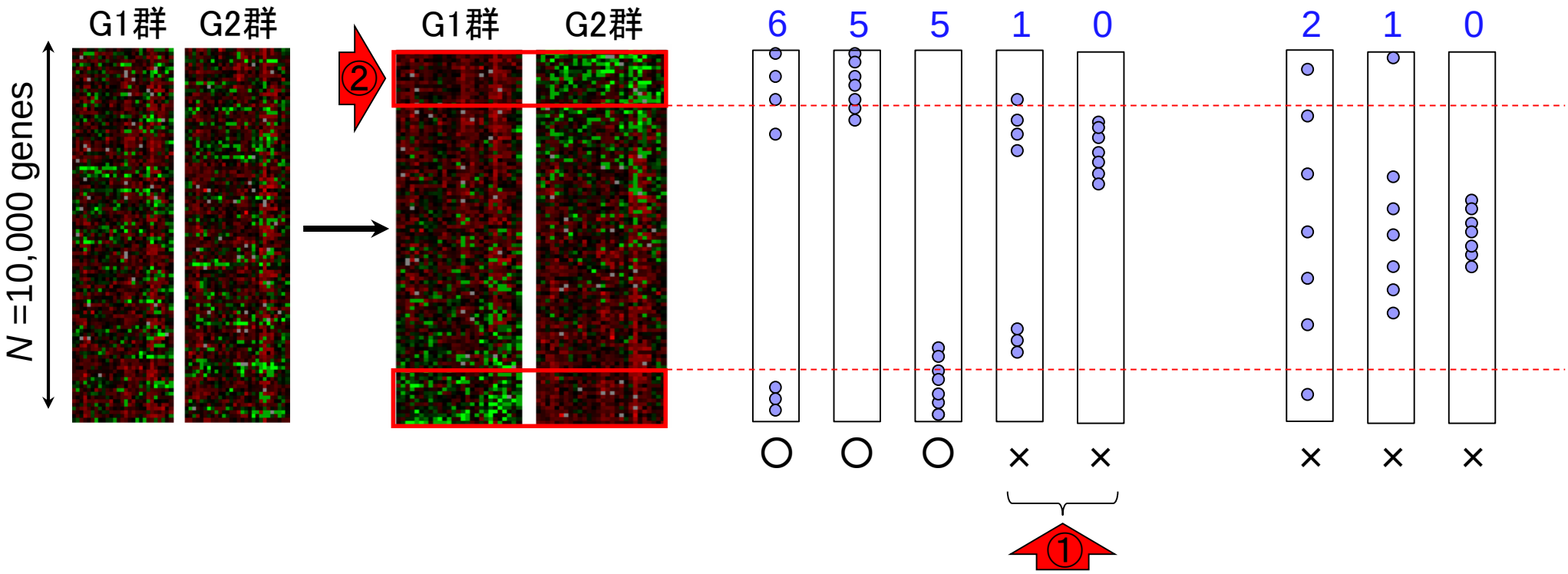
- Over-Representation Analysis (ORA)
  - GenMAPP (Dahlquist et al., *Nature Genet.,* **31**: 19-20, 2002) ①
  - FatiGO (Al-Shahrour et al., *Bioinformatics*, **20**: 578-580, 2004)
  - GOstat (Beissbarth et al., *Bioinformatics*, **20**: 1464-1465, 2004)
  - GOFFA (Sun et al., *BMC Bioinformatics*, **7 Suppl 2**: S23, 2006)
  - agriGO (Du et al., *Nucleic Acids Res.*, **38**: W64-W70, 2010)
  - …

第1世代（ORA）の短所

①全体的には動いているものの、個々の発現変動の度合いが弱い場合に検出困難。②上位X個のX次第で結果が変わる。③情報量低下（発現変動の度合い→カウント情報）

③

	XXX	XXX以外	計
non-DEG数	n-x	$(N$-$n)$-$(X$-$x)$	N-X
DEG数	x	X-x	X
計	n	N-n	N



もちろん分割表ベースの方法(ORA)ではない第2世代以降の方法があります

第2世代（FCS）

- Functional Class Scoring (FCS)
  1. 遺伝子ごとの統計量を算出（発現変動の度合いを数値化）
     例：*t*-統計量、log（G2/G1）、相関係数、…
  2. 目的の遺伝子セットXXX（=酸化的リン酸化関連遺伝子）の偏りを何らかの方法で評価
     - *t*検定（XXX中の遺伝子群の統計量 vs. それ以外の遺伝子群の統計量）
     - Wilcoxon rank sum test （XXX中の遺伝子群の発現変動の順位 vs. それ以外）
     - XXX中の*n*個の遺伝子群の何らかの要約統計量S_{XXX}を計算しておき、*N*個の全遺伝子の中からランダムに*n*個を抽出して同じ統計量を計算する（例えば10万回）。10万回のうちS_{XXX}「以上」（大きければ大きいほど発現変動していることを意味する場合；その逆のときは「以下」）だった回数（例えばj回）に基づいてp値（= j / 100,000）を算出（いわゆるgene set permutationというアプローチ）
     - 本来のG1群 vs. G2群のラベル情報を用いて得られたXXX中の*n*個の遺伝子群の何らかの要約統計量S_{XXX}を計算しておく。ランダムにラベル情報を入れ替えて、同じ統計量を計算することを何回も繰り返して*p*値を算出（いわゆるPhenotype permutationというアプローチ）

第2世代（FCS）

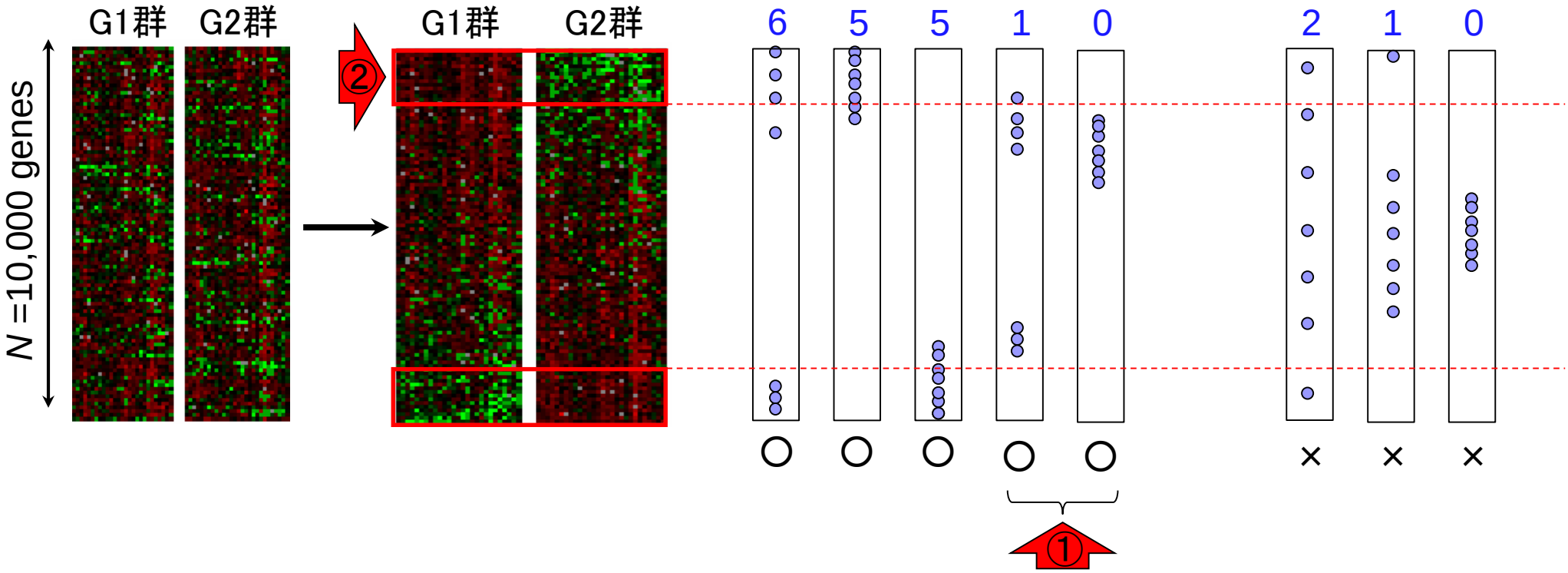
- 第一世代(ORA)の欠点が改善

遺伝子ごとのlog比で考えると、遺伝子を等価に取り扱うのではなく、log比そのものを足し込むことで、発現変動の度合いが大きいほどより大きな重みをかけるようなイメージ

①全体的には動いているものの、個々の発現変動の度合いが弱い場合に検出困難。②上位X個のX次第で結果が変わる。③情報量低下（発現変動の度合い→カウント情報）

③

	XXX	XXX以外	計
non-DEG数	n-x	$(N$-$n)$-$(X$-$x)$	N-X
DEG数	x	X-x	X
計	n	N-n	N



第2世代(FCS)

- Functional Class Scoring (FCS)
  - **GSEA** (Subramanian et al., *PNAS*, **102**: 15545–15550, 2005) ①
  - PAGE (Kim and Volsky, *BMC Bioinformatics*, **6**: 144, 2005)
  - sigPathway (Tian et al., *PNAS*, **102**: 13544–13549, 2005)
  - GSA (Efron and Tibshirani, *Ann. Appl. Stat.*, **1**: 107–129, 2007)
  - GeneTrail (Backes et al., *Nucleic Acids Res.*, **35**: W186–W192, 2007)
  - SAM-GS (Dinu et al., *BMC Bioinformatics*, **8**: 242, 2007)
  - …

最も有名なのが①GSEA。ここでリストアップされているのは、基本的にマイクロアレイデータ解析用なので情報としては古い。よって、(Rで)塩基配列解析ではほとんどリストアップしていない

遺伝子セット解析の課題

突っ込みどころは満載だが、ネガティブなことばかりいってもしょうがないし、この種の機能解析が目的の場合も多い

- （知識ベースの解析法なので）解析対象がアノテーションの情報の豊富な生物種に限定
  - それ以外の生物種は、まずは地道にアノテーション情報を増やしていくことが先決（ではないだろうか）
  - アノテーションの解像度を上げる努力も大事
- アノテーション情報の信頼度が高いとはいえない
  - なんらかのGO termがついていたとしても、その大部分のevidence codeが自動でつけられたもの（IEA, inferrred from electronic annotations）である…
- 遺伝子セット間の独立性の問題
  - 「数百個程度の遺伝子セットの中から、比較するサンプル間で動いている遺伝子セットはどれか？」という解析を遺伝子セット間の独立性を仮定して調べるが、そもそも独立ではない（GO term間の親子関係などから明らか）
    → いくつくらいの遺伝子セットが動いているのか？という問いに答えるすべがない
- 評価に用いられる「よく研究されているデータセット」は答えが完全に分かっているものではない（the actual biology is never fully known!）
  - “感度が高い”と謳っているだけの方法は…（全部の遺伝子セットが動いている → 感度100%）

Maciejewski H., Brief Bioinform., **14**: 504–518, 2014

Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

GO解析用

①遺伝子セットとしてGOの情報を用いる場合。①のリンク先が赤枠。

(Rで)塩基配列解析

(last modified 2019/07/17, since 2010)

このウェブページのR関連部分
必要なパッケージをインスト
Macintosh2019.03.12版)で
しました。(2018/07/18)

What's new? (過去のお知

- 「解析 | 発現変動 | 3群間
TCC(Sun 2013)」で内部
たからです(山本裕一郎 氏

- 解析 | 機能解析 | について (last modified 2018/06/24)
- 解析 | 機能解析 | GMTファイル取得 | について (last modified 2018/07/17)
- 解析 | 機能解析 | GMTファイル取得 | EGSEAdata(Alhamdoosh 2017) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | GeneSetDB(Araki 2012) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian 2005) (last modified 2018/06/25)
- 解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan 2018) (last modified 2018/06/25)
- 解析 | 機能解析 | 遺伝子セット解析 | GSVA(Hänzelmann 2013)(last modified 2018/08/10)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について ① t modified 2019/05/12)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | SeqGSEA(Wang 2014)(last modified 2018/06/25)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | GSVA(Hänzelmann 2013)(last modified 2018/06/26)
- 解析 | 機能解析 | 遺伝子オン
- 解析 | 機能解析 | パスウェイ
- 解析 | 機能解析 | パスウェイ
- 解析 | 機能解析 | パスウェイ
- 解析 | 機能解析 | パスウェイ
- 解析 | 分類 | について(last m

解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について

RNA-seqなどのタグカウントデータから遺伝子オントロジー(GO)解析を行うためのパッケージもいくつか出ています。 遺伝子セット解析(Gene Set Analysis; GSA)という枠組みではGO解析もパスウェイ解析も同じなので、そちらもチェックするといいかもしれません。

R用:

- GAGE : Luo et al., BMC Bioinformatics, 2009
- goseq : Young et al., Genome Biol., 2010
- GOSemSim : Yu et al., Bioinformatics, 2010
- Rスクリプト : Gao et al., Bioinformatics, 2011
- clusterProfiler : Yu et al., OMICS., 2012
- RamiGO : Schröder et al., Bioinformatics, 2013

念のため…

赤枠の全体像がこれ。GO解析用のカテゴリにリストアップしているものでも、パスウェイ解析用にも使えるものが含まれますのでご注意ください。

解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について

RNA-seqなどのタグカウントデータから遺伝子オントロジー(GO)解析を行うためのパッケージもいくつか出ています。 遺伝子セット解析(Gene Set Analysis; GSA)という枠組みではGO解析もパスウェイ解析も同じなので、そちらもチェックするといいかもしれません。

R用:

- GAGE : Luo et al., BMC Bioinformatics, 2009
- goseq : Young et al., Genome Biol., 2010
- GOSemSim : Yu et al., Bioinformatics, 2010
- Rスクリプト : Gao et al., Bioinformatics, 2011
- clusterProfiler : Yu et al., OMICS., 2012
- RamiGO : Schröder et al., Bioinformatics, 2013
- GSVA : Hänzelmann et al., BMC Bioinformatics, 2013
- SeqGSEA(各群5反復以上を要求) : Wang et al., Bioinformatics, 2014
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- GOplot(Visualization用) : Walter et al., Bioinformatics, 2015
- RNA-Enrich(論文のsuppl) : Lee et al., Bioinformatics, 2015
- GOexpress : Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- rgsepd : Stamm et al., BMC Bioinformatics, 2019

R以外:

- Enrichr(web tool; gene listが入力) : Chen et al., BMC Bioinformatics, 2013
- RNA-Enrich(web tool) : Lee et al., Bioinformatics, 2015
- NET-GE(ヒト専用) : Bovo et al., Bioinformatics, 2016

Review、ガイドライン、パイプライン系:

- 手法比較 : Rahmatallah et al., BMC Bioinformatics, 2014
- ガイドライン : Rahmatallah et al., Brief Bioinform., 2015

最もお手軽なのは…

解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について

RNA-seqなどのタグカウントデータから遺伝子オントロジー(GO)解析を行うためのパッケージもいくつか出ています。 遺伝子セット解析(Gene Set Analysis; GSA)という枠組みではGO解析もパスウェイ解析も同じなので、そちらもチェックするといいかもしれません。

R用:

- GAGE : Luo et al., BMC Bioinformatics, 2009
- goseq : Young et al., Genome Biol., 2010
- GOSemSim : Yu et al., Bioinformatics, 2010
- Rスクリプト : Gao et al., Bioinformatics, 2011
- clusterProfiler : Yu et al., OMICS., 2012
- RamiGO : Schröder et al., Bioinformatics, 2013
- GSVA : Hänzelmann et al., BMC Bioinformatics, 2013
- SeqGSEA(各群5反復以上を要求) : Wang et al., Bioinformatics, 2014
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- GOplot(Visualization用) : Walter et al., Bioinformatics, 2015
- RNA-Enrich(論文のsuppl) : Lee et al., Bioinformatics, 2015
- GOexpress : Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- rgsepd : Stamm et al., BMC Bioinformatics, 2019

R以外:

- Enrichr(web tool; gene listが入力) : Chen et al., BMC Bioinformatics, 2013
- RNA-Enrich(web tool) : Lee et al., Bioinformatics, 2015
- NET-GE(ヒト専用) : Bovo et al., Bioinformatics, 2016

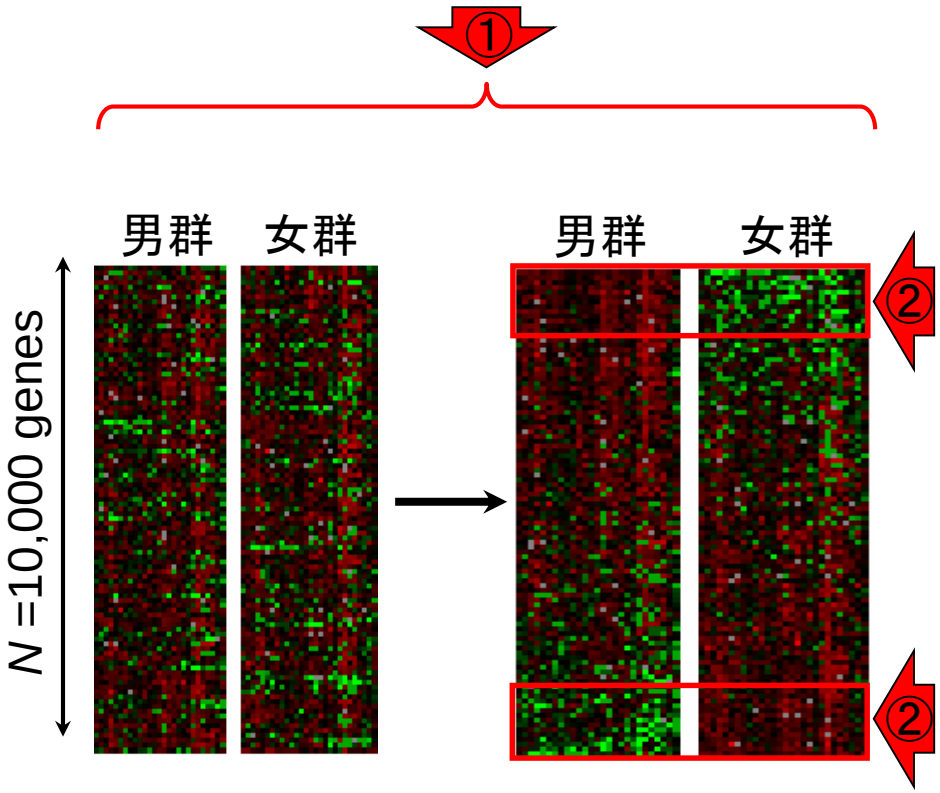
Review、ガイドライン、パイプライン系:

- 手法比較 : Rahmatallah et al., BMC Bioinformatics, 2014
- ガイドライン : Rahmatallah et al., Brief Bioinform., 2015

最もお手軽に遺伝子セット解析（エンリッチメント解析とも呼ばれる）が行えて、且つ利用実績も多いのは、おそらく①Enrichr。②ウェブツールなのでお手軽。③入力は（第3回で解説した）発現変動している遺伝子のリスト。

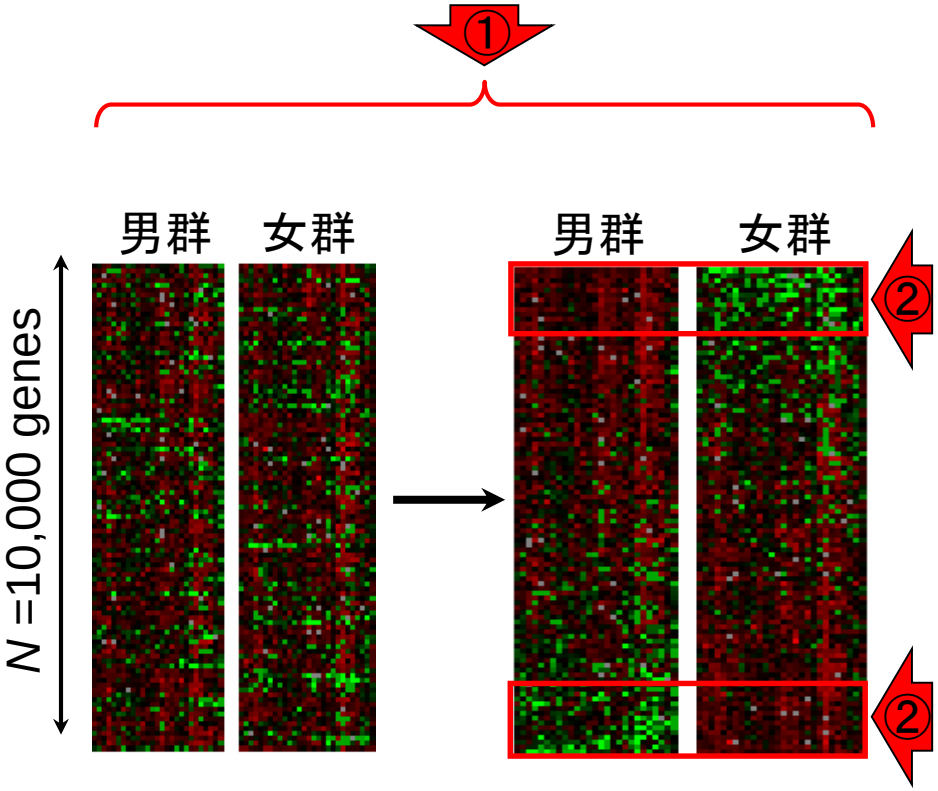
Enrichr利用の概念図1

①男 vs. 女で2群間比較を行い、②赤枠内の発現変動遺伝子群をリストアップしたところ。これがEnrichrの入力。

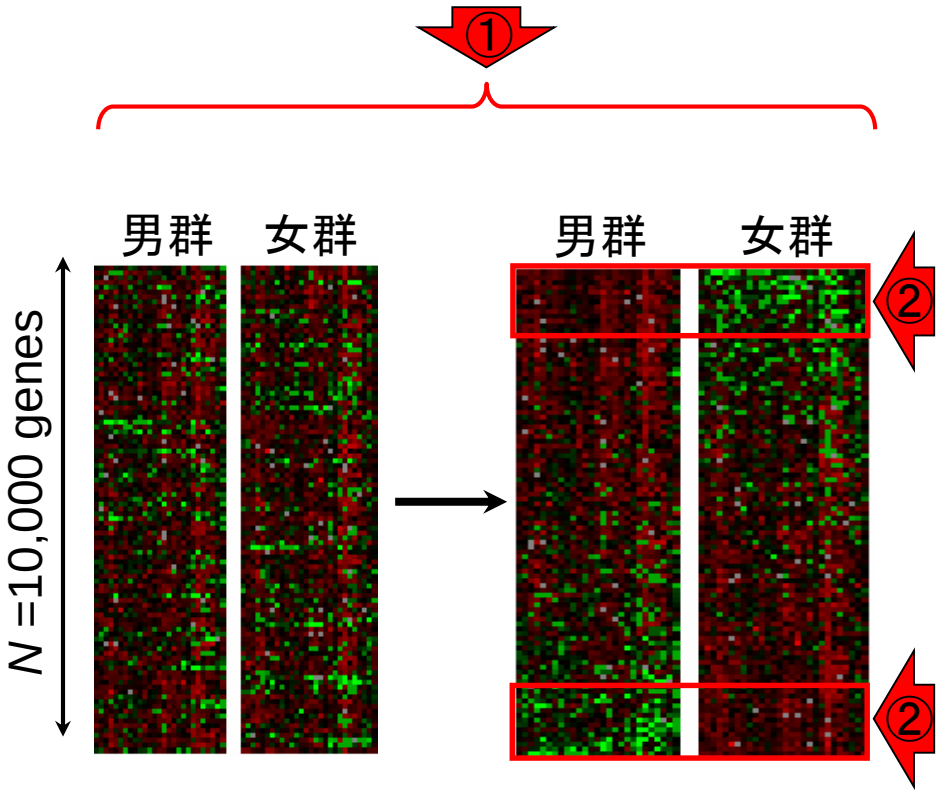


Enrichr利用の概念図2

①男 vs. 女で2群間比較を行い、②赤枠内の発現変動遺伝子群をリストアップしたところ。これがEnrichrの入力。遺伝子セット解析はGO解析やパスウェイ解析でなくてもよい。例えば、ヒトの1番染色体上にある遺伝子セットが変動しているかなども調べることができる。



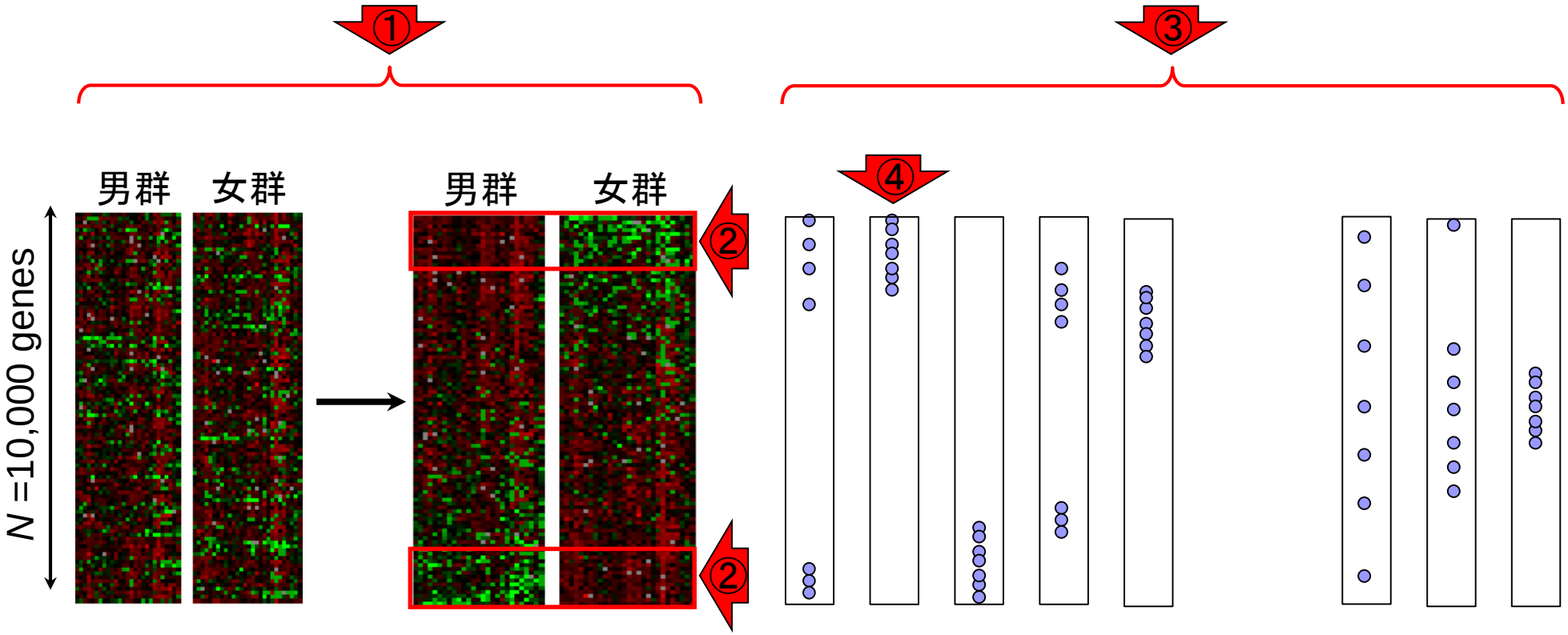
Enrichr利用の概念図2



①男 vs. 女で2群間比較を行い、②赤枠内の発現変動遺伝子群をリストアップしたところ。これがEnrichrの入力。遺伝子セット解析はGO解析やパスウェイ解析でなくてもよい。例えば、ヒトの1番染色体上にある遺伝子セットが変動しているかなども調べることができる。それゆえ、ヒトの様々な染色体の遺伝子セットも解析対象に含めていれば、男だけに存在するy染色体の遺伝子セットが有意だという判定結果が得られるというイメージです。

Enrichr利用の概念図3

Enrichr内部で③様々な遺伝子セットを評価している概念図。ここでは8つの遺伝子セット調べている。y染色体上にある遺伝子セットは男群のみで高発現パターンとなるので、④のような感じになる。これだけ偏りがあると有意だと判定される。



• 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について 参考

GO解析用

①SeqGSEAは、手順が煩雑な上、ものすごく計算時間がかかる。2015.03.05の講習会資料作成当時の個人の感想です。

解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について

RNA-seqなどのタグカウントデータから遺伝子オントロジー(GO)解析を行うためのパッケージもいくつか出ています。 遺伝子セット解析(Gene Set Analysis; GSA)という枠組みではGO解析もパスウェイ解析も同じなので、そちらもチェックするといいかもしれません。

R用:

- GAGE : Luo et al., BMC Bioinformatics, 2009
- goseq : Young et al., Genome Biol., 2010
- GOSemSim : Yu et al., Bioinformatics, 2010
- Rスクリプト : Gao et al., Bioinformatics, 2011
- clusterProfiler : Yu et al., OMICS., 2012
- RamiGO : Schröder et al., Bioinformatics, 2013
- GSVA : Hänzelmann et al., BMC Bioinformatics, 2013
- SeqGSEA(各群5反復以上を要求) : Wang et al., Bioinformatics, 2014 ①
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- GOplot(Visualization用) : Walter et al., Bioinformatics, 2015
- RNA-Enrich(論文のsuppl) : Lee et al., Bioinformatics, 2015
- GOexpress : Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- rgsepd : Stamm et al., BMC Bioinformatics, 2019

R以外:

- Enrichr(web tool; gene listが入力) : Chen et al., BMC Bioinformatics, 2013
- RNA-Enrich(web tool) : Lee et al., Bioinformatics, 2015
- NET-GE(ヒト専用) : Bovo et al., Bioinformatics, 2016

Review、ガイドライン、パイプライン系:

- 手法比較 : Rahmatallah et al., BMC Bioinformatics, 2014
- ガイドライン : Rahmatallah et al., Brief Bioinform., 2015

• 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について 参考

GO解析用

解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について

RNA-seqなどのタグカウントデータから遺伝子オントロジー(GO)解析を行うためのパッケージもいくつか出ています。 遺伝子セット解析(Gene Set Analysis; GSA)という枠組みではGO解析もパスウェイ解析も同じなので、そちらもチェックするといいかもしれません。

R用:

- GAGE : Luo et al., BMC Bioinformatics, 2009
- goseq : Young et al., Genome Biol., 2010
- GOSemSim : Yu et al., Bioinformatics, 2010
- Rスクリプト : Gao et al., Bioinformatics, 2011
- clusterProfiler : Yu et al., OMICS., 2012
- RamiGO : Schröder et al., Bioinformatics, 2013
- GSVA : Hänzelmann et al., BMC Bioinformatics, 2013
- SeqGSEA(各群5反復以上を要求) : Wang et al., Bioinformatics, 2014
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- GOplot(Visualization用) : Walter et al., Bioinformatics, 2015
- RNA-Enrich(論文のsuppl) : Lee et al., Bioinformatics, 2015
- GOexpress : Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- rgsepd : Stamm et al., BMC Bioinformatics, 2019

R以外:

- Enrichr(web tool; gene listが入力) : Chen et al., BMC Bioinformatics, 2013
- RNA-Enrich(web tool) : Lee et al., Bioinformatics, 2015
- NET-GE(ヒト専用) : Bovo et al., Bioinformatics, 2016

Review、ガイドライン、パイプライン系:

- 手法比較 : Rahmatallah et al., BMC Bioinformatics, 2014
- ガイドライン : Rahmatallah et al., Brief Bioinform., 2015

①SeqGSAはマニュアルがないに等しい。②GSARは、発現変動遺伝子セットを探すというよりは、興味ある遺伝子セットを与えてネットワーク図を描き、どの遺伝子がhub（hub genes）かを返すのがメイン

EGSEA

①EGSEAは、②複数のツールを組み合わせるやり方。様々な分野でこの種の戦略がよいことは実証されており、おそらく妥当。しかしその分だけ依存関係が複雑になるため、私もまだ試してはいない

解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について

RNA-seqなどのタグカウントデータから遺伝子オントロジー(GO)解析を行うためのパッケージもいくつか出ています。 遺伝子セット解析(Gene Set Analysis; GSA)という枠組みではGO解析もパスウェイ解析も同じなので、そちらもチェックするといいかもしれません。

R用:

- GAGE : Luo et al., BMC Bioinformatics, 2009
- goseq : Young et al., Genome Biol., 2010
- GOSemSim : Yu et al., Bioinformatics, 2010
- Rスクリプト : Gao et al., Bioinformatics, 2011
- clusterProfiler : Yu et al., OMICS., 2012
- RamiGO : Schröder et al., Bioinformatics, 2013
- GSVA : Hänzelmann et al., BMC Bioinformatics, 2013
- SeqGSEA(各群5反復以上を要求) : Wang et al., Bioinformatics, 2014
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- GOplot(Visualization用) : Walter et al., Bioinformatics, 2015
- RNA-Enrich(論文のsuppl) : Lee et al., Bioinformatics, 2015
- GOexpress : Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- rgsepd : Stamm et al., BMC Bioinformatics, 2019

R以外:

- Enrichr(web tool; gene listが入力) : Chen et al., BMC Bioinformatics, 2013
- RNA-Enrich(web tool) : Lee et al., Bioinformatics, 2015
- NET-GE(ヒト専用) : Bovo et al., Bioinformatics, 2016

Review、ガイドライン、パイプライン系:

- 手法比較 : Rahmatallah et al., BMC Bioinformatics, 2014
- ガイドライン : Rahmatallah et al., Brief Bioinform., 2015

Bioinformatics. 2017 Feb 1;33(3):414-424. doi: 10.1093/bioinformatics/btw623.

② Combining multiple tools outperforms individual methods in gene set enrichment analyses.

Alhamdoosh M[1], Ng M[1], Wilson NJ[1], Sheridan JM[2,3], Huynh H[1], Wilson MJ[1], Ritchie ME[4,5].

Author information

Abstract

MOTIVATION: Gene set enrichment (GSE) analysis allows researchers to efficiently extract biological insight from long lists of differentially expressed genes by interrogating them at a systems level. In recent years, there has been a proliferation of GSE analysis methods and hence it has become increasingly difficult for researchers to select an optimal GSE tool based on their particular dataset. Moreover, the majority of GSE analysis methods do not allow researchers to simultaneously compare gene set level results between multiple experimental conditions.

RESULTS: The ensemble of genes set enrichment analyses (EGSEA) is a method developed for RNA-sequencing data that combines results from twelve algorithms and calculates collective gene set scores to improve the biological relevance of the highest ranked gene sets. EGSEA's gene set database contains around 25 000 gene sets from sixteen collections. It has multiple visualization capabilities that allow researchers to view gene sets at various levels of granularity. EGSEA has been tested on simulated data and on a number of human and mouse datasets and, based on biologists' feedback, consistently outperforms the individual tools that have been combined. Our evaluation demonstrates the superiority of the ensemble approach for GSE analysis, and its utility to effectively and efficiently extrapolate biological functions and potential involvement in disease processes from lists of differentially regulated genes.

AVAILABILITY AND IMPLEMENTATION: EGSEA is available as an R package at http://www.bioconductor.org/packages/EGSEA/ . The gene sets collections are available in the R package EGSEAdata from http://www.bioconductor.org/packages/EGSEAdata/ .

CONTACTS: monther.alhamdoosh@csl.com.au mritchie@wehi.edu.au.

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics online.

PMID: 27694195 PMCID: PMC5408797 DOI: 10.1093/bioinformatics/btw623

GSVA

①GSVAは、②EGSEAでも利用されている。また③引用回数も多い(≒使いやすい)ので、後半はこれをベースに説明。

解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について

RNA-seqなどのタグカウントデータから遺伝子オントロジー(GO)解析を行うためのパッケージもいくつか出ています。 遺伝子セット解析(Gene Set Analysis; GSA)という枠組みではGO解析もパスウェイ解析も同じなので、そちらもチェックするといいかもしれません。

R用:

- GAGE : Luo et al., BMC Bioinformatics, 2009
- goseq : Young et al., Genome Biol., 2010
- GOSemSim : Yu et al., Bioinformatics, 2010
- Rスクリプト : Gao et al., Bioinformatics, 2011
- clusterProfiler : Yu et al., OMICS., 2012
- RamiGO : Schröder et al., Bioinformatics, 2013
- GSVA : Hänzelmann et al., BMC Bioinformatics, 2013 ①
- SeqGSEA(各群5反復以上を要求) : Wang et al., Bioinformatics, 2014
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- GOplot(Visualization用) : Walter et al., Bioinformatics, 2015
- RNA-Enrich(論文のsuppl) : Lee et al., Bioinformatics, 2015
- GOexpress : Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 201
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017 ②
- AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- rgsepd : Stamm et al., BMC Bioinformatics, 2019

R以外:

- Enrichr(web tool; gene listが入力) : Chen et al., BMC Bioinformatics, 2013
- RNA-Enrich(web tool) : Lee et al., Bioinformatics, 2015
- NET-GE(ヒト専用) : Bovo et al., Bioinformatics, 2016

Review、ガイドライン、パイプライン系:

- 手法比較 : Rahmatallah et al., BMC Bioinformatics, 2014
- ガイドライン : Rahmatallah et al., Brief Bioinform., 2015

BMC Bioinformatics. 2013 Jan 16;14:7. doi: 10.1186/1471-2105-14-7.

GSVA: gene set variation analysis for microarray and RNA-seq data.

Hänzelmann S[1], Castelo R, Guinney J.

Author information

Abstract

BACKGROUND: Gene set enrichment (GSE) analysis is a popular framework for condensing information from gene expression profiles into a pathway or signature summary. The strengths of this approach over single gene analysis include noise and dimension reduction, as well as greater biological interpretability. As molecular profiling experiments move beyond simple case-control studies, robust and flexible GSE methodologies are needed that can model pathway activity within highly heterogeneous data sets.

RESULTS: To address this challenge, we introduce Gene Set Variation Analysis (GSVA), a GSE method that estimates variation of pathway activity over a sample population in an unsupervised manner. We demonstrate the robustness of GSVA in a comparison with current state of the art sample-wise enrichment methods. Further, we provide examples of its utility in differential pathway activity and survival analysis. Lastly, we show how GSVA works analogously with data from both microarray and RNA-seq experiments.

CONCLUSIONS: GSVA provides increased power to detect subtle pathway activity changes over a sample population in comparison to corresponding methods. While GSE methods are generally regarded as end points of a bioinformatic analysis, GSVA constitutes a starting point to build pathway-centric models of biology. Moreover, GSVA contributes to the current need of GSE methods for RNA-seq data. GSVA is an open source software package for R which forms part of the Bioconductor project and can be downloaded at http://www.bioconductor.org.

PMID: 23323831 PMCID: PMC3618321 DOI: 10.1186/1471-2105-14-7

[Indexed for MEDLINE] Free PMC Article

Images from this publication. See all images (7) Free text

Read free full text at BMC

PMC FREE Full text

Save items

Add to Favorites

Similar articles

A flexible count data model to fit the wid [BMC Bioinformatics. 2013]

Importing ArrayExpress datasets into R/Biocon [Bioinformatics. 2009]

Combining multiple tools outperforms i [Bioinformatics. 2017]

Review Comparing bioinformatic ge [Med Sci Monit Basic Res. 2014]

Review Open source software for the analysis ([Biotechniques. 2003]

See reviews...

See all...

③

Cited by over 100 PubMed Central articles

A modular transcriptional signature identifies pher [Nat Commun. 2018]

MEGF10, a Glioma Survival-Associated Mol [Dis Markers. 2018]

Association between angiogenesis and cyto [Onco Targets Ther. 2018]

パスウェイ解析用

①遺伝子セットとしてKEGG Pathway やReactomeなどの情報を用いる場合

(Rで)塩基配列解析

(last modified 2019/07/17, since 2010)

このウェブページのR関連部分
必要なパッケージをインスト
Macintosh2019.03.12版)で
しました。(2018/07/18)

What's new? (過去のお知

- 「解析 | 発現変動 | 3群間
TCC(Sun 2013)」で内部
たからです(山本裕一郎 氏

- 解析 | 機能解析 | について (last modified 2018/06/24)
- 解析 | 機能解析 | GMTファイル取得 | について (last modified 2018/07/17)
- 解析 | 機能解析 | GMTファイル取得 | EGSEAdata(Alhamdoosh 2017) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | GeneSetDB(Araki 2012) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian 2005) (last modified 2018/06/25)
- 解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan 2018) (last modified 2018/06/25)
- 解析 | 機能解析 | 遺伝子セット解析 | GSVA(Hänzelmann 2013)(last modified 2018/08/10)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について (last modified 2019/05/12)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | SeqGSEA(Wang 2014)(last modified 2018/06/25)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | GSVA(Hänzelmann 2013)(last modified 2018/06/26)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | GOseq(Young 2010) (last modified 2010/11/26)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | について ① t modified 2019/05/31)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | GSAR (Rahmatallah 2017)(last modified 2017/03/17)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | SeqGSEA(Wang 2014) (last modified 2015/02/27)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | GSVA(Hänzelmann 2013)(last modified 2018/06/26)
- 解析 | 分類 | について(last modified 2018/06/27)

パスウェイ解析用

GO解析もできるのに、ここにしか記載していないものもいくつかあるはずですのでご注意ください。

解析 | 機能解析 | パスウェイ(Pathway)解析 | について

RNA-seqなどのタグカウントデータからパスウェイ(Pathway)解析を行うためのパッケージもいくつか出ています。 入力のカウントデータファイルのgene IDは、Ensembl (Zerbino et al., Nucleic Acids Res., 2018) が多いようです。 遺伝子セット解析(Gene Set Analysis; GSA)という枠組みではGO解析もパスウェイ解析も同じなので、そちらもチェックするといいかもしれません。

R用:

- KEGGgraph : Zhang et al., Bioinformatics, 2009
- GAGE : Luo et al., BMC Bioinformatics, 2009
- clusterProfiler : Yu et al., OMICS., 2012
- GSVA : Hänzelmann et al., BMC Bioinformatics, 2013
- Pathview : Luo et al., Bioinformatics, 2013
- GSNCA法(GSARで提供されている) : Rahmatallah et al., Bioinformatics, 2014
- SeqGSEA : Wang et al., Bioinformatics, 2014
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- seq2pathway : Wang et al., Bioinformatics, 2015
- ToPASeq : Ihnatova and Budinska, BMC Bioinformatics, 2015
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- pathDESeq : Dona et al., Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- NEArender : Jeggari and Alexeyenko, BMC Bioinformatics, 2017
- tcgsaseq(時系列の遺伝子セット解析) : Agniel et al., Biostatistics, 2017

R以外:

- CCS : Schissler et al., Bioinformatics, 2016
- NET-GE(webtool; ヒト専用) : Bovo et al., Bioinformatics, 2016
- ContextTRAP(時系列解析用) : Lee et al., BMC Bioinformatics, 2016
- MrGSEA(MATLAB) : Zyla et al., BMC Bioinformatics, 2017
- NFPscanner(webtool) : Xu et al., BMC Bioinformatics, 2017
- EviNet : Jeggari et al., Nucleic Acids Res., 2018

Review、ガイドライン、パイプライン系:

- 手法比較 : Rahmatallah et al., BMC Bioinformatics, 2014

Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

遺伝子セット情報取得

①遺伝子セット情報はGMTという形式(拡張子が.gmt)で提供されており、②3つの手段で取得可能です。

(Rで)塩基配列解析

(last modified 2019/07/17, since 2010)

このウェブページの
必要なパッケージを
Macintosh2019.0
しました。(2018/

What's new? (

- 「解析 | 発現変
  TCC(Sun 2013
  たからです(山本

- 解析 | 機能解析 | について (last modified 2018/06/24)
- 解析 | 機能解析 | GMTファイル取得 | について (last modified 2018/07/17)
- 解析 | 機能解析 | GMTファイル取得 | EGSEAdata(Alhamdoosh 2017) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | GeneSetDB(Araki 2012) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian 2005) (last modified 2018/06/25)
- 解析 | 機能解析 | GMTファイル
- 解析 | 機能解析 | 遺伝子セット解
- 解析 | 機能解析 | 遺伝子オントロ
- 解析 | 機能解析 | 遺伝子オントロ
- 解析 | 機能解析 | 遺伝子オントロ
- 解析 | 機能解析 | 遺伝子オントロ
- 解析 | 機能解析 | パスウェイ(Pa
- 解析 | 機能解析 | パスウェイ(Pa
- 解析 | 機能解析 | パスウェイ(Pa
- 解析 | 機能解析 | パスウェイ(Pa
- 解析 | 分類 | について(last mod

①

解析 | 機能解析 | GMTファイル取得 | について

Gene Set Enrichment Analysis (GSEA)に代表される遺伝子セット解析を行うためには、(発現情報と)**遺伝子セット情報**が必要です。遺伝子セット情報の取得場所として最も有名なのは、Molecular Signatures Database (MSigDB)であり、Gene Matrix Transposed (GMT)形式で提供されています。 遺伝子セット情報は、MSigDBを含む以下からも提供されています：

- MSigDB(ウェブサイト)
- GeneSetDB(ウェブサイト)
- EGSEAdata(Rパッケージ)
- gskb(Rパッケージ)
- Enrichr(web server)

②

- Gene Set Enrichment Analysis (GSEA) : Mootha et al., Nat Genet., 2003
- Gene Set Enrichment Analysis (GSEA) : Subramanian et al., PNAS, 2005
- Molecular Signatures Database (MSigDB) : Subramanian et al., PNAS, 2005
- Molecular Signatures Database (MSigDB) : Liberzon et al., Bioinformatics, 2011
- EGSEAdata : Alhamdoosh et al., F1000Res., 2017
- GeneSetDB : Araki et al., FEBS Open Bio., 2012
- Enrichr(web server) : Kuleshov et al., Nucleic Acids Res., 2016

MSigDB ver. 6.1

①最も有名なのはMSigDB。②2018年6月のver. 6.1では、8個の主要なコレクションが提供されている。2019年7月現在はver. 6.2ですが、ほとんど変化がないのでver. 6.1でやります。

(Rで)塩基配列解析

(last modified 2019/07/17, since 2010)

このウェブページの
必要なパッケージを
Macintosh2019.0
しました。(2018/

What's new? (

- 「解析 | 発現変
  TCC(Sun_2013
  たからです(山本

- 解析 | 機能解析 | について (last modified 2018/06/24)
- 解析 | 機能解析 | GMTファイル取得 | について (last modified 2018/07/17)
- 解析 | 機能解析 | GMTファイル取得 | EGSEAdata(Alhamdoosh_2017) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | GeneSetDB(Araki_2012) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian_2005) (last modified 2018/06/25)
- 解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan_2018) (last modified 2018/06/25)
- 解析 | 機能解析 | 遺伝子セット解析 | GSVA(Hänzelmann_2013)(last modified 2018/08/10)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について (last modified 2019/05/12)
- 解析 | 機
- 解析 | 機
- 解析 | 機
- 解析 | 機
- 解析 | 機
- 解析 | 機
- 解析 | 機
- 解析 | 分

①

解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian_2005) NEW

MSigDBから遺伝子セット情報を含むGMTファイル(.gmt)を取得するやり方を示します。2018年6月現在は、MSigDB version 6.1です。Release Notesにもありますが、MSigDB v6.0以降、提供データがCreative Commons Attribution 4.0 International Licenseになったようです（KEGG、BioCarta、AAAS/STKE Cell Signaling Database dataの3種類を除く）。MSigDB v6.1では、以下に示す8個の主要なコレクション(8 major collections)が提供されています。ときどきMSigDBからgmtファイルをダウンロードできない事態に遭遇しますので、このサイト上でも**提供可能なものはダウンロードできる状態**にしています。但し、**このサイト上からダウロードした場合は、register し、MSigDBに対して仁義を果たして下さい。**MSigDBがfunding agenciesに利用者情報を報告するために必要です。また、MSigDBの論文だけでなく、例えば 1番目のコレクションである**H hallmark gene sets**を利用する場合はLiberzon et al., 2015を、そして 2番目のコレクションであるC2(curated gene sets)に含まれる**CP:Reactomeの gene sets**を利用する場合はJoshi-Tope et al., 2005などを適切に引用しましょう。

②

1. H: hallmark gene sets(50 gene sets)
   - gene symbols(h.all.v6.1.symbols.gmt)
   - entrez genes ids(h.all.v6.1.entrez.gmt)
2. C1: positional gene sets(326 gene sets)

MSigDB ver. 6.1

発現変動と関連するKEGGパスウェイを調べたいときは、①3番目のc2というカテゴリーに属する、②CP:KEGGというところの186個の遺伝子セットが含まれる.gmtファイルを予めダウンロードしておく。

1. H: hallmark gene sets (50 gene sets)
2. c1: positional gene sets (326 gene sets)
   - ヒト染色体の位置ごとの遺伝子セットリストファイル (326 gene sets)
3. ① c2: curated gene sets (4,738 gene sets)
   - **CGP**: chemical and genetic perturbations (3,409 gene sets)
   - **CP**: canonical pathways (1,329 gene sets)
   - **CP:BIOCARTA**: BioCarta gene sets (217 gene sets)
   - **CP:KEGG**: KEGG gene sets (186 gene sets) ②
   - **CP:REACTOME**: Reactome gene sets (674 gene sets)
4. c3: motif gene sets (836 gene sets)
   - **MIR**: microRNA targets (221 gene sets)
   - **TFT**: transcription factor targets (615 gene sets)
5. c4: computational gene sets (858 gene sets)
   - **CGM**: cancer gene neighborhoods (427 gene sets)
   - **CM**: cancer modules (431 gene sets)
6. c5: gene ontology (GO) gene sets (5,917 gene sets)
   - **BP**: biological process (4,436 gene sets)
   - **CC**: cellular component (580 gene sets)
   - **MF**: molecular function (901 gene sets)
7. c6: oncogenic signatures gene sets (189 gene sets)
8. c7: immunologic signatures gene sets (4,872 gene sets)

Subramanian et al., *PNAS*, **102**: 15545–15550, 2005

MSigDB ver. 6.1

発現変動と関連するGOのbiological processを調べたいときは、①6番目のc5というカテゴリーに属する、②BPというところの4,436個の遺伝子セットが含まれる.gmtファイルを予めダウンロードしておく。

1. H: hallmark gene sets (50 gene sets)
2. c1: positional gene sets (326 gene sets)
   - ヒト染色体の位置ごとの遺伝子セットリストファイル (326 gene sets)
3. c2: curated gene sets (4,738 gene sets)
   - **CGP**: chemical and genetic perturbations (3,409 gene sets)
   - **CP**: canonical pathways (1,329 gene sets)
   - **CP:BIOCARTA**: BioCarta gene sets (217 gene sets)
   - **CP:KEGG**: KEGG gene sets (186 gene sets)
   - **CP:REACTOME**: Reactome gene sets (674 gene sets)
4. c3: motif gene sets (836 gene sets)
   - **MIR**: microRNA targets (221 gene sets)
   - **TFT**: transcription factor targets (615 gene sets)
5. c4: computational gene sets (858 gene sets)
   - **CGM**: cancer gene neighborhoods (427 gene sets)
   - **CM**: cancer modules (431 gene sets)
6. ① c5: gene ontology (GO) gene sets (5,917 gene sets)
   - **BP**: biological process (4,436 gene sets) ②
   - **CC**: cellular component (580 gene sets)
   - **MF**: molecular function (901 gene sets)
7. c6: oncogenic signatures gene sets (189 gene sets)
8. c7: immunologic signatures gene sets (4,872 gene sets)

MSigDB ver. 6.1

講義で利用する解析プログラムGSVAの検証用としては、①326遺伝子セットが最適なので、これをダウンロードしておきます。

1. H: hallmark gene sets (50 gene sets)
2. c1: positional gene sets (326 gene sets)
   - ヒト染色体の位置ごとの遺伝子セットリストファイル (326 gene sets) ①
3. c2: curated gene sets (4,738 gene sets)
   - **CGP**: chemical and genetic perturbations (3,409 gene sets)
   - **CP**: canonical pathways (1,329 gene sets)
   - **CP:BIOCARTA**: BioCarta gene sets (217 gene sets)
   - **CP:KEGG**: KEGG gene sets (186 gene sets)
   - **CP:REACTOME**: Reactome gene sets (674 gene sets)
4. c3: motif gene sets (836 gene sets)
   - **MIR**: microRNA targets (221 gene sets)
   - **TFT**: transcription factor targets (615 gene sets)
5. c4: computational gene sets (858 gene sets)
   - **CGM**: cancer gene neighborhoods (427 gene sets)
   - **CM**: cancer modules (431 gene sets)
6. c5: gene ontology (GO) gene sets (5,917 gene sets)
   - **BP**: biological process (4,436 gene sets)
   - **CC**: cellular component (580 gene sets)
   - **MF**: molecular function (901 gene sets)
7. c6: oncogenic signatures gene sets (189 gene sets)
8. c7: immunologic signatures gene sets (4,872 gene sets)

ダウンロード

①MSigDB本家からもダウンロードできますが、一気にやると迷惑をかけるので、②からc1.all.v6.1.entrez.gmtをダウンロードしておいてください。Entrez gene IDのヒト染色体の位置ごとの遺伝子セットリストファイル (326 gene sets)です。
③ c1.all.v6.1.symbols.gmtもダウンロードしておきましょう。

解析 | 機能解析 | GMTファイル取得

①

MSigDBから遺伝子セット情報を含むGMTファイル(.gmt)
MSigDB version 6.1です。Release Notesにもありますが、MSigDB v6.0以降、提供データがCreative Commons Attribution 4.0 International Licenseになったようです（KEGG、BioCarta、AAAS/STKE Cell Signaling Database dataの3種類を除く）。MSigDB v6.1では、以下に示す8個の主要なコレクション(8 major collections)が提供されています。ときどきMSigDBからgmtファイルをダウンロードできない事態に遭遇しますので、このサイト上でも**提供可能なものはダウンロードできる状態**にしています。但し、**このサイト上からダウロードした場合は、registerし、MSigDBに対して仁義を果たして下さい。**MSigDBがfunding agenciesに利用者情報を報告するために必要です。
また、MSigDBの論文だけでなく、例えば 1番目のコレクションである**H hallmark gene sets**を利用する場合はLiberzon et al., 2015を、そして 2番目のコレクションであるC2(curated gene sets)に含まれる**CP:Reactomeのgene sets**を利用する場合はJoshi-Tope et al., 2005などを適切に引用しましょう。

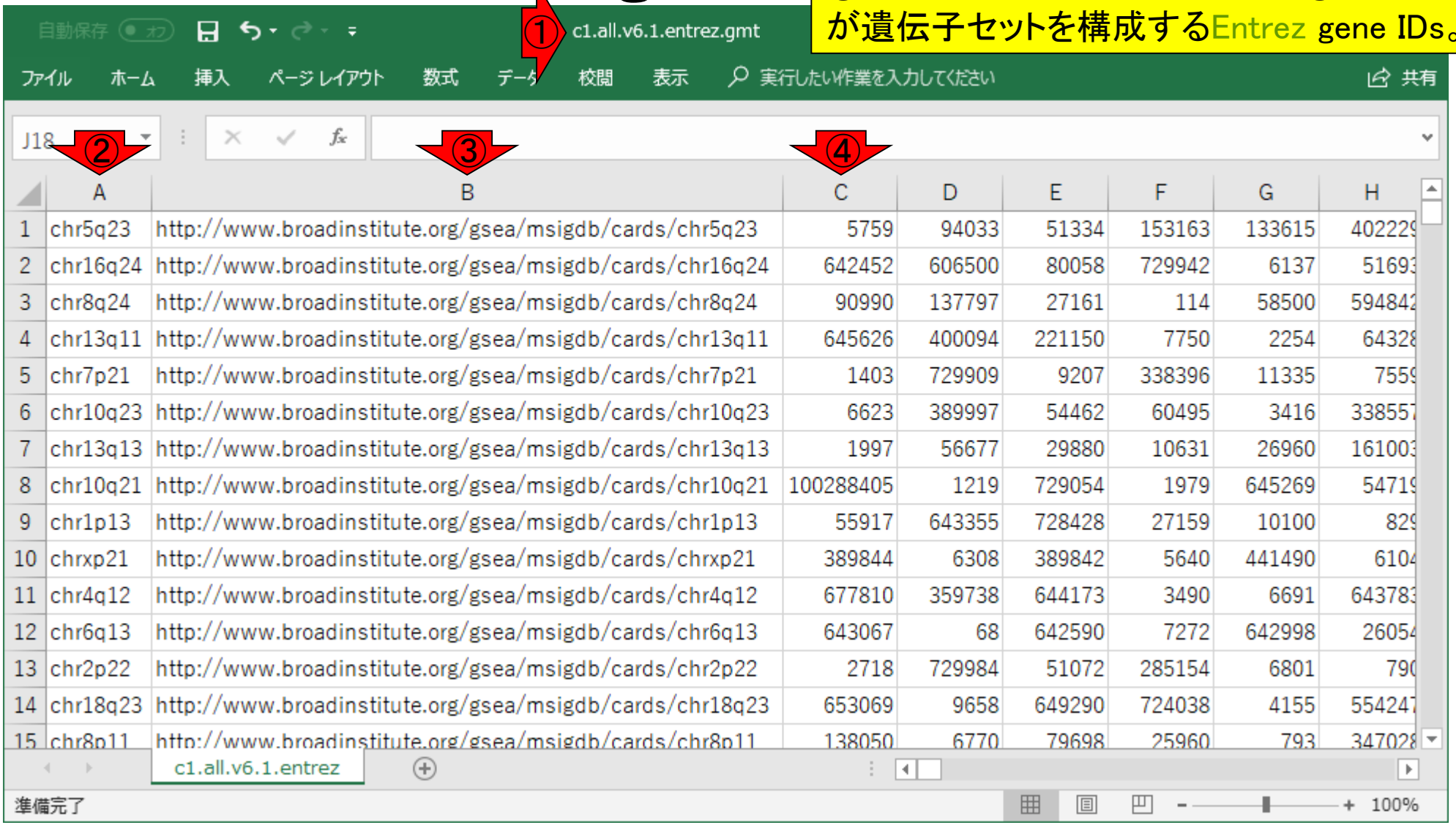
1. H: hallmark gene sets(50 gene sets)
    - gene symbols(h.all.v6.1.symbols.gmt)
    - entrez genes ids(h.all.v6.1.entrez.gmt)
2. C1: positional gene sets(326 gene sets)
    - gene symbols(c1.all.v6.1.symbols.gmt)
    - entrez genes ids(c1.all.v6.1.entrez.gmt)
3. C2: curated gene sets(4738 gene sets)
    - gene symbols(c2.all.v6.1.symbols.gmt)
    - entrez genes ids(c2.all.v6.1.entrez.gmt)
    - CGP: chemical and genetic perturbations(3409 gene sets)
        - gene symbols(c2.cgp.v6.1.symbols.gmt)
        - entrez genes ids(c2.cgp.v6.1.entrez.gmt)
    - CP: Canonical pathways(1329 gene sets)
        - gene symbols(c2.cp.v6.1.symbols.gmt)
        - entrez genes ids(c2.cp.v6.1.entrez.gmt)

③

②

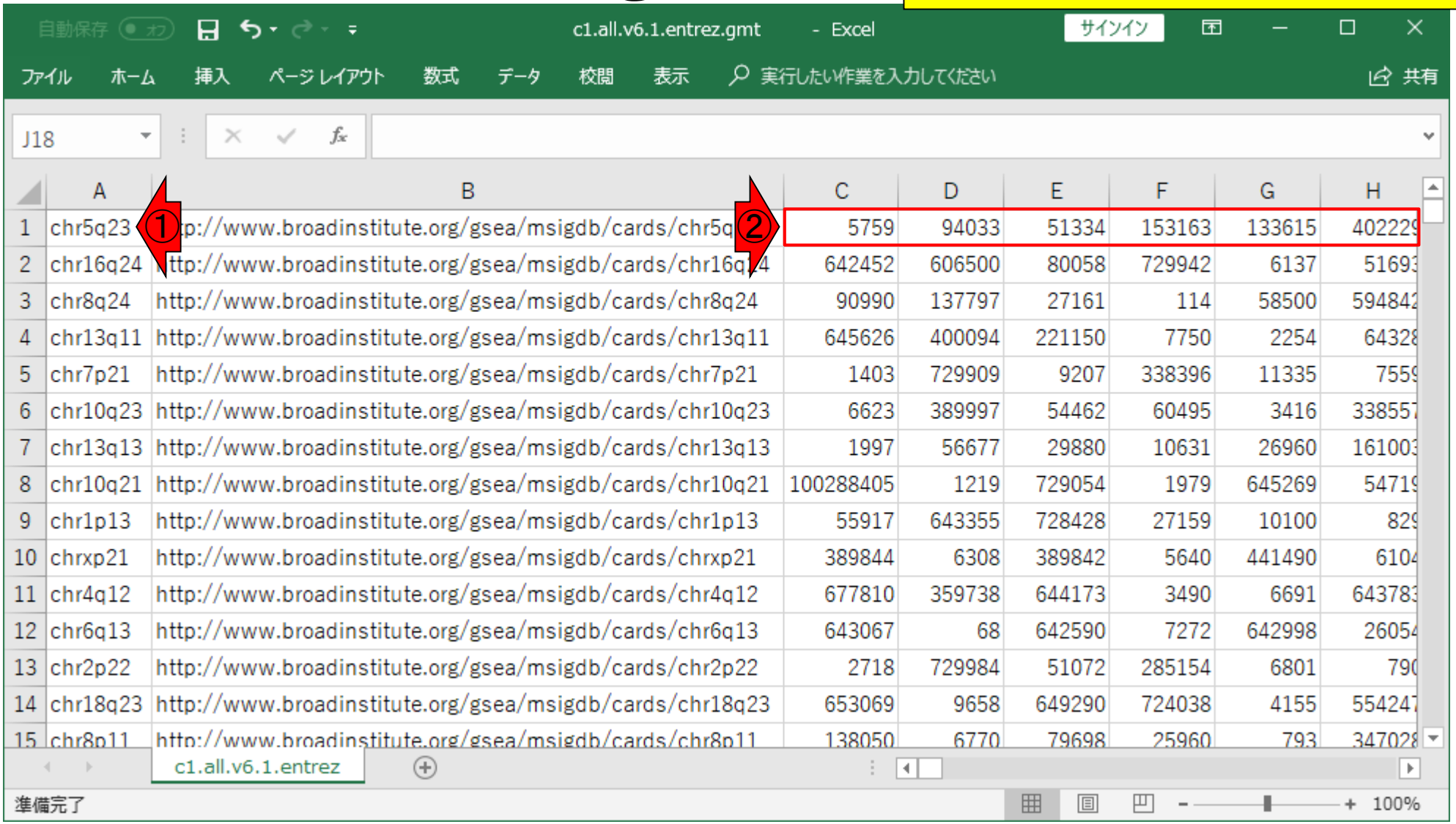
c1.all.v6.1.entrez.gmt

①c1.all.v6.1.entrez.gmtをExcelで眺めるとこんな感じ。GMT形式は、②1列目がgene set名、③2列目が情報源のURL、そして④3列目以降が遺伝子セットを構成するEntrez gene IDs。



c1.all.v6.1.entrez.gmt

①chr5q23が1番目のデータセット。②赤枠部分がchr5q23という遺伝子セットに含まれる遺伝子群のEntrez gene IDs。



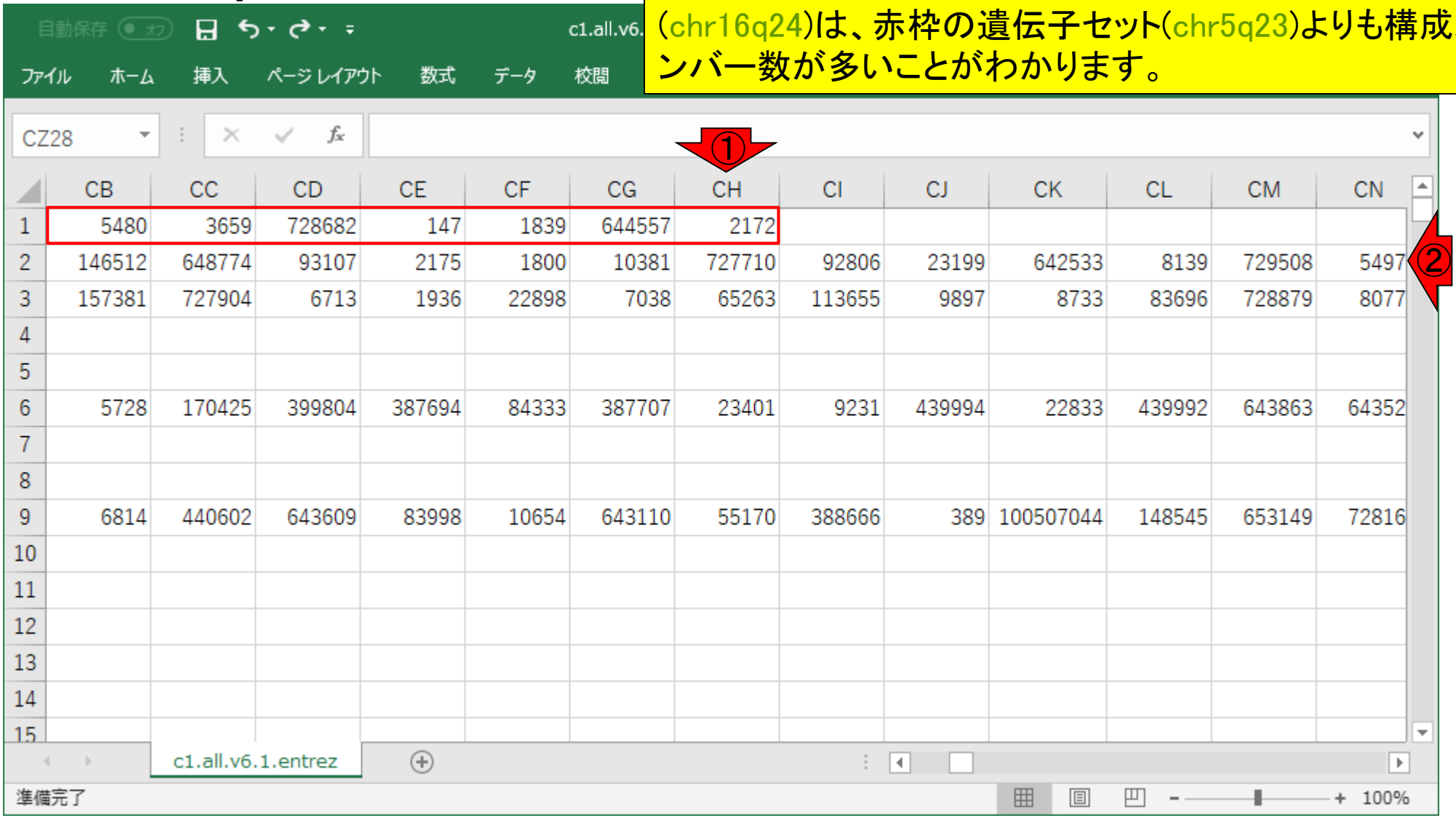
c1.all.v6.1.entrez.gmt - Excel

	A	B	C	D	E	F	G	H
1	chr5q23	ttp://www.broadinstitute.org/gsea/msigdb/cards/chr5q	5759	94033	51334	153163	133615	40222
2	chr16q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr16q24	642452	606500	80058	729942	6137	51693
3	chr8q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr8q24	90990	137797	27161	114	58500	594842
4	chr13q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q11	645626	400094	221150	7750	2254	64328
5	chr7p21	http://www.broadinstitute.org/gsea/msigdb/cards/chr7p21	1403	729909	9207	338396	11335	7559
6	chr10q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q23	6623	389997	54462	60495	3416	338557
7	chr13q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q13	1997	56677	29880	10631	26960	161003
8	chr10q21	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q21	100288405	1219	729054	1979	645269	54719
9	chr1p13	http://www.broadinstitute.org/gsea/msigdb/cards/chr1p13	55917	643355	728428	27159	10100	829
10	chrxp21	http://www.broadinstitute.org/gsea/msigdb/cards/chrxp21	389844	6308	389842	5640	441490	6104
11	chr4q12	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q12	677810	359738	644173	3490	6691	643783
12	chr6q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr6q13	643067	68	642590	7272	642998	26054
13	chr2p22	http://www.broadinstitute.org/gsea/msigdb/cards/chr2p22	2718	729984	51072	285154	6801	790
14	chr18q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr18q23	653069	9658	649290	724038	4155	554247
15	chr8p11	http://www.broadinstitute.org/gsea/msigdb/cards/chr8p11	138050	6770	79698	25960	793	347028

• 解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramar

chr5q23

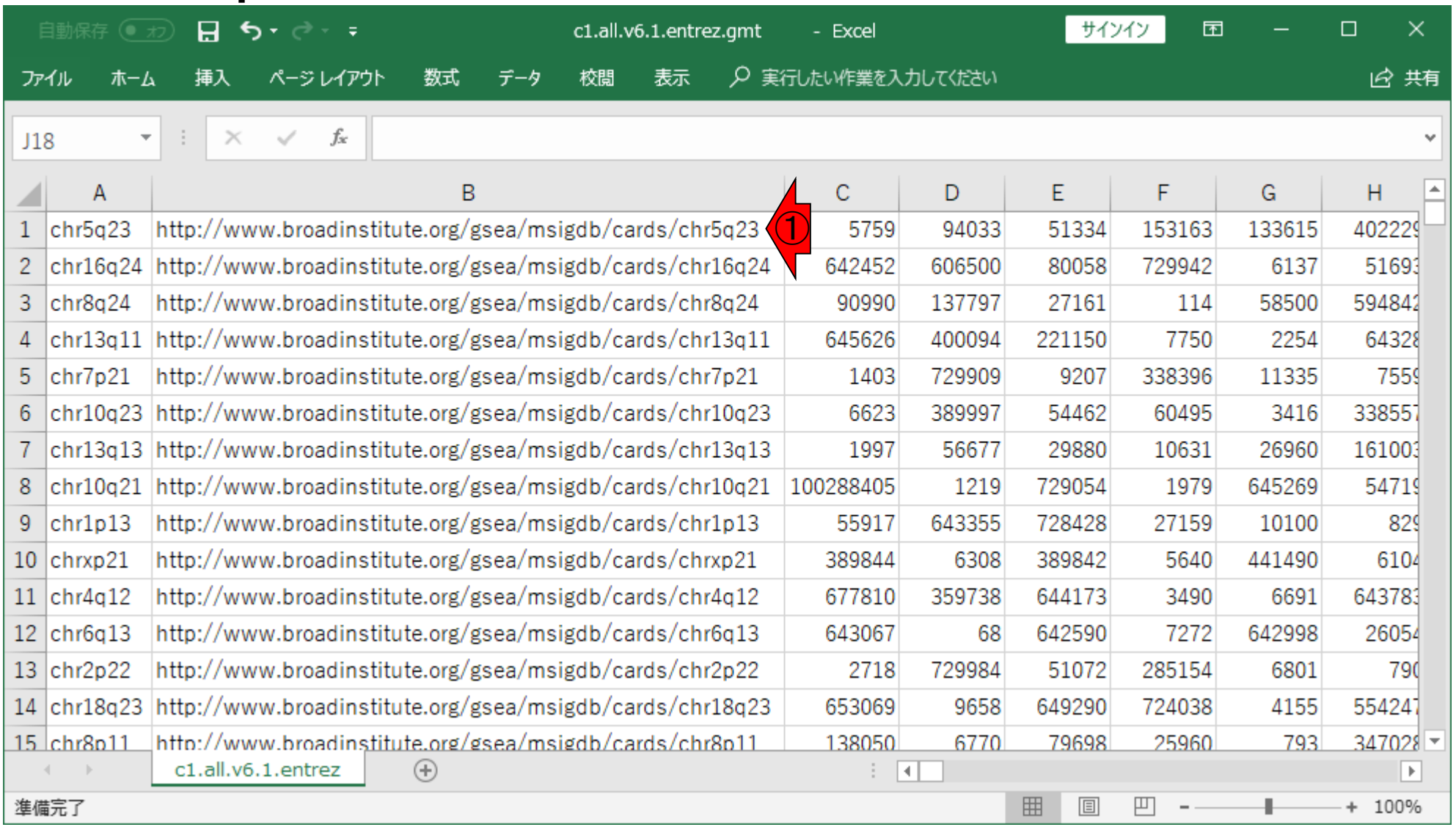
①最後のEntrez gene IDである2172まで、ずーっと右のほうに移動したところ。遺伝子セットによって、構成メンバー数が異なることがわかります。②2番目の遺伝子セット(chr16q24)は、赤枠の遺伝子セット(chr5q23)よりも構成メンバー数が多いことがわかります。



	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN
1	5480	3659	728682	147	1839	644557	2172						
2	146512	648774	93107	2175	1800	10381	727710	92806	23199	642533	8139	729508	5497
3	157381	727904	6713	1936	22898	7038	65263	113655	9897	8733	83696	728879	8077
4													
5													
6	5728	170425	399804	387694	84333	387707	23401	9231	439994	22833	439992	643863	64352
7													
8													
9	6814	440602	643609	83998	10654	643110	55170	388666	389	100507044	148545	653149	72816
10													
11													
12													
13													
14													
15													

①2列目の情報源のURLを眺めると…

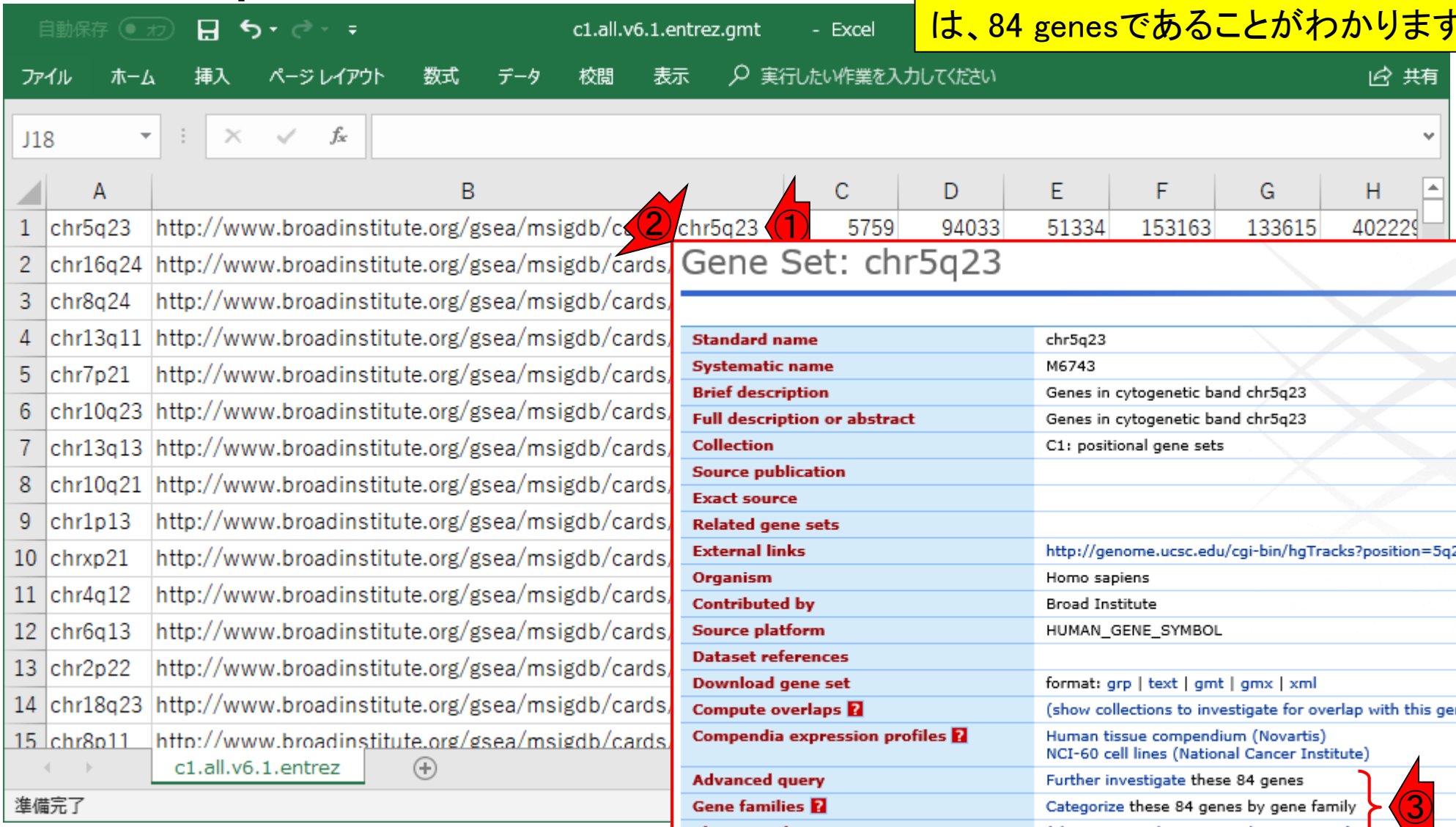
chr5q23



	A	B	C	D	E	F	G	H
1	chr5q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr5q23	5759	94033	51334	153163	133615	40222
2	chr16q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr16q24	642452	606500	80058	729942	6137	5169
3	chr8q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr8q24	90990	137797	27161	114	58500	59484
4	chr13q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q11	645626	400094	221150	7750	2254	6432
5	chr7p21	http://www.broadinstitute.org/gsea/msigdb/cards/chr7p21	1403	729909	9207	338396	11335	755
6	chr10q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q23	6623	389997	54462	60495	3416	33855
7	chr13q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q13	1997	56677	29880	10631	26960	16100
8	chr10q21	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q21	100288405	1219	729054	1979	645269	5471
9	chr1p13	http://www.broadinstitute.org/gsea/msigdb/cards/chr1p13	55917	643355	728428	27159	10100	82
10	chrxp21	http://www.broadinstitute.org/gsea/msigdb/cards/chrxp21	389844	6308	389842	5640	441490	610
11	chr4q12	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q12	677810	359738	644173	3490	6691	64378
12	chr6q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr6q13	643067	68	642590	7272	642998	2605
13	chr2p22	http://www.broadinstitute.org/gsea/msigdb/cards/chr2p22	2718	729984	51072	285154	6801	790
14	chr18q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr18q23	653069	9658	649290	724038	4155	55424
15	chr8p11	http://www.broadinstitute.org/gsea/msigdb/cards/chr8p11	138050	6770	79698	25960	793	34702

chr5q23

①2列目の情報源のURLを眺めると…
②こんな感じで詳細情報が得られます。③遺伝子セットchr5q23のメンバー数は、84 genesであることがわかります。



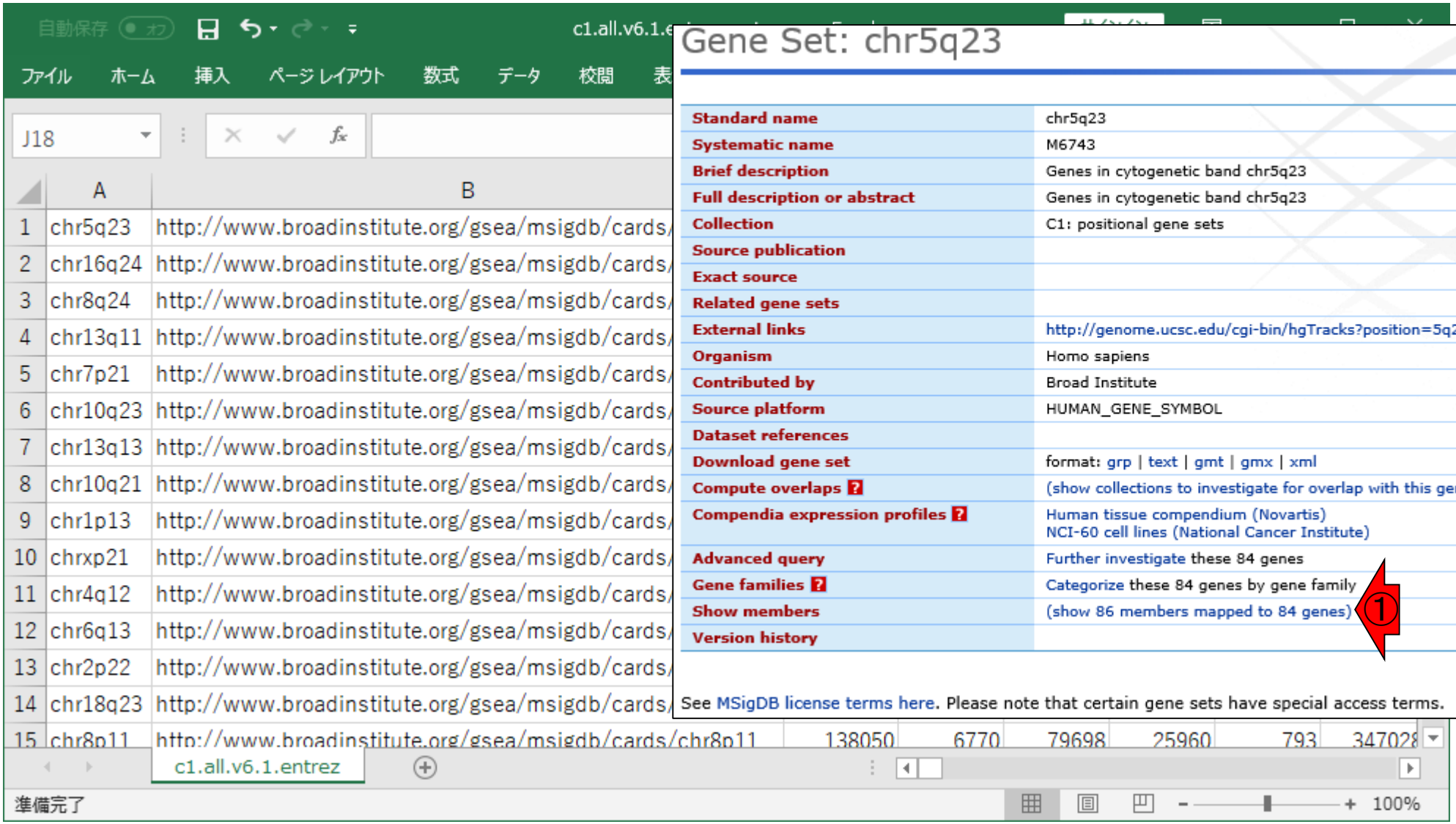
Gene Set: chr5q23

Standard name	chr5q23
Systematic name	M6743
Brief description	Genes in cytogenetic band chr5q23
Full description or abstract	Genes in cytogenetic band chr5q23
Collection	C1: positional gene sets
Source publication	
Exact source	
Related gene sets	
External links	http://genome.ucsc.edu/cgi-bin/hgTracks?position=5q23
Organism	Homo sapiens
Contributed by	Broad Institute
Source platform	HUMAN_GENE_SYMBOL
Dataset references	
Download gene set	format: grp | text | gmt | gmx | xml
Compute overlaps	(show collections to investigate for overlap with this gene set)
Compendia expression profiles	Human tissue compendium (Novartis) NCI-60 cell lines (National Cancer Institute)
Advanced query	Further investigate these 84 genes
Gene families	Categorize these 84 genes by gene family
Show members	(show 86 members mapped to 84 genes)
Version history	

③

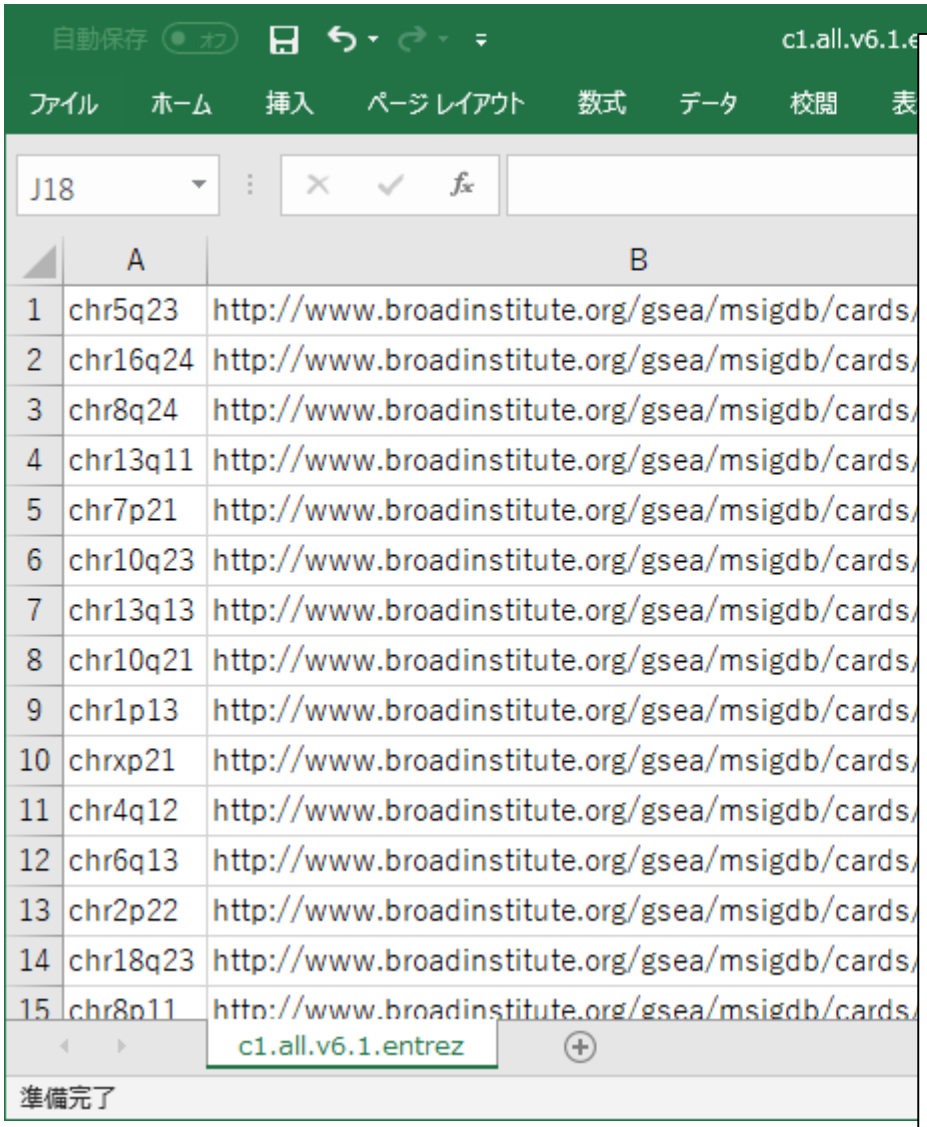
①をクリックすると…

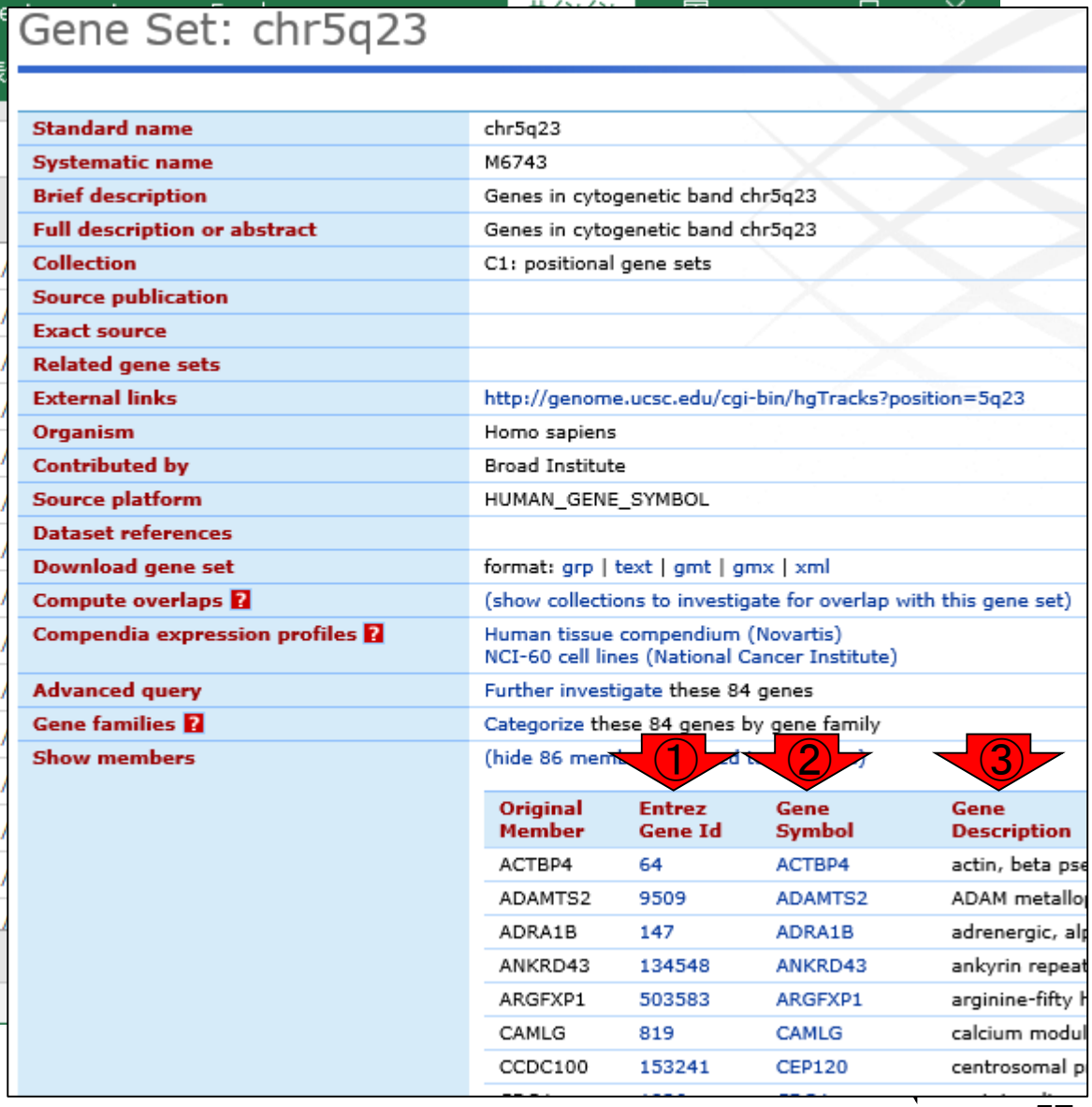
Show membersで…



Show membersで…

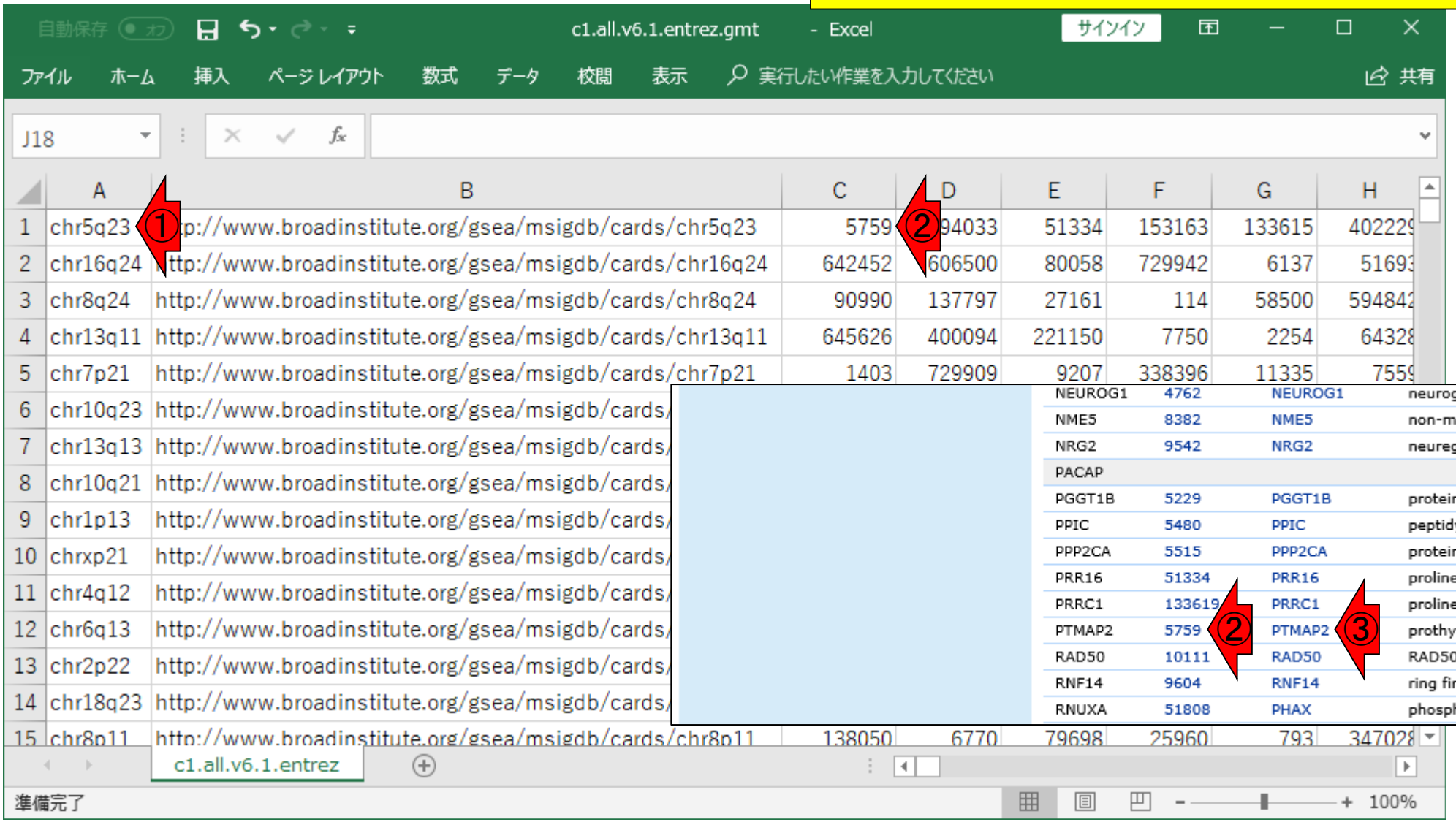
こんな感じになって、①Entrez gene IDに対応する②gene symbolや③gene descriptionが見られます





5759はPTMAP2

①1番目のデータセットchr5q23に含まれるメンバーで、Entrez gene IDが②5759の、③gene symbolはPTMAP2であることがわかります。

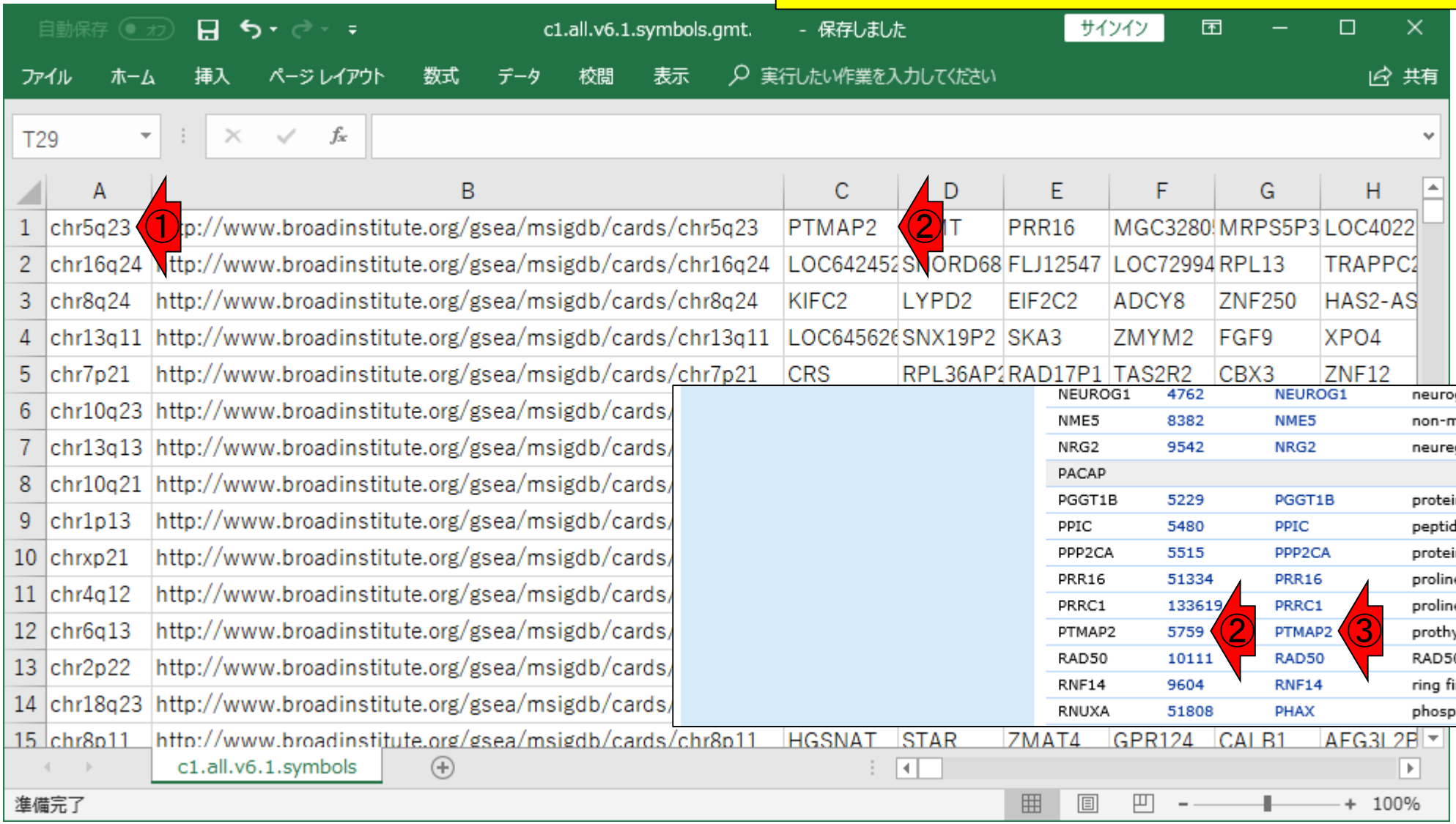


• 解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian 2005)

5759はPTMAP2

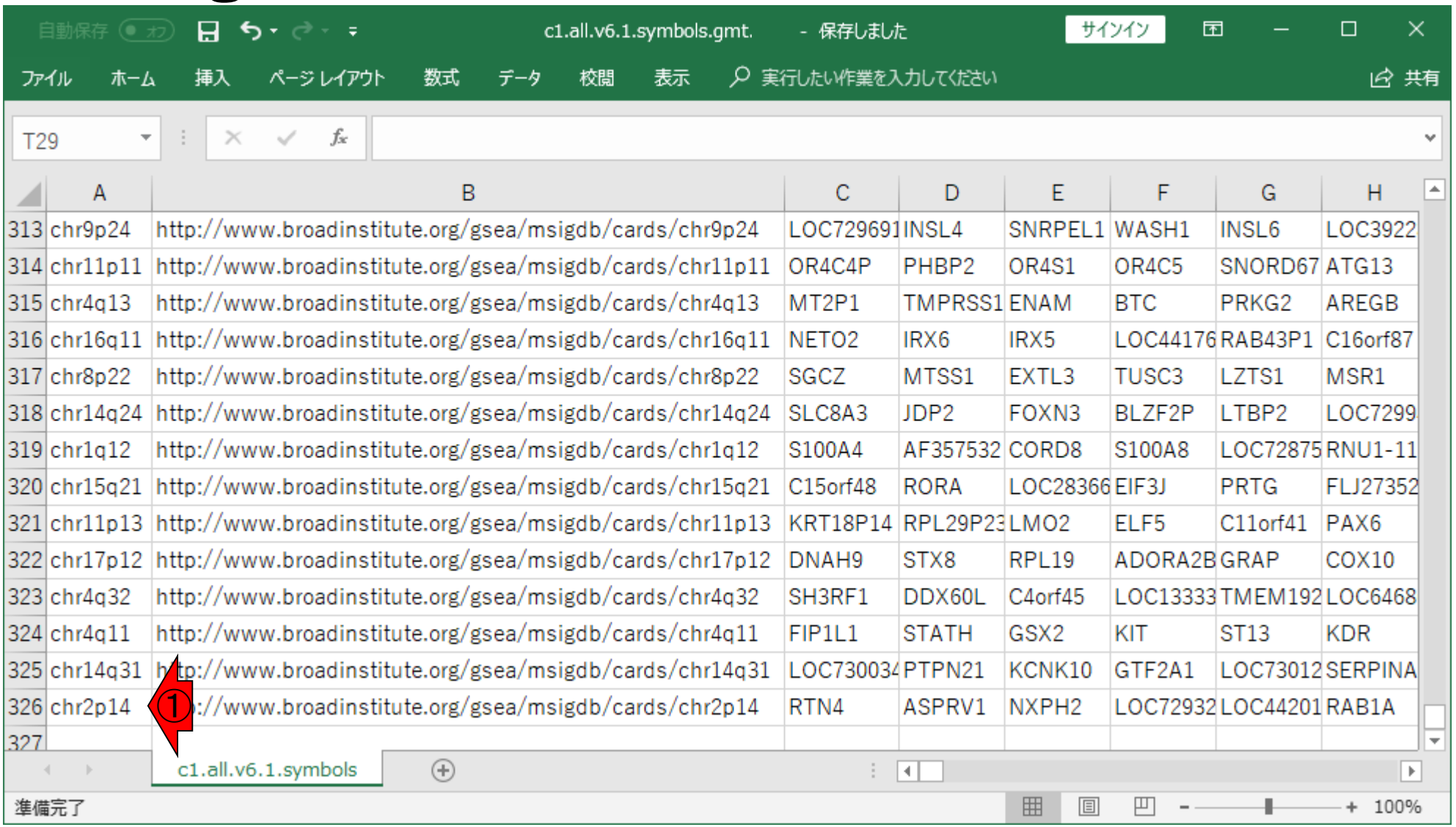
参考

c1.all.v6.1.symbols.gmtだとこんな感じになります。このファイルは遺伝子セット情報をgene symbolsで提供しているものなので妥当ですね。



①326行目までで終わりです。

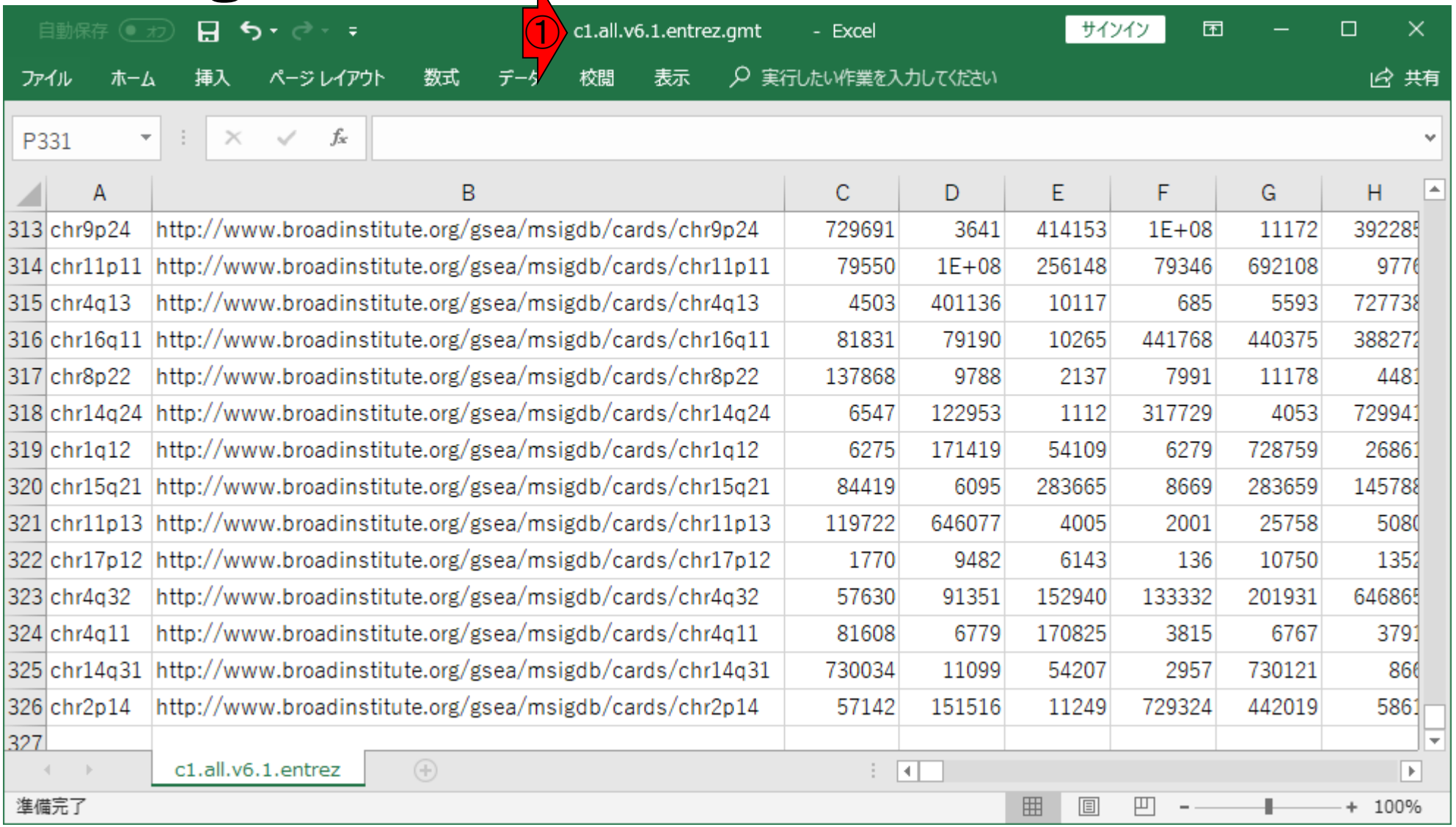
326 gene setsなので…



	A	B	C	D	E	F	G	H
313	chr9p24	http://www.broadinstitute.org/gsea/msigdb/cards/chr9p24	LOC729691	INSL4	SNRPEL1	WASH1	INSL6	LOC3922
314	chr11p11	http://www.broadinstitute.org/gsea/msigdb/cards/chr11p11	OR4C4P	PHBP2	OR4S1	OR4C5	SNORD67	ATG13
315	chr4q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q13	MT2P1	TMPRSS1	ENAM	BTC	PRKG2	AREGB
316	chr16q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr16q11	NETO2	IRX6	IRX5	LOC44176	RAB43P1	C16orf87
317	chr8p22	http://www.broadinstitute.org/gsea/msigdb/cards/chr8p22	SGCZ	MTSS1	EXTL3	TUSC3	LZTS1	MSR1
318	chr14q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr14q24	SLC8A3	JDP2	FOXN3	BLZF2P	LTBP2	LOC7299
319	chr1q12	http://www.broadinstitute.org/gsea/msigdb/cards/chr1q12	S100A4	AF357532	CORD8	S100A8	LOC72875	RNU1-11
320	chr15q21	http://www.broadinstitute.org/gsea/msigdb/cards/chr15q21	C15orf48	RORA	LOC28366	EIF3J	PRTG	FLJ27352
321	chr11p13	http://www.broadinstitute.org/gsea/msigdb/cards/chr11p13	KRT18P14	RPL29P23	LMO2	ELF5	C11orf41	PAX6
322	chr17p12	http://www.broadinstitute.org/gsea/msigdb/cards/chr17p12	DNAH9	STX8	RPL19	ADORA2B	GRAP	COX10
323	chr4q32	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q32	SH3RF1	DDX60L	C4orf45	LOC13333	TMEM192	LOC6468
324	chr4q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q11	FIP1L1	STATH	GSX2	KIT	ST13	KDR
325	chr14q31	http://www.broadinstitute.org/gsea/msigdb/cards/chr14q31	LOC730034	PTPN21	KCNK10	GTF2A1	LOC73012	SERPINA
326	chr2p14	http://www.broadinstitute.org/gsea/msigdb/cards/chr2p14	RTN4	ASPRV1	NXPH2	LOC72932	LOC44201	RAB1A
327								

326 gene setsなので…

①c1.all.v6.1.entrez.gmt
でも当然同じです。



	A	B	C	D	E	F	G	H
313	chr9p24	http://www.broadinstitute.org/gsea/msigdb/cards/chr9p24	729691	3641	414153	1E+08	11172	39228
314	chr11p11	http://www.broadinstitute.org/gsea/msigdb/cards/chr11p11	79550	1E+08	256148	79346	692108	9776
315	chr4q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q13	4503	401136	10117	685	5593	72773
316	chr16q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr16q11	81831	79190	10265	441768	440375	38827
317	chr8p22	http://www.broadinstitute.org/gsea/msigdb/cards/chr8p22	137868	9788	2137	7991	11178	448
318	chr14q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr14q24	6547	122953	1112	317729	4053	72994
319	chr1q12	http://www.broadinstitute.org/gsea/msigdb/cards/chr1q12	6275	171419	54109	6279	728759	2686
320	chr15q21	http://www.broadinstitute.org/gsea/msigdb/cards/chr15q21	84419	6095	283665	8669	283659	14578
321	chr11p13	http://www.broadinstitute.org/gsea/msigdb/cards/chr11p13	119722	646077	4005	2001	25758	508
322	chr17p12	http://www.broadinstitute.org/gsea/msigdb/cards/chr17p12	1770	9482	6143	136	10750	135
323	chr4q32	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q32	57630	91351	152940	133332	201931	64686
324	chr4q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q11	81608	6779	170825	3815	6767	379
325	chr14q31	http://www.broadinstitute.org/gsea/msigdb/cards/chr14q31	730034	11099	54207	2957	730121	86
326	chr2p14	http://www.broadinstitute.org/gsea/msigdb/cards/chr2p14	57142	151516	11249	729324	442019	586
327								

Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

MSigDB提供ファイル

①MSigDBは、Entrez gene IDs(.entrez.gmt)とGene symbols(.symbols.gmt)の2種類を提供。
当然、本家もそうなっています。

解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian_2005) NEW

MSigDBから遺伝子セット情報を含むGMTファイル(.gmt)を取得するやり方を示します。2018年6月現在は、MSigDB version 6.1です。Release Notesにもありますが、MSigDB v6.0以降、提供データがCreative Commons Attribution 4.0 International Licenseになったようです(KEGG、BioCarta、AAAS/STKE Cell Signaling Database dataの3種類を除く)。MSigDB v6.1では、以下に示す8個の主要なコレクション(8 major collections)が提供されています。ときどきMSigDBからgmtファイルをダウンロードできない事態に遭遇しますので、このサイト上でも**提供可能なものはダウンロードできる状態**にしています。但し、**このサイト上からダウロードした場合は、register**し、**MSigDBに対して仁義を果たして下さい。**MSigDBがfunding agenciesに利用者情報を報告するために必要です。また、MSigDBの論文だけでなく、例えば1番目のコレクションである**H hallmark gene sets**を利用する場合はLiberzon et al., 2015を、そして2番目のコレクションであるC2(curated gene sets)に含まれる**CP:Reactomeのgene sets**を利用する場合はJoshi-Tope et al., 2005などを適切に引用しましょう。

1. H: hallmark gene sets(50 gene sets)
   - gene symbols(h.all.v6.1.symbols.gmt)
   - entrez genes ids(h.all.v6.1.entrez.gmt)
2. C1: positional gene sets(326 gene sets)
   - gene symbols(c1.all.v6.1.symbols.gmt)
   - entrez genes ids(c1.all.v6.1.entrez.gmt)
3. C2: curated gene sets(4738 gene sets)
   - gene symbols(c2.all.v6.1.symbols.gmt)
   - entrez genes ids(c2.all.v6.1.entrez.gmt)
   - CGP: chemical and genetic perturbations(3409 gene sets)
     - gene symbols(c2.cgp.v6.1.symbols.gmt)
     - entrez genes ids(c2.cgp.v6.1.entrez.gmt)
   - CP: Canonical pathways(1329 gene sets)
     - gene symbols(c2.cp.v6.1.symbols.gmt)
     - entrez genes ids(c2.cp.v6.1.entrez.gmt)

機能解析の全体像

①機能解析の全体像に関する説明では、②のあたりで記載しています。

解析 | 機能解析 | について NEW

①

多少間違えているかもしれませんが、とりあえず2018年6月現在の私の理解に基づいて、全貌をざっくりと書きます。機能解析という項目ですが、実質的にはGene Set Enrichment Analysis (GSEA)を意識した内容です。GSEAは、発現変動解析の枠組みに属するもして、その後の**たものが多く濃**子群がDEGの

Mootha et al.,20
はGene Ontolog
する遺伝子群の
唱しています。
Subramanian et
解析(Gene Set
した機能解析を
トとしてKE
以下に示すよう

1. GOやKE
2. マイクロ
3. それらを
4. エンドユ
を提供(

MSigDBは、様々な遺伝子セットの情報を含むデータベースです。GSEAのサイトの右上にある図中で、Gene set Databaseと書かれているものに相当します。ユーザは、MSigDBから自分が調べたい遺伝子セット情報を含むGMTファイル(.gmt)を予めダウンロードしておく必要があります。従って、入力ファイルとして必要なものは2種類(マイクロアレイやRNA-seqで得られた発現行列のファイルと.gmtファイル)になります。これらを入力として、GSEAプログラムそのものや、その後提案された様々な遺伝子セット解析用プログラムを実行するのです。発現変動に関連した機能解析用プログラムの中には、パスウェイ解析に特化したもの、GO解析もできるもの、遺伝子セット解析全般ができるものなどいろいろあります。

現実問題として、**エンドユーザが特に手元にあるRNA-seqの発現データを用いてプログラムを実行する障壁は非常に高い**です。理由は、発現情報ファイル中のfeature IDと.gmtファイル中のIDとの対応付けを行う部分が厄介だからです。featureという曖昧な用語を用いているのは、発現行列の各行が仮にgeneを指し示すIDに限定されていたとしても、Ensembl gene ID、Entrez gene ID、gene symbolsなどが現実にあり得ます。また、exonやtranscriptを指し示すIDかもしれませんし、マイクロアレイデータの場合は各メーカーによって異なる独自のID (例えばAffymetrixのID)になります。それゆえ、**featureという曖昧な表現がよく使われる**のです。

②

現在MSigDBでは、Entrez gene ID(ファイル名の最後のほうが.entrez.gmt)とgene symbols(ファイル名の最後のほうが.symbols.gmt)の2種類が提供されています。それゆえ、手元の発現データファイル中のfeature IDがもしEntrez gene IDなら、.entrez.gmtを利用することになります。feature IDがもしEntrez gene ID以外なら、(多くの場合はgene symbolsとの対応付けは行える状況にあるので)予めfeature IDをgene symbolsにどうにかして変換してから、.symbols.gmtを利用してプログラムを実行することになります。

但し、**発現行列データ側の前処理として、同一feature IDsの重複除去**を行っておく必要もあります。有意な発現変動遺伝子セットを検出する際に、複数個存在する同一feature IDsの情報が過大評価されないようにするのが目的です。これも、実際に重複除去をやろうとすると色々と厄介です。例えば、発現変動遺伝子セット解析用パッケージのGSVAは、**発現行列データの格納形式としてExpressionSetオブジェクトを利用**しています。そして、前処理として重複除去を行う際にgenefilterパッケージ内のnsFilter関数(入力がExpressionSetオブジェクト)を利用しています。ExpressionSetオブジェクトは、特にユーザに意識させることなく(Rで)マイクロアレイデータ解析上でも使っていましたが、このようなデータ形式を取り扱うスキルもきっちり身につけていく必要があります。尚、マイクロアレイデータの頃はExpressionSetオブジェクトがよく使われていましたが、RNA-seqカウントデータの現在はSummarizedExperimentやRangedSummarizedExperimentがよく使われます。

機能解析の全体像

①手元の発現データのgene IDの種類に応じて、利用する遺伝子セットファイル(.entrez.gmt or .symbols.gmt)を切り替えます。この後で実行するGSVAパッケージの例題の発現データはEntrez gene IDなので、c1.all.v6.1.entrez.gmtをダウンロードしたのです。

解析 | 機能解析 | について NEW

多少間違えているかもしれませんが、とりあえず2018年6月現在
析という項目ですが、実質的にはGene Set Enrichment Analysis (GSEA)を意識した内容です。GSEAは、発現変動解析の枠
組みに属するも
して、その後の
たものが多く濃
子群がDEGの
Mootha et al.,20
はGene Ontolog
する遺伝子群の
唱しています。
Subramanian et
解析(Gene Set
した機能解析を
トとしてKEGG
以下に示すよう

1. GOや
2. マイク
3. それらを
4. エンドユ
   を提供(

MSigDBは、様々な遺伝子セットの情報を含むデータベースです。GSEAのサイトの右上にある図中で、Gene set Databaseと書かれているものに相当します。ユーザは、MSigDBから自分が調べたい遺伝子セット情報を含むGMTファイル(.gmt)を予めダウンロードしておく必要があります。従って、入力ファイルとして必要なものは2種類(マイクロアレイやRNA-seqで得られた発現行列のファイルと.gmtファイル)になります。これらを入力として、GSEAプログラムそのものや、その後提案された様々な遺伝子セット解析用プログラムを実行するのです。発現変動に関連した機能解析用プログラムの中には、パスウェイ解析に特化したもの、GO解析もできるもの、遺伝子セット解析全般ができるものなどいろいろあります。

現実問題として、**エンドユーザが特に手元にあるRNA-seqの発現データを用いてプログラムを実行する障壁は非常に高い**です。理由は、発現情報ファイル中のfeature IDと.gmtファイル中のIDとの対応付けを行う部分が厄介だからです。featureという曖昧な用語を用いているのは、発現行列の各行が仮にgeneを指し示すIDに限定されていたとしても、Ensembl gene ID、Entrez gene ID、gene symbolsなどが現実にあり得ます。また、exonやtranscriptを指し示すIDかもしれませんし、マイクロアレイデータの場合は各メーカーによって異なる独自のID (例えばAffymetrixのID)になります。それゆえ、**featureという曖昧な表現がよく使われる**のです。

現在MSigDBでは、Entrez gene ID(ファイル名の最後のほうが.entrez.gmt)とgene symbols(ファイル名の最後のほうが.symbols.gmt)の2種類が提供されています。それゆえ、手元の発現データファイル中のfeature IDがもしEntrez gene IDなら、.entrez.gmtを利用することになります。feature IDがもしEntrez gene ID以外なら、(多くの場合はgene symbolsとの対応付けは行える状況にあるので)予めfeature IDをgene symbolsにどうにかして変換してから、.symbols.gmtを利用してプログラムを実行することになります。

但し、**発現行列データ側の前処理として、同一feature IDsの重複除去**を行っておく必要もあります。有意な発現変動遺伝子セットを検出する際に、複数個存在する同一feature IDsの情報が過大評価されないようにするのが目的です。これも、実際に重複除去をやろうとすると色々と厄介です。例えば、発現変動遺伝子セット解析用パッケージのGSVAは、**発現行列データの格納形式としてExpressionSetオブジェクトを利用**しています。そして、前処理として重複除去を行う際にgenefilterパッケージ内のnsFilter関数(入力がExpressionSetオブジェクト)を利用しています。ExpressionSetオブジェクトは、特にユーザに意識させることなく(Rで)マイクロアレイデータ解析上でも使っていましたが、このようなデータ形式を取り扱うスキルもきっちり身につけていく必要があります。尚、マイクロアレイデータの頃はExpressionSetオブジェクトがよく使われていましたが、RNA-seqカウントデータの現在はSummarizedExperimentやRangedSummarizedExperimentがよく使われます。

• 解析 | 機能解析 | について　参考

gene symbolsの場合

①手元の発現データがEntrez gene ID以外であり、例えばEnsembl gene IDだった場合は、Ensembl gene IDとgene symbolsの対応情報を取得しておきます。そして、発現データのほうをgene symbolsに変換しておいて、.symbols.gmtファイルを用いて遺伝子セット解析を行います。

解析 | 機能解析 | について NEW

多少間違えているかもしれませんが、とりあえず2018年6月現在の私の理
析という項目ですが、実質的にはGene Set Enrichment Analysis (GSEA)
組みに属するも
して、その後の
たものが多く濃
子群がDEGの
Mootha et al.,20
はGene Ontolog
する遺伝子群の
唱しています。
Subramanian et
解析(Gene Set
した機能解析を
トとしてKEGG
以下に示すよう

1. GOやKE
2. マイクロ
3. それらを
4. エンドユ
   を提供(

MSigDBは、様々な遺伝子セットの情報を含むデータベースです。GSEAのウェブサイトの右上にある図中で、Gene set Databaseと書かれているものに相当します。ユーザは、MSigDBから自分が調べたい遺伝子セット情報を含むGMTファイル(.gmt)を予めダウンロードしておく必要があります。従って、入力ファイルとして必要なものは2種類(マイクロアレイやRNA-seqで得られた発現行列のファイルと.gmtファイル)になります。これらを入力として、GSEAプログラムそのものや、その後提案された様々な遺伝子セット解析用プログラムを実行するのです。発現変動に関連した機能解析用プログラムの中には、パスウェイ解析に特化したもの、GO解析もできるもの、遺伝子セット解析全般ができるものなどいろいろあります。

現実問題として、**エンドユーザが特に手元にあるRNA-seqの発現データを用いてプログラムを実行する障壁は非常に高い**です。理由は、発現情報ファイル中のfeature IDと.gmtファイル中のIDとの対応付けを行う部分が厄介だからです。featureという曖昧な用語を用いているのは、発現行列の各行が仮にgeneを指し示すIDに限定されていたとしても、Ensembl gene ID、Entrez gene ID、gene symbolsなどが現実にあり得ます。また、exonやtranscriptを指し示すIDかもしれませんし、マイクロアレイデータの場合は各メーカーによって異なる独自のID (例えばAffymetrixのID)になります。それゆえ、**featureという曖昧な表現がよく使われる**のです。

現在MSigDBでは、Entrez gene ID(ファイル名の最後のほうが.entrez.gmt)とgene symbols(ファイル名の最後のほうが.symbols.gmt)の2種類が提供されています。それゆえ、手元の発現データファイル中のfeature IDがもしEntrez gene IDなら、.entrez.gmtを利用することになります。feature IDがもしEntrez gene ID以外なら、(多くの場合はgene symbolsとの対応付けは行える状況にあるので)予めfeature IDをgene symbolsにどうにかして変換してから、.symbols.gmtを利用してプログラムを実行することになります。

但し、**発現行列データ側の前処理として、同一feature IDsの重複除去**を行っておく必要もあります。有意な発現変動遺伝子セットを検出する際に、複数個存在する同一feature IDsの情報が過大評価されないようにするのが目的です。これも、実際に重複除去をやろうとすると色々と厄介です。例えば、発現変動遺伝子セット解析用パッケージのGSVAは、**発現行列データの格納形式としてExpressionSetオブジェクトを利用**しています。そして、前処理として重複除去を行う際にgenefilterパッケージ内のnsFilter関数(入力がExpressionSetオブジェクト)を利用しています。ExpressionSetオブジェクトは、特にユーザに意識させることなく(Rで)マイクロアレイデータ解析上でも使っていましたが、このようなデータ形式を取り扱うスキルもきっちり身につけていく必要があります。尚、マイクロアレイデータの頃はExpressionSetオブジェクトがよく使われていましたが、RNA-seqカウントデータの現在はSummarizedExperimentやRangedSummarizedExperimentがよく使われます。

gene symbolsの場合

様々なIDの変換を行う際によく用いられるBioMartというデータベースをR経由で実行可能なbiomaRtというRパッケージもあります。が話がそれていくので深入りしません。

(Rで)塩基配列解析

(last modified 2018/06/26, since 2010)

このウェブページのR関連部分は、インストール | についての 推奨手順 (Windows
という前提で記述しています。初心者の 方は 基本的な利用法(Windows2015.04.0
(2015/04/03)

What's new?

- 「解析 | 機能解
- 「解析 | 機能解
- 「解析 | 機能解
- 「解析 | 機能解

- イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu 2013) (last modified 2015/02/24)
- イントロ | NGS | 配列取得 | シミュレーションデータ | について (last modified 2015/01/18)
- イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成から (last modified 2015/01/18)
- イントロ | NGS | アノテーション情報取得 | について (last modified 2014/03/26)
- イントロ | NGS | アノテーション情報取得 | GFF/GTF形式ファイル (last modified 2018/03/29)
- イントロ | NGS | アノテーション情報取得 | refFlat形式ファイル (last modified 2013/09/25)
- イントロ | NGS | アノテーション情報取得 | biomaRt(Durinck 2009) ① modified 2013/09/26)
- イントロ | NGS | アノテーション情報取得 | TxDb | について (last modified 2014/03/28)
- イントロ | NGS | アノテーション情報取得 | TxDb | TxDb.*から (last modified 2015/02/19)
- イントロ | NGS | アノテーション情報取得 | TxDb | GenomicFeatures(Lawrence 2013) (last modified 2015/02/19)推奨
- イントロ | NGS | アノテーション情報取得 | TxDb | GFF/GTF形式ファイルから (last modified 2016/02/09)
- イントロ | NGS | 読み込み | BSgenome | 基本情報を取得 (last modified 2016/04/22)
- イントロ | NGS | 読み込み | FASTA形式 | 基本情報を取得 (last modified 2016/04/22)
- イントロ | NGS | 読み込み | FASTA形式 | description行の記述を整形 (last modified 2014/04/05)
- イントロ | NGS | 読み込み | FASTQ形式 | 基礎 (last modified 2015/07/26)

Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

検証用データ

以前のスライドで遺伝子セット解析を実行するプログラムとして②GSVAを採択した根拠を示した際に、赤枠内のプログラムについては言及したが、③については言及しなかった。

(Rで)塩基配列解析

(last modified 2019/07/17, since 2010)

このウェブペ
必要なパッケ
Macintosh20
しました。(2

What's new

- 「解析 | 発
  TCC(Sun
  たからです

- 解析 | 機能解析 | について (last modified 2018/06/24)
- 解析 | 機能解析 | GMTファイル取得 | について (last modified 2018/07/17)
- 解析 | 機能解析 | GMTファイル取得 | EGSEAdata(Alhamdoosh 2017) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | GeneSetDB(Araki 2012) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian 2005) (last modified 2018/06/25)
- 解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan 2018) (
- 解析 | 機能解析 | 遺伝子セット解析 | GSVA(Hänzelmann 2
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | SeqGSEA(
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | GSVA(Hänzelma
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | GOseq(Young 2
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | について (last modif
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | GSAR (Rahmatallah
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | SeqGSEA(Wang 20
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | GSVA(Hänzelmann
- 解析 | 分類 | について(last modified 2018/06/27)

①

R用:

- GAGE : Luo et al., BMC Bioinformatics, 2009
- goseq : Young et al., Genome Biol., 2010
- GOSemSim : Yu et al., Bioinformatics, 2010
- Rスクリプト : Gao et al., Bioinformatics, 2011
- clusterProfiler : Yu et al., OMICS., 2012
- RamiGO : Schröder et al., Bioinformatics, 2013
- GSVA : Hänzelmann et al., BMC Bioinformatics, 2013
- SeqGSEA(各群5反復以上を要求) : Wang et al., Bioinformatics, 2014
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- GOplot(Visualization用) : Walter et al., Bioinformatics, 2015
- RNA-Enrich(論文のsuppl) : Lee et al., Bioinformatics, 2015
- GOexpress : Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- rgsepd : Stamm et al., BMC Bioinformatics, 2019

②

③

Pickrell data

講義資料作成当時の私は、説明しやすい検証用データセットも探していた。①AbsFilterGSEA論文中の、②のあたりで、③Pickrell dataという2群間比較用（29 males vs. 40 females）のカウントデータが存在することを発見。

Comparison of GSEA methods for RNA-seq data ②

The performances of GSEA methods were compared for published RNA-s
eral aspects. First, two RNA-seq datasets denoted by Pickrell and Li data, respectively, were
anal d for comparing power and accuracy as follows:
③ The Pickrell data were generated from the lymphoblastoid cell lines of 69 unrelated Nigerian individuals (29 male and 40 female) [44]. To analyze the chromosomal differences in expression between male and female, MSigDB C1 (cytogenetic band gene-sets) [45–47] was used for analysis. The GSEA-SP with SNR gene score was applied for the total dataset which resulted in two significant gene-sets 'chryq11' (FDR = 0.00143) and 'chrxp22' (FDR = 0.0514) both of which were sex-specific. These two gene-sets were significantly up-regulated in male and female groups, respectively. Since the GSEA-SP controls the false positives well, these two gene-sets were regarded as true positives. Then, five samples were randomly selected from

- nformatics, 2009
- me Biol., 2010
- nformatics, 2010
- oinformatics, 2011
- MICS., 2012
- RamiGO : Schroder et al., Bioinformatics, 2013
- GSVA : Hänzelmann et al., BMC Bioinformatics, 2013
- SeqGSEA(各群5反復以上を要求) : Wang et al., Bioinformatics, 2014
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- GOplot(Visualization用) : Walter et al., Bioinformatics, 2015
- RNA-Enrich(論文のsuppl) : Lee et al., Bioinformatics, 2015
- GOexpress : Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016 ③
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- rgsepd : Stamm et al., BMC Bioinformatics, 2019

AbsFilterGSEA(Yoon et al., PLoS One, 11: e0165919, 2016)

MSigDB C1

①このPickrell dataと、②MSigDB C1カテゴリーの発現変動遺伝子セット解析結果として、③chryq11がおそらく最上位に近い有意性を示し、④male群で高発現であることまで分かった。この段階で、c1.all.v6.1.entrez.gmtまたはc1.all.v6.1.symbols.gmtが有望株だと認識。

Comparison of GSEA methods for RNA-seq data

The performances of GSEA methods were compared for published R
eral aspects. First, two RNA-seq datasets denoted by Pickrell and Li
ana d for comparing power and accuracy as follows:
The Pickrell data were generated from the lymphoblastoid cell lines of 69 unrelated Nigerian individuals (29 male and 40 f e) [44]. To analyze the chromosomal differences in expression between male and female, MSigDB C1 (cytogenetic band gene-sets) [45–47] was used for analysis. The GSEA-SP w SNR gene score was applied for the total dataset which resulted in two significant gene sets 'chryq11' (FDR = 43) and 'chrxp22' (FDR = 0.0514) both of which were sex-specific. These two gene-sets were significantly up-regulated in male and female groups, respectively. Since the GSEA-SP controls the false positives well, these two gene-sets were regarded as true positives. Then, five samples were randomly selected from

- oinformatics, 2009
- me Biol., 2010
- nformatics, 2010
- oinformatics, 2011
- OMICS., 2012
- RamiGO : Schröder et al., Bioinformatics, 2013
- GSVA : Hänzelmann et al., BMC Bioinformatics, 2013
- SeqGSEA(各群5反復以上を要求) : Wang et al., Bioinformatics, 2014
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- GOplot(Visualization用) : Walter et al., Bioinformatics, 2015
- RNA-Enrich(論文のsuppl) : Lee et al., Bioinformatics, 2015
- GOexpress : Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- rgsepd : Stamm et al., BMC Bioinformatics, 2019

AbsFilterGSEA

①AbsFilterGSEAが使いやすければよいが…
実際はそうではなかった（個人の感想です）。

R用:

- GAGE : Luo et al., BMC Bioinformatics, 2009
- goseq : Young et al., Genome Biol., 2010
- GOSemSim : Yu et al., Bioinformatics, 2010
- Rスクリプト : Gao et al., Bioinformatics, 2011
- clusterProfiler : Yu et al., OMICS., 2012
- RamiGO : Schröder et al., Bioinformatics, 2013
- GSVA : Hänzelmann et al., BMC Bioinformatics, 2013
- SeqGSEA(各群5反復以上を要求) : Wang et al., Bioinformatics, 2014
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- GOplot(Visualization用) : Walter et al., Bioinformatics, 2015
- RNA-Enrich(論文のsuppl) : Lee et al., Bioinformatics, 2015
- GOexpress : Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017
- ① AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- rgsepd : Stamm et al., BMC Bioinformatics, 2019

AbsFilterGSEA

一般に、①CRAN提供パッケージは、Bioconductor提供パッケージに比べて利用法の解読が難解です。実際私は②のReference manualをみてガッカリし、他にわかりやすいパッケージはないか探した結果としてGSVAを眺め、採用しました。

https://cran.r-project.org/web/packages/AbsFilterGSEA/

CRAN - Package AbsFilterG

①

AbsFilterGSEA: Improved False Positive Control of Gene-Permuting GSEA with Absolute Filtering

Gene-set enrichment analysis (GSEA) is popularly used to assess the enrichment of differential signal in a pre-defined gene-set without using a cutoff threshold for differential expression. The significance of enrichment is evaluated through sample- or gene-permutation method. Although the sample-permutation approach is highly recommended due to its good false positive control, we must use gene-permuting method if the number of samples is small. However, such gene-permuting GSEA (or preranked GSEA) generates a lot of false positive gene-sets as the inter-gene correlation in each gene set increases. These false positives can be successfully reduced by filtering with the one-tailed absolute GSEA results. This package provides a function that performs gene-permuting GSEA calculation with or without the absolute filtering. Without filtering, users can perform (original) two-tailed or one-tailed absolute GSEA.

Version:	1.5.1
Imports:	Rcpp, Biobase, stats, DESeq, limma
LinkingTo:	Rcpp, RcppArmadillo
Published:	2017-09-21
Author:	Sora Yoon
Maintainer:	Sora Yoon <yoonsora at unist.ac.kr>
License:	GPL-2
NeedsCompilation:	yes
CRAN checks:	AbsFilterGSEA results

Downloads:

②

Reference manual:	AbsFilterGSEA.pdf
Package source:	AbsFilterGSEA_1.5.1.tar.gz
Windows binaries:	r-devel: AbsFilterGSEA_1.5.1.zip, r-release: AbsFilterGSEA_1.5.1.zip, r-oldrel: AbsFilterGSEA_1.5.1.zip
OS X binaries:	r-release: AbsFilterGSEA_1.5.1.tgz, r-oldrel: AbsFilterGSEA_1.5.1.tgz
Old sources:	AbsFilterGSEA archive

Linking:

Please use the canonical form https://CRAN.R-project.org/package=AbsFilterGSEA to link to this page.

Bioinformatics, 2009
nome Biol., 2010
ioinformatics, 2010
Bioinformatics, 2011
, OMICS., 2012
., Bioinformatics, 2013
al., BMC Bioinformatics, 2013
を要求) : Wang et al., Bioinformatics, 2014
l., Sci Rep., 2014
: Walter et al., Bioinformatics, 2015
) : Lee et al., Bioinformatics, 2015
ht et al., BMC Bioinformatics, 2016
nd ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
al., Bioinformatics, 2017
licates用) : Yoon et al., PLoS One, 2016
al., BMC Bioinformatics, 2017
Data Min., 2017
BMC Bioinformatics, 2019

GSVA

②GSVAは引用回数も多く、③Bioconductorのパッケージであったことが決め手。さらに解説PDFを読むと、Pickrellデータも例題として使われていることが判明。③をクリック

- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について

(Rで)塩基配列解析

(last modified 2019/07/17, since 2010)

- 解析 | 機能解析 | について (last modified 2018/06/24)
- 解析 | 機能解析 | GMTファイル取得 | について (last modified 2018/07/17)
- 解析 | 機能解析 | GMTファイル取得 | EGSEAdata(Alhamdoosh_2017) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | GeneSetDB(Araki_2012) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian_2005) (last modified 2018/06/25)
- 解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan_2018) (
- 解析 | 機能解析 | 遺伝子セット解析 | GSVA(Hänzelmann_2
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | SeqGSEA(
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | GSVA(Hänzelma
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | GOseq(Young_2
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | について (last modif
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | GSAR (Rahmatallah
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | SeqGSEA(Wang_20
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | GSVA(Hänzelmann_
- 解析 | 分類 | について(last modified 2018/06/27)

①

R用:

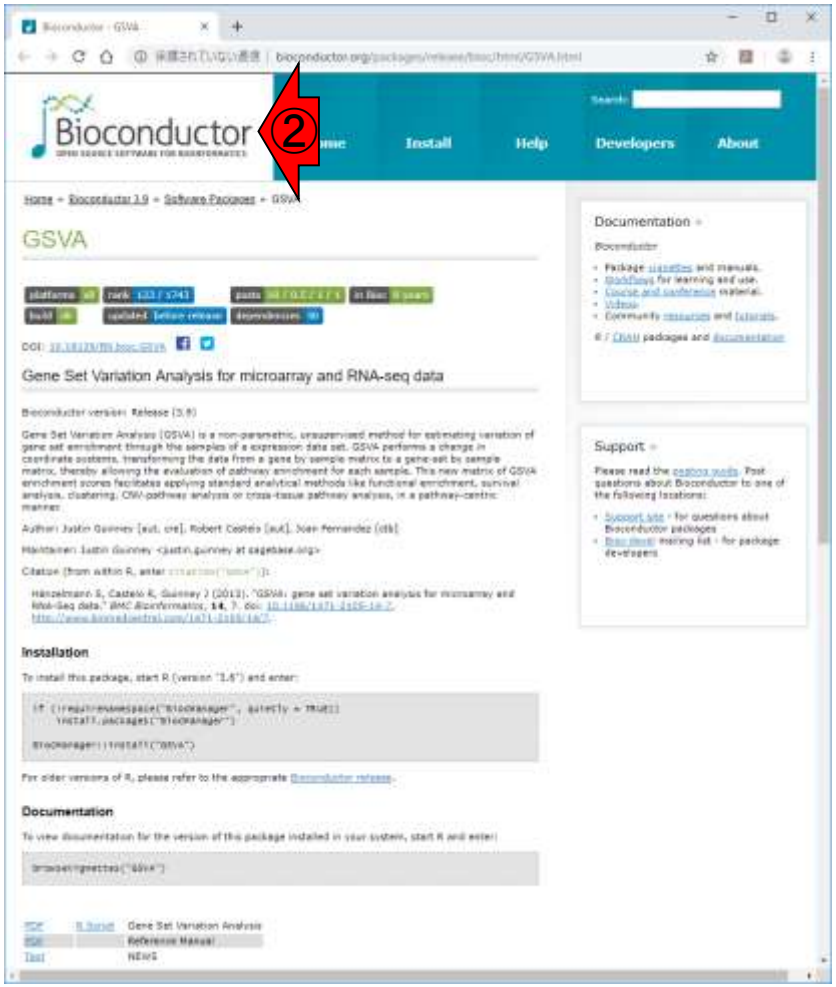
- GAGE : Luo et al., BMC Bioinformatics, 2009
- goseq : Young et al., Genome Biol., 2010
- GOSemSim : Yu et al., Bioinformatics, 2010
- Rスクリプト : Gao et al., Bioinformatics, 2011
- clusterProfiler : Yu et al., OMICS., 2012
- RamiGO : Schröder et al., Bioinformatics, 2013
- GSVA : Hänzelmann et al., BMC Bioinformatics, 2013
- SeqGSEA(各群5反復以上を要求) : Wang et al., Bioinformatics, 2014
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- GOplot(Visualization用) : Walter et al., Bioinformatics, 2015
- RNA-Enrich(論文のsuppl) : Lee et al., Bioinformatics, 2015
- GOexpress : Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- rgsepd : Stamm et al., BMC Bioinformatics, 2019

③

②

GSVA

①をクリックすると、②Bioconductorから提供されているパッケージの場合はこんな感じになります。



R用:

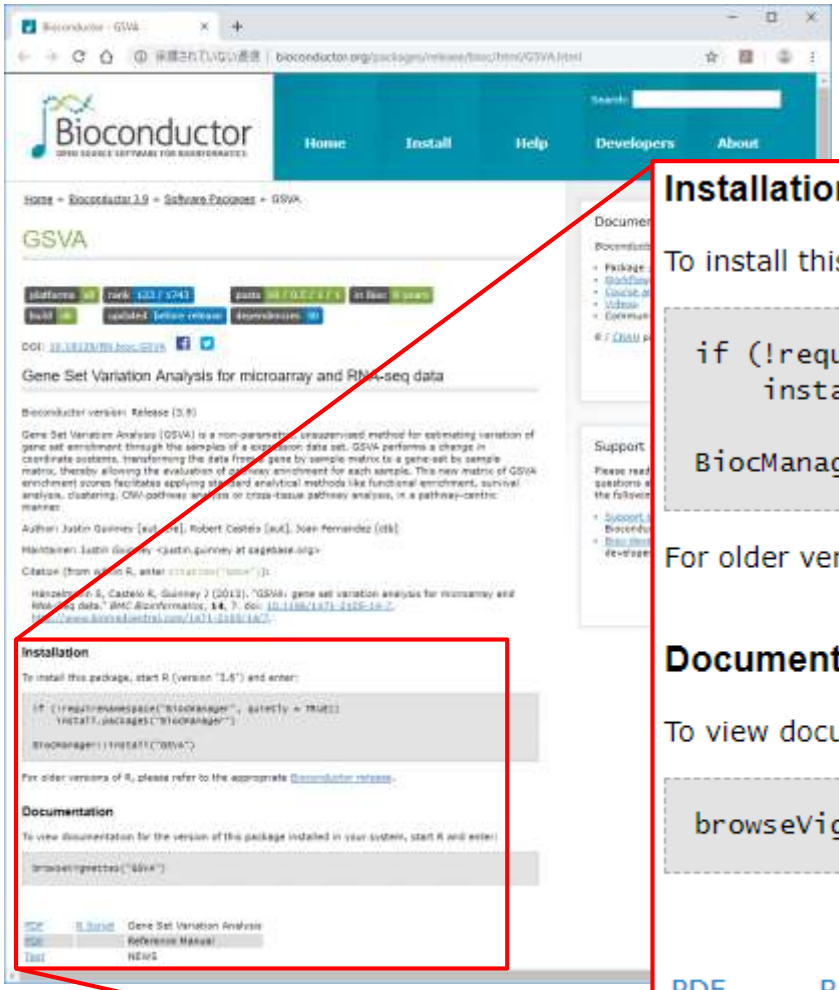
- GAGE : Luo et al., BMC Bioinformatics, 2009
- goseq : Young et al., Genome Biol., 2010
- GOSemSim : Yu et al., Bioinformatics, 2010
- Rスクリプト : Gao et al., Bioinformatics, 2011
- clusterProfiler : Yu et al., OMICS., 2012
- RamiGO : Schröder et al., Bioinformatics, 2013
- ① GSVA : Hänzelmann et al., BMC Bioinformatics, 2013
- SeqGSEA(各群5反復以上を要求) : Wang et al., Bioinformatics, 2014
- GSAASeqSP : Xiong et al., Sci Rep., 2014
- GOplot(Visualization用) : Walter et al., Bioinformatics, 2015
- RNA-Enrich(論文のsuppl) : Lee et al., Bioinformatics, 2015
- GOexpress : Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA) : Hundt et al., BMC Bioinformatics, 2016
- EGSEA : Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用) : Yoon et al., PLoS One, 2016
- GSAR : Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA : Ren et al., BioData Min., 2017
- rgsepd : Stamm et al., BMC Bioinformatics, 2019

Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

左側の赤枠部分を拡大表示。

GSVA



Installation

To install this package, start R (version "3.6") and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("GSVA")
```

For older versions of R, please refer to the appropriate Bioconductor release.

Documentation

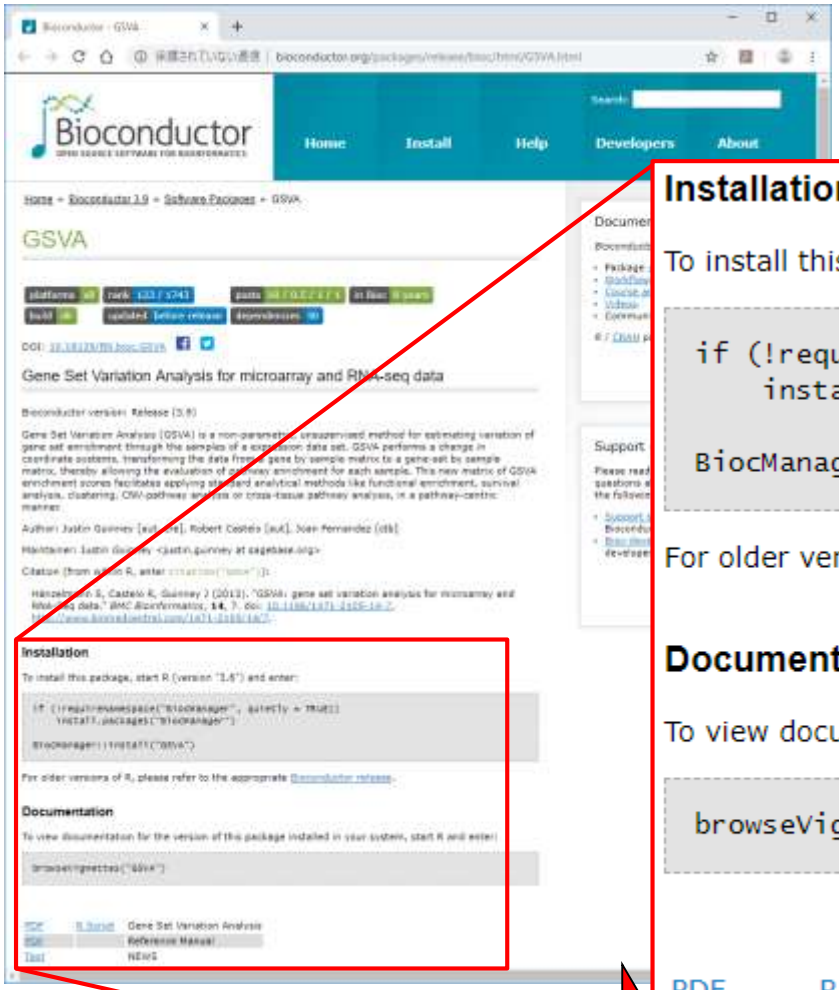
To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("GSVA")
```

PDF	R Script	Gene Set Variation Analysis
PDF		Reference Manual
Text		NEWS

GSVA

①GSVAのインストール手順の説明。
②Reference ManualのPDF。これは
CRAN（しーらん）にもあるやつです。



Installation

To install this package, start R (version "3.6") and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("GSVA")
```

①

For older versions of R, please refer to the appropriate Bioconductor release.

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

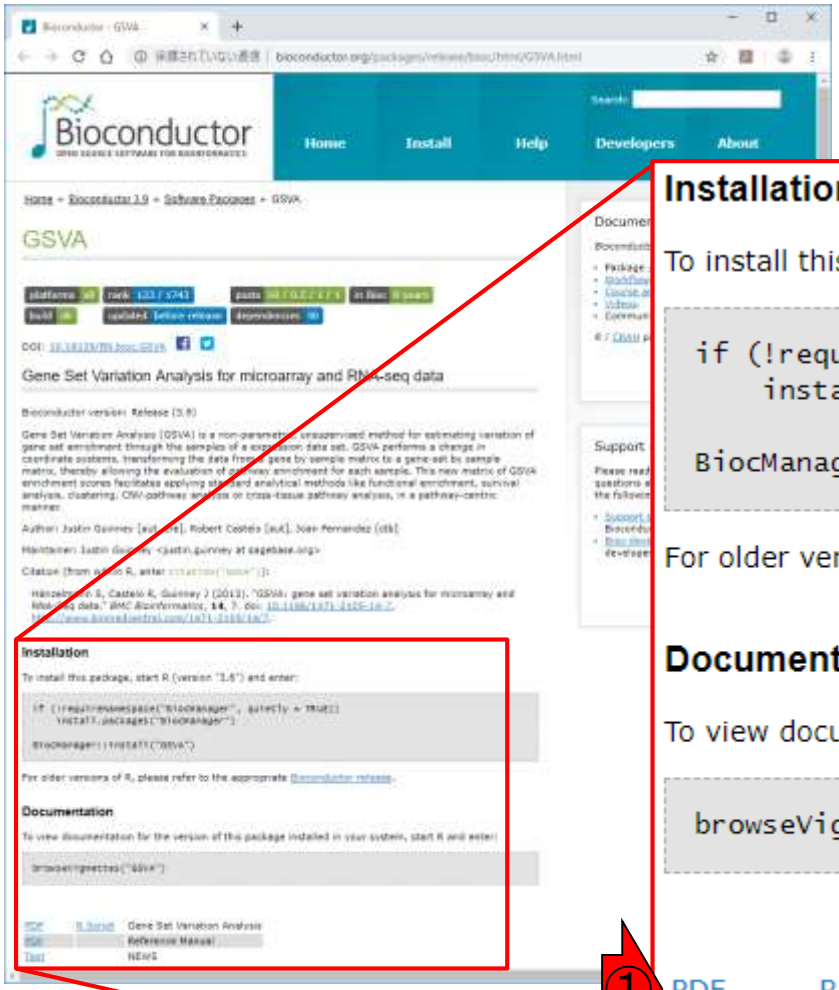
```
browseVignettes("GSVA")
```

PDF	R Script	Gene Set Variation Analysis
PDF		Reference Manual
Text		NEWS

②

GSVAの解説PDF

①GSVAの解説PDF(vignettes;ビニェット)。これはBioconductorのサイトに存在するものです。GSVAインストール後は、R起動直後に、②のコピペでも(手元のPC内に存在する)解説PDFを開けます。



Installation

To install this package, start R (version "3.6") and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("GSVA")
```

For older versions of R, please refer to the appropriate Bioconductor release.

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("GSVA")
```

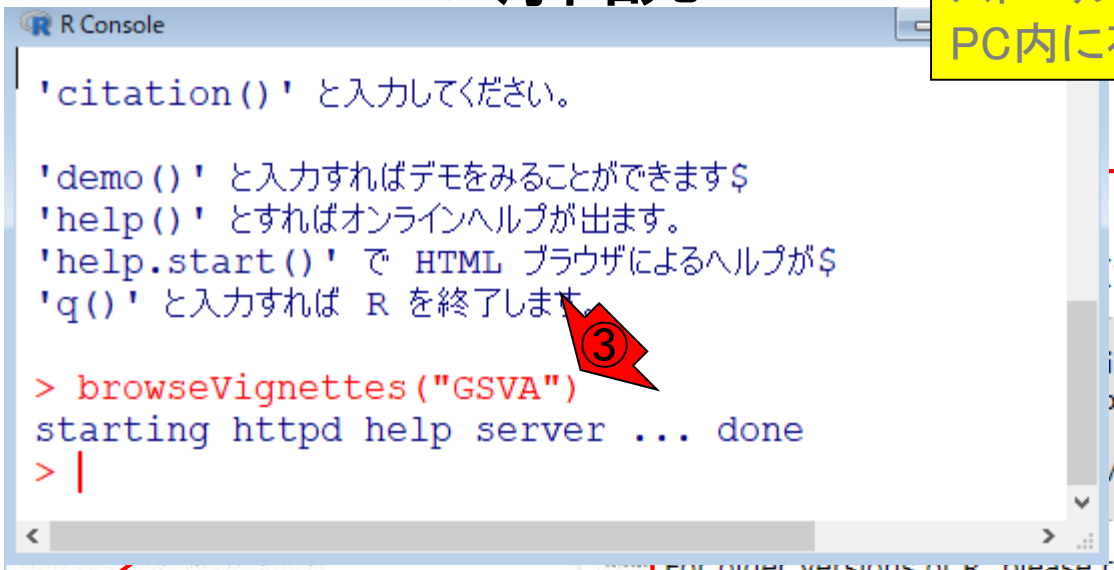
②

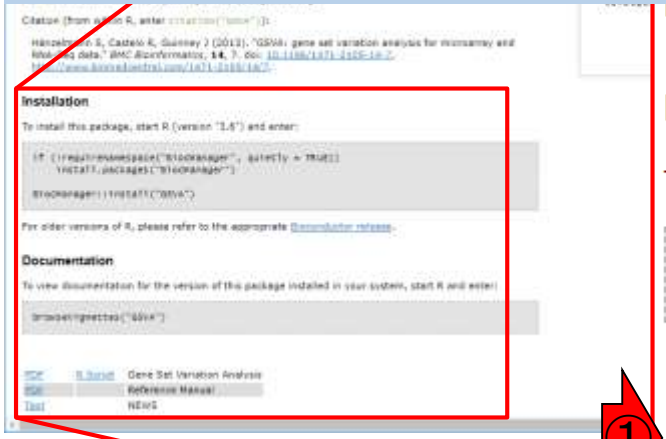
①

PDF	R Script	Gene Set Variation Analysis
PDF		Reference Manual
Text		NEWS

GSVAの解説PDF

①GSVAの解説PDF(vignettes;ビニェット)。これはBioconductorのサイトに存在するものです。GSVAインストール後は、R起動直後に、②のコピペでも(手元のPC内に存在する)解説PDFを開けます。③こんな感じ。





version "3.6") and enter:

```
iocManager", quietly = TRUE))
ocManager")

/A")
```

For older versions of R, please refer to the appropriate Bioconductor release.

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("GSVA")
```

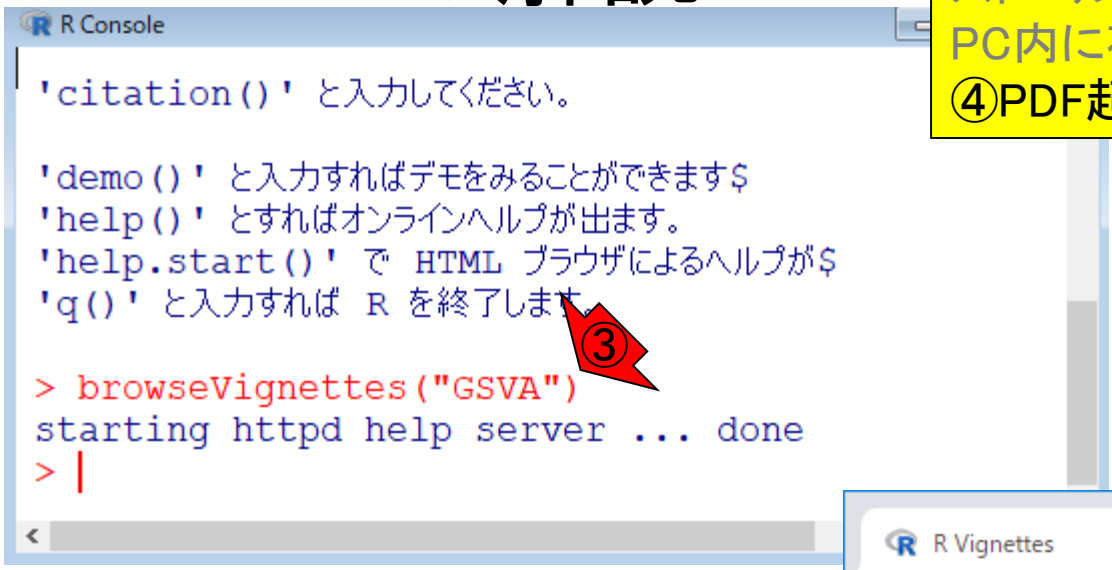
②

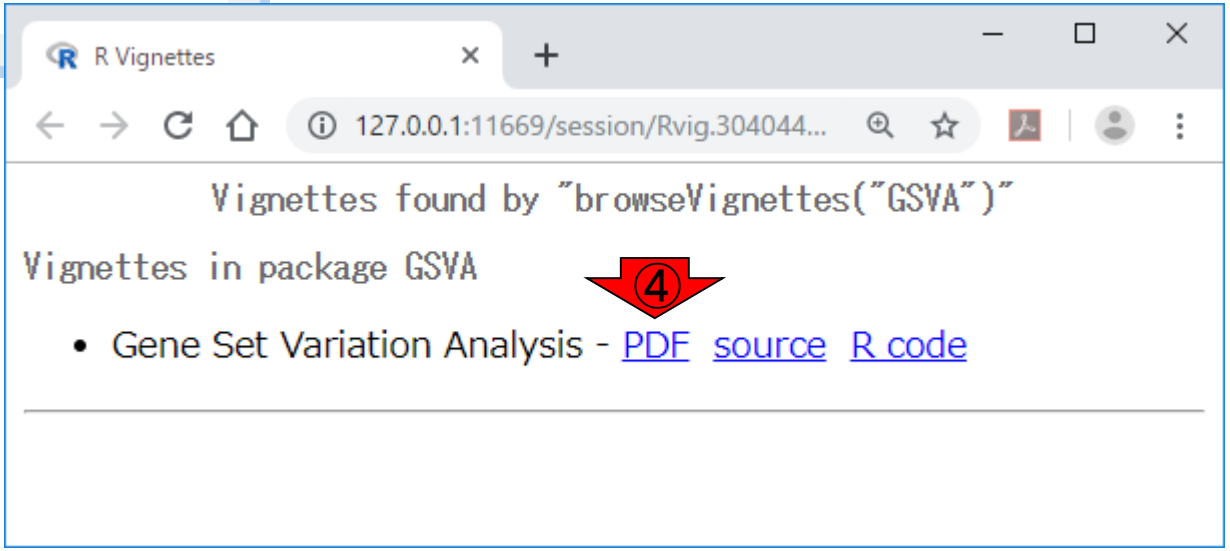
①

PDF	R Script	Gene Set Variation Analysis
PDF		Reference Manual
Text		NEWS

GSVAの解説PDF

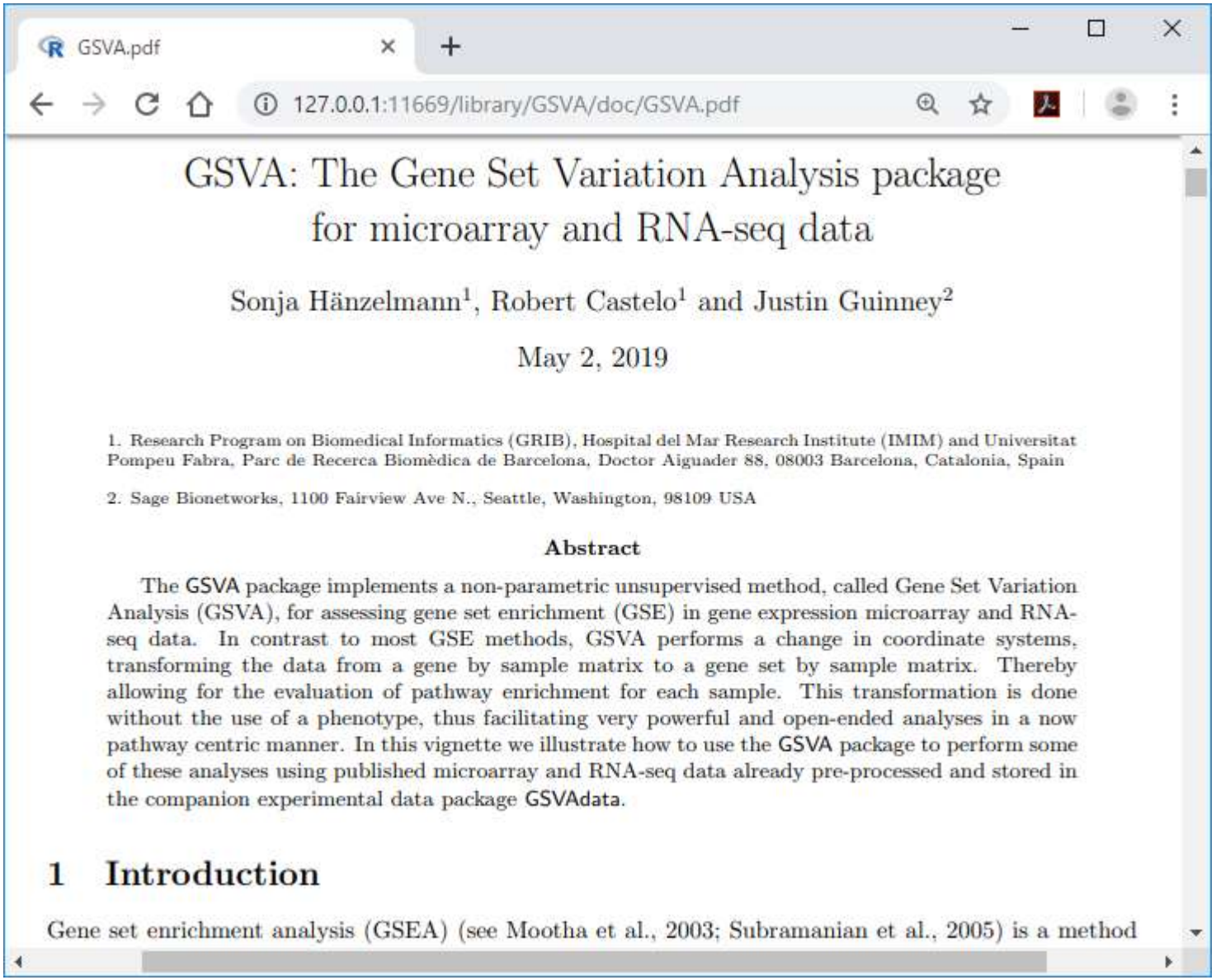
①GSVAの解説PDF(vignettes;ビニェット)。これはBioconductorのサイトに存在するものです。GSVAインストール後は、R起動直後に、②のコピペでも(手元のPC内に存在する)解説PDFを開けます。③こんな感じ。
④PDF起動して、利用法を学んでいく。





GSVAの解説PDF

こんな感じになります。基本的には、順番にざっと読んで、このパッケージの使用法のノリに慣れていきます。

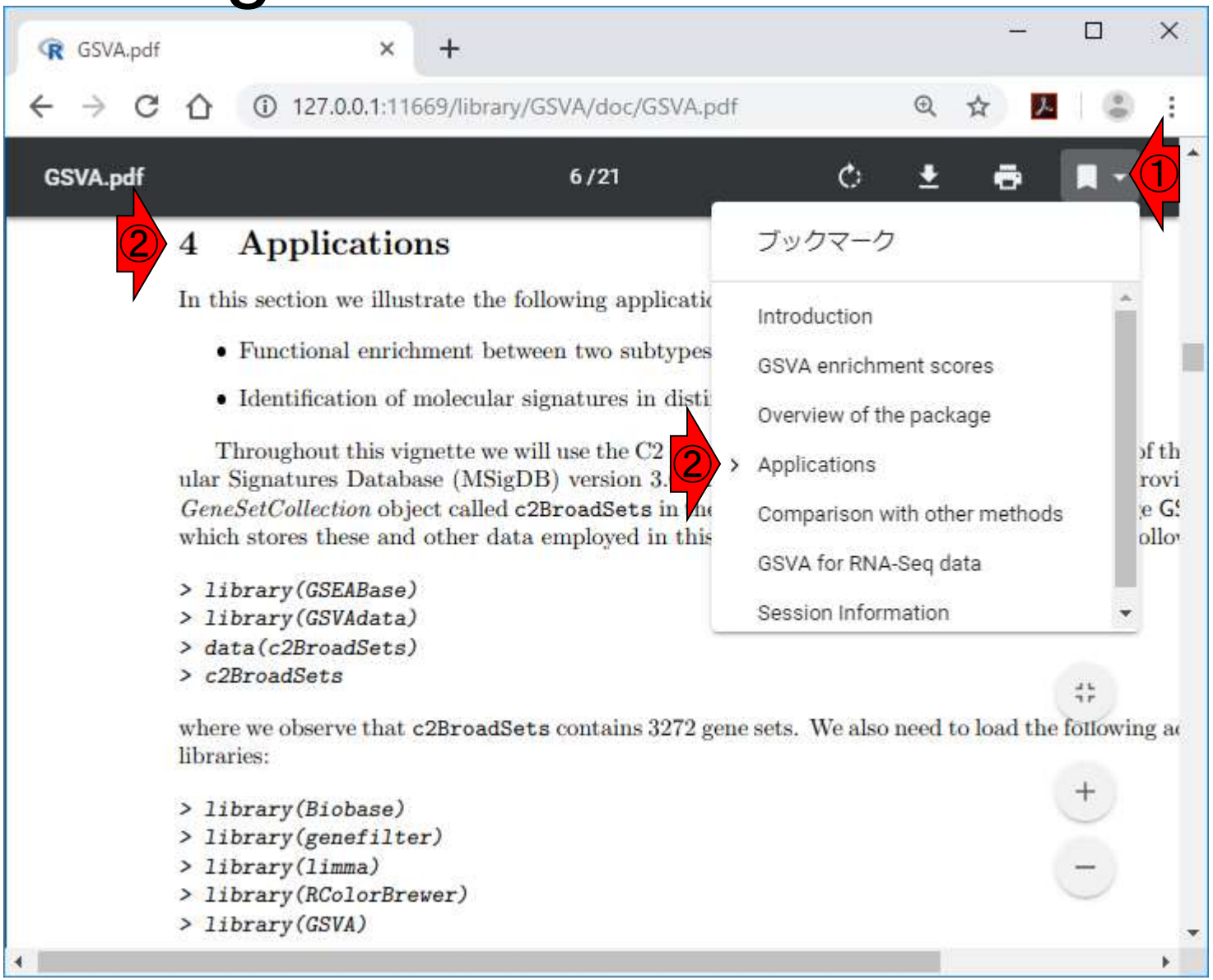


Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

MSigDBのC2

①ブックマークの、②Applicationを選択した結果。単に②までページ下部に移動しているだけです。



MSigDBのC2

①ブックマークの、②Applicationを選択した結果。単に②までページ下部に移動しているだけです。③このパッケージは、遺伝子セット情報としてMSigDBのC2コレクションを利用していることがわかる。

GSVA.pdf

127.0.0.1:11669/library/GSVA/doc/GSVA.pdf

4 Applications

In this section we illustrate the following applications of GSVA:

- Functional enrichment between two subtypes of leukemia.
- Identification of molecular signatures in distinct glioblastoma subtypes.

Throughout this vignette we will use the C2 collection of curated gene sets that form part of the Molecular Signatures Database (MSigDB) version 3.0. This particular collection of gene sets is provided as a *GeneSetCollection* object called `c2BroadSets` in the accompanying experimental data package GSVAdata, which stores these and other data employed in this vignette. These data can be loaded as follows:

③

```
> library(GSEABase)
> library(GSVAdata)
> data(c2BroadSets)
> c2BroadSets
```

where we observe that `c2BroadSets` contains 3272 gene sets. We also need to load the following additional libraries:

```
> library(Biobase)
> library(genefilter)
> library(limma)
> library(RColorBrewer)
```

MSigDBのC2

①それはGSVAdataというパッケージに含まれる、②c2BroadSetsという名前の、③GeneSetCollectionオブジェクトだということがわかる。

GSVA.pdf

127.0.0.1:11669/library/GSVA/doc/GSVA.pdf

4 Applications

In this section we illustrate the following applications of GSVA:

- Functional enrichment between two subtypes of leukemia.
- Identification of molecular signatures in distinct glioblastoma subtypes.

Throughout this vignette we will use the C2 collection of curated gene sets that form part of the Molecular Signatures Database (MSigDB) version 3.0. This particular collection of gene sets is provided as a *GeneSetCollection* object called `c2BroadSets` in the accompanying experimental data package GSVAdata, which stores these and other data employed in this vignette. These data can be loaded as follow

③ ② ①

```
> library(GSEABase)
> library(GSVAdata)
> data(c2BroadSets)
> c2BroadSets
```

where we observe that `c2BroadSets` contains 3272 gene sets. We also need to load the following additional libraries:

```
> library(Biobase)
> library(genefilter)
> library(limma)
> library(RColorBrewer)
```

MSigDBのC2

①これをコピーして、Rコンソール画面上で「コマンドのみペースト」すれば確認できる

GSVA.pdf

127.0.0.1:11669/library/GSVA/doc/GSVA.pdf

4 Applications

In this section we illustrate the following applications of GSVA:

- Functional enrichment between two subtypes of leukemia.
- Identification of molecular signatures in distinct glioblastoma subtypes.

Throughout this vignette we will use the C2 collection of curated gene sets that form part of the Molecular Signatures Database (MSigDB) version 3.0. This particular collection of gene sets is provided as a *GeneSetCollection* object called `c2BroadSets` in the accompanying experimental data package GSVAdata, which stores these and other data employed in this vignette. These data can be loaded as follows:

```
> library(GSEABase)
> library(GSVAdata)
> data(c2BroadSets)
> c2BroadSets
```

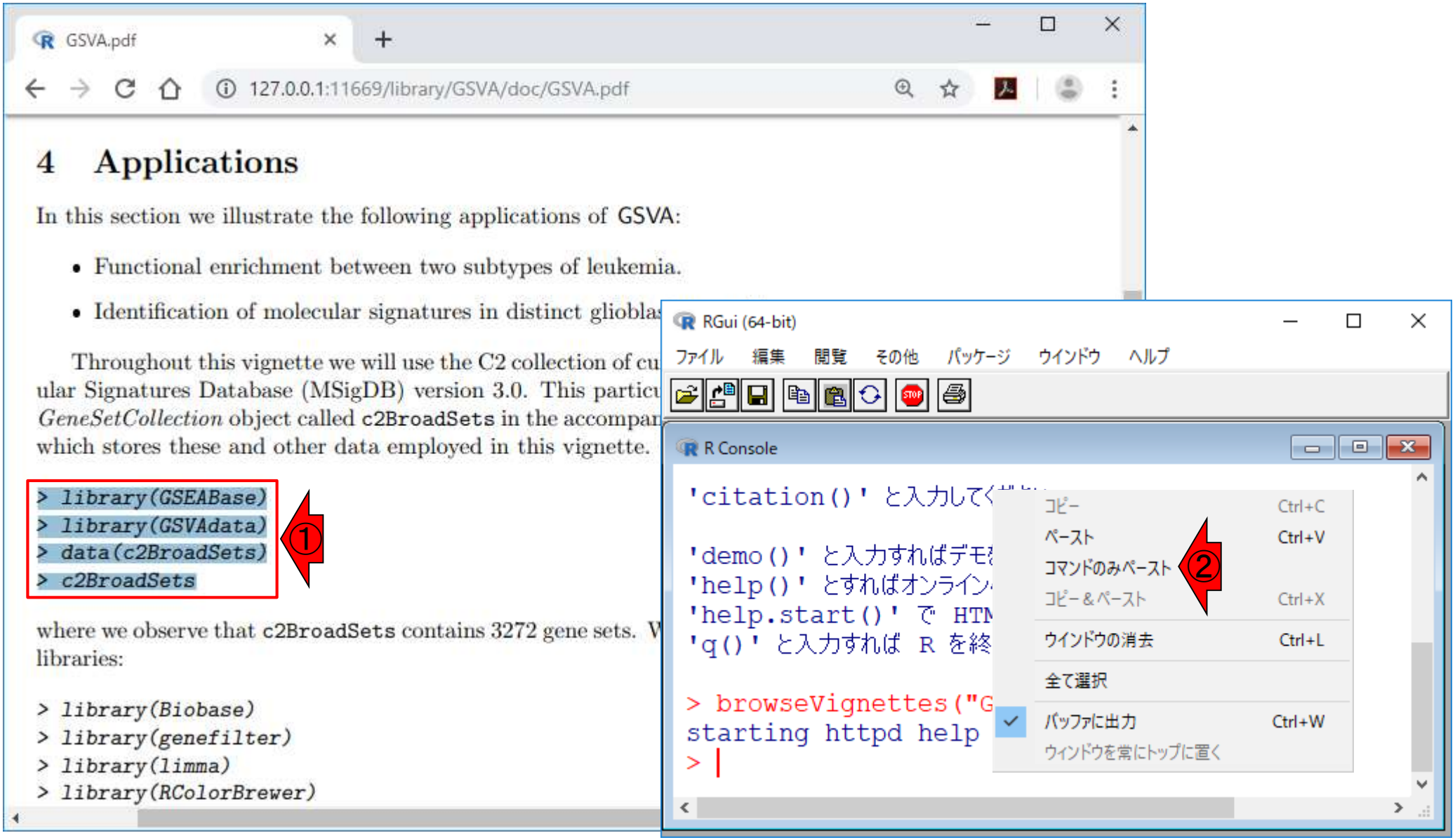
①

where we observe that `c2BroadSets` contains 3272 gene sets. We also need to load the following additional libraries:

```
> library(Biobase)
> library(genefilter)
> library(limma)
> library(RColorBrewer)
```

c2BroadSetsの確認

①こんな感じでコピーして、
②コマンドのみペースト



c2BroadSetsの確認

こんな感じになり、①GeneSetCollectionという形式の②c2BroadSetsを無事読み込めました

GSVA.pdf

127.0.0.1:11669/library/GSVA/doc/GSVA.pdf

4 Applications

In this section we illustrate the following applications

- Functional enrichment between two subtypes of
- Identification of molecular signatures in distinc

Throughout this vignette we will use the C2 collect
ular Signatures Database (MSigDB) version 3.0. Thi
GeneSetCollection object called `c2BroadSets` in the a
which stores these and other data employed in this v

```
> library(GSEABase)
> library(GSVAdata)
> data(c2BroadSets)
> c2BroadSets
```

where we observe that `c2BroadSets` contains 3272 gen
libraries:

```
> library(Biobase)
> library(genefilter)
> library(limma)
> library(RColorBrewer)
```

RGui (64-bit)

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

R Console

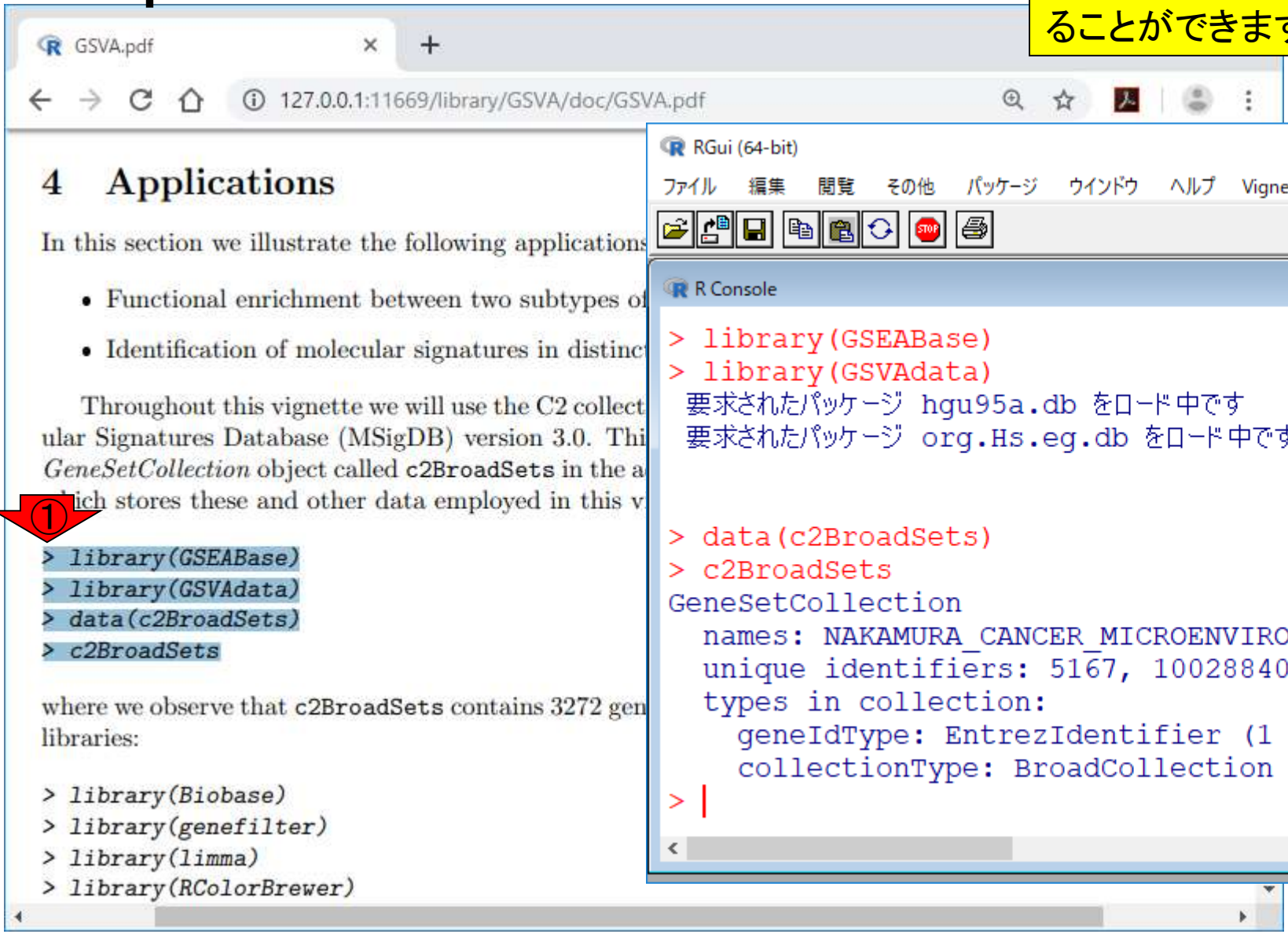
```
> library(GSEABase)
> library(GSVAdata)
 要求されたパッケージ hgu95a.db をロード中です
 要求されたパッケージ org.Hs.eg.db をロード中です


> data(c2BroadSets)
> c2BroadSets
GeneSetCollection
  names: NAKAMURA_CANCER_MICROENVIRONMENT_UP, NAKAM$
  unique identifiers: 5167, 100288400, ..., 57191 ($
  types in collection:
    geneIdType: EntrezIdentifier (1 total)
    collectionType: BroadCollection (1 total)
>
```

①

②

Tips:コマンドのみペースト

「コマンドのみペースト」とすることで、①「> 」などの余分な文字が誤ってコマンドとして認識されないようにすることができます。

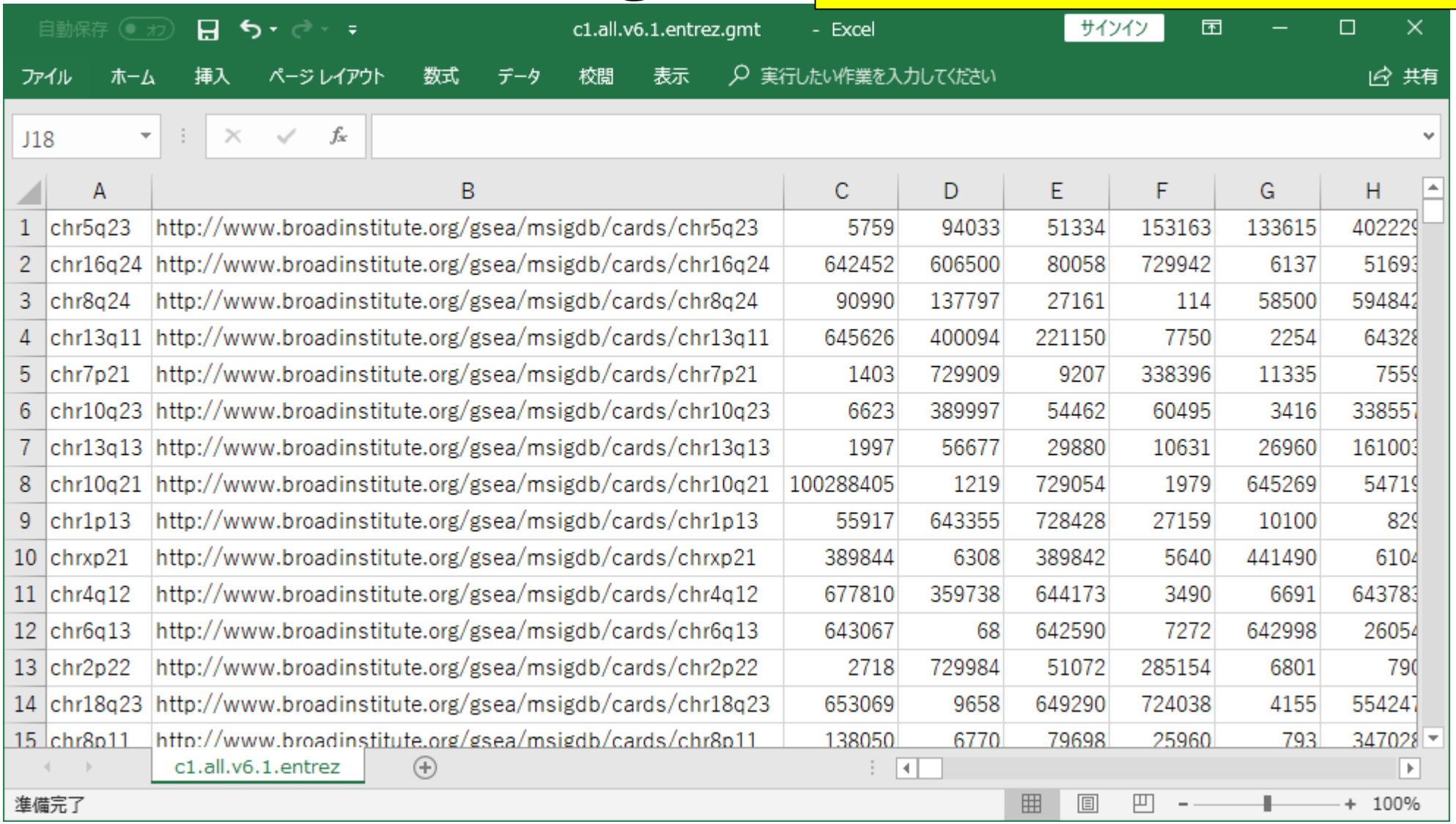


Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

c1.all.v6.1.entrez.gmt

①今手元にあるのは、c1.all.v6.1.entrez.gmt。gmtファイルを読み込ませ、遺伝子セット情報を自在に変えて解析できるようになりたい！



	A	B	C	D	E	F	G	H
1	chr5q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr5q23	5759	94033	51334	153163	133615	40222
2	chr16q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr16q24	642452	606500	80058	729942	6137	5169
3	chr8q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr8q24	90990	137797	27161	114	58500	59484
4	chr13q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q11	645626	400094	221150	7750	2254	6432
5	chr7p21	http://www.broadinstitute.org/gsea/msigdb/cards/chr7p21	1403	729909	9207	338396	11335	755
6	chr10q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q23	6623	389997	54462	60495	3416	33855
7	chr13q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q13	1997	56677	29880	10631	26960	16100
8	chr10q21	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q21	100288405	1219	729054	1979	645269	5471
9	chr1p13	http://www.broadinstitute.org/gsea/msigdb/cards/chr1p13	55917	643355	728428	27159	10100	82
10	chrxp21	http://www.broadinstitute.org/gsea/msigdb/cards/chrxp21	389844	6308	389842	5640	441490	610
11	chr4q12	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q12	677810	359738	644173	3490	6691	64378
12	chr6q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr6q13	643067	68	642590	7272	642998	2605
13	chr2p22	http://www.broadinstitute.org/gsea/msigdb/cards/chr2p22	2718	729984	51072	285154	6801	79
14	chr18q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr18q23	653069	9658	649290	724038	4155	55424
15	chr8p11	http://www.broadinstitute.org/gsea/msigdb/cards/chr8p11	138050	6770	79698	25960	793	34702

?c2BroadSets

①?c2BroadSetsと打ち込んで、このデータがどのような手順で作成されたのか手がかりを探る。

GSVA.pdf

127.0.0.1:11669/library/GSVA/doc/GSVA.pdf

4 Applications

In this section we illustrate the following applications

- Functional enrichment between two subtypes of
- Identification of molecular signatures in distinct

Throughout this vignette we will use the C2 collect
ular Signatures Database (MSigDB) version 3.0. Thi
GeneSetCollection object called `c2BroadSets` in the a
which stores these and other data employed in this v

```
> library(GSEABase)
> library(GSVAdata)
> data(c2BroadSets)
> c2BroadSets
```

where we observe that `c2BroadSets` contains 3272 gen
libraries:

```
> library(Biobase)
> library(genefilter)
> library(limma)
> library(RColorBrewer)
```

RGui (64-bit)

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

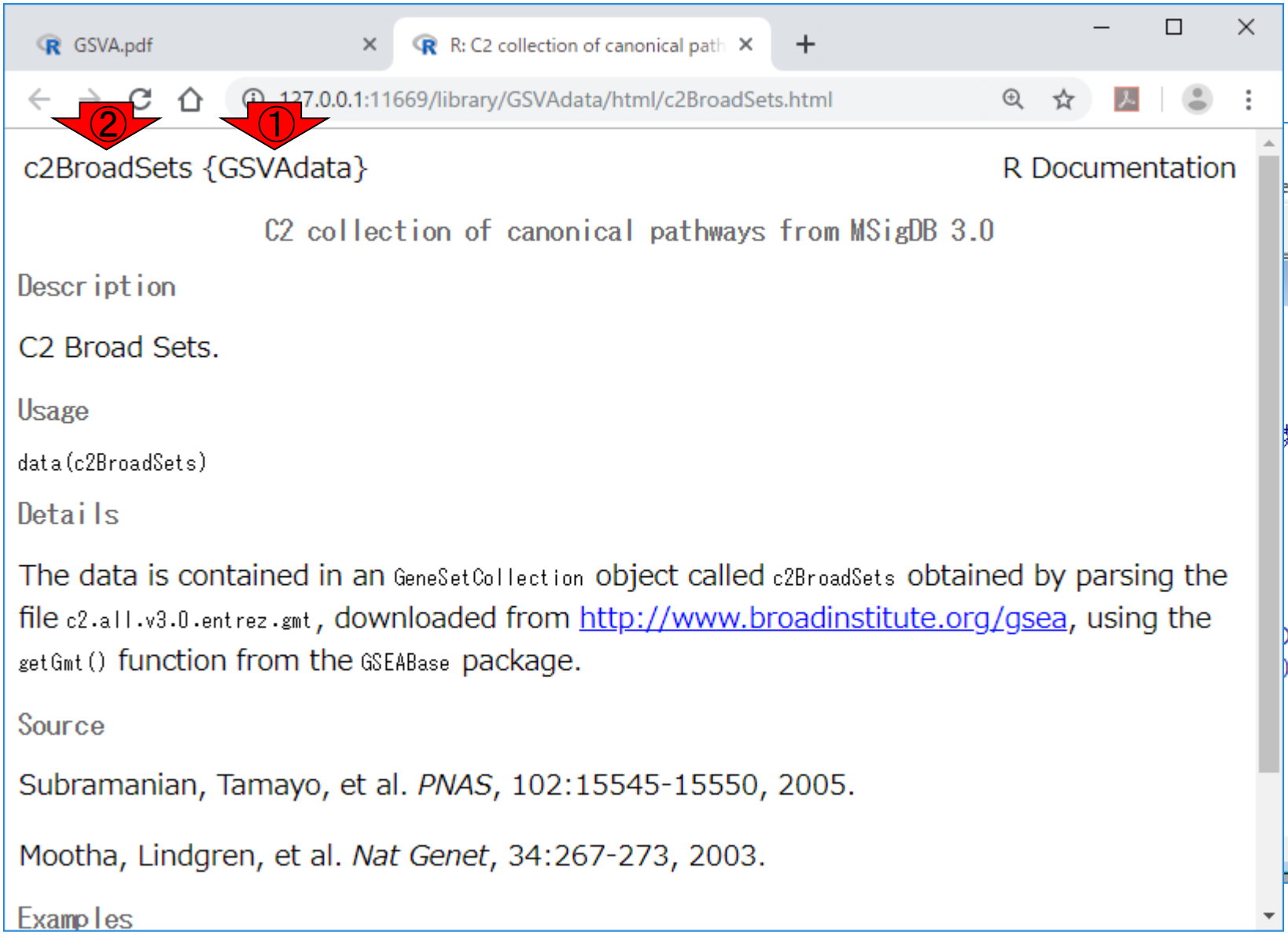
R Console

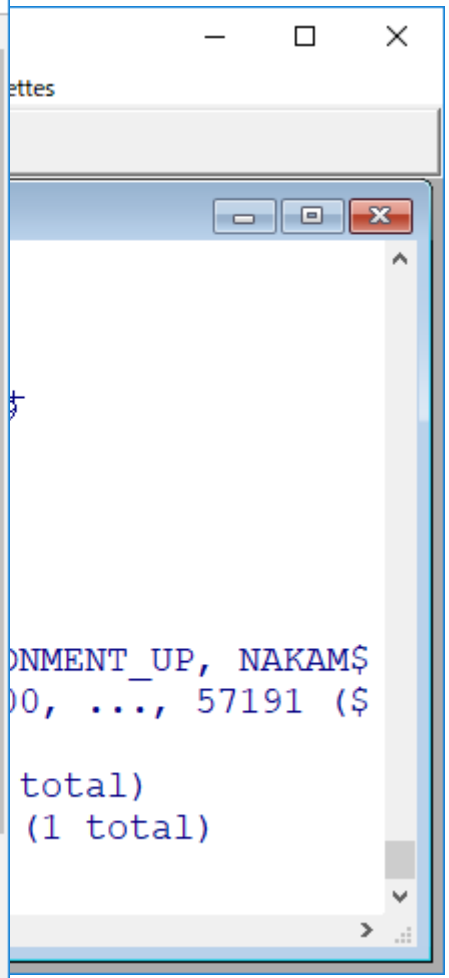
```
> library(GSEABase)
> library(GSVAdata)
 要求されたパッケージ hgu95a.db をロード中です
 要求されたパッケージ org.Hs.eg.db をロード中です


> data(c2BroadSets)
> c2BroadSets
GeneSetCollection
  names: NAKAMURA_CANCER_MICROENVIRONMENT_UP, NAKAM$
  unique identifiers: 5167, 100288400, ..., 57191 ($
  types in collection:
    geneIdType: EntrezIdentifier (1 total)
    collectionType: BroadCollection (1 total)
> ?c2BroadSets
```

①

?c2BroadSets

こんな感じになります。①GSVAdataパッケージ中の、②c2BroadSetsの説明のページという意味。





?c2BroadSets

①c2BroadSetsという名前の、②GeneSetCollection形式のオブジェクトは、③GSEABaseパッケージ内の、④getGmt関数を用いて、⑤c2.all.v3.0.entrez.gmtファイルを読み込んで得られたものだということが分かる。

GSVA.pdf | R: C2 collection of canonical pa

127.0.0.1:11669/library/GSVAdata/html/c2BroadSets.html

c2BroadSets {GSVAdata} R Documentation

C2 collection of canonical pathways from MSigDB 3.0

Description

C2 Broad Sets.

Usage

```
data(c2BroadSets)
```

Details

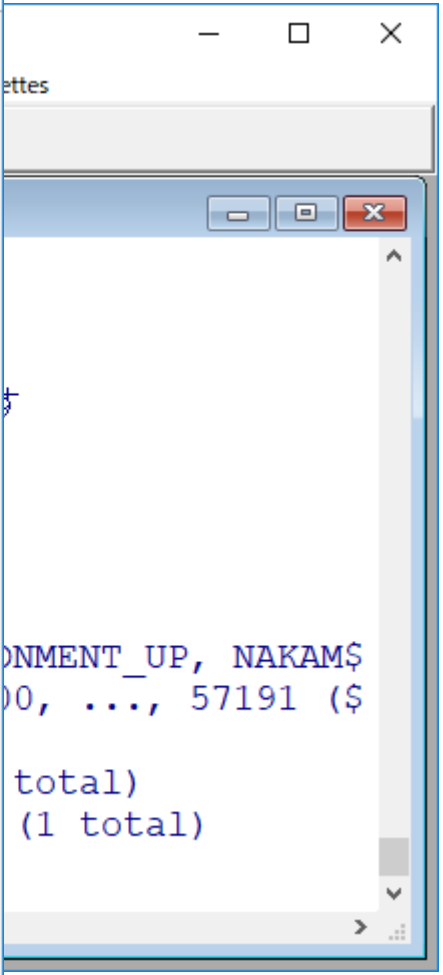
The data is contained in an `GeneSetCollection` object called `c2BroadSets` obtained by parsing the file `c2.all.v3.0.entrez.gmt`, downloaded from http://www.broadinstitute.org/gsea, using the `getGmt()` function from the `GSEABase` package.

Source

Subramanian, Tamayo, et al. *PNAS*, 102:15545-15550, 2005.

Mootha, Lindgren, et al. *Nat Genet*, 34:267-273, 2003.

Examples



gmtファイルの読込

②例題1の、gmtファイル読込の基本形を実行してみましょう。

(Rで)塩基配列解析

(last mod

このウェ
います。

What's

- 「解析

- 解析 | 機能解析 | について (last modified 2018/06/24) NEW
- 解析 | 機能解析 | GMTファイル取得 | について (last modified 2018/06/27) NEW
- 解析 | 機能解析 | GMTファイル取得 | EGSEAdata(Alhamdoosh_2017) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | GeneSetDB(Araki_2012) (last modified 2018/06/27) NEW
- 解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian_2005) (last modified 2018/06/25) NE
- 解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan_2018) (last modified 2018/06/25) NEW
- 解析 | 機能解析 | 遺伝子セット 解析 | GSVA(Hänzelmann_2013)(last modified 2018/06/26) NEW
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について (last modified 2018/06/29) NEW
- 解析 | 機能解析
- 解析 | 機能解析
- 解析 | 機能解析
- 解析 | 機能解析

①

解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan_2018) NEW

GSEABaseを用いて.gmtファイルを読み込むやり方を示します。GeneSetCollectionという形式で情報が格納されています。
「ファイル」ー「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。

②

1. MSigDBから得られた326 gene setsからなるc

基本形です。最後のgenesetオブジェクトの確認
NullCollectionになっているのが分かります。

```
in_f <- "c1.all.v6.1.symbols.gmt"

#必要なパッケージをロード
library(GSEABase)

#入力ファイルの読み込み
geneset <- getGmt(in_f)
geneset
```

RGui (64-bit)

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

R Console

```
> getwd()
[1] "C:/Users/kojik/Desktop"
> list.files()
[1] "c1.all.v6.1.entrez.gmt"
[2] "c1.all.v6.1.symbols.gmt"
[3] "desktop.ini"
[4] "share"
>
```

例題1：基本形

コピペ実行結果。①無事GeneSetCollection形式のgenesetオブジェクトが得られていることがわかります。②確かにC1コレクションは326遺伝子セットでした。③1つめの遺伝子セットがchr5q23で、④2つめがchr16q24となっています。

解析 | 機能解析 | GMTファイル読込 | GSEA

GSEABaseを用いて.gmtファイルを読み込むやり方を示します。GeneSetCollectionという形式で情報が格納されています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. MSigDBから得られた326 gene setsからなるc1.all.v6.1.symb

基本形です。最後のgenesetオブジェクトの確認で、geneIdTyp
NullCollectionになっているのが分かります。

```
in_f <- "c1.all.v6.1.symbols.gmt"        #入力フ

#必要なパッケージをロード
library(GSEABase)                        #パッケ

#入力ファイルの読み込み
geneset <- getGmt(in_f)                  #in_fで
geneset                                  #確認し
```

RGui (64-bit)

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

R Console

```
[4] "share"
> in_f <- "c1.all.v6.1.symbols.gmt"        #入力ファイ$
>
> #必要なパッケージをロード
> library(GSEABase)                        #パッケージ$
>
> #入力ファイルの読み込み
> geneset <- getGmt(in_f)                  #in_fで指定$
> geneset                                  #確認してる$
GeneSetCollection
  names: chr5q23, chr16q24, ..., chr2p14 (326 total)
  unique id[illegible]tifiers: [illegible]TMAP2, FTMT, ..., GCA (30010 t[illegible]
  types in collection:
    geneIdType: NullIdentifier (1 total)
    collectionType: NullCollection (1 total)
>
```

①

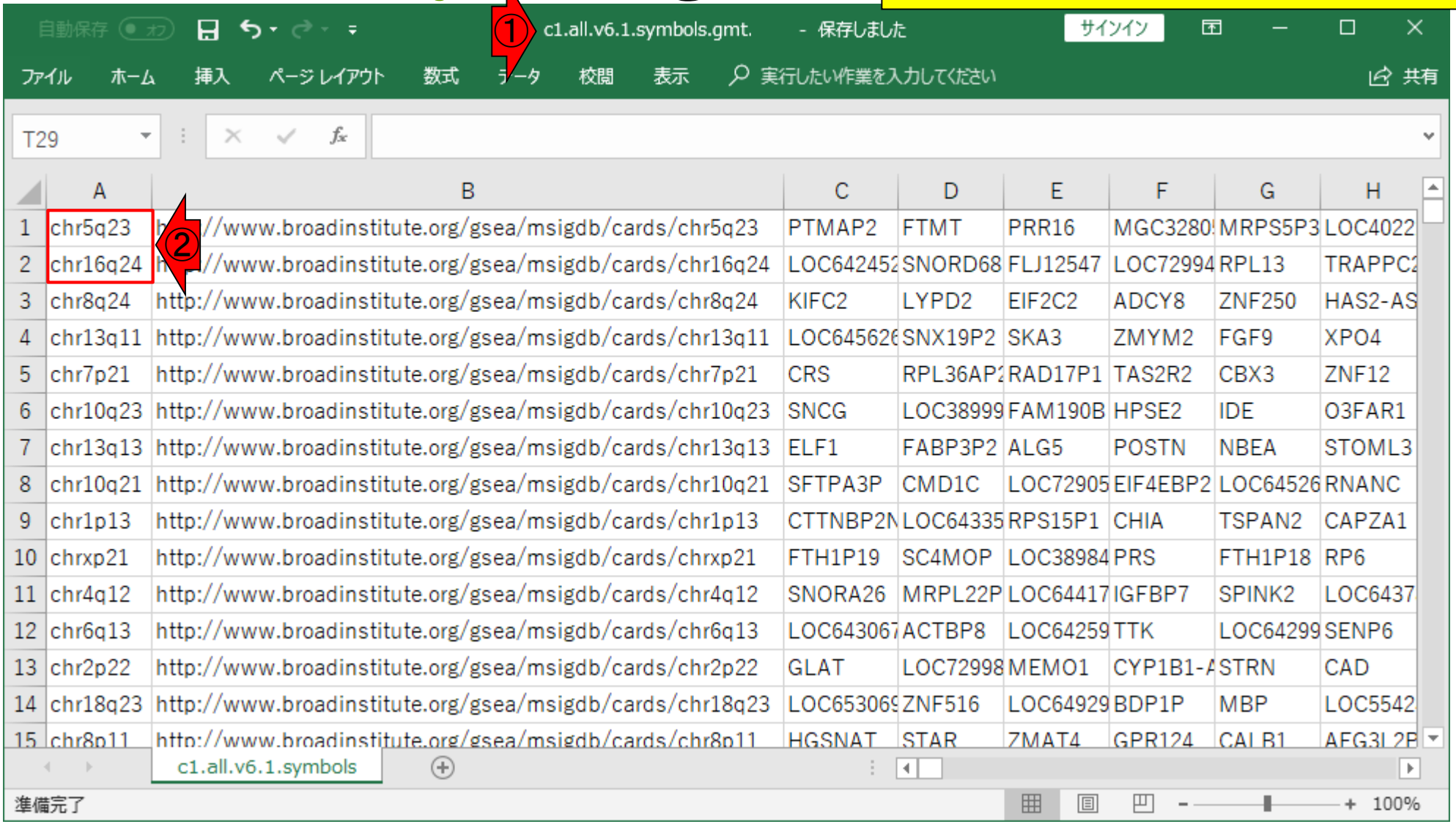
②

③

④

c1.all.v6.1.symbols.gmt

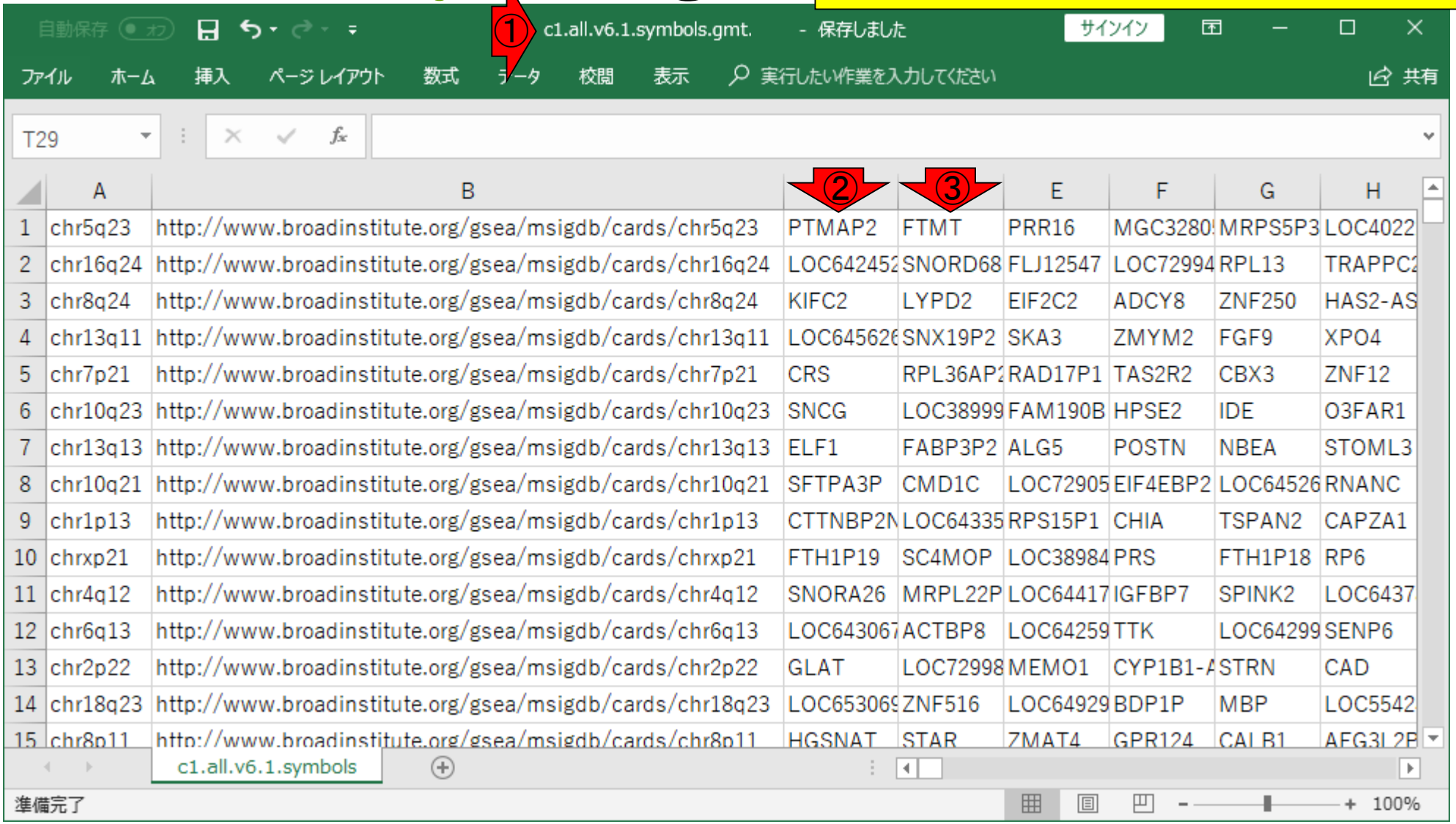
①入力ファイル(c1.all.v6.1.symbols.gmt)をExcelで眺めたところ。②最初の2行の遺伝子セット名と同じであり、妥当ですね。



	A	B	C	D	E	F	G	H
1	chr5q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr5q23	PTMAP2	FTMT	PRR16	MGC3280	MRPS5P3	LOC4022
2	chr16q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr16q24	LOC642452	SNORD68	FLJ12547	LOC72994	RPL13	TRAPPC2
3	chr8q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr8q24	KIFC2	LYPD2	EIF2C2	ADCY8	ZNF250	HAS2-AS
4	chr13q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q11	LOC645626	SNX19P2	SKA3	ZMYM2	FGF9	XPO4
5	chr7p21	http://www.broadinstitute.org/gsea/msigdb/cards/chr7p21	CRS	RPL36AP2	RAD17P1	TAS2R2	CBX3	ZNF12
6	chr10q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q23	SNCG	LOC38999	FAM190B	HPSE2	IDE	O3FAR1
7	chr13q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q13	ELF1	FABP3P2	ALG5	POSTN	NBEA	STOML3
8	chr10q21	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q21	SFTPA3P	CMD1C	LOC72905	EIF4EBP2	LOC64526	RNANC
9	chr1p13	http://www.broadinstitute.org/gsea/msigdb/cards/chr1p13	CTTNBP2N	LOC64335	RPS15P1	CHIA	TSPAN2	CAPZA1
10	chrxp21	http://www.broadinstitute.org/gsea/msigdb/cards/chrxp21	FTH1P19	SC4MOP	LOC38984	PRS	FTH1P18	RP6
11	chr4q12	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q12	SNORA26	MRPL22P	LOC64417	IGFBP7	SPINK2	LOC6437
12	chr6q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr6q13	LOC643067	ACTBP8	LOC64259	TTK	LOC64299	SENP6
13	chr2p22	http://www.broadinstitute.org/gsea/msigdb/cards/chr2p22	GLAT	LOC72998	MEMO1	CYP1B1-A	STRN	CAD
14	chr18q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr18q23	LOC653069	ZNF516	LOC64929	BDP1P	MBP	LOC5542
15	chr8p11	http://www.broadinstitute.org/gsea/msigdb/cards/chr8p11	HGSNAT	STAR	ZMAT4	GPR124	CALB1	AEG3L2P

c1.all.v6.1.symbols.gmt

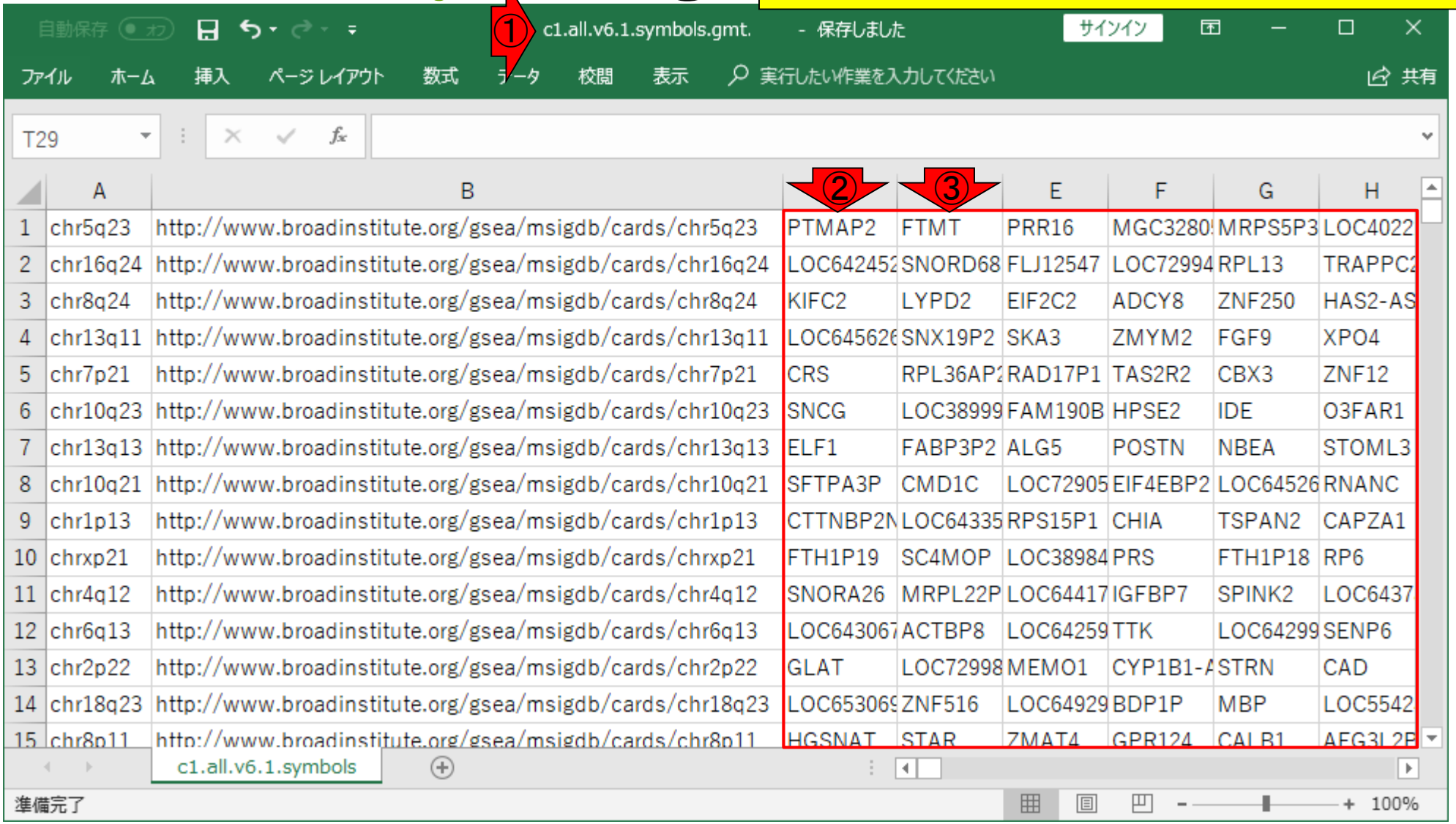
①入力がc1.all.v6.1.symbols.gmtなので、②PTMAP2や③FTMTのようなgene symbolsで遺伝子セット情報が記載されている。



	A	B	②	③	E	F	G	H
1	chr5q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr5q23	PTMAP2	FTMT	PRR16	MGC3280	MRPS5P3	LOC4022
2	chr16q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr16q24	LOC642452	SNORD68	FLJ12547	LOC72994	RPL13	TRAPPC2
3	chr8q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr8q24	KIFC2	LYPD2	EIF2C2	ADCY8	ZNF250	HAS2-AS
4	chr13q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q11	LOC645626	SNX19P2	SKA3	ZMYM2	FGF9	XPO4
5	chr7p21	http://www.broadinstitute.org/gsea/msigdb/cards/chr7p21	CRS	RPL36AP2	RAD17P1	TAS2R2	CBX3	ZNF12
6	chr10q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q23	SNCG	LOC38999	FAM190B	HPSE2	IDE	O3FAR1
7	chr13q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q13	ELF1	FABP3P2	ALG5	POSTN	NBEA	STOML3
8	chr10q21	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q21	SFTPA3P	CMD1C	LOC72905	EIF4EBP2	LOC64526	RNANC
9	chr1p13	http://www.broadinstitute.org/gsea/msigdb/cards/chr1p13	CTTNBP2N	LOC64335	RPS15P1	CHIA	TSPAN2	CAPZA1
10	chrxp21	http://www.broadinstitute.org/gsea/msigdb/cards/chrxp21	FTH1P19	SC4MOP	LOC38984	PRS	FTH1P18	RP6
11	chr4q12	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q12	SNORA26	MRPL22P	LOC64417	IGFBP7	SPINK2	LOC6437
12	chr6q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr6q13	LOC643067	ACTBP8	LOC64259	TTK	LOC64299	SENP6
13	chr2p22	http://www.broadinstitute.org/gsea/msigdb/cards/chr2p22	GLAT	LOC72998	MEMO1	CYP1B1-A	STRN	CAD
14	chr18q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr18q23	LOC653069	ZNF516	LOC64929	BDP1P	MBP	LOC5542
15	chr8p11	http://www.broadinstitute.org/gsea/msigdb/cards/chr8p11	HGSNAT	STAR	ZMAT4	GPR124	CALB1	AFG3L2P

c1.all.v6.1.symbols.gm

①c1.all.v6.1.symbols.gmt内にある、gene symbolsは、②PTMAP2や③FTMTを含めて全部で何種類あるのだろうか？そのあたりの情報は…



	A	B	②	③	E	F	G	H
1	chr5q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr5q23	PTMAP2	FTMT	PRR16	MGC3280	MRPS5P3	LOC4022
2	chr16q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr16q24	LOC642452	SNORD68	FLJ12547	LOC72994	RPL13	TRAPPC2
3	chr8q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr8q24	KIFC2	LYPD2	EIF2C2	ADCY8	ZNF250	HAS2-AS
4	chr13q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q11	LOC645626	SNX19P2	SKA3	ZMYM2	FGF9	XPO4
5	chr7p21	http://www.broadinstitute.org/gsea/msigdb/cards/chr7p21	CRS	RPL36AP2	RAD17P1	TAS2R2	CBX3	ZNF12
6	chr10q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q23	SNCG	LOC38999	FAM190B	HPSE2	IDE	O3FAR1
7	chr13q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q13	ELF1	FABP3P2	ALG5	POSTN	NBEA	STOML3
8	chr10q21	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q21	SFTPA3P	CMD1C	LOC72905	EIF4EBP2	LOC64526	RNANC
9	chr1p13	http://www.broadinstitute.org/gsea/msigdb/cards/chr1p13	CTTNBP2N	LOC64335	RPS15P1	CHIA	TSPAN2	CAPZA1
10	chrxp21	http://www.broadinstitute.org/gsea/msigdb/cards/chrxp21	FTH1P19	SC4MOP	LOC38984	PRS	FTH1P18	RP6
11	chr4q12	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q12	SNORA26	MRPL22P	LOC64417	IGFBP7	SPINK2	LOC6437
12	chr6q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr6q13	LOC643067	ACTBP8	LOC64259	TTK	LOC64299	SENP6
13	chr2p22	http://www.broadinstitute.org/gsea/msigdb/cards/chr2p22	GLAT	LOC72998	MEMO1	CYP1B1-A	STRN	CAD
14	chr18q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr18q23	LOC653069	ZNF516	LOC64929	BDP1P	MBP	LOC5542
15	chr8p11	http://www.broadinstitute.org/gsea/msigdb/cards/chr8p11	HGSNAT	STAR	ZMAT4	GPR124	CALB1	AEG3L2P

例題1：基本形

①c1.all.v6.1.symbols.gmt内には、②PTMAP2や③FTMTを含めて、④全部で30,010種類あるのだろう。

解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan_2018) NEW

GSEABaseを用いて.gmtファイルを読み込むやり方を示します。GeneSetCollectionという形式で情報が格納されています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. MSigDBから得られた326 gene setsからなるc1.all.v6.1.symb

基本形です。最後のgenesetオブジェクトの確認で、geneIdTyp
NullCollectionになっているのが分かります。

```
in_f <- "c1.all.v6.1.symbols.gmt"  ①  #入力フ

#必要なパッケージをロード
library(GSEABase)                      #パッケ

#入力ファイルの読み込み
geneset <- getGmt(in_f)                #in_fで
geneset                                #確認し
```

RGui (64-bit)

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

R Console

```
[4] "share"
> in_f <- "c1.all.v6.1.symbols.gmt"  ①   #入力ファイ$
>
> #必要なパッケージをロード
> library(GSEABase)                      #パッケージ$
>
> #入力ファイルの読み込み
> geneset <- getGmt(in_f)                #in_fで指定$
> geneset                                #確認してる$
GeneSetCollection
  names: chr5q23, chr16q[illegible], ..., chr2p14 (326 t[illegible])
  unique identifiers: PTMAP2, FTMT, ..., GCA (30010 t$
  types in collection:
    geneIdType: NullIdentifier (1 total)
    collectionType: NullCollection (1 total)
>
```

②　③　④

まだ不十分か?!

①赤下線部分に着目！この部分がNullIdentifierやNullCollectionとなっている。GSVAdataパッケージの②c2BroadSetsではそうなっていなかった。

解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan_2018) NEW

GSEABaseを用いて.gmtファイルを読み込むやり方を示します。GeneSetCollectionという形式で情報が格納されています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. MSigDBから得られた326 gene setsからなるc1.all.v6.1.symb

基本形です。最後のgenesetオブジェクトの確認で、geneIdTyp
NullCollectionになっているのが分かります。

```
in_f <- "c1.all.v6.1.symbols.gmt"        #入力フ

#必要なパッケージをロード
library(GSEABase)                        #パッケ

#入力ファイルの読み込み
geneset <- getGmt(in_f)                  #in_fで
geneset                                  #確認し
```

RGui (64-bit)

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

R Console

```
[4] "share"
> in_f <- "c1.all.v6.1.symbols.gmt"        #入力ファイ$
>
> #必要なパッケージをロード
> library(GSEABase)                        #パッケージ$
>
> #入力ファイルの読み込み
> geneset <- getGmt(in_f)                  #in_fで指定$
> geneset                                  #確認してる$
GeneSetCollection
  names: chr5q23, chr16q24, ..., chr2p14 (326 total)
  unique identifiers: PTMAP2, FTMT, ..., GCA (30010 t$
  types in collection:
    geneIdType: NullIdentifier (1 total)
    collectionType: NullCollection (1 total)
> c2BroadSets
```

②

①

両者を比較

①gmtファイルから読み込んだgenesetでは、②NullIdentifierやNullCollectionとなっている。その一方で、GSVAdataパッケージの③c2BroadSetsでは、④EntrezIdentifierやBroadCollectionとなっている。ここまでやっておく必要性については今のところ不明ではあるが、念のためやったのが例題3。

解析 | 機能解析 | GMTファイル読込 | GS

GSEABaseを用いて.gmtファイルを読み込むやり方を示します。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. MSigDBから得られた326 gene setsからなるc1.all.v6.1.symb

基本形です。最後のgenesetオブジェクトの確認で、geneIdTyp
NullCollectionになっているのが分かります。

```
in_f <- "c1.all.v6.1.symbols.gmt"        #入力フ

#必要なパッケージをロード
library(GSEABase)                        #パッケ

#入力ファイルの読み込み
geneset <- getGmt(in_f)                  #in_f
geneset                                  #確認し
```

RGui (64-bit)

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

R Console

```
> geneset <- getGmt(in_f)                #in_fで指定$
> geneset                                #確認してる$
GeneSetCollection
  names: chr5q23, chr16q24, ..., chr2p14 (326 total)
  unique identifiers: PTMAP2, FTMT, ..., GCA (30010 t$
  types in collection:
    geneIdType: NullIdentifier (1 total)
    collectionType: NullCollection (1 total)
> c2BroadSets
GeneSetCollection
  names: NAKAMURA_CANCER_MICROENVIRONMENT_UP, NAKAMUR$
  unique identifiers: 5167, 100288400, ..., 57191 (29$
  types in collection:
    geneIdType: EntrezIdentifier (1 total)
    collectionType: BroadCollection (1 total)
>
```

②例題3です。

例題3

(Rで)塩基配列解析

(last mod

このウェ
います。

What's

- 「解析

- 解析 | 機能解析 | について (last modified 2018/06/24) NEW
- 解析 | 機能解析 | GMTファイル取得 | について (last modified 2018/06/27) NEW
- 解析 | 機能解析 | GMTファイル取得 | EGSEAdata(Alhamdoosh_2017) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | GeneSetDB(Araki_2012) (last modified 2018/06/27) NEW
- 解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian_2005) (last modified 2018/06/25) NE
- 解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan_2018) ① modified 2018/06/25) NEW
- 解析 | 機能解析 | 遺伝子セット 解析 | GSVA(Hänzelmann_2013)(las modified 2018/06/26) NEW
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について (last modified 2018/06/29) NEW
- 解析 | 機能解析
- 解析 | 機能解析
- 解析 | 機能解析
- 解析 | 機能解析

解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan_2018) NEW

GSEABaseを用いて.gmtファイルを読み込むやり方を示します。GeneSetCollectionという形式で情報が格納されています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. MSigDBから得られた326 gene setsからなるc1.all.v6.1.symbols.gmtの場合:

基本形です。最後のgenesetオブジェクトの確認で、geneIdTypeのところがNullIdentifierに、そしてcollectionTypeのところがNullCollectionになっているのが分かります。

```
in_f <- "c1.all.v6.1.symbols.gmt"        #入力ファイル名を指定してin_fに格納

#必要
library(

#入力ファ
geneset
geneset
```

② **3. MSigDBから得られた326 gene setsからなるc1.all.v6.1.entrez.gmtの場合:**

例題2とは入力ファイルが異なります。このファイルはgene ID情報がEntrez gene IDsですので、その部分のみ例題2とは異なります。

```
in_f <- "c1.all.v6.1.entrez.gmt"        #入力ファイル名を指定してin_fに格納

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

#入力ファイルの読み込み
geneset <- getGmt(in_f, geneIdType=EntrezIdentifier(),#in_fで指定したファイルの読み込み
                  collectionType=BroadCollection(category="c1"))#in_fで指定したファイルの読み込
geneset                                 #確認してるだけです
```

①例題3の、②入力はc1.all.v6.1.entrez.gmt。

例題3

3. MSigDBから得られた326 gene setsからなるc1.all.v6.1.entrez.gmtの場合: ①

例題2とは入力ファイルが異なります。このファイルはgene ID情報がEntrez gene IDsですので、その部分のみ例題2とは異なります。

```
in_f <- "c1.all.v6.1.entrez.gmt"  ②    #入力ファイル名を指定してin_fに格納

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

#入力ファイルの読み込み
geneset <- getGmt(in_f, geneIdType=EntrezIdentifier(),#in_fで指定したファイルの読み込み
                  collectionType=BroadCollection(category="c1"))#in_fで指定したファイルの読み込
geneset                                  #確認してるだけです
```

例題3

①getGmt関数実行時に、②geneIdTypeと③collectionTypeオプションを与えて、Entrez gene IDであることや、Broad institute提供の④C1コレクションであることを明示しておけば…

3. MSigDBから得られた326 gene setsからなるc1.all.v6.1.entrez.gmtの場合:

例題2とは入力ファイルが異なります。このファイルはgene ID情報がEntrez gene IDsですので、その部分のみ例題2とは異なります。

④

```
in_f <- "c1.all.v6.1.entrez.gmt"        #入力ファイル名を指定してin_fに格納

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

①                    ②
#入力ファイルの読み込み
geneset <- getGmt(in_f, geneIdType=EntrezIdentifier(),#in_fで指定したファイルの読み込み
                  collectionType=BroadCollection(category="c1"))#in_fで指定したファイルの読み込
geneset                                 #確認してるだけです
          ③                                          ④
```

例題3

① c1.all.v6.1.entrez.gmtを読み込んで得られた、②genesetオブジェクトの中身が③EntrezIdentifierやBroadCollectionになります。これで見た目上は、GSVAdataパッケージのc2BroadSets同じような見栄えになりました。実際問題としてここまでやっておく必要があるかどうかはわかりません。ここまででgmtファイルを読み込んでGeneSetCollectionオブジェクトを作成するところまで完了。

3. MSigDBから得られた326

例題2とは 入力ファイルが異
なります。

```
in_f <- "c1.all.v6.1.entrez.gmt"

#必要なパッケージをロード
library(GSEABase)

#入力ファイルの読み込み
geneset <- getGmt(in_f, geneIdType=Ent
                  collectionType=Broad
geneset
```

①

RGui (64-bit)

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

R Console

```
> in_f <- "c1.all.v6.1.entrez.gmt"      #入力ファイ$
>
> #必要なパッケージをロード
> library(GSEABase)                     #パッケージ$
>
> #入力ファイルの読み込み
> geneset <- getGmt(in_f, geneIdType=EntrezIdentifier$
+                   collectionType=BroadCollection(ca$
> geneset                               #確認してる$
GeneSetCollection
  names: chr5q23, chr16q24, ..., chr2p14 (326 total)
  unique identifiers: 5759, 94033, ..., 25801 (30012 $
  types in collection:
    geneIdType: EntrezIdentifier (1 total)
    collectionType: BroadCollection (1 total)
>
```

①

②

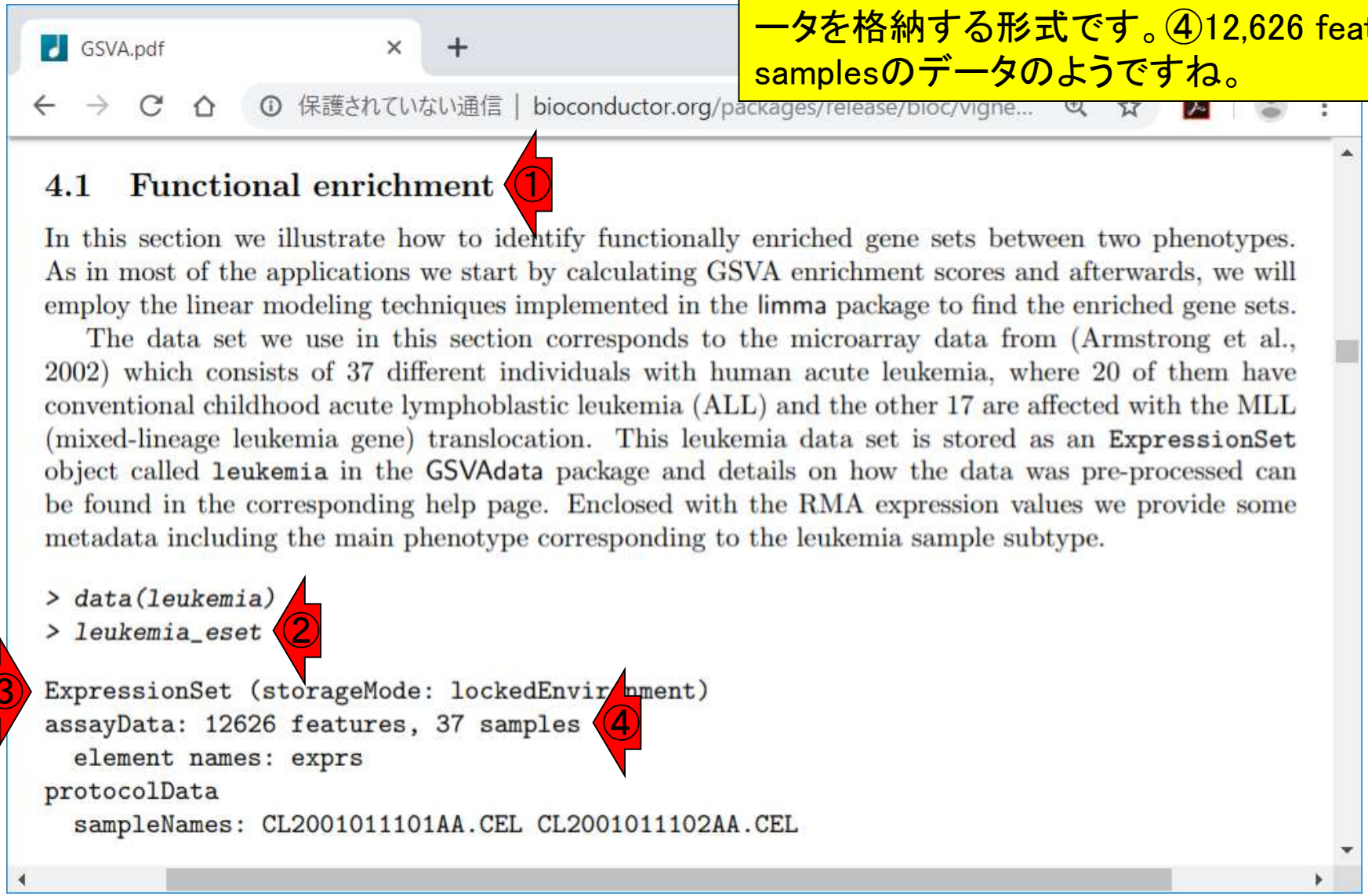
③

Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

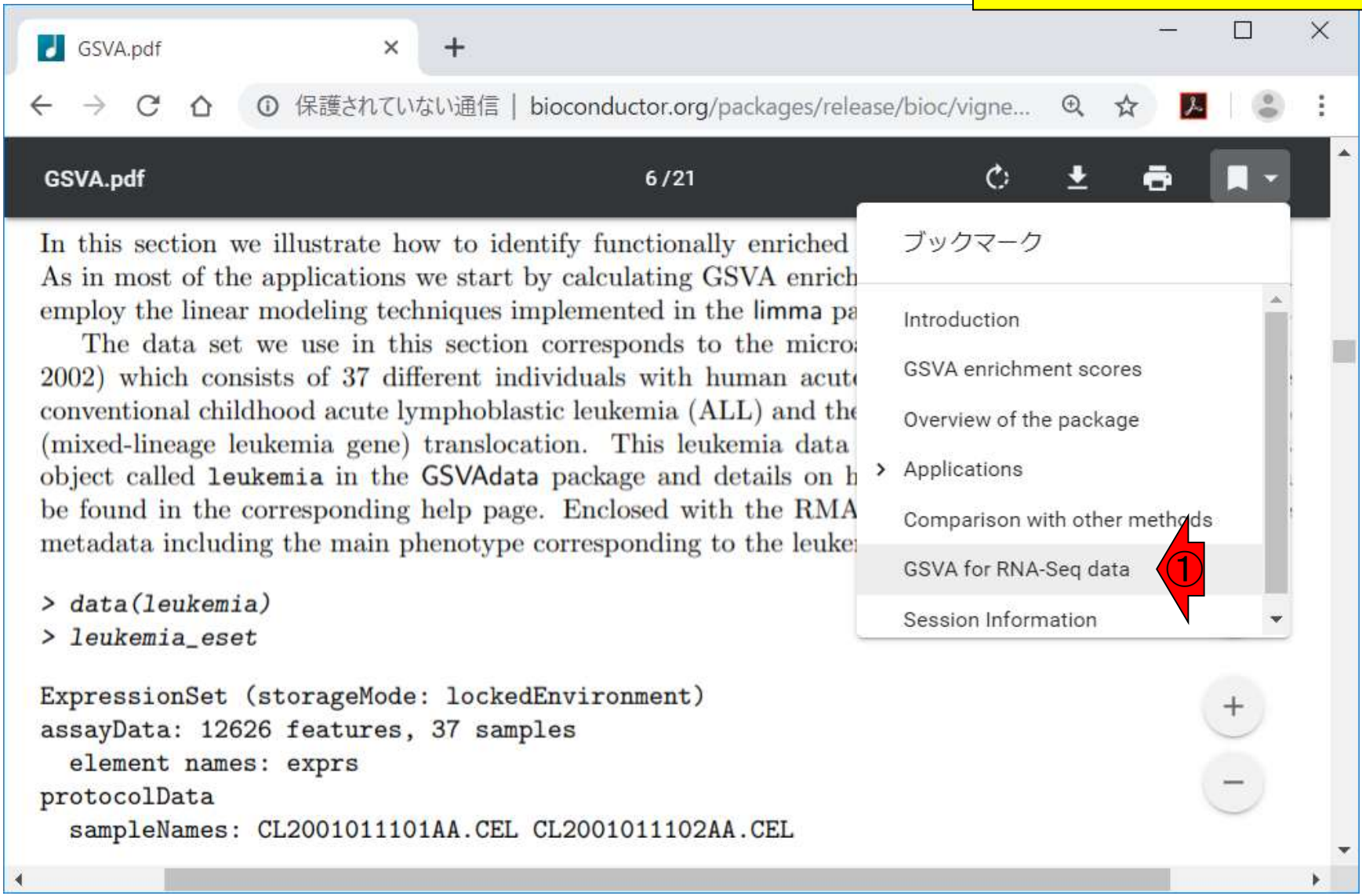
GSVAの解説PDF

①4.1 Functional enrichmentのところ。最初に発現データとして、②マイクロアレイデータのleukemia_esetを見せている。これは③ExpressionSetという発現データを格納する形式です。④12,626 features × 37 samplesのデータのようですね。



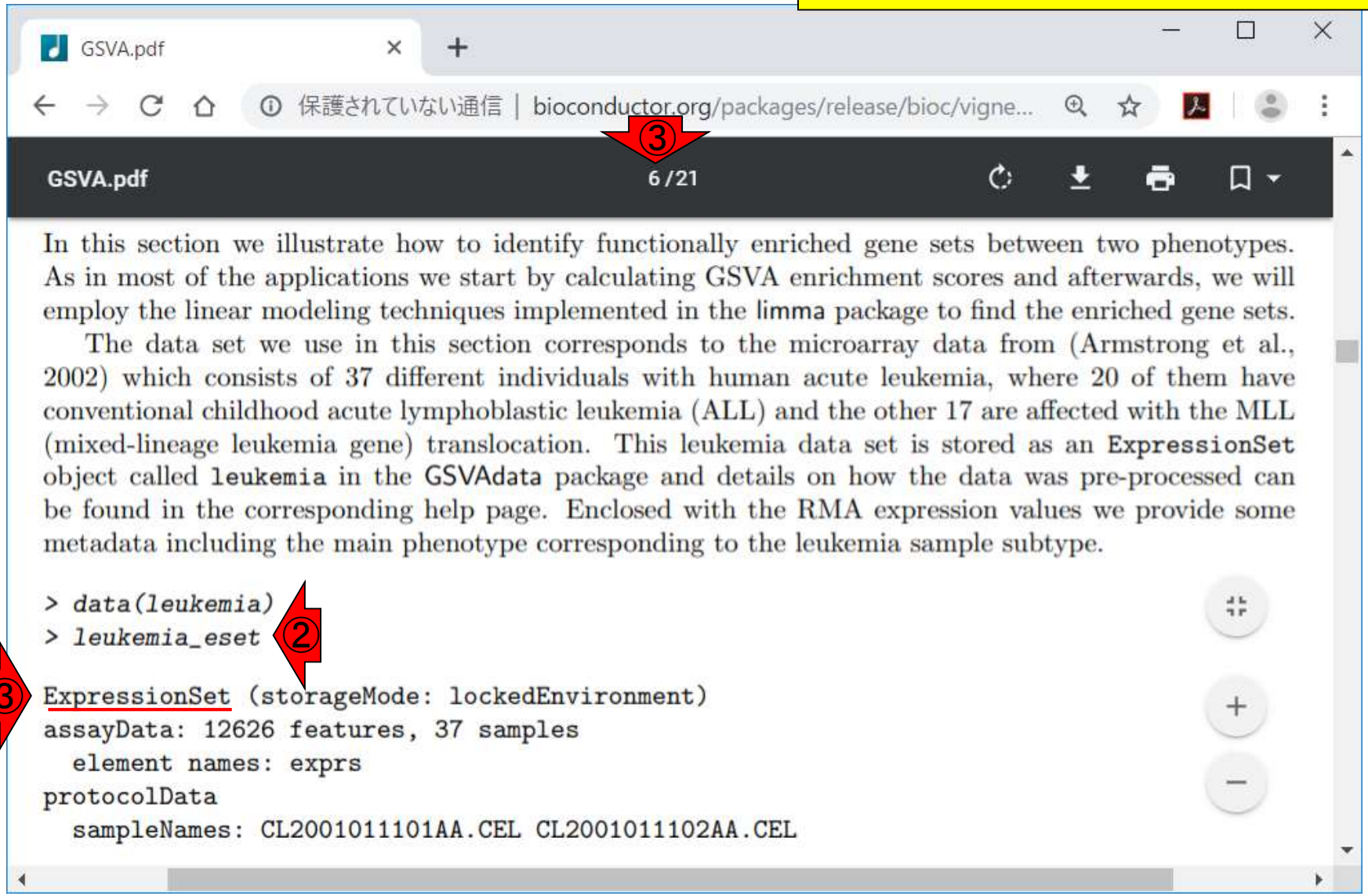
GSVAの解説PDF

GSVAパッケージでは、①RNA-seqカウントデータも、ExpressionSet形式になっています。まだクリックしない！



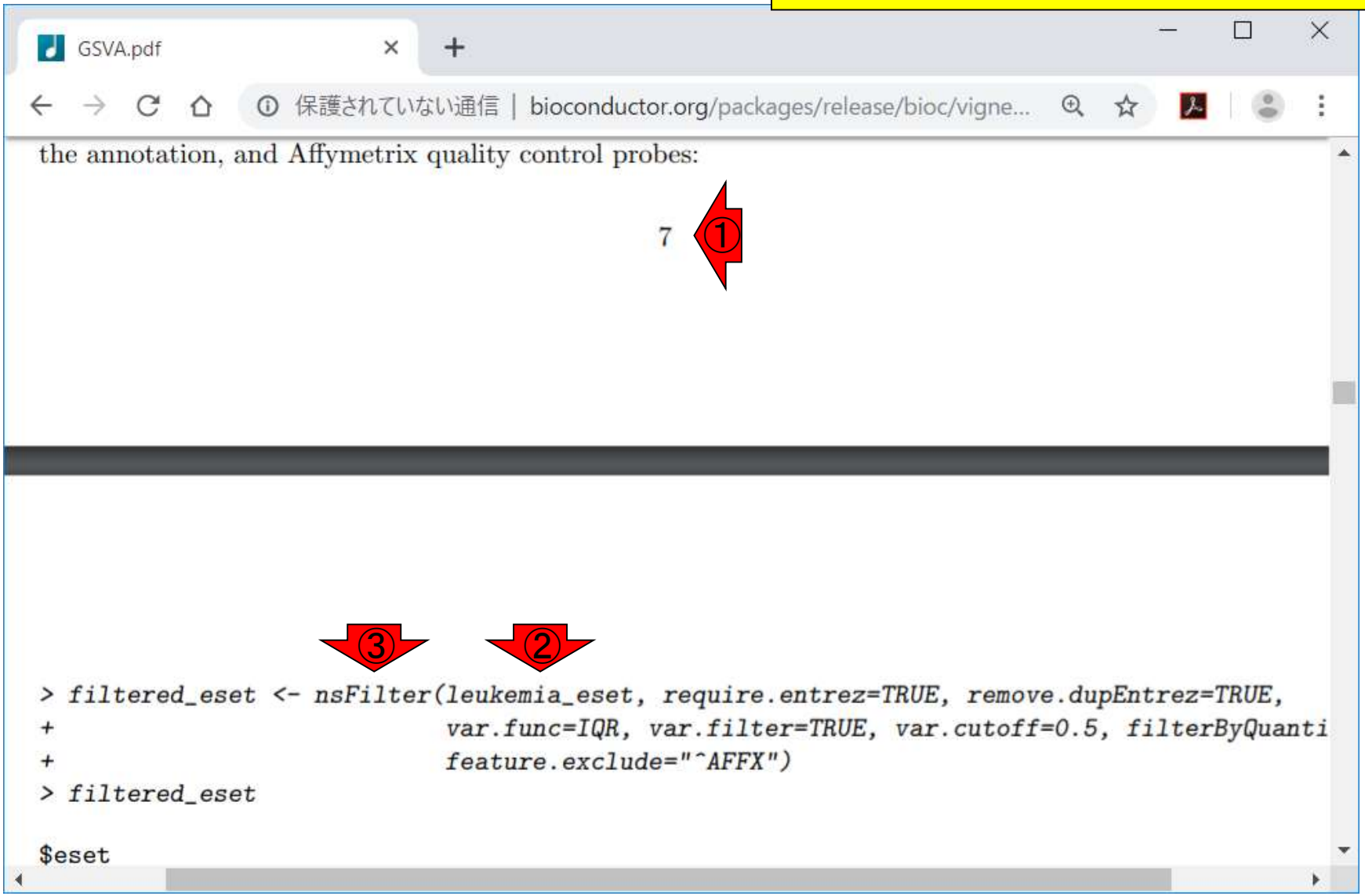
重複除去時の入力

このあと行う同一gene IDの重複除去時の入力として、①ExpressionSet形式の、②leukemia_esetが与えられています。③今は6ページのあたり。



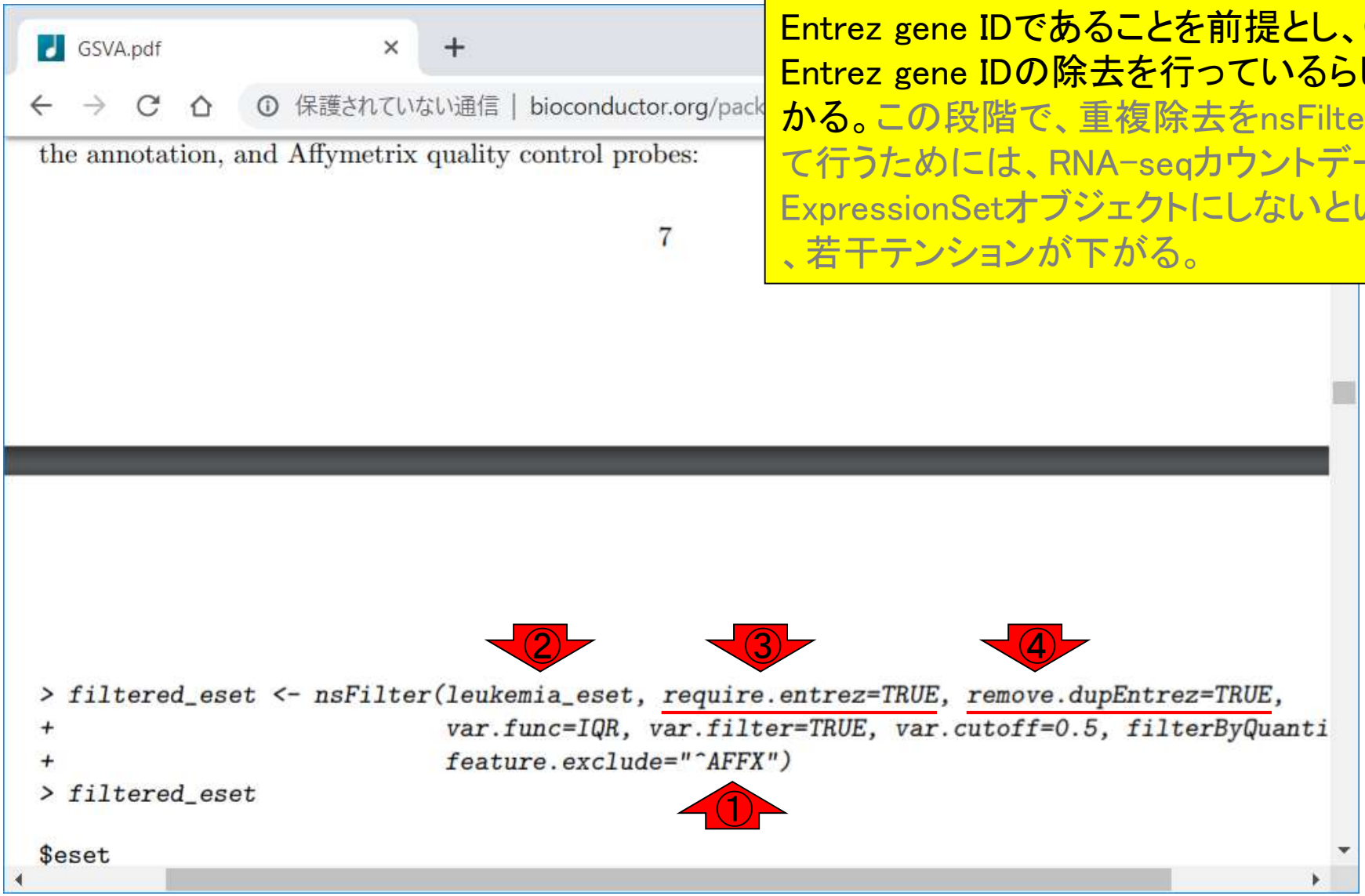
重複除去時の入力

①7～8ページにかけて、②ExpressionSetオブジェクトのleukemia_esetを入力として、③nsFilter関数を用いた重複除去が行われています。



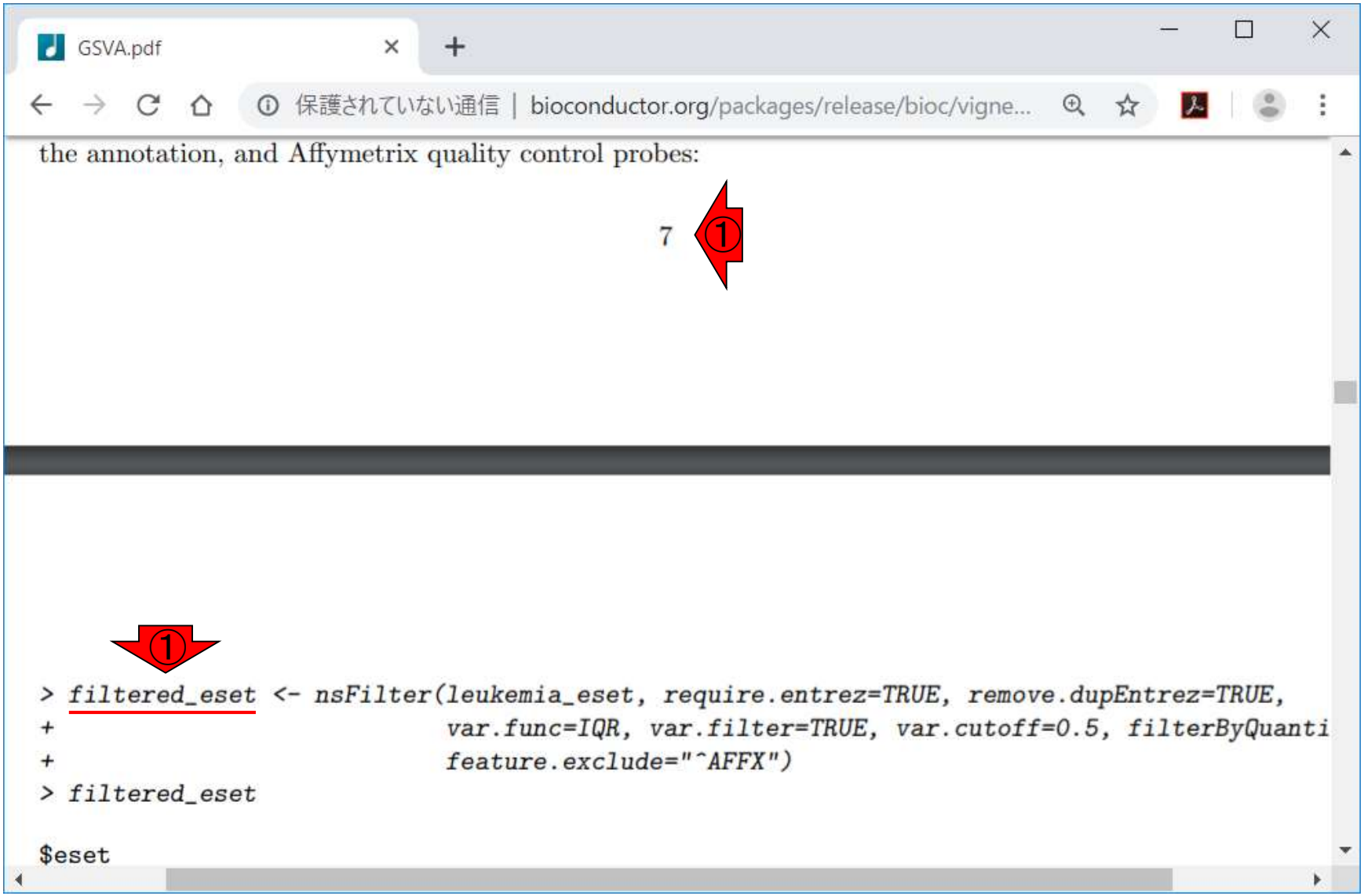
nsFilterで重複除去

マイクロアレイ時代を知るヒトは、①AFFXという文字のみで、②leukemia_esetがAffymetrix GeneChipデータであることがわかる。また、③の記述からEntrez gene IDであることを前提とし、④で重複したEntrez gene IDの除去を行っているらしいことがわかる。この段階で、重複除去をnsFilter関数を用いて行うためには、RNA-seqカウントデータの場合もExpressionSetオブジェクトにしないといけないので、若干テンションが下がる。



GSVAの入力

重複除去の実行結果は、①filtered_eset。
②8ページ目の上のほうです。



Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

GSVAの入力

重複除去の実行結果は、①filtered_eset。
8ページ目の下のほうです。

GSVA.pdf

保護されていない通信 | bioconductor.org/packages/release/bioc/vigne...

```
[1] 19
```

①

```
> leukemia_filtered_eset <- filtered_eset$eset
```

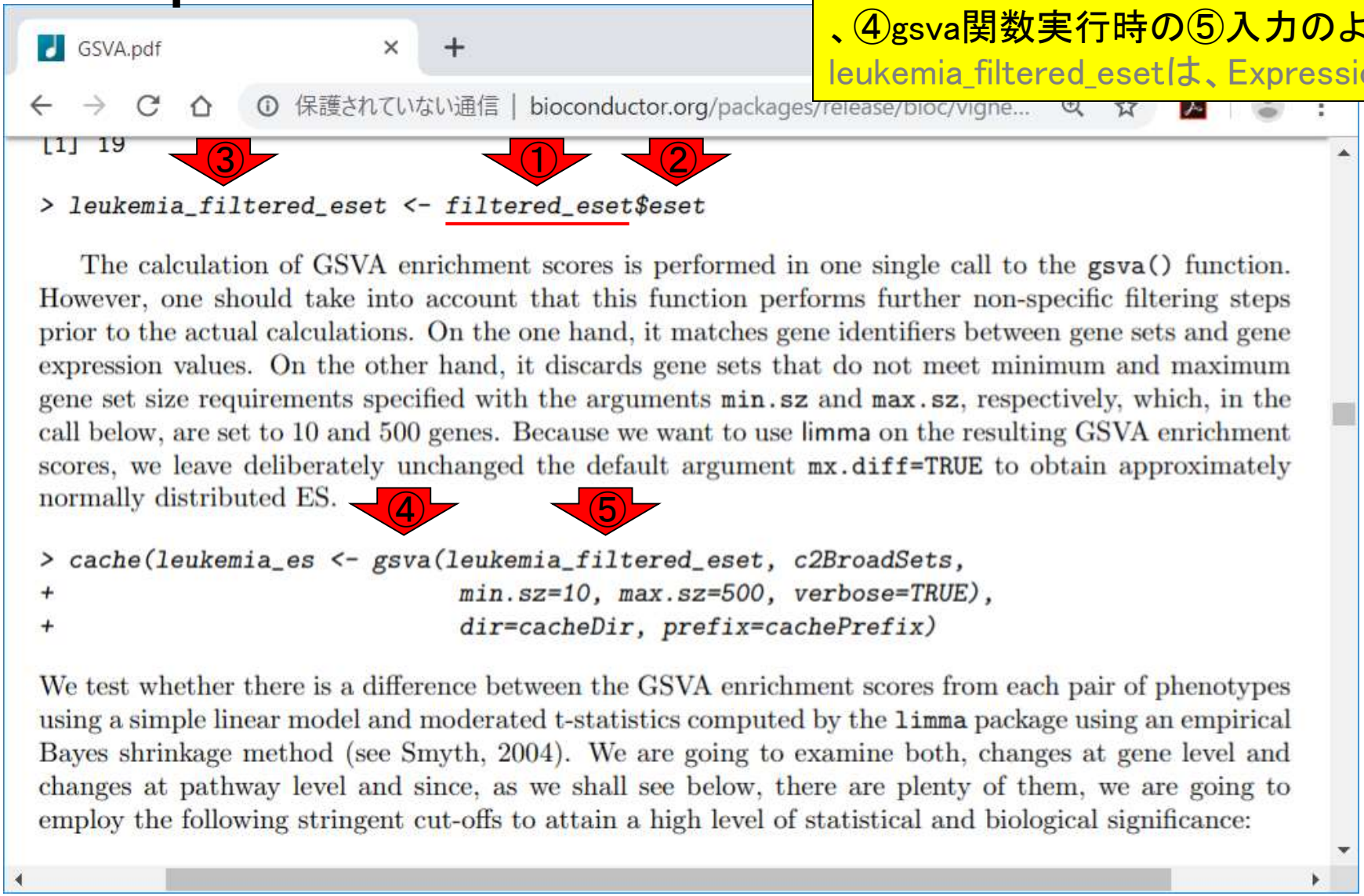
The calculation of GSVA enrichment scores is performed in one single call to the `gsva()` function. However, one should take into account that this function performs further non-specific filtering steps prior to the actual calculations. On the one hand, it matches gene identifiers between gene sets and gene expression values. On the other hand, it discards gene sets that do not meet minimum and maximum gene set size requirements specified with the arguments `min.sz` and `max.sz`, respectively, which, in the call below, are set to 10 and 500 genes. Because we want to use limma on the resulting GSVA enrichment scores, we leave deliberately unchanged the default argument `mx.diff=TRUE` to obtain approximately normally distributed ES.

```
> cache(leukemia_es <- gsva(leukemia_filtered_eset, c2BroadSets,
+                           min.sz=10, max.sz=500, verbose=TRUE),
+                           dir=cacheDir, prefix=cachePrefix)
```

We test whether there is a difference between the GSVA enrichment scores from each pair of phenotypes using a simple linear model and moderated t-statistics computed by the `limma` package using an empirical Bayes shrinkage method (see Smyth, 2004). We are going to examine both, changes at gene level and changes at pathway level and since, as we shall see below, there are plenty of them, we are going to employ the following stringent cut-offs to attain a high level of statistical and biological significance:

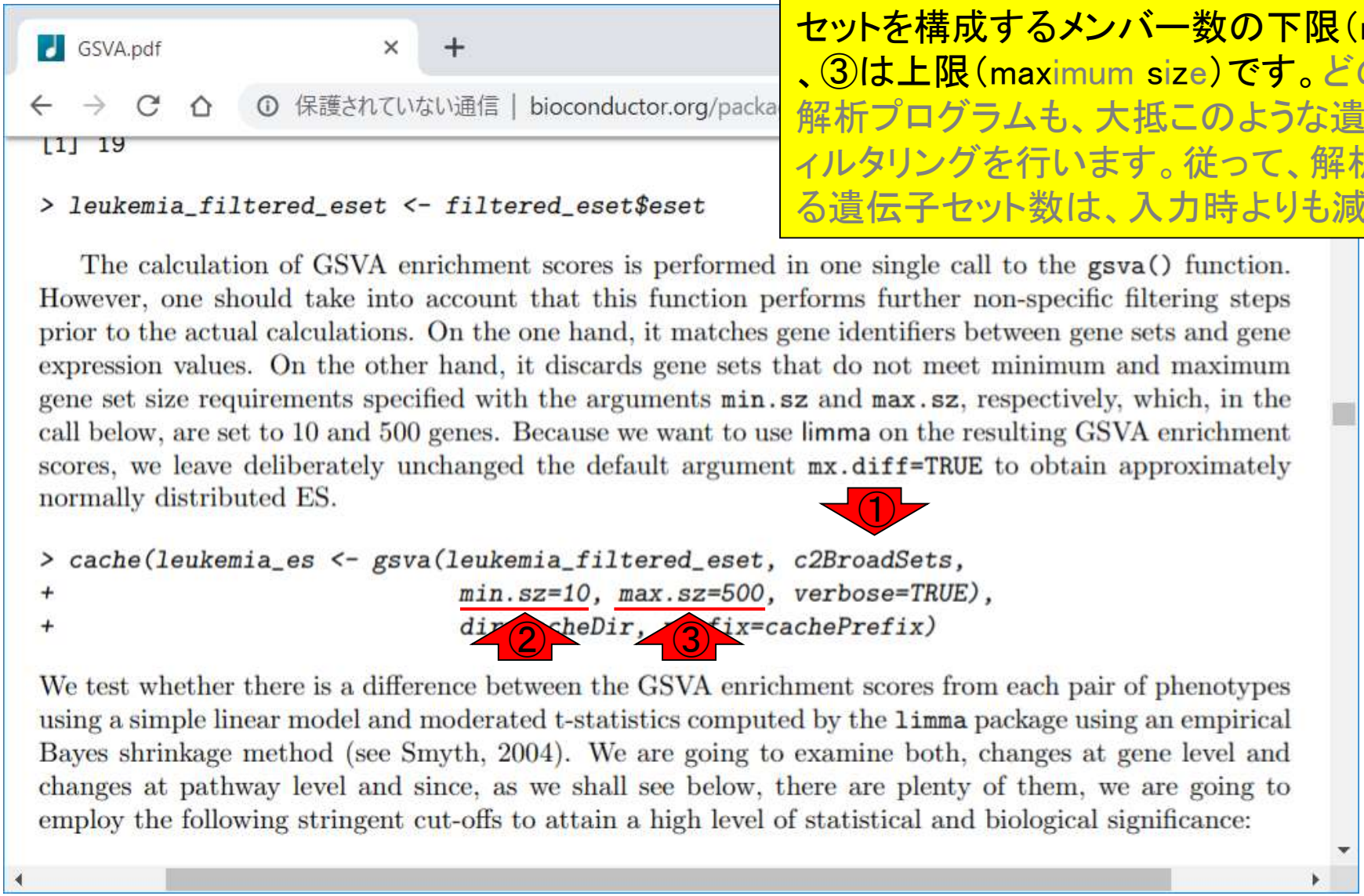
ExpressionSet形式

重複除去の実行結果は、①filtered_eset。①filtered_esetオブジェクト中の、②esetという部分の情報を抜き出した、③leukemia_filtered_esetが、④gsva関数実行時の⑤入力のようです。③leukemia_filtered_esetは、ExpressionSet形式です



GeneSetCollection

①c2BroadSetsは、GeneSetCollectionという形式の遺伝子セット情報です。②と③で解析する遺伝子セットのフィルタリングを指定しています。②は遺伝子セットを構成するメンバー数の下限（minimum size）、③は上限（maximum size）です。どの遺伝子セット解析プログラムも、大抵このような遺伝子セットのフィルタリングを行います。従って、解析結果で見られる遺伝子セット数は、入力時よりも減るのが普通。



GSVAの入力形式

①gsvaの入力が、②ExpressionSet、および③GeneSetCollectionという形式に限定されているかを、④?gsvaで確認。

GSVA.pdf

保護されていない通信 | bioconductor.org/packages/release/bioc/vigne...

```
[1] 19

> leukemia_filtered_eset <- filtered_eset$eset
```

The calculation of GSVA enrichment scores is performed in one single call to the `gsva()` function. However, one should take into account that this function performs further non-specific filtering steps prior to the actual calculations. On the one hand, it matches gene identifiers between gene sets and gene expression values. On the other hand, it discards gene sets that do not meet minimum and maximum gene set size requirements specified with the arguments `min.sz` and `max.sz`, respectively, which, in the call below, are set to 10 and 500 genes. Because we want to use limma on the resulting GSVA enrichment scores, we leave deliberately unchanged the default argument `mx.diff=TRUE` to obtain approximately normally distributed ES.

①　②　③

```
> cache(leukemia_es <- gsva(leukemia_filtered_eset, c2BroadSets,
+                           min.sz=10, max.sz=500, verbose=TRUE),
+                           dir=cacheDir, prefix=cachePrefix)
```

We test whether there is a difference between the GSVA enrichment scores from e
using a simple linear model and moderated t-statistics computed by the `limma` pa
Bayes shrinkage method (see Smyth, 2004). We are going to examine both, ch
changes at pathway level and since, as we shall see below, there are plenty of t
employ the following stringent cut-offs to attain a high level of statistical and biol

④

R Console

```
> library(GSVA)
> ?gsva
```

?gsva

こんな感じになります。①GSVAパッケージ中の、②gsva関数の説明のページという意味。

?gsva

何を書いてるのか（S4 methodって何よ?とか…）分かりづらいだろうが、①と②の比較から…

http://127.0.0.1:16111/library/GSVA/html/gsva.html

検索...

R: Gene Set Variation Analysis

```
## S4 method for signature 'ExpressionSet,list'
gsva(expr, gset.idx.list, annotation,
    method=c("gsva", "ssgsea", "zscore", "plage"),
    kcdf=c("Gaussian", "Poisson", "none"),
    rnaseq=FALSE,
    abs.ranking=FALSE,
    min.sz=1,
    max.sz=Inf,
    no.bootstraps=0,
    bootstrap.percent = .632,
    parallel.sz=0,
    parallel.type="SOCK",
    mx.diff=TRUE,
    tau=switch(method, gsva=1, ssgsea=0.25, NA),
    kernel=TRUE,
    ssgsea.norm=TRUE,
    verbose=TRUE,
    return.old.value=FALSE)
## S4 method for signature 'ExpressionSet,GeneSetCollection'
gsva(expr, gset.idx.list, annotation,
    method=c("gsva", "ssgsea", "zscore", "plage"),
    kcdf=c("Gaussian", "Poisson", "none"),
    rnaseq=FALSE,
    abs.ranking=FALSE,
    min.sz=1,
    max.sz=Inf,
    no.bootstraps=0,
```

①

②

?gsva

何を書いてるのか（S4 methodって何よ?とか…）分かりづらいだろうが、①と②の比較から、遺伝子セット情報は③GeneSetCollection形式以外に、④list形式でもよいのだろう、ということがわかる。

http://127.0.0.1:16111/library/GSVA/html/gsva.html

R: Gene Set Variation Analysis

④ ①

```
## S4 method for signature 'ExpressionSet,list'
gsva(expr, gset.idx.list, annotation,
    method=c("gsva", "ssgsea", "zscore", "plage"),
    kcdf=c("Gaussian", "Poisson", "none"),
    rnaseq=FALSE,
    abs.ranking=FALSE,
    min.sz=1,
    max.sz=Inf,
    no.bootstraps=0,
    bootstrap.percent = .632,
    parallel.sz=0,
    parallel.type="SOCK",
    mx.diff=TRUE,
    tau=switch(method, gsva=1, ssgsea=0.25, NA),
    kernel=TRUE,
    ssgsea.norm=TRUE,
    verbose=TRUE,
    return.old.value=FALSE)
```

③ ②

```
## S4 method for signature 'ExpressionSet,GeneSetCollection'
gsva(expr, gset.idx.list, annotation,
    method=c("gsva", "ssgsea", "zscore", "plage"),
    kcdf=c("Gaussian", "Poisson", "none"),
    rnaseq=FALSE,
    abs.ranking=FALSE,
    min.sz=1,
    max.sz=Inf,
    no.bootstraps=0,
```

?gsva

発現情報もまた、①ExpressionSet以外に、②matrix形式でもよいことがわかる。この結果から、RNA-seqカウントデータの入力が通常のタブ区切りテキストファイルの場合は、基本そのまま読み込むのでよい（正確にはas.matrixしないといけない）と判断する。

http://127.0.0.1:16111/library/GSVA/html/gsva.html

R: Gene Set Variation Analysis

```
    verbose=TRUE,
    return.old.value=FALSE)
```

①

```
## S4 method for signature 'ExpressionSet,GeneSetCollection'
gsva(expr, gset.idx.list, annotation,
    method=c("gsva", "ssgsea", "zscore", "plage"),
    kcdf=c("Gaussian", "Poisson", "none"),
    rnaseq=FALSE,
    abs.ranking=FALSE,
    min.sz=1,
    max.sz=Inf,
    no.bootstraps=0,
    bootstrap.percent = .632,
    parallel.sz=0,
    parallel.type="SOCK",
    mx.diff=TRUE,
    tau=switch(method, gsva=1, ssgsea=0.25, NA),
    kernel=TRUE,
    ssgsea.norm=TRUE,
    verbose=TRUE,
    return.old.value=FALSE)
```

②

```
## S4 method for signature 'matrix,GeneSetCollection'
gsva(expr, gset.idx.list, annotation,
    method=c("gsva", "ssgsea", "zscore", "plage"),
    kcdf=c("Gaussian", "Poisson", "none"),
    rnaseq=FALSE,
    abs.ranking=FALSE,
    min.sz=1,
```

参考

おまけ

遺伝子セットのフィルタリングは、デフォルトでは行わない設定になっていることがわかる。①は遺伝子セットを構成するメンバー数の下限（minimum size）が1、上限（maximum size）がInfになっているからです。Infは無限大の意味です

http://127.0.0.1:16111/library/GSVA/html/gsva.html

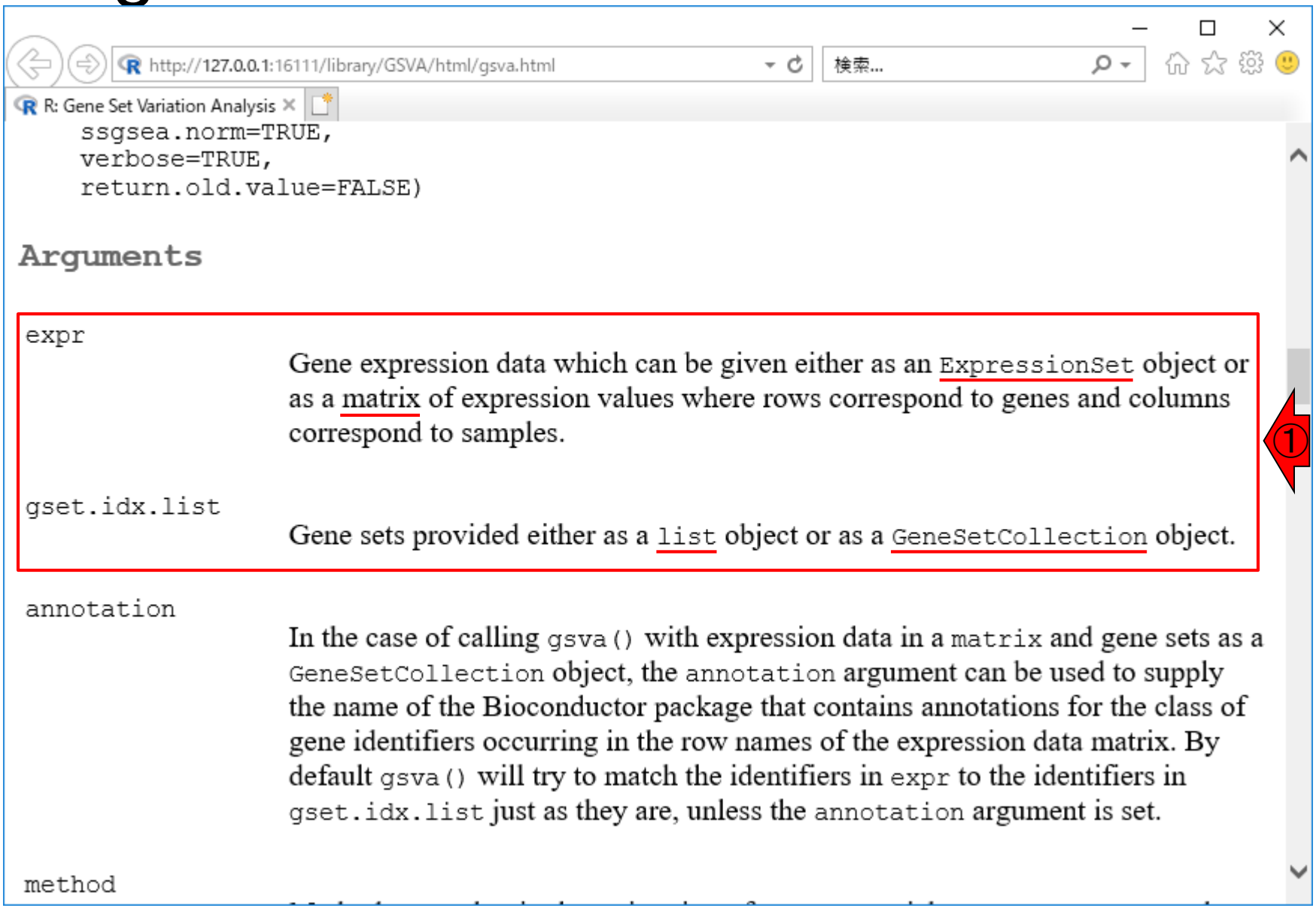
R: Gene Set Variation Analysis

```
    verbose=TRUE,
    return.old.value=FALSE)
## S4 method for signature 'ExpressionSet,GeneSetCollection'
gsva(expr, gset.idx.list, annotation,
    method=c("gsva", "ssgsea", "zscore", "plage"),
    kcdf=c("Gaussian", "Poisson", "none"),
    rnaseq=FALSE,
    abs.ranking=FALSE,
    min.sz=1,
    max.sz=Inf,
    no.bootstraps=0,
    bootstrap.percent = .632,
    parallel.sz=0,
    parallel.type="SOCK",
    mx.diff=TRUE,
    tau=switch(method, gsva=1, ssgsea=0.25, NA),
    kernel=TRUE,
    ssgsea.norm=TRUE,
    verbose=TRUE,
    return.old.value=FALSE)
## S4 method for signature 'matrix,GeneSetCollection'
gsva(expr, gset.idx.list, annotation,
    method=c("gsva", "ssgsea", "zscore", "plage"),
    kcdf=c("Gaussian", "Poisson", "none"),
    rnaseq=FALSE,
    abs.ranking=FALSE,
    min.sz=1,
```

①

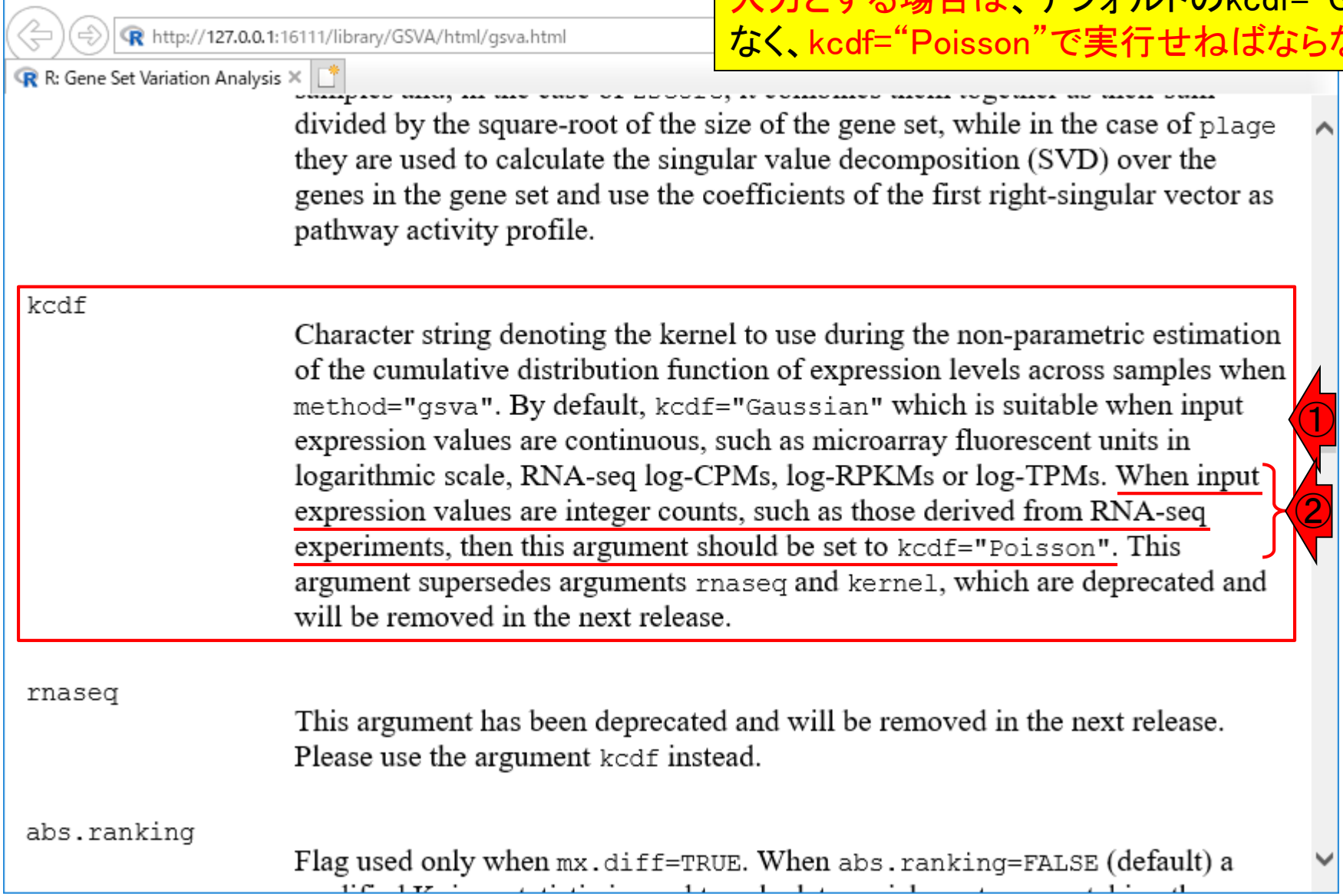
①このあたりにもちゃんと書いてますね。

?gsva



?gsva

①kcdfオプションは、この後のGSVA for RNA-seq dataの記述を見てから気づくのが実際のところかもしれない。結論のみ述べると、②RNA-seqのカウントデータを入力とする場合は、デフォルトのkcdf="Gaussian"ではなく、kcdf="Poisson"で実行せねばならない。



Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

おさらい

②GSVAの解説PDFには、Pickrellデータも例題として使われています。実際にPickrellデータを眺めます。

(Rで)塩基配列解析

(last modified 2018/06/29, since 2010)

このウェブページの
います。初心者の方

What's new?

• 「解析 | 一般 | Seq

• 解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan 2018) (last modified 2018/06
• 解析 | 機能解析 | 遺伝子セット 解析 | GSVA(Hänzelmann 20
• 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について ① modified 2018/06/27)
• 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | SeqGSEA(ang 2014)(last modified
• 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | GSVA(Hänzelmann 2013)(last modifi
• 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 |
• 解析 | 機能解析 | パスウェイ(Pathway)解析 |
• 解析 | 機能解析 | パスウェイ(Pathway)解析 |
• 解析 | 機能解析 | パスウェイ(Pathway)解析 |

解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について NEW

RNA-seqなどのタグカウントデータから遺伝子オントロジー(GO)解析を行うためのパッケージもいくつか出ています。遺伝子セット解析(Gene Set Analysis; GSA)という枠組みではGO解析もパスウェイ解析も同じなので、そちらもチェックするといいかもしれません。

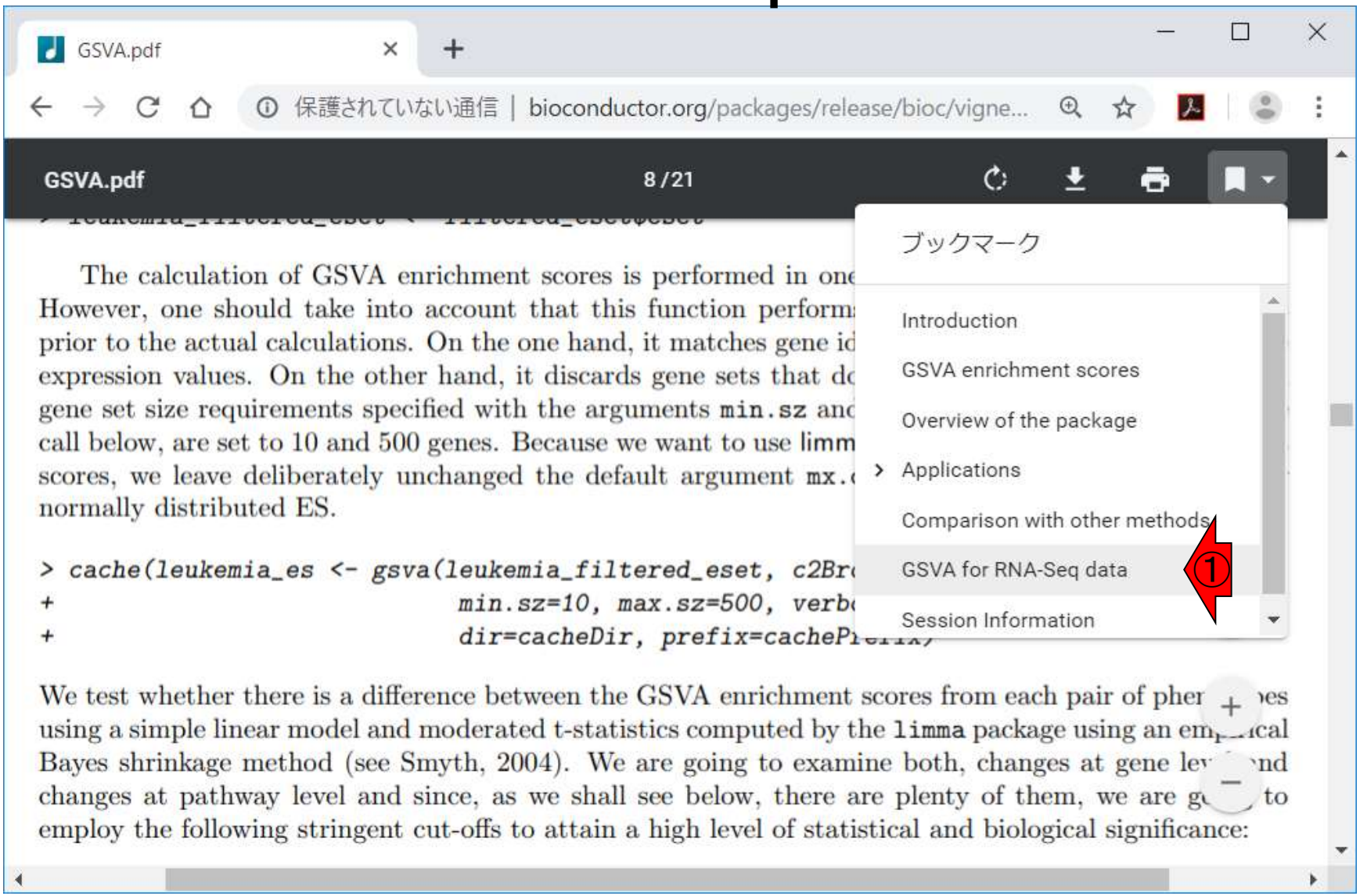
R用:

- GAGE: Luo et al., BMC Bioinformatics, 2009
- goseq: Young et al., Genome Biol., 2010
- GOSemSim: Yu et al., Bioinformatics, 2010
- Rスクリプト: Gao et al., Bioinformatics, 2011
- RamiGO: Schröder et al., Bioinformatics, 2013
- GSVA: Hänzelmann et al., BMC Bioinformatics, 2013 ②
- SeqGSEA(各群5反復以上を要求): Wang et al., Bioinfo atics, 2014
- GSAASeqSP: Xiong et al., Sci Rep., 2014
- GOplot(Visualization用): Walter et al., Bioinformatics, 2015
- GOexpress: Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA): Hundt et al., BMC Bioinformatics, 2016
- EGSEA: Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用): Yoon et al., PLoS One, 2016
- GSAR: Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA: Ren et al., BioData Min., 2017

③

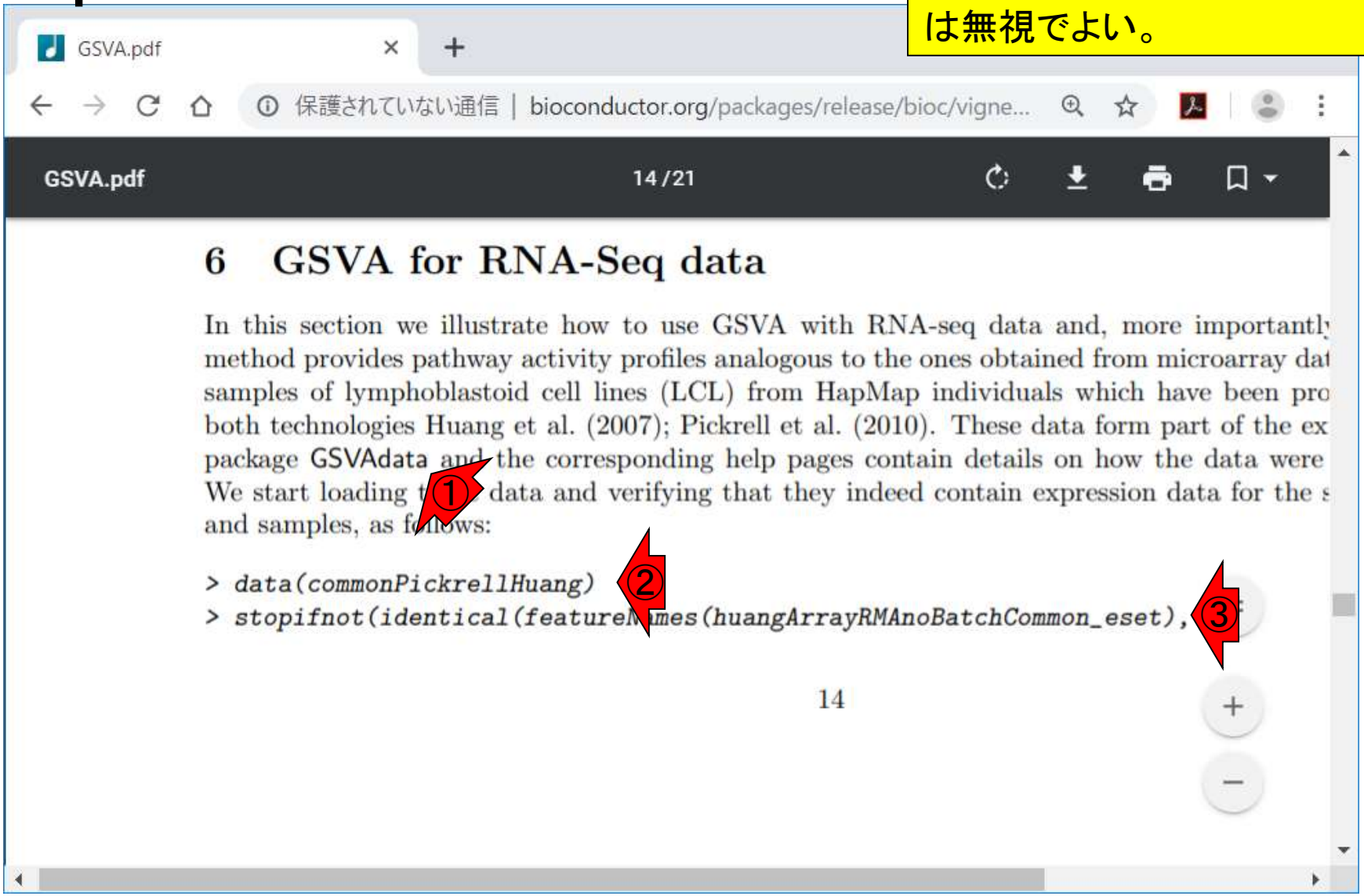
①GSVA for RNA-seq dataの項目へ。

GSVA for RNA-seq data



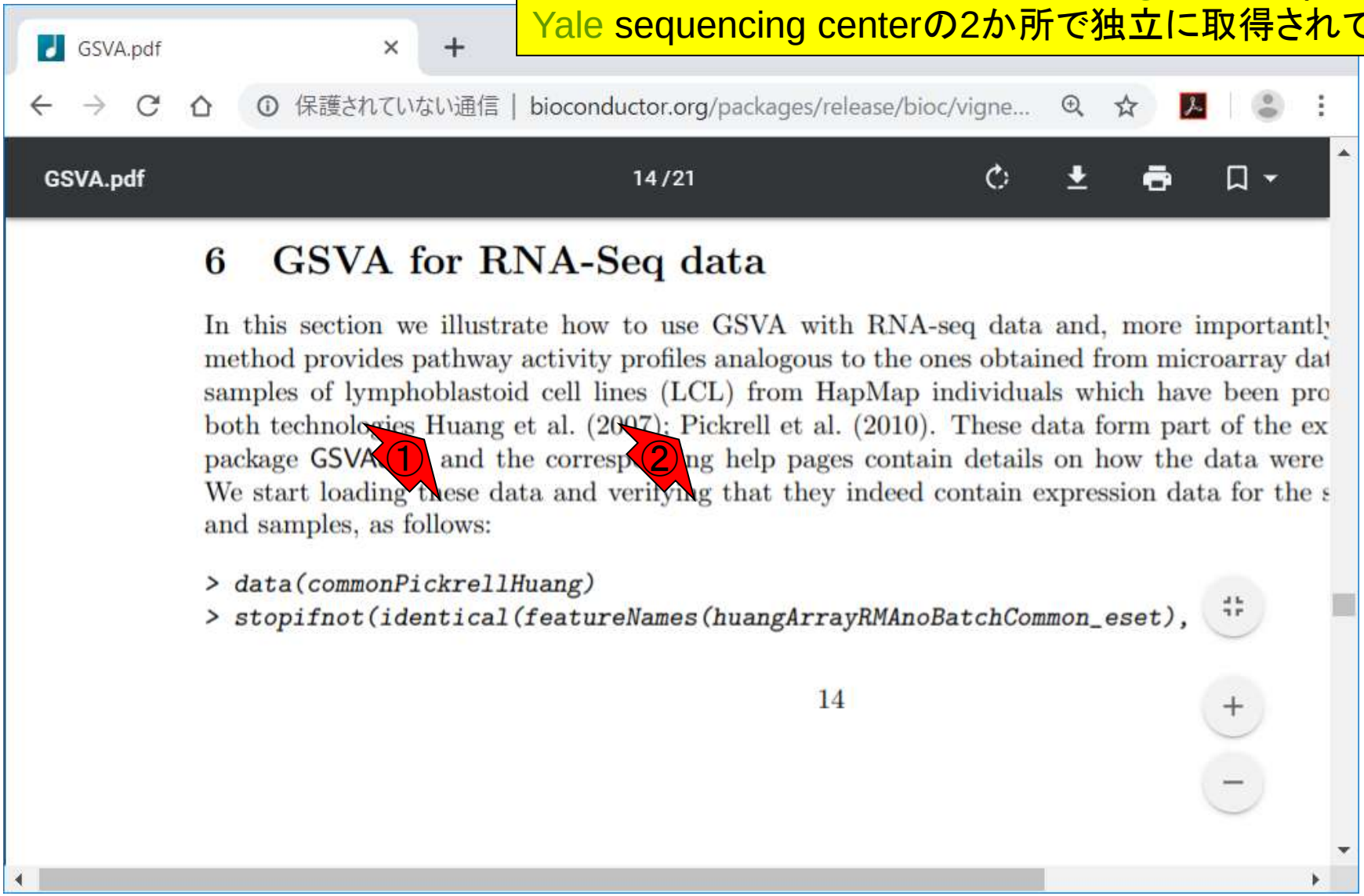
p14

こんな感じになります。この画面内で実際に行うのは、①GSVAdataパッケージをロードした後、②をコピペ実行することのみ。③は無視でよい。



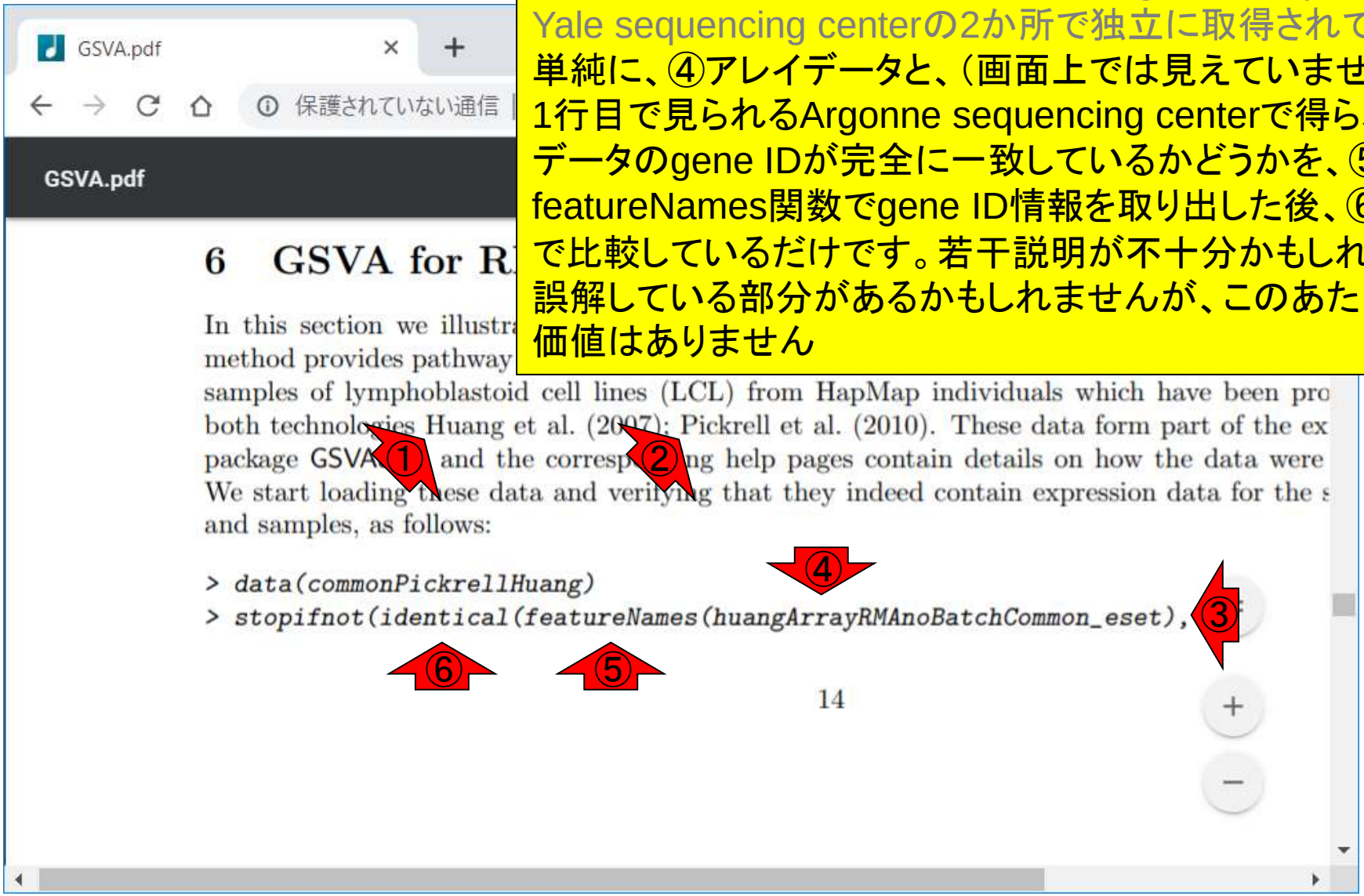
補足説明

①の論文ではマイクロアレイデータが、そして②の論文ではRNA-seqデータが取得されており、両者は比較可能な状態にあります。そして、②のRNA-seqデータはさらに、Argonne sequencing centerとYale sequencing centerの2か所で独立に取得されています。



補足説明

①の論文ではマイクロアレイデータが、そして②の論文ではRNA-seqデータが取得されており、両者は比較可能な状態にあります。そして、②のRNA-seqデータはさらに、Argonne sequencing centerとYale sequencing centerの2か所で独立に取得されています。③は単純に、④アレイデータと、(画面上では見えていませんが…)p16の1行目で見られるArgonne sequencing centerで得られたRNA-seqデータのgene IDが完全に一致しているかどうかを、⑤featureNames関数でgene ID情報を取り出した後、⑥identical関数で比較しているだけです。若干説明が不十分かもしれませんし私も誤解している部分があるかもしれませんが、このあたりは深入りする価値はありません



p15～16

①のあたりがp16の最初のほうの記述内容です。これもさらっと流すところ。

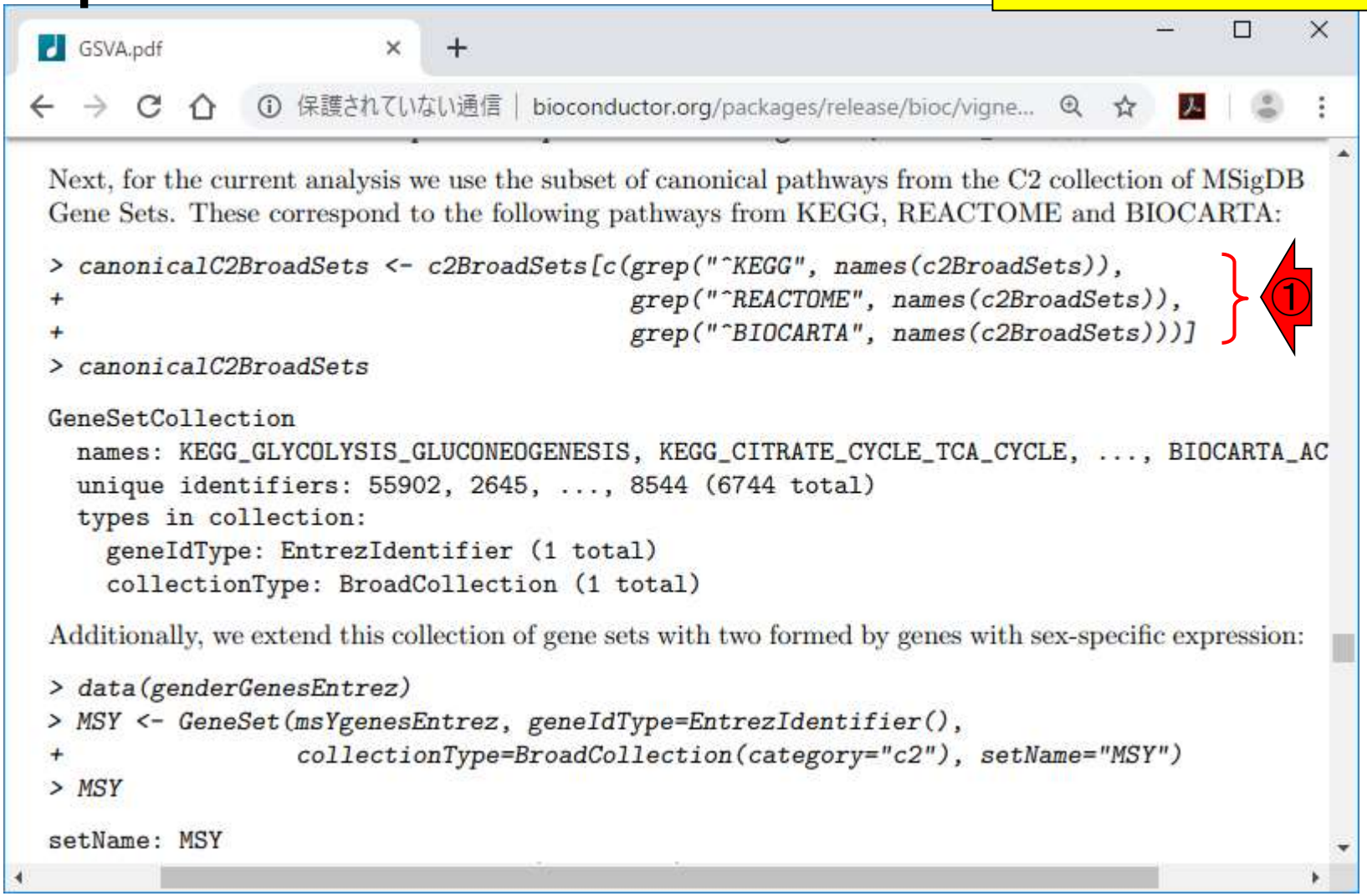
Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

参考

p16の上のほう

①のあたりもc2BroadSetsの中から、一部を抜き取っているだけ。個人的には、なぜこれをやる必要があるのか理解不能。



参考

C2ではなくC1

①c2BroadSetsの中から一部を抜き取って作成したcanonicalC2BroadSetsは、あくまでもC2コレクションのデータの一部。その一方で、②発現データのPickrell dataとMSigDB C1コレクションの遺伝子セット解析で有意な発現変動を示したのはchryq11であった。

GSVA.pdf

保護されていない通信 | bioconductor.org/packages/release/bioc/vignette...

Next, for the current analysis we use the subset of canonical pathways from the C2 collection of MSigDB Gene Sets. These correspond to the following pathways from KEGG, REACTOME and BIOCARTA:

```
> canonicalC2BroadSets <- c2BroadSets[c(grep("^KEGG", names(c2BroadSets)),
+                                       grep("^REACTOME", names(c2BroadSets)),
+                                       grep("^BIOCARTA", names(c2BroadSets)))]
> canonicalC2BroadSets
```

①

```
GeneSetCollection
  names: KEGG_GLYCOLYSIS_GLUCONEOGENESIS, KEGG_CITRATE_CYCLE_TCA_CYCLE, ..., BIOCARTA_AC
  unique identifiers: 55902, 2645, ..., 8544 (6744 total)
  types in collection:
    geneIdType: EntrezIdentifier (1 total)
    collectionType: BroadCollection (1 total)
```

Additionally, we extend this collection of gene sets with two formed by genes with sex-specific expression:

```
> data(genderGenesEntrez)
> MSY <- GeneSet(msYgenesEntrez, geneIdType=EntrezIdentifier(),
+                collectionType=BroadCollection(category="c2"), setName="MSY")
> MSY

setName: MSY
```

②

AbsFilterGSEA(Yoon et al., PLoS One, 11: e0165919, 2016)

参考

性特異的セットを追加

①canonicalC2BroadSetsのセットだけではPickrell dataの解析結果を示しづらいので、②2つのsex-specific expressionを示す遺伝子セットを追加すると書いています。

参考

性特異的セット1つめ

①1つめのsex-specific expressionを示す遺伝子セットをMSYというセット名で作成したところ。

GSVA.pdf

保護されていない通信 | bioconductor.org/packages/release/bioc/vigne...

Next, for the current analysis we use the subset of canonical pathways from the C2 collection of MSigDB Gene Sets. These correspond to the following pathways from KEGG, REACTOME and BIOCARTA:

```
> canonicalC2BroadSets <- c2BroadSets[c(grep("^KEGG", names(c2BroadSets)),
+                                       grep("^REACTOME", names(c2BroadSets)),
+                                       grep("^BIOCARTA", names(c2BroadSets)))]
> canonicalC2BroadSets

GeneSetCollection
  names: KEGG_GLYCOLYSIS_GLUCONEOGENESIS, KEGG_CITRATE_CYCLE_TCA_CYCLE, ..., BIOCARTA_AC
  unique identifiers: 55902, 2645, ..., 8544 (6744 total)
  types in collection:
    geneIdType: EntrezIdentifier (1 total)
    collectionType: BroadCollection (1 total)
```

Additionally, we extend this collection of gene sets with two formed by genes with sex-specific expression:

```
> data(genderGenesEntrez)
> MSY <- GeneSet(msYgenesEntrez, geneIdType=EntrezIdentifier(),
+                collectionType=BroadCollection(category="c2"), setName="MSY")
> MSY

setName: MSY
```

①

参考

性特異的セット2つめ

①1つめのsex-specific expressionを示す遺伝子セットをMSYというセット名で作成したところ。
②2つめのsex-specific expressionを示す遺伝子セットをXiEというセット名で作成したところ。

GSVA.pdf

保護されていない通信 | bioconductor.org/packages/release/bioc/vigne...

```
> data(genderGenesEntrez)
> MSY <- GeneSet(msYgenesEntrez, geneIdType=EntrezIdentifier(),
+                collectionType=BroadCollection(category="c2"), setName="MSY")
> MSY
```

①

```
setName: MSY
geneIds: 266, 84663, ..., 353513 (total: 34)
geneIdType: EntrezId
collectionType: Broad
  bcCategory: c2 (Curated)
  bcSubCategory: NA
details: use 'details(object)'
```

```
> XiE <- GeneSet(XiEgenesEntrez, geneIdType=EntrezIdentifier(),
+                collectionType=BroadCollection(category="c2"), setName="XiE")
> XiE
```

②

```
setName: XiE
geneIds: 293, 8623, ..., 1121 (total: 66)
geneIdType: EntrezId
collectionType: Broad
  bcCategory: c2 (Curated)
```

参考

性特異的セットマージ

②2つめのsex-specific expressionを示す遺伝子セットをXiEというセット名で作成したところ。③既存のcanonicalC2BroadSetsに対して、作成した2つのsex-specific expressionを示す④MSYと⑤XiEを、⑥GeneSetCollectionという形式でマージした結果を、⑦新たなcanonicalC2BroadSetsとする

GSVA.pdf

保護されていない通信 | bioconductor.org/packages/rele

```
> XiE <- GeneSet(XiEgenesEntrez, geneIdType=EntrezIdentifier(),
+                collectionType=BroadCollection(category="c2"), setName="XiE")
> XiE
```

②

```
setName: XiE
geneIds: 293, 8623, ..., 1121 (total: 66)
geneIdType: EntrezId
collectionType: Broad
  bcCategory: c2 (Curated)
  bcSubCategory: NA
details: ⑦ details(object)'

> canonicalC2BroadSets <- GeneSetCollection(c(canonicalC2BroadSets, MSY, XiE))
> canonicalC2BroadSets

GeneSetCollection
  names: KEGG_GLYCOLYSIS_GLUCONEOGENESIS, KEGG_CITRATE_CYCLE_TCA_CYCLE, ..., XiE (835 to
  unique identifiers: 55902, 2645, ..., 1121 (6810 total)
  types in collection:
    geneIdType: EntrezIdentifier (1 total)
    collectionType: BroadCollection (1 total)

We calculate now GSVA enrichment scores for these gene sets using first the microarray data and then
```

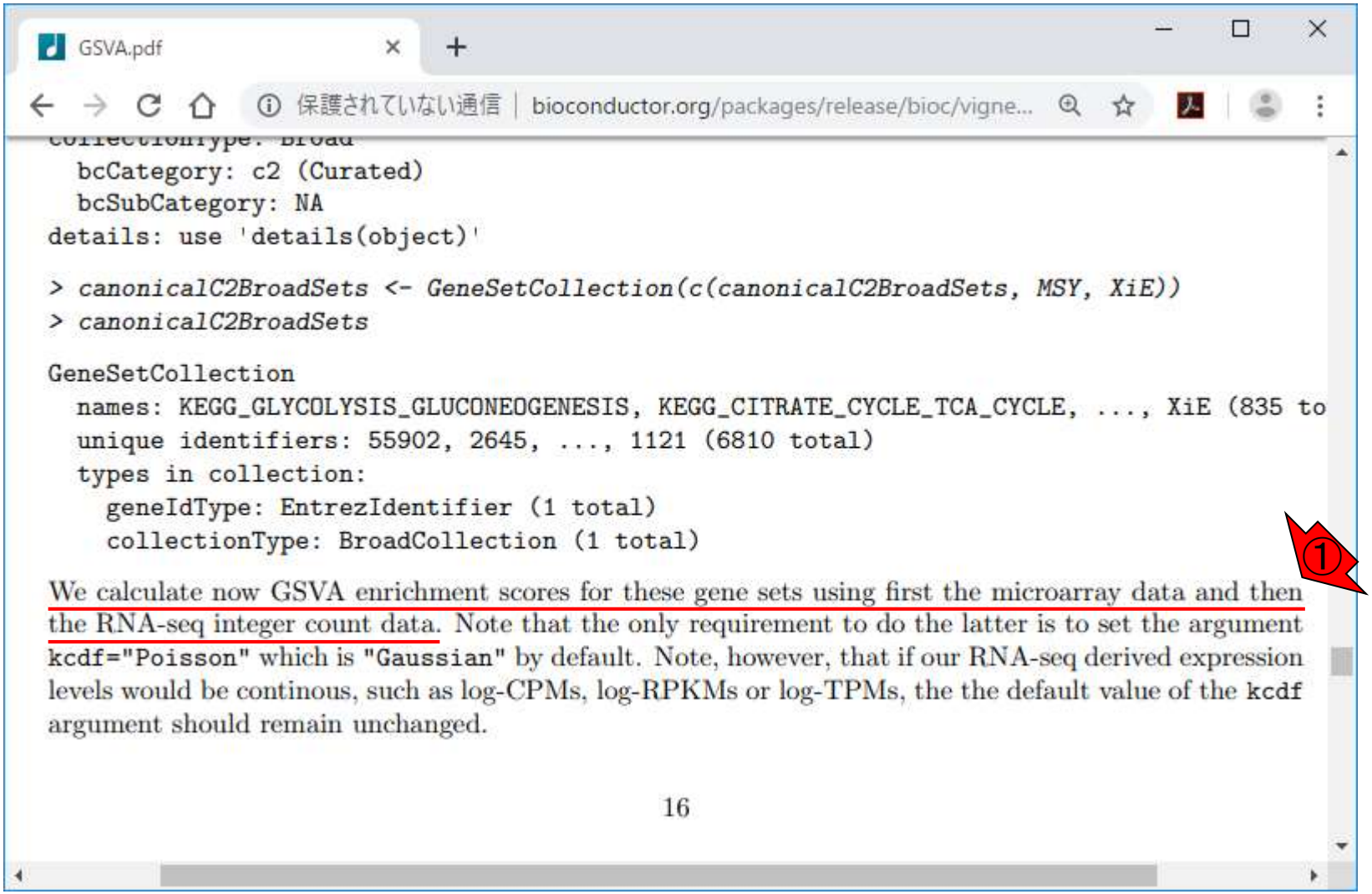
⑥ ③ ④ ⑤

Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

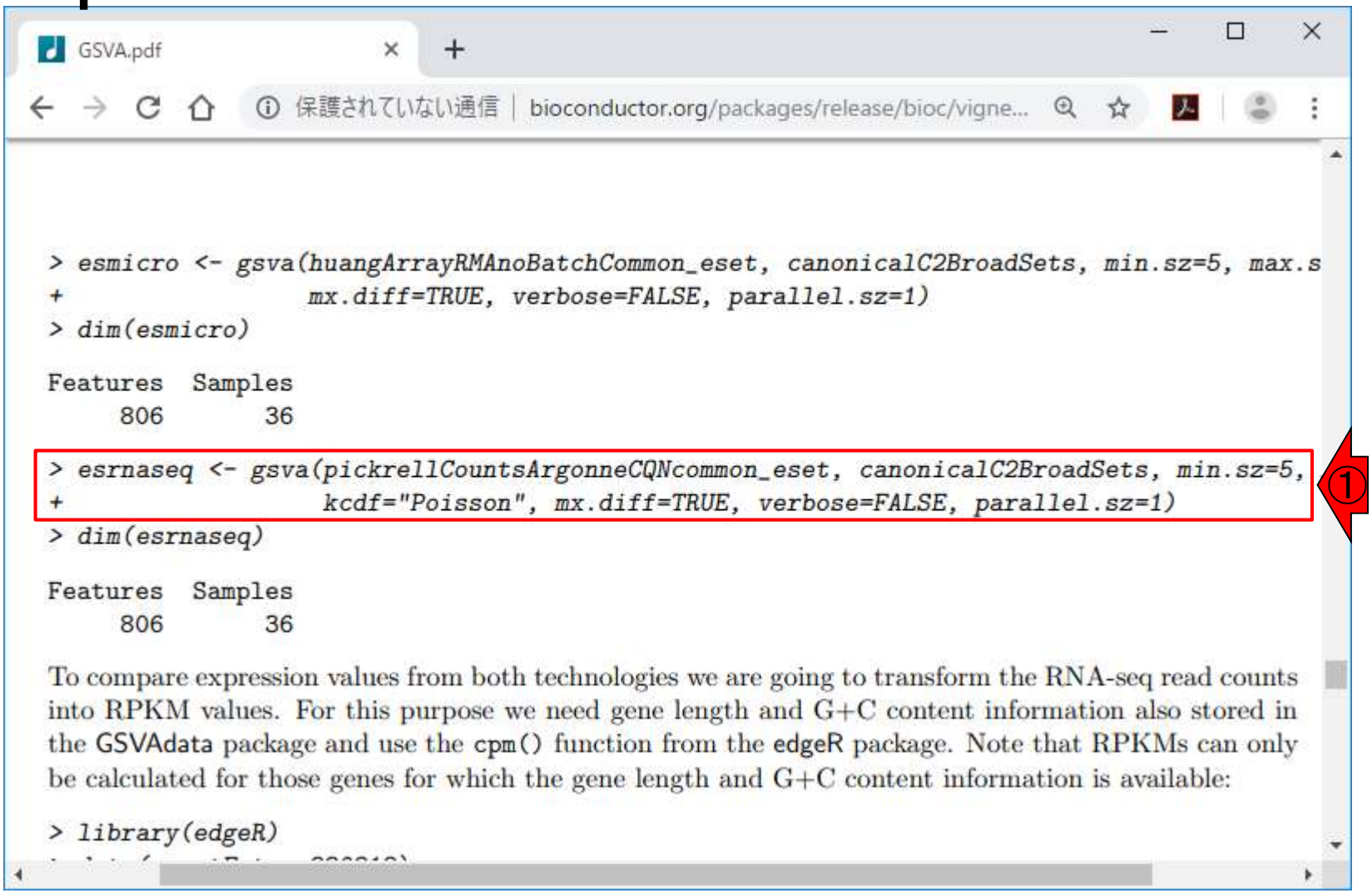
①やっとGSVAを実行できる状況になったようです。

GSVAの実行準備完了



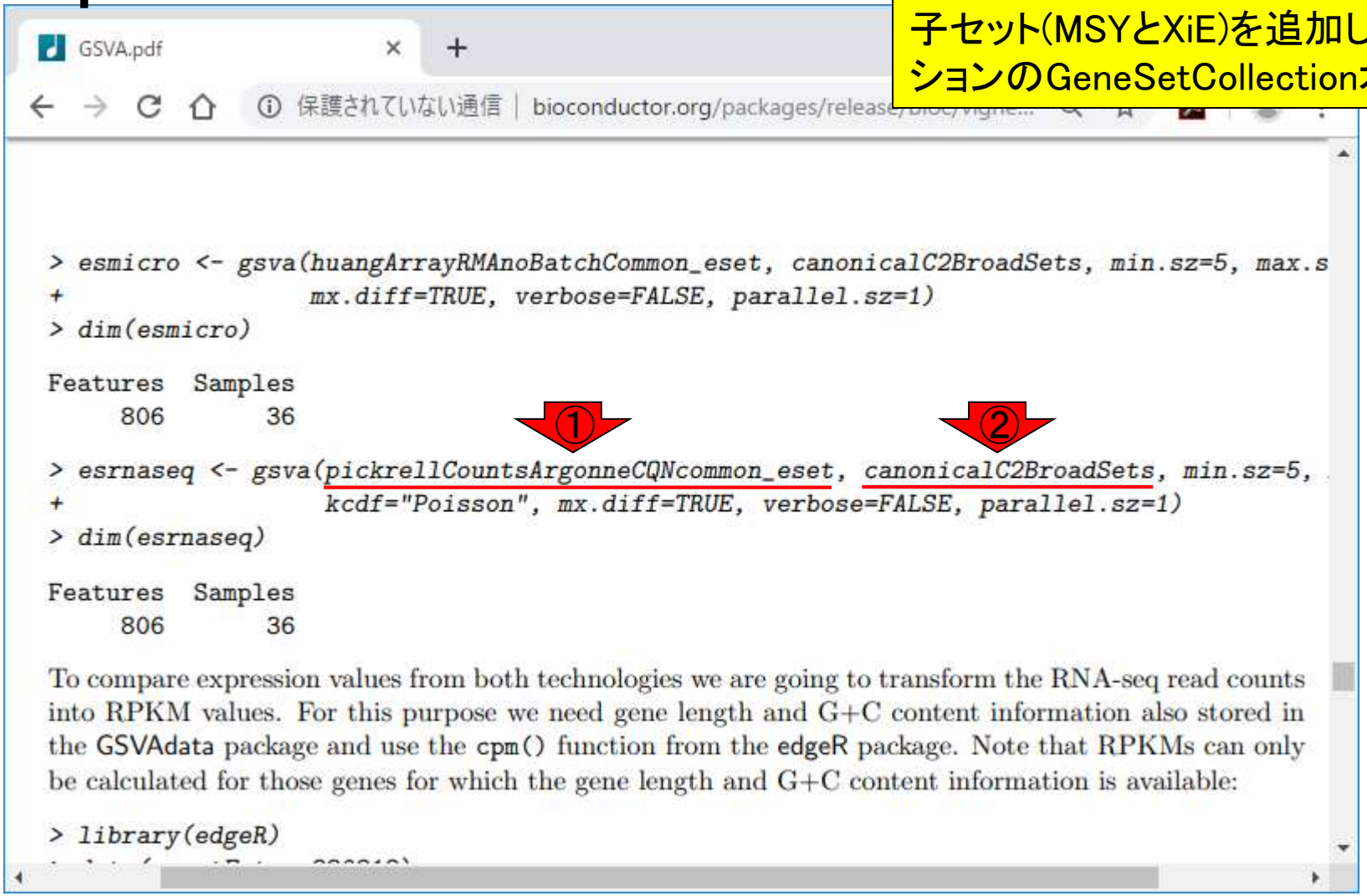
①がRNA-seqデータのほうのGSVA実行コマンドです。

p17の上のほう



p17の上のほう

①Argonne sequence centerで取得されたRNA-seqカウントデータのExpressionSetオブジェクトと、②sex-specific expressionを示す2つの遺伝子セット(MSYとXiE)を追加したMSigDB C2コレクションのGeneSetCollectionオブジェクトが入力。



p17の上のほう

①GSVA実行結果であるesrnaseqは、806行×36列からなるデータ。②この後は、RNA-seqカウントデータをRPKM値に変換してから、③マイクロアレイデータの結果であるesmicroとの比較を行って「ほら似た結果になってるでしょ」でオシマイ。sex-specific expressionを示す2つの遺伝子セット(MSYとXiE)については、一応サンプルごとに算出したEnrichment Scores (GSVA scores)の散布図を示している。そして男女間でサンプルごとのスコアが確かに異なっており、その傾向はマイクロアレイデータでもRPKMデータでも同じですね、ということは述べられている。しかしながら、有意な発現変動遺伝子セットはどれかを調べる枠組みはガイドラインも示されておらず残念。

GSVA.pdf

保護されていない通信 | bioconduc

③

```
> esmicro <- gsva(huangArrayRMAnoBatchCor
+                 mx.diff=TRUE, verbose=F
> dim(esmicro)

Features  Samples
     806       36

> esrnaseq <- gsva(pickrellCountsArgonneCQNcommon_eset, canonicalC2BroadSets, min.sz=5,
+                  kcdf="Poisson", mx.diff=TRUE, verbose=FALSE, parallel.sz=1)
> dim(esrnaseq)

Features  Samples
     806       36
```

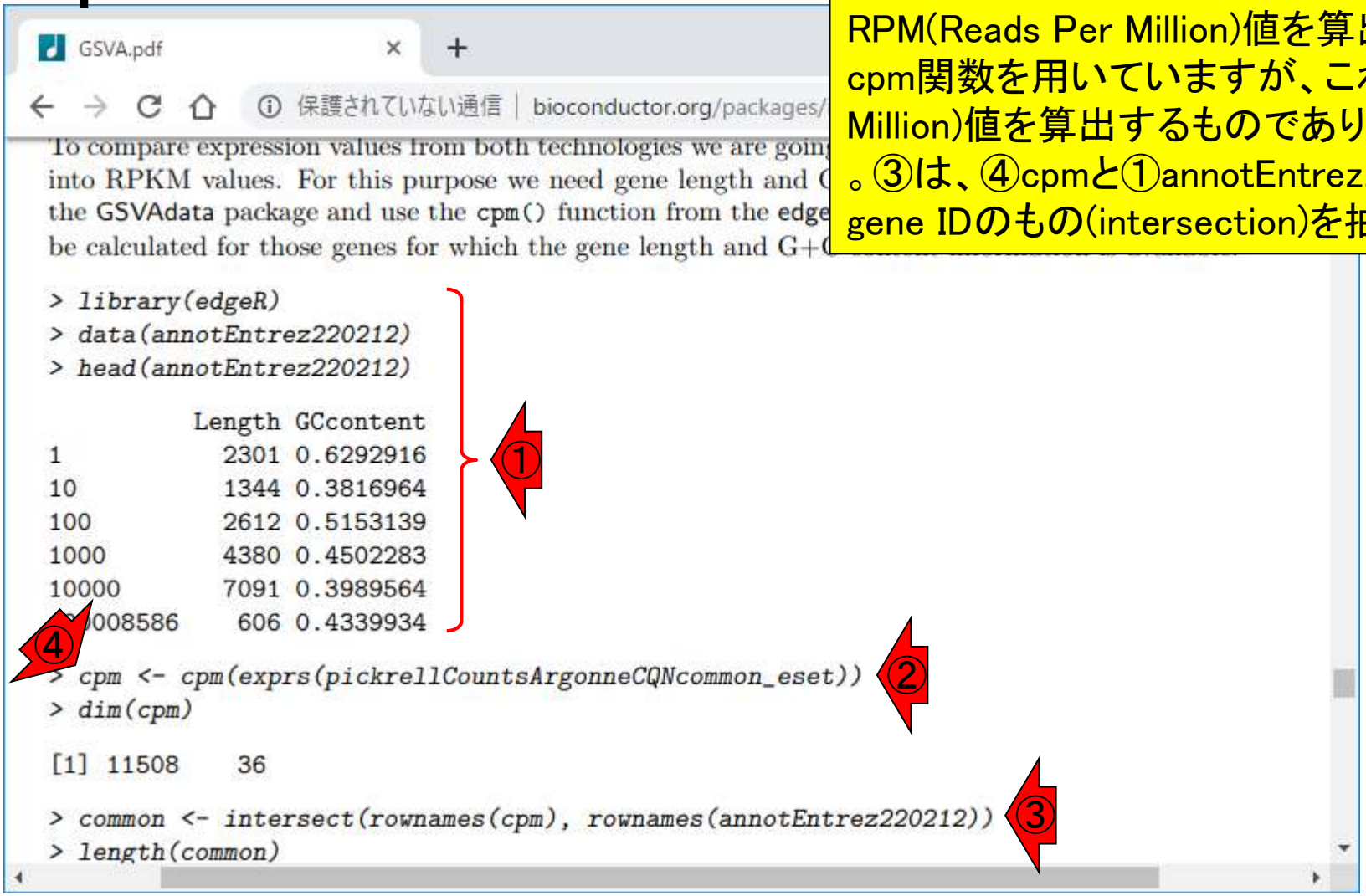
②

To compare expression values from both technologies we are going to transform the RNA-seq read counts into RPKM values. For this purpose we need gene length and G+C content information also stored in the GSVAdata package and use the `cpm()` function from the edgeR package. Note that RPKMs can only be calculated for those genes for which the gene length and G+C content information is available:

```
> library(edgeR)
```

p17の中央あたり

RPKM値を算出する最初のほうを示しているところ。①のあたりで、RPKM値への変換に必要な配列長情報を含むannoEntrez220212を取得。②がまずRPM(Reads Per Million)値を算出しているところ。cpm関数を用いていますが、これは(Counts Per Million)値を算出するものであり、実質的に同じです。③は、④cpmと①annotEntrez220212で共通のgene IDのもの(intersection)を抽出しているだけです



Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

順を追って説明をしますが、①の、②例題7を実行

発現行列データを整形して取得

(Rで)塩基配列解析

(last modified 2018/06/29, since 2010)

このウェブ
います。初

What's ne

• 「解析 |

- マップ後 | カウント情報取得 | トランスクリプトーム | BEDファイルから (last modified 2014/06/21)
- カウント情報取得 | リアルデータ | について (last modified 2019/05/16)
- カウント情報取得 | リアルデータ | SRP061240 | recount(Collado-Torres_2017) (last modified 2018/08/28)
- カウント情報取得 | リアルデータ | SRP056295 | recount(Collado-Torres_2017) (last modified 2018/08/29)
- カウント情報取得 | リアルデータ | SRP056146 | recount(Collado-Torres_2017) (last modified 2018/10/25)
- カウント情報取得 | リアルデータ | SRP035988 | recount(Collado-Torres_2017) (last modified 2018/08/25)
- カウント情報取得 | リアルデータ | SRP026126 | recount(Collado-Torres_2017) (last modified 2018/08/30)
- カウント情報取得 | リアルデータ | SRP018853 | recount(Collado-Torres_2017) (last modified 2018/08/26)
- カウント情報取得 | リアルデータ | SRP012167 | recount(Collado-Torres_2017) (last modified 2018/08/24)
- カウント情報取得 | リアルデータ | SRP012167 | parathyroidSE(Haglund_2012) (last modified 2018/08/19)
- カウント情報取得 | リアルデータ | SRP001558 | recount(Collado-Torres_2017) (last modified 2018/08/08)
- カウント情報取得 | リアルデータ | SRP001540 | recount(Collado-Torres_2017) (last modified 2018/08/10)
- カウント情報取得 | リアルデータ | SRP001540 | GSVAdata(Hänzelmann_2013) (last modified 2018/07/03)
- カウント情報取得 | リアルデータ | ERP000546 | recount(Collado-Torres_2017) (last modified 2019/07/03) NEW
- カウント情報取得 | シミュレーションデータ | RNA-seq | について (last modified 2019/04/05)
- カウント情報取得 | シミュレーションデータ | RNA-seq | Technical rep (ポアソン分布) (last modified 2018/07/22)
- カウント情

①

カウント情報取得 | リアルデータ | SRP001540 | GSVAdata(Hänzelmann_2013)

GSVAdataパッケージを用いて、 SRP001540(Pickrell et al., Nature, 2010 ; ブラウザはIE以外を推奨) のカウント情報を含むExpressionSetオブジェクトという形式のデータセット(commonPickrellHuangという名前で格納されています)をdata関数を用いてロードしたり、カウントデータの数値行列にした状態で保存するやり方を示します。 原著論文では、ヒトの69個体(40 female samples and 29 male samples)のカウントデータを取得しています。 このデータはAbsFilterGSEA (Yoon et al., PLoS One, 2016)、 およびGSVA (Hänzelmann et al., BMC Bioinformatics, 2013)中で、 検証用データとして用いられています。具体的には、MSigDB C1で、 2つのsex-specificな遺伝子セット(chryq11とchrxp22)が発現変動しているという結果を得ているようです。

順を追って説明をしますが、①の、②例題7を実行

発現行列データを整形して取得

カウント情報取得 | リアルデータ | SRP001540 | GSVAdata(Hänzelmann_2013)

GSVAdataパッケージを用いて、 SRP001540(Pickrell et al., Nature, 2010 ; ブラウザはIE以外を推奨) のカウント情報を含むExpressionSetオブジェクトという形式のデータセット(commonPickrellHuangという名前で格納されています)をdata関数を用いてロードしたり、カウントデータの数値行列にした状態で保存するやり方を示します。 原著論文では、ヒトの69個体(40 female samples and 29 male samples)のカウントデータを取得しています。 このデータはAbsFilterGSEA (Yoon et al., PLoS One, 2016)、 およびGSVA (Hänzelmann et al., BMC Bioinformatics, 2013)中で、 検証用データとして用いられています。具体的には、MSigDB C1で、 2つのsex-specificな遺伝子セット(chryq11とchrxp22)が発現変動しているという結果を得ているようです。
「ファイル」－「ディレクトリの変更」でダウン

②

1. pickrellCountsArgonneCQNcommon_esetの

原著論文(Pickrell et al., Nature, 2010)中で記載さ
データのExpressionSetオブジェクトです。

```
#必要なパッケージをロード
library(GSVAdata)                          #パ

#本番(目的のデータセットをロード)
data(commonPickrellHuang)                  #pa
ls()                                       #利
eset <- pickrellCountsArgonneCQNcommon_ese
eset                                       #確
```

7. pickrellCountsYaleCQNcommon_esetの場合：

例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニークにしています。
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルはhoge7.txtです。

```
out_f <- "hoge7.txt"                       #出力ファイル名を指定してout_fに格納

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

#読込(目的のデータセットをロード)
data(commonPickrellHuang)                  #paramで指定したデータセットのロード
ls()                                       #利用可能なオブジェクト名を表示
eset <- pickrellCountsYaleCQNcommon_eset#esetとして取り扱う
dim(pData(eset))                           #確認してるだけです
head(pData(eset))                          #確認してるだけです
pData(eset)$Gender                         #確認してるだけです
table(pData(eset)$Gender)                  #確認してるだけです

#本番(重複除去)
length(rownames(exprs(eset)))              #重複除去前の遺伝子数を確認
length(unique(rownames(exprs(eset))))      #重複除去前のユニークな遺伝子数を確認
hoge <- nsFilter(eset, var.filter=F,       #同一IDの重複除去のみを実行
                 remove.dupEntrez=T)       #同一IDの重複除去のみを実行
eset <- hoge$eset                          #esetとして取り扱う
length(rownames(exprs(eset)))              #重複除去前の遺伝子数を確認
```

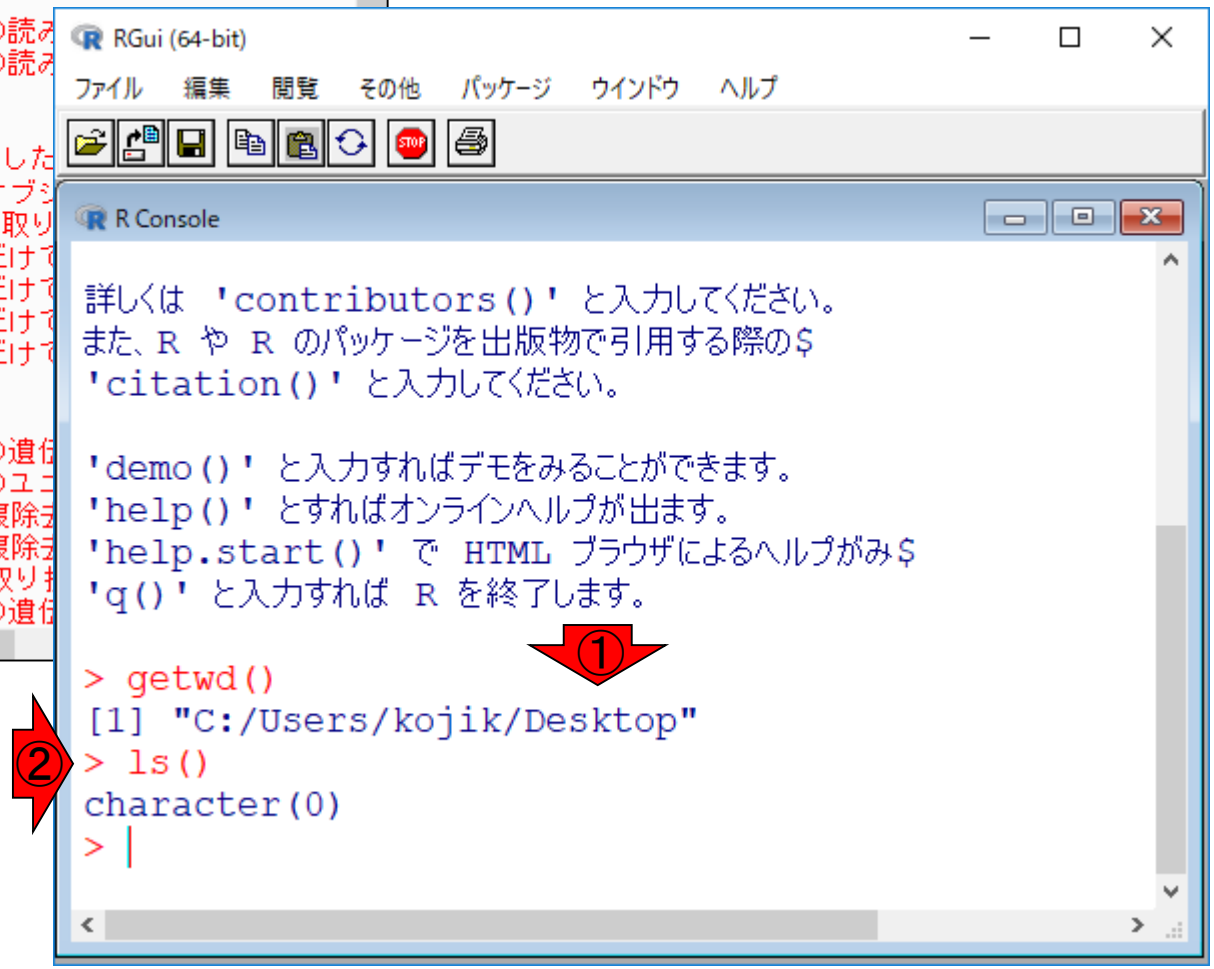
例題7実行の下準備

話がややこしくなるので、一旦Rを再起動し、作業ディレクトリを①Desktopに変更してから、②ls()を実行。これは利用可能なオブジェクトが何もないことを確認しているだけです。単に全員の環境を揃えているだけ。「rm(list = ls())」でオブジェクトの全消去と同じことをしたいだけです。

7. pickrellCountsYaleCQNcommon_esetの場合:

例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニークにし
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルはhoge7.txtで

```
out_f <- "hoge7.txt"                          #出力ファイル名を指定して

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

#読込(目的のデータセットをロード)
data(commonPickrellHuang)                     #paramで指定した
ls()                                          #利用可能なオブジ
eset <- pickrellCountsYaleCQNcommon_eset#esetとして取り
dim(pData(eset))                              #確認してるだけで
head(pData(eset))                             #確認してるだけで
pData(eset)$Gender                            #確認してるだけで
table(pData(eset)$Gender)                     #確認してるだけで

#本番(重複除去)
length(rownames(exprs(eset)))                 #重複除去前の遺伝
length(unique(rownames(exprs(eset))))         #重複除去前のユニ
hoge <- nsFilter(eset, var.filter=F,          #同一IDの重複除去
                 remove.dupEntrez=T)          #同一IDの重複除去
eset <- hoge$eset                             #esetとして取り扱
length(rownames(exprs(eset)))                 #重複除去前の遺伝
```



例題7

①赤枠内をコピペ。GSVAdataパッケージ中にある発現データcommonPickrellHuangをロードし、②どのようなオブジェクトが利用可能かを見るところまでです。commonPickrellHuangには、3種類の発現データが含まれています。③アレイデータ、および④Argonne sequencing centerと⑤Yale sequencing centerで取得されたRNA-seqカウントデータです

7. pickrellCountsYaleCQNcommon_esetの場合:

例題6をベースとして、重複したEntrez gene IDのものが存在するの
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力フ

```
out_f <- "hoge7.txt"                          #出力ファ

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

#読込(目的のデータセットをロード)
data(commonPickrellHuang)                     #paramで指定した
ls()                                          #利用可能なオブジ
eset <- pickrellCountsYaleCQNcommon_eset#esetとして取り
dim(pData(eset))                              #確認してるだけで
head(pData(eset))                             #確認してるだけで
pData(eset)$Gender                            #確認してるだけで
table(pData(eset)$Gender)                     #確認してるだけで

#本番(重複除去)
length(rownames(exprs(eset)))                 #重複除去前の遺伝
length(unique(rownames(exprs(eset))))         #重複除去前のユニ
hoge <- nsFilter(eset, var.filter=F,          #同一IDの重複除去
                 remove.dupEntrez=T)          #同一IDの重複除去
eset <- hoge$eset                             #esetとして取り
length(rownames(exprs(eset)))                 #重複除去前の
```

①

②

RGui (64-bit)

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

R Console

```
 警告メッセージ:
1:  パッケージ 'annotate' はバージョン 3.4.4 $
2:  パッケージ 'XML' はバージョン 3.4.4 の R $
> library(genefilter)                         #パッ$
>
> #読込(目的のデータセットをロード)
> data(commonPickrellHuang)                   #param$
> ls()                                        #利用$
[1] "huangArrayRMAnoBatchCommon_eset"
[2] "out_f"
[3] "pickrellCountsArgonneCQNcommon_eset"
[4] "pickrellCountsYaleCQNcommon_eset"
>
```

③ ④ ⑤

①赤枠内をコピペ。

列の並びはバラバラ

7. pickrellCountsYaleCQNcommon_esetの場合:

例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニークにしています。
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルはhoge7.txtです。

```
out_f <- "hoge7.txt"                    #出力ファイル名を指定してout_fに格納

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

#読込(目的のデータセットをロード)
data(commonPickrellHuang)               #paramで指定した
ls()                                    #利用可能なオブジ
eset <- pickrellCountsYaleCQNcommon_eset#esetとして取り
dim(pData(eset))                        #確認してるだけで
head(pData(eset))                       #確認してるだけで
pData(eset)$Gender                      #確認してるだけで
table(pData(eset)$Gender)               #確認してるだけで

#本番(重複除去)
length(rownames(exprs(eset)))           #重複除去前の遺伝
length(unique(rownames(exprs(eset))))   #重複除去前のユニ
hoge <- nsFilter(eset, var.filter=F,    #同一IDの重複除去
                 remove.dupEntrez=T)    #同一IDの重複除去
eset <- hoge$eset                       #esetとして取り扱
length(rownames(exprs(eset)))           #重複除去前の遺伝
```

①

RGui (64-bit)

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

R Console

```
NA19137             Y043-2                mother
NA18861             Y024-2                mother
NA19116             Y060-2                mother
> pData(eset)$Gender                          #確認$
 [1] "Female" "Female" "Male"   "Female"
 [5] "Female" "Female" "Male"   "Female"
 [9] "Male"   "Female" "Male"   "Female"
[13] "Male"   "Male"   "Male"   "Female"
[17] "Male"   "Female" "Male"   "Female"
[21] "Female" "Female" "Female" "Female"
[25] "Female" "Female" "Female" "Male"
[29] "Female" "Male"   "Female" "Male"
[33] "Female" "Female" "Female" "Male"
> |
```

列の並びはバラバラ

①Yale sequencing centerで取得されたRNA-seqカウントデータのオブジェクトをesetという名前で取り扱い、②発現行列の列（phenotypeという意味でpData）に関するメタデータ情報を確認したところ。Gender列が各サンプルの性別情報を含んでいると学習し、そこを表示させた結果が③赤枠の最後のコマンド。列の並びは性別できれいに分離されていないことがわかる。この段階でGSVA実行結果を見やすくするために、予めソートしておこうと考える。

7. pickrellCountsYaleCQNcommon_esetの場合:

例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニークI
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルはhoge7.

```
out_f <- "hoge7.txt"                          #出力ファイル名を指定し

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

#読込(目的のデータセットをロード)
data(commonPickrellHuang)                     #paramで指定した
ls()                                          #利用可能なオブジ
eset <- pickrellCountsYaleCQNcommon_eset      #esetとして取り
dim(pData(eset))                              #確認してるだけで
head(pData(eset))                             #確認してるだけで
pData(eset)$Gender                            #確認してるだけで
table(pData(eset)$Gender)                     #確認してるだけで

#本番(重複除去)
length(rownames(exprs(eset)))                 #重複除去前の遺
length(unique(rownames(exprs(eset))))         #重複除去前のユニ
hoge <- nsFilter(eset, var.filter=F,          #同一IDの重複除去
                 remove.dupEntrez=T)          #同一IDの重複除去
eset <- hoge$eset                             #esetとして取り扱
length(rownames(exprs(eset)))                 #重複除去前の遺伝
```

①

②

③

R Console

```
NA19137          Y043-2          mother
NA18861          Y024-2          mother
NA19116          Y060-2          mother
> pData(eset)$Gender                          #確認$
 [1] "Female" "Female" "Male"   "Female"
 [5] "Female" "Female" "Male"   "Female"
 [9] "Male"   "Female" "Male"   "Female"
[13] "Male"   "Male"   "Male"   "Female"
[17] "Male"   "Female" "Male"   "Female"
[21] "Female" "Female" "Female" "Female"
[25] "Female" "Female" "Female" "Male"
[29] "Female" "Male"   "Female" "Male"
[33] "Female" "Female" "Female" "Male"
>
```

女23人、男13人

①table関数で内訳を調査。Femaleが23人、Maleが13人。「G1群23サンプル vs. G2群13サンプル」の2群間比較データだと読み替えてもよい。

7. pickrellCountsYaleCQNcommon_esetの場合:

例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニークにしています。
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルはhoge7.txtです。

```
out_f <- "hoge7.txt"                    #出力ファイル名を指定してout_fに格納

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

#読込(目的のデータセットをロード)
data(commonPickrellHuang)               #paramで指定した
ls()                                    #利用可能なオブジ
eset <- pickrellCountsYaleCQNcommon_eset#esetとして取り
dim(pData(eset))                        #確認してるだけで
head(pData(eset))                       #確認してるだけで
pData(eset)$Gender                      #確認してるだけで
table(pData(eset)$Gender)               #確認してるだけで

#本番(重複除去)
length(rownames(exprs(eset)))           #重複除去前の遺伝
length(unique(rownames(exprs(eset))))   #重複除去前のユニ
hoge <- nsFilter(eset, var.filter=F,    #同一IDの重複除去
                 remove.dupEntrez=T)    #同一IDの重複除去
eset <- hoge$eset                       #esetとして取り扱
length(rownames(exprs(eset)))           #重複除去前の遺伝
```

①

RGui (64-bit)

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

R Console

```
 [1] "Female" "Female" "Male"   "Female"
 [5] "Female" "Female" "Male"   "Female"
 [9] "Male"   "Female" "Male"   "Female"
[13] "Male"   "Male"   "Male"   "Female"
[17] "Male"   "Female" "Male"   "Female"
[21] "Female" "Female" "Female" "Female"
[25] "Female" "Female" "Female" "Male"
[29] "Female" "Male"   "Female" "Male"
[33] "Female" "Female" "Female" "Male"
> table(pData(eset)$Gender)                    #確認$

Female   Male
    23     13
>
```

コード下部に移動

①一番下までカーソル移動して最後の部分まで表示。

7. pickrellCountsYaleCQNcommon_esetの場合:

例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニークにしています。
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルはhoge7.txtです。

```
pData(eset)$Gender                          #確認してるだけです
table(pData(eset)$Gender)                   #確認してるだけです

#本番(重複除去)
length(rownames(exprs(eset)))               #重複除去前の遺伝子数を確認
length(unique(rownames(exprs(eset))))       #重複除去前のユニークな遺伝子数を確認
hoge <- nsFilter(eset, var.filter=F,        #同一IDの重複除去のみを実行
                 remove.dupEntrez=T)        #同一IDの重複除去のみを実行
eset <- hoge$eset                           #esetとして取り扱う
length(rownames(exprs(eset)))               #重複除去前の遺伝子数を確認

#後処理(列名を変更し、列をソート)
data <- exprs(eset)                         #dataとして取り扱う
colnames(data) <- pData(eset)$Gender        #列名を変更
head(data[1:2,])                            #確認してるだけです
data <- data[, order(colnames(data))]       #列名でソート
head(data[1:2,])                            #確認してるだけです

#ファイルに保存
tmp <- cbind(rownames(data), data)          #保存したい情報をtmpに格納
write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)#tmpの中
```

①

重複gene IDの除去

①オリジナルの発現行列exprs(eset)に対して、その行名からなるベクトルrownames(exprs(eset))の要素数と、行名をユニークにした後のベクトルの要素数を表示。このデータの場合は、11508 − 11482 = 26個分の重複があったことがわかる。この重複を除去してユニークなgene IDにしておかないと、保存(hoge7.txtの作成)自体はできるが、次にそれを(私のウェブページ上の通常のやり方で)読み込むときに「重複したrow.namesは許されない!」と言われる。

7. pickrellCountsYaleCQNcommon_esetの場合:

例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニーク
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルはhoge7

```
pData(eset)$Gender                          #確認してるだけです
table(pData(eset)$Gender)                   #確認してるだけです

#本番(重複除去)
length(rownames(exprs(eset)))               #重複除去前の遺伝
length(unique(rownames(exprs(eset))))       #重複除去前のユニ
hoge <- nsFilter(eset, var.filter=F,        #同一IDの重複除去
                 remove.dupEntrez=T)        #同一IDの重複除去
eset <- hoge$eset                           #esetとして取り出
length(rownames(exprs(eset)))               #重複除去前の遺伝

#後処理(列名を変更し、列をソート)
data <- exprs(eset)                         #dataとして取り出
colnames(data) <- pData(eset)$Gender        #列名を変更
head(data[1:2,])                            #確認してるだけで
data <- data[, order(colnames(data))]       #列名でソート
head(data[1:2,])                            #確認してるだけで

#ファイルに保存
tmp <- cbind(rownames(data), data)          #保存したい情報を
write.table(tmp, out_f, sep="\t", append=F, quote=F, r
```

①

R Console

```
[25] "Female" "Female" "Female" "Male"
[29] "Female" "Male"   "Female" "Male"
[33] "Female" "Female" "Female" "Male"
> table(pData(eset)$Gender)                 #確認$

Female   Male
    23     13
> dim(exprs(eset))
[1] 11508    36
> length(rownames(exprs(eset)))             #重複$
[1] 11508
> length(unique(rownames(exprs(eset))))     #重複$
[1] 11482
>
```

①

nsFilter関数の実行

①重複除去を行うnsFilter関数を実行。②var.filter=Fオプションがついているが、こうしておかないと重複除去のみにならないからです。③思い描いた通りの結果になっています。実際問題としては重複したEntrez gene IDsがあった場合にどのような取り扱いを行っているか（平均値を得ているのかどうかなど）までは調べきれていません。

7. pickrellCountsYaleCQNcommon_esetの場合:

例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニーク
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルはhoge7

```
pData(eset)$Gender                    #確認してるだけです
table(pData(eset)$Gender)             #確認してるだけです

#本番(重複除去)
length(rownames(exprs(eset)))         #重複除去前の遺伝
length(unique(rownames(exprs(eset)))) #重複除去前のユニ
hoge <- nsFilter(eset, var.filter=F,  #同一IDの重複除去
                 remove.dupEntrez=T)  #同一IDの重複除去
eset <- hoge$eset                     #esetとして取り扱
length(rownames(exprs(eset)))         #重複除去前の遺伝

#後処理(列名を変更し、列をソート)
data <- exprs(eset)                   #dataとして取り扱
colnames(data) <- pData(eset)$Gender  #列名を変更
head(data[1:2,])                      #確認してるだけで
data <- data[, order(colnames(data))] #列名でソート
head(data[1:2,])                      #確認してるだけで

#ファイルに保存
tmp <- cbind(rownames(data), data)    #保存したい情報を
write.table(tmp, out_f, sep="\t", append=F, quote=F, r
```

①

RGui (64-bit)

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

R Console

```
Female   Male
    23     13
> dim(exprs(eset))
[1] 11508    36
> length(rownames(exprs(eset)))          #重複$
[1] 11508
> length(unique(rownames(exprs(eset))))  #重複$
[1] 11482
> hoge <- nsFilter(eset, var.filter=F,   #同一I$
+                  remove.dupEntrez=T)   #同一I$
> eset <- hoge$eset                      #eset$
> length(rownames(exprs(eset)))          #重複$
[1] 11482
>
```

②

③

列名変更の必要性

①の後処理として列名変更をしなかった場合は、出力ファイル中の列名が②の実行結果のようになってしまいます。今は男女間での発現変動遺伝子セット解析を行いたいので、列名はFemale or Maleで充分

7. pickrellCountsYaleCQNcommon_esetの場合:

例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニークにしています。
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルは hoge7.txt です。

```
pData(eset)$Gender                    #確認してるだけです
table(pData(eset)$Gender)             #確認してるだけです

#本番(重複除去)
length(rownames(exprs(eset)))         #重複除去前の遺伝
length(unique(rownames(exprs(eset)))) #重複除去前のユニ
hoge <- nsFilter(eset, var.filter=F,  #同一IDの重複除去
                 remove.dupEntrez=T)  #同一IDの重複除去
eset <- hoge$eset                     #esetとして取り扱
length(rownames(exprs(eset)))         #重複除去前の遺伝

#後処理(列名を変更し、列をソート)
data <- exprs(eset)
colnames(data) <- pData(eset)$Gender  #列名を変更
head(data[1:2,])                      #確認してるだけで
data <- data[, order(colnames(data))] #列名でソート
head(data[1:2,])                      #確認してるだ

#ファイルに保存
tmp <- cbind(rownames(data), data)    #保存したい情報を
write.table(tmp, out_f, sep="\t", append=F, quote=F, r
```

①

②

RGui (64-bit)

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

R Console

```
> eset <- hoge$eset                          #eset$
> length(rownames(exprs(eset)))              #重複$
[1] 11482
> colnames(exprs(eset))
 [1] "NA19099" "NA18523" "NA19144" "NA19137"
 [5] "NA18861" "NA19116" "NA19130" "NA19131"
 [9] "NA19119" "NA19152" "NA19153" "NA19140"
[13] "NA19138" "NA18522" "NA19192" "NA19193"
[17] "NA19239" "NA19238" "NA19210" "NA19201"
[21] "NA19172" "NA18870" "NA18858" "NA18852"
[25] "NA19159" "NA18855" "NA19204" "NA18501"
[29] "NA19127" "NA19098" "NA19093" "NA18856"
[33] "NA18912" "NA18517" "NA18502" "NA19171"
>
```

列名変更後

①赤枠内コピペ実行後。確かに列名変更できていることがわかります。ただ、以前pData(eset)$Genderでも確認したように、FemaleとMaleが性別順に並んでいないことがわかります。

7. pickrellCountsYaleCQNcommon_esetの場合:

例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニークにしています。
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルはhoge7.txtです。

```
pData(eset)$Gender                    #確認してるだけです
table(pData(eset)$Gender)             #確認してるだけです

#本番(重複除去)
length(rownames(exprs(eset)))         #重複除去前の遺伝
length(unique(rownames(exprs(eset)))) #重複除去前のユニ
hoge <- nsFilter(eset, var.filter=F,  #同一IDの重複除去
                 remove.dupEntrez=T)  #同一IDの重複除去
eset <- hoge$eset                     #esetとして取り出
length(rownames(exprs(eset)))         #重複除去前の遺伝

#後処理(列名を変更し、列をソート)
data <- exprs(eset)
colnames(data) <- pData(eset)$Gender  #列名を変更
head(data[1:2,])                      #確認してるだけで
data <- data[, order(colnames(data))] #列名でソート
head(data[1:2,])                      #確認してるだけで

#ファイルに保存
tmp <- cbind(rownames(data), data)    #保存したい情報を
write.table(tmp, out_f, sep="\t", append=F, quote=F, r
```

①

RGui (64-bit)

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

R Console

```
100  723    840  164    873  957    892 1109
    Male Male Female Male Female Male Female
1     39   13     97   29     40   63     46
100  373  617    458  709    703  411    828
    Female Female Female Female Female Female
1       37    112     55     94     31     73
100    789    911    880   1043    466    551
    Female Male Female Male Female Male Female
1       41   45    120   28      6   61    131
100    584  403    680  635    410  470    664
    Female Female Male
1       44     42   22
100   1072    869  609
> |
```

列名でソート

①赤枠内コピペ実行後。確かに列名でソートできていることがわかります。②order関数を用いた行列のソートテクは、よく利用します。

7. pickrellCountsYaleCQNcommon_esetの場合:

例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニークにしています。
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルはhoge7.txtです。

```
pData(eset)$Gender                    #確認してるだけです
table(pData(eset)$Gender)             #確認してるだけです

#本番(重複除去)
length(rownames(exprs(eset)))         #重複除去前の遺伝
length(unique(rownames(exprs(eset)))) #重複除去前のユニ
hoge <- nsFilter(eset, var.filter=F,  #同一IDの重複除去
                 remove.dupEntrez=T)  #同一IDの重複除去
eset <- hoge$eset                     #esetとして取り扱
length(rownames(exprs(eset)))         #重複除去前の遺伝

#後処理(列名を変更し、列をソート)
data <- exprs(eset)                   #dataとして取り扱
colnames(data) <- pData(eset)$Gender  #列名を変更
head(data[1:2,])                      #確認してるだけで
data <- data[, order(colnames(data))] #列名でソート
head(data[1:2,])                      #確認してるだけで

#ファイルに保存
tmp <- cbind(rownames(data), data)    #保存したい情報を
write.table(tmp, out_f, sep="\t", append=F, quote=F, r
```

②

①

RGui (64-bit)

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

R Console

```
100    873    892    458    703    828    789
    Female Female Female Female Female Female
1      112     55     94     31     73     41
100    911    880   1043    466    551    584
    Female Female Female Female Female Male
1      120      6    131     44     42     38
100    680    410    664   1072    869   621
    Male Male Male Male Male Male Male Male
1      9   15   33   17   39   13   29   63
100  723  164  957 1109  373  617  709  411
    Male Male Male Male
1     45   28   61   22
100  403  635  470  609
> |
```

①hoge7.txtが作成されます。

例題7を最後までコピペ後

7. pickrellCountsYaleCQNcommon_esetの場合:

例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニークにしています。
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルはhoge7.txtです。

```
out_f <- "hoge7.txt"                    #出力ファイル名を指定してout_fに格納

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

#読込(目的のデータセットをロード)
data(commonPickrellHuang)               #paramで指定したデータセットのロード
ls()                                    #利用可能なオブジェクト名を表示
eset <- pickrellCountsYaleCQNcommon_eset#esetとして取り扱う
dim(pData(eset))                        #確認してるだけです
head(pData(eset))                       #確認してるだけです
pData(eset)$Gender                      #確認してるだけです
table(pData(eset)$Gender)               #確認してるだけです

#本番(重複除去)
length(rownames(exprs(eset)))           #重複除去前の遺伝子数を確認
length(unique(rownames(exprs(eset))))   #重複除去前のユニークな遺伝子数を確認
hoge <- nsFilter(eset, var.filter=F,    #同一IDの重複除去のみを実行
                 remove.dupEntrez=T)    #同一IDの重複除去のみを実行
eset <- hoge$eset                       #esetとして取り扱う
length(rownames(exprs(eset)))           #重複除去前の遺伝子数を確認
```

hoge7.txt

これが①Entrez gene IDの重複除去を行って、性別ごとに列をソートした後の発現行列データ。技術的なことに終始している印象を受けるかもしれませんが、ほとんどのヒトは入力ファイルをうまく読み込ませる段階でコケマス。ここあたりのことを正しく理解しておかねば、自身のデータ解析に活かすことができません。

G1群：23 females　　G2群：13 males

11,482 gene IDs

	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Male	Male	Male	Male	Male	Male	Male	Male	Male	Male	Male	Male
1	47	10	53	14	17	54	40	29	97	40	46	37	112	55	94	31	73	41	120	6	131	44	42	38	9	15	33	17	39	13	29	63	45	28	61	22
100	679	677	765	479	745	840	873	892	458	703	828	789	911	880	##	466	551	584	680	410	664	##	869	621	723	164	957	##	373	617	709	411	403	635	470	609
1000	4	2	3	12	11	9	15	1	8	14	3	13	8	31	2	1	2	2	3	11	3	5	2	4	2	1	2	2	2	1	4	36	1	21	2	3
10000	252	66	93	16	152	124	253	96	163	151	200	90	130	218	33	201	214	27	137	158	179	110	159	226	174	45	188	236	165	160	173	220	151	316	169	147
10001	252	86	175	151	269	260	276	239	178	241	257	186	156	194	238	196	207	157	200	221	247	277	234	268	232	53	183	216	240	208	268	357	234	308	255	195
10003	20	2	15	17	3	4	5	2	4	6	9	24	4	2	2	3	3	3	4	6	7	4	21	3	3	2	2	2	34	1	4	24	16	18	4	3
…																																				
9988	260	163	222	111	539	334	443	421	588	352	405	342	308	394	394	180	383	266	529	215	452	423	310	371	368	195	366	367	201	422	328	528	300	354	420	237
9989	202	91	218	98	254	156	212	158	145	248	202	153	109	157	283	110	116	232	115	167	163	296	266	177	262	56	200	143	220	199	167	128	207	256	206	159
999	11	6	5	17	265	12	13	4	90	37	46	169	4	9	276	31	29	214	21	13	195	29	47	29	100	22	457	59	46	433	2	14	13	14	27	3
9990	270	161	127	97	568	260	325	269	401	417	745	567	320	347	429	248	396	342	297	486	703	395	450	338	345	180	845	781	233	476	295	271	222	446	378	233
9991	921	239	557	536	688	710	839	876	444	748	895	468	511	758	581	490	612	444	707	586	498	881	714	720	##	303	919	726	699	563	910	716	655	##	762	734
9993	285	141	244	138	325	240	342	284	210	315	288	249	207	305	371	284	220	259	202	257	264	414	290	331	321	90	281	305	364	276	309	296	332	368	396	170

①

Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行

C1コレクションでGSVA

②遺伝子セット情報は、MSigDB C1コレクションのc1.all.v6.1.entrez.gmtを利用します。これは326遺伝子セットからなります。

(Rで)塩基配列解析

(last modified 2018/06/29, since 2010)

このウェブ
います。初

What's new

• 「解析 | –

- 解析 | 機能解析 | について (last modified 2018/06/24)
- 解析 | 機能解析 | GMTファイル取得 | について (last modified 2018/07/17)
- 解析 | 機能解析 | GMTファイル取得 | EGSEAdata(Alhamdoosh_2017) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | GeneSetDB(Araki_2012) (last modified 2018/06/27)
- 解析 | 機能解析 | GMTファイル取得 | MSigDB(Subramanian_2005) (last modified 2019/07/19) NEW
- 解析 | 機能解析 | GMTファイル読込 | GSEABase(Morgan_2018) (last modified 2018/06/25)
- 解析 | 機能解析 | 遺伝子セット解析 | GSVA(Hänzelmann_2013) (① modified 2018/08/10)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について (last modified 2019/05/12)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | SeqGSEA(Wang_2014)(last modified 2018/06/25)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | GSVA(Hänzelmann_2013)(last modified 2018/06/26)
- 解析 | 機能解析 | 遺伝子オント
- 解析 | 機能解析 | パスウェイ(P

解析 | 機能解析 | 遺伝子セット解析 | GSVA(Hänzelmann_2013) NEW

GSVAを用いて遺伝子セット解析を行うやり方を示します。このデータはAbsFilterGSEA (Yoon et al., PLoS One, 2016)、およびGSVA (Hänzelmann et al., BMC Bioinformatics, 2013)中で、検証用データとして用いられています。具体的には、MSigDBのC1というコレクションに含まれる2つのsex-specificな遺伝子セット(chryq11とchrxp22)が発現変動しているという結果を得ているようです。従って、ここではGSVA実行に必要な2つのファイルのうち、gmtファイルをMSigDBから得られた**326** gene setsからなるc1.all.v6.1.entrez.gmtに固定して、いくつかの例題を示します。尚、GSVA自体はエンリッチメントスコア(Enrithment Score)をサンプルごとに② た結果を返すだけなので、GSVAの実行のみの場合はどのサンプルがどの群に属しているかのグループラ 情報を与える必要はありません。そのため、GSVA実行結果であるoutオブジェクト(フィルタリング後の遺伝子セット数×サンプル数)を入力として、**とりあえず**non-parametricのシンプルなwilcox.testを実行して得られたp-value情報を取得した結果も出力しています。
「ファイル」–「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. 11,482 genes × 36 samplesからなるカウントデータファイル(SRP001540_23_13.txt)の場合:

「カウント情報取得 | リアルデータ | SRP001540 | GSVAdata(Hänzelmann_2013)」の例題7を実行して得られた、23 females vs. 13 malesの2群間比較用データと同じものです。同一gene IDsを重複除去した後のデータです。遺伝子セットchryq11のp値が最も低くなっており、妥当ですね。

```
in_f1 <- "SRP001540_23_13.txt"          #入力ファイル名を指定してin_f1に格納(発現ファイル)
```

①発現データファイルの中身は…

C1コレクションでGSVA

1. 11,482 genes × 36 samplesからなるカウントデータファイル(SRP001540_23_13.txt)の場合:

「カウント情報取得 | リアルデータ | SRP001540 | GSVAdata(Hänzelmann 2013)」の例題7を実行して得られた、23 females vs. 13 malesの2群間比較用データと同じものです。同一gene IDsを重複除去した後のデータです。遺伝子セットchryq11のp値が最も低くなっており、妥当ですね。

```
in_f1 <- "SRP001540_23_13.txt"                  #入力ファイル名を指定してin_f1に格納(発現ファイル)
in_f2 <- "c1.all.v6.1.entrez.gm                 #入力ファイル名を指定してin_f2に格納(gmtファイル)
out_f <- "hoge1.txt"                            #出力ファイル名を指定してout_fに格納
param_G1 <- 23                                  #G1群のサンプル数を指定
param_G2 <- 13                                  #G2群のサンプル数を指定

#必要なパッケージをロード
library(GSVA)                                   #パッケージの読み込み
library(GSEABase)                               #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_f1, header=TRUE, row.names=1, sep="\t", quote="")#in_f1で指定した
geneset <- getGmt(in_f2, geneIdType=EntrezIdentifier(),#in_f2で指定したファイルの読み込
                  collectionType=BroadCollection(category="c1"))#in_f2で指定したファイ
geneset                                         #確認してるだけです

#本番(GSVAの実行)
data <- as.matrix(data)                         #データの型をmatrixにしている
out <- gsva(data, geneset,                      #GSVAの実行
            min.sz=5, max.sz=500, kcdf="Poisson",#GSVAの実行(遺伝子セットのメンバー数が5
            mx.diff=T, verbose=F, parallel.sz=1) #GSVAの実行
dim(out)                                        #確認してるだけです
```

SRP001540_23_13.txt

①Entrez gene IDの重複除去を行い、性別ごとに列をソートした後の発現行列データ

G1群：23 females　　G2群：13 males

11,482 gene IDs

	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Female	Male	Male	Male	Male	Male	Male	Male	Male	Male	Male	Male	Male
1	47	10	53	14	17	54	40	29	97	40	46	37	112	55	94	31	73	41	120	6	131	44	42	38	9	15	33	17	39	13	29	63	45	28	61	22
100	679	677	765	479	745	840	873	892	458	703	828	789	911	880	##	466	551	584	680	410	664	##	869	621	723	164	957	##	373	617	709	411	403	635	470	609
1000	4	2	3	12	11	9	15	1	8	14	3	13	8	31	2	1	2	2	3	11	3	5	2	4	2	1	2	2	2	1	4	36	1	21	2	3
10000	252	66	93	16	152	124	253	96	163	151	200	90	130	218	33	201	214	27	137	158	179	110	159	226	174	45	188	236	165	160	173	220	151	316	169	147
10001	252	86	175	151	269	260	276	239	178	241	257	186	156	194	238	196	207	157	200	221	247	277	234	268	232	53	183	216	240	208	268	357	234	308	255	195
10003	20	2	15	17	3	4	5	2	4	6	9	24	4	2	2	3	3	3	4	6	7	4	21	3	3	2	2	2	34	1	4	24	16	18	4	3
…																																				
9988	260	163	222	111	539	334	443	421	588	352	405	342	308	394	394	180	383	266	529	215	452	423	310	371	368	195	366	367	201	422	328	528	300	354	420	237
9989	202	91	218	98	254	156	212	158	145	248	202	153	109	157	283	110	116	232	115	167	163	296	266	177	262	56	200	143	220	199	167	128	207	256	206	159
999	11	6	5	17	265	12	13	4	90	37	46	169	4	9	276	31	29	214	21	13	195	29	47	29	100	22	457	59	46	433	2	14	13	14	27	3
9990	270	161	127	97	568	260	325	269	401	417	745	567	320	347	429	248	396	342	297	486	703	395	450	338	345	180	845	781	233	476	295	271	222	446	378	233
9991	921	239	557	536	688	710	839	876	444	748	895	468	511	758	581	490	612	444	707	586	498	881	714	720	##	303	919	726	699	563	910	716	655	##	762	734
9993	285	141	244	138	325	240	342	284	210	315	288	249	207	305	371	284	220	259	202	257	264	414	290	331	321	90	281	305	364	276	309	296	332	368	396	170

①

コピペ実行

コード全体をコピペ実行後。全部で326個の遺伝子セットの発現変動解析を行うべく、①入力として与えたが、②遺伝子セットのメンバー数が5以上500以下という条件でフィルタリングすると、③298個の遺伝子セットになったようです

1. 11,482 genes × 36 samplesからなるカウントデータファイル(SRP001540_23_13.txt)の

「カウント情報取得 | リアルデータ | SRP001540 | GSVAdata(Hänzelmann 2013)」の例題 females vs. 13 malesの2群間比較用データと同じものです。同一gene IDsを重複除去したセット chryq11のp値が最も低くなっており、妥当ですね。

```
in_f1 <- "SRP001540_23_13.txt"          #入力ファイル名を指定してin_f1に格納(発現ファイル)
in_f2 <- "c1.all.v6.1.entrez.gmt"       #入力ファイル名を指定してin_f2に格納(gmtファイル)
out_f <- "hoge1.txt"                    #出力ファイル名を
param_G1 <- 23                          #G1群のサンプル数
param_G2 <- 13                          #G2群のサンプル数

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

#入力ファイルの読み込み
data <- read.table(in_f1, header=TRUE, row.names=1, se
geneset <- getGmt(in_f2, geneIdType=EntrezIdentifier()
                  collectionType=BroadCollection(categ
geneset                                 #確認してるだけで

#本番(GSVAの実行)
data <- as.matrix(data)                 #データの型をmat
out <- gsva(data, geneset,              #GSVAの実行
            min.sz=5, max.sz=500, kcdf="Poisson",#GSVA
            mx.diff=T, verbose=F, parallel.sz=1) #GSVA
dim(out)                                #確認してるだけで
```

①

②

RGui (64-bit)

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

R Console

③

```
[1] 298  36
>
> #後処理(wilcox.testを実行)
> data.cl <- c(rep(1, param_G1), rep(2, param_G$
> pvalue <- NULL                          #pvalu$
> for(i in 1:nrow(out)){                  #遺伝$
+   pvalue <- c(pvalue, wilcox.test(out[i, data$
+ }
>
> #ファイルに保存
> tmp <- cbind(rownames(out), out, pvalue)#保存$
> tmp <- tmp[order(pvalue),]              #pvalu$
> write.table(tmp, out_f, sep="\t", append=F, q$
>
```

実行結果ファイル

実行結果ファイルの①hoge1.txt
をエクセルで眺めてみましょう。

1. 11,482 genes × 36 samplesからなるカウントデータファイル(SRP001540_23_13.txt)の場合:

「カウント情報取得 | リアルデータ | SRP001540 | GSVAdata(Hänzelmann 2013)」の例題7を実行して得られた、23 females vs. 13 malesの2群間比較用データと同じものです。同一gene IDsを重複除去した後のデータです。遺伝子セットchryq11のp値が最も低くなっており、妥当ですね。

```
in_f1 <- "SRP001540_23_13.txt"          #入力ファイル名を指定してin_f1に格納(発現ファイル)
in_f2 <- "c1.all.v6.1_entrez.gmt"       #入力ファイル名を指定してin_f2に格納(gmtファイル)
out_f <- "hoge1.txt"                    #出力ファイル名を指定してout_fに格納
param_G1 <- 23                          #G1群のサンプル数を指定
param_G2 <- 13                          #G2群のサンプル数を指定

#必要なパッケージをロード
library(GSVA)                           #パッケージの読み込み
library(GSEABase)                       #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_f1, header=TRUE, row.names=1, sep="\t", quote="")#in_f1で指定した
geneset <- getGmt(in_f2, geneIdType=EntrezIdentifier(),#in_f2で指定したファイルの読み込
                  collectionType=BroadCollection(category="c1"))#in_f2で指定したファイ
geneset                                 #確認してるだけです

#本番(GSVAの実行)
data <- as.matrix(data)                 #データの型をmatrixにしている
out <- gsva(data, geneset,              #GSVAの実行
            min.sz=5, max.sz=500, kcdf="Poisson",#GSVAの実行(遺伝子セットのメンバー数が5
            mx.diff=T, verbose=F, parallel.sz=1) #GSVAの実行
dim(out)                                #確認してるだけです
```

hoge1.txt

GSVA自体はEnrichment scoreを返すだけのプログラム。そのスコアからなる数値ベクトルを入力として、ノンパラメトリックなWilcoxon rank sum test (Mann–Whitney U testと同じもの)で得られた、①p値が最も低い発現変動遺伝子セットは、②chryq11でした。この結果は…

自動保存 オフ

ファイル ホーム 挿入 ページ レイアウト 数式 データ 校閲 表示

AO26

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL
1		Female	Female.1	Female.2	Female.3	Female.4	Female.5	Female.6	Female.7	Female.8	Female.9	Female.10	Female.11	Female.12	Female.13	Female.14	Female.15	Female.16	Female.17	Female.18	Female.19	Female.20	Female.21	Female.22	Male	Male.1	Male.2	Male.3	Male.4	Male.5	Male.6	Male.7	Male.8	Male.9	Male.10	Male.11	Male.12	pvalue
2	chryq11	② 1	-0	0	-1	-1	-1	-1	0	-1	-1	0	0	-1	-1	-0	-0	-0	-0	-0	-0	-1	-1	1	1	1	1	1	1	1	1	0	1	1	1	1	2.60E-08 ①
3	chr5p13	0	-0	-0	0	0	-0	-0	-0	-0	-0	0	-0	-0	-0	-0	-0	-0	-0	0	-0	-0	-0	0	0	0	0	0	0	0	0	0	0	0	0	-0	0	0.00014
4	chrxp22	0	0	0	0	0	0	0	0	0	0	-0	-0	0	0	-0	-0	0	-0	0	0	0	0	-0	-0	-0	-0	-0	-0	0	-0	-0	0	-0	-0	-0	-0	0.00036
5	chrxq26	-0	-0	-0	0	-0	-0	-0	0	0	0	-0	-0	-0	-0	-0	0	0	-0	-0	-0	-0	-0	0	0	0	-0	0	0	0	0	0	0	-0	0	-0	-0	0.00115
6	chr10q26	-0	-0	-0	0	-0	-0	0	-0	0	-0	0	-0	-0	-0	-0	-0	0	-0	0	-0	0	-0	-0	0	0	0	0	0	0	0	0	0	-0	-0	0	0	0.00221
7	chr8q22	-0	-0	0	0	0	-0	-0	-0	-0	0	-0	-0	-0	-0	-0	-0	-0	-0	0	0	-0	-0	-0	0	0	-0	0	0	0	-0	0	-0	0	-0	0	-0	0.00406
8	chr4p16	-0	0	0	-0	0	0	-0	-0	-0	-0	0	0	0	-0	0	0	0	0	-0	-0	0	-0	0	-0	-0	0	-0	-0	-0	-0	-0	-0	0	-0	0	-0	0.00572
9	chr17q23	0	-0	-0	0	0	-0	-0	-0	-0	-0	0	-0	-0	0	0	0	-0	-0	-0	-0	-0	0	0	0	-0	-0	0	0	-0	0	0	0	0	0	0	0	0.00713
10	chr5q11	0	-0	-0	0	0	0	-0	-0	-0	0	-0	0	-0	-0	-0	-0	-0	-0	-0	0	0	-0	-0	0	0	-0	-0	0	0	-0	0	0	0	0	0	-0	0.00795
11	chr15q14	0	-0	-0	0	-0	-0	-0	-0	-0	0	0	-0	0	0	0	-0	-0	0	-0	-0	-0	0	-0	0	0	0	0	0	0	0	0	-0	-0	0	-0	0	0.00982
12	chrxq12	-0	0	0	-0	-0	-0	0	-0	0	0	-0	0	0	-0	-0	0	-0	0	0	0	0	-0	1	-1	-0	-0	0	-0	-0	-0	-0	0	-0	-0	-0	0	0.01623
13	chr18q22	0	-0	0	0	-0	0	-0	-0	0	-0	-0	-0	0	-0	-0	-0	-0	0	-0	0	-0	-0	-0	-0	0	0	0	0	-0	-0	-0	-0	0	1	0	0	0.01623

hoge1

準備完了 100%

• 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | について

同じ結果が得られた

①AbsFilterGSEAの論文の結果と同じですね。②chryq11がおそらく最上位だと思われます。③解析データがちょっと異なりますが、それでも結果は同じ。今回の我々の実行結果も、確かにmaleで高発現になっています。

Comparison of GSEA methods for RNA-seq data

The performances of GSEA methods were compared for published R
eral aspects. First, two RNA-seq datasets denoted by Pickrell and Li data, respectively, were analyzed for comparing power and accuracy as follows:

The Pickrell data were generated from the lymphoblastoid cell lines of 69 unrelated Nigerian individuals (29 male and 40 female) [44]. To analyze the chromosomal differences in expression between male and female, MSigDB C1 (cytogenetic band gene-sets) [45–47] was used for analysis. The GSEA-SP with SNR gene score was applied for the total dataset which resulted in two significant gene-sets 'chryq11' (FDR = 0.00143) and 'chrxp22' (FDR = 0.0514) both of which were sex-specific. These gene-sets were significantly up-regulated in male and female groups, respectively. Since the GSEA-SP controls the false po s well, these two gene-sets were regarded as true positives. Then, five samples were randomly selected from

③ ② ④

ントロジー(GO)解析 | について NEW

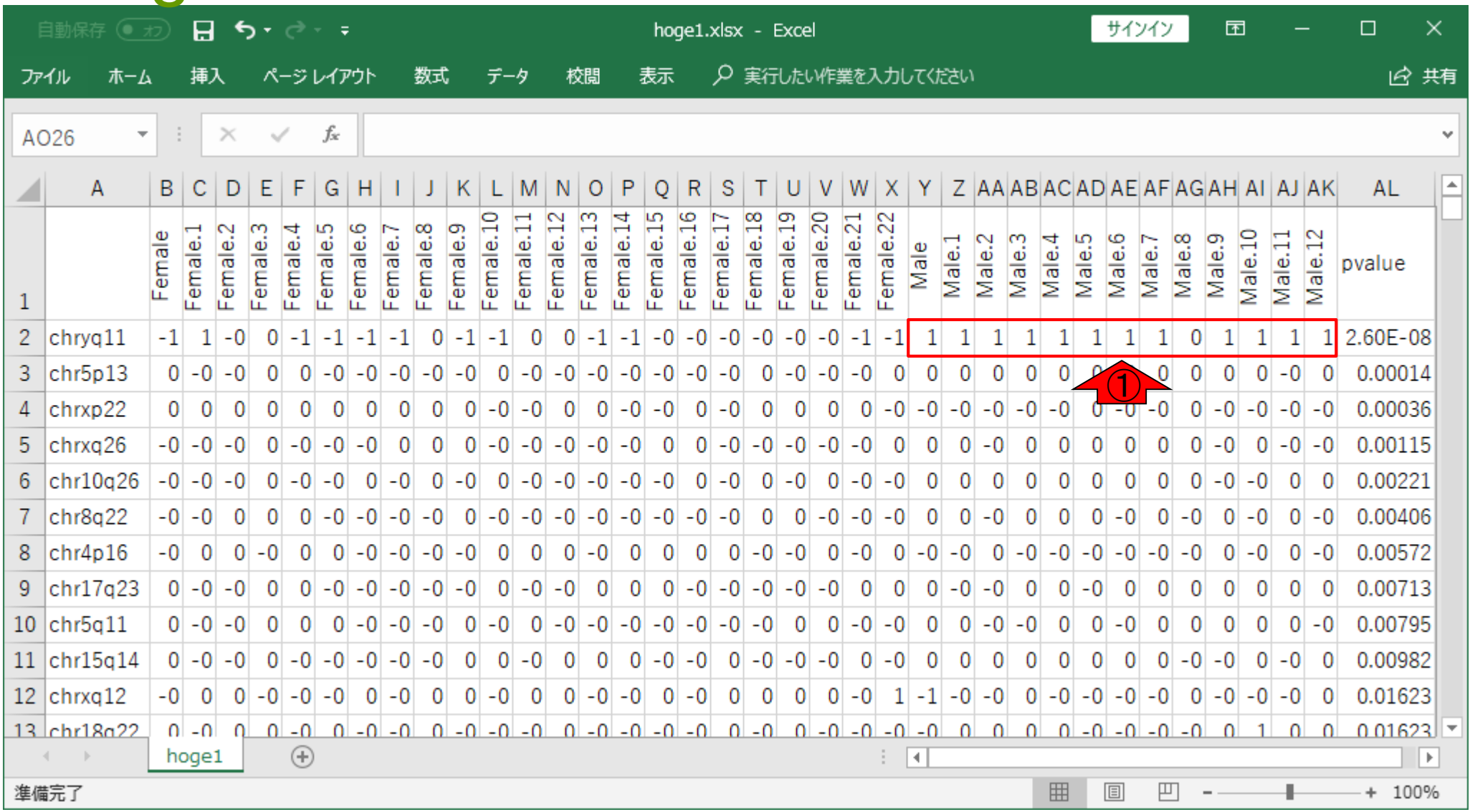
伝子オントロジー(GO)解析を行うためのパッケージもいく
Set Analysis; GSA)という枠組みではGO解析もパスウェ
といいかもしれません。

- GAGE: Luo et al., BMC Bioinformatics, 2009
- goseq: Young et al., Genome Biol., 2010
- GOSemSim: Yu et al., Bioinformatics, 2010
- Rスクリプト: Gao et al., Bioinformatics, 2011
- RamiGO: Schröder et al., Bioinformatics, 2013
- GSVA: Hänzelmann et al., BMC Bioinformatics, 2013
- SeqGSEA(各群5反復以上を要求): Wang et al., Bioinformatics, 2014
- GSAASeqSP: Xiong et al., Sci Rep., 2014
- GOplot(Visualization用): Walter et al., Bioinformatics, 2015
- GOexpress: Rue-Albrecht et al., BMC Bioinformatics, 2016
- rapidGSEA(cudaGSEA and ompGSEA): Hundt et al., BMC Bioinformatics, 2016
- EGSEA: Alhamdoosh et al., Bioinformatics, 2017
- AbsFilterGSEA(small replicates用): Yoon et al., PLoS One, 2016 ①
- GSAR: Rahmatallah et al., BMC Bioinformatics, 2017
- SeqGSA: Ren et al., BioData Min., 2017

AbsFilterGSEA(Yoon et al., PLoS One, 11: e0165919, 2016)

maleで高発現というのは、①の部分のEnrichment scoreがmale群で高いという理解で正しいはずです。

hoge1.txt



hoge1.xlsx - Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL
1		Female	Female.1	Female.2	Female.3	Female.4	Female.5	Female.6	Female.7	Female.8	Female.9	Female.10	Female.11	Female.12	Female.13	Female.14	Female.15	Female.16	Female.17	Female.18	Female.19	Female.20	Female.21	Female.22	Male	Male.1	Male.2	Male.3	Male.4	Male.5	Male.6	Male.7	Male.8	Male.9	Male.10	Male.11	Male.12	pvalue
2	chryq11	-1	1	-0	0	-1	-1	-1	-1	0	-1	-1	0	0	-1	-1	-0	-0	-0	-0	-0	-0	-1	-1	1	1	1	1	1	1	1	1	0	1	1	1	1	2.60E-08
3	chr5p13	0	-0	-0	0	0	-0	-0	-0	-0	-0	0	-0	-0	-0	-0	-0	-0	-0	0	-0	-0	-0	0	0	0	0	0	0	0	0	0	0	0	0	-0	0	0.00014
4	chrxp22	0	0	0	0	0	0	0	0	0	0	-0	-0	0	0	-0	-0	0	-0	0	0	0	0	-0	-0	-0	-0	-0	-0	0	-0	-0	0	-0	-0	-0	-0	0.00036
5	chrxq26	-0	-0	-0	0	-0	-0	-0	0	0	0	-0	-0	-0	-0	-0	0	0	-0	-0	-0	-0	-0	0	0	0	-0	0	0	0	0	0	0	-0	0	-0	-0	0.00115
6	chr10q26	-0	-0	-0	0	-0	-0	0	-0	0	-0	0	-0	-0	-0	-0	-0	0	-0	0	-0	0	-0	-0	0	0	0	0	0	0	0	0	0	-0	-0	0	0	0.00221
7	chr8q22	-0	-0	0	0	0	-0	-0	-0	-0	0	-0	-0	-0	-0	-0	-0	-0	-0	0	0	-0	-0	-0	0	0	-0	0	0	0	-0	0	-0	0	-0	0	-0	0.00406
8	chr4p16	-0	0	0	-0	0	0	-0	-0	-0	-0	0	0	0	-0	0	0	0	0	-0	-0	0	-0	0	-0	-0	0	-0	-0	-0	-0	-0	-0	0	-0	0	-0	0.00572
9	chr17q23	0	-0	-0	0	0	-0	-0	-0	-0	-0	0	-0	-0	0	0	0	-0	-0	-0	-0	-0	0	0	0	-0	-0	0	0	-0	0	0	0	0	0	0	0	0.00713
10	chr5q11	0	-0	-0	0	0	0	-0	-0	-0	0	-0	0	-0	-0	-0	-0	-0	-0	-0	0	0	-0	-0	0	0	-0	-0	0	0	-0	0	0	0	0	0	-0	0.00795
11	chr15q14	0	-0	-0	0	-0	-0	-0	-0	-0	0	0	-0	0	0	0	-0	-0	0	-0	-0	-0	0	-0	0	0	0	0	0	0	0	0	-0	-0	0	-0	0	0.00982
12	chrxq12	-0	0	0	-0	-0	-0	0	-0	0	0	-0	0	0	-0	-0	0	-0	0	0	0	0	-0	1	-1	-0	-0	0	-0	-0	-0	-0	0	-0	-0	-0	0	0.01623
13	chr18q22	0	-0	0	0	-0	0	-0	-0	0	-0	-0	-0	0	-0	-0	-0	-0	0	-0	0	-0	-0	-0	-0	0	0	0	0	-0	-0	-0	-0	0	1	0	0	0.01623

Contents

- 機能解析（発現変動遺伝子セット解析）
  - 全体像、基本的な考え方と解析戦略の変遷、様々なプログラム
  - 遺伝子セット情報の取得（gmtファイルの取得）
  - 発現データ情報と遺伝子セット情報のIDの対応付け
  - 検証用RNA-seqカウントデータセットPickrell data（なぜGSVAにしたか）
  - GSVAの解説PDFを読み解く（手元のc1.all.v6.1.entrez.gmt をどう読み込ませるか）
    - GSVAdataパッケージ提供の、MSigDB c2コレクションであるc2BroadSetsを理解する
    - 手元のgmtファイルを読み込ませて、GeneSetCollection形式で取り扱えるようにする
  - GSVAの解説PDFを読み解く（手元の発現データファイルをどう取り扱うか）
    - ExpressionSetの取り扱い、nsFilter関数を用いた同一IDの重複除去
    - メインプログラムgsva関数が入力として受け付けるデータ形式（ExpressionSetとMatrix）
    - 検証用RNA-seqカウントデータセットPickrell dataのイントロ、スルーしていいところ
    - MSigDB c2コレクションに2つの性特異的遺伝子セットを追加したものでGSVAを実行
  - ユニークなEntrez gene IDで、グループごとに分離させた発現データファイル作成
  - 整形後の発現データファイルとc1.all.v6.1.entrez.gmtを入力としてGSVAを実行
  - 最後に…

最後に

講義日程（2019年度）

1. 2019年07月01日 (PC使用)
   講義資料PDF(最終更新：2019.07.05)
   (Rで)塩基配列解析
   QuasR：Gaidatzis et al., Bioinformatics, 2015
   HTSeq：Anders et al., Bioinformatics, 2015
   hoge10.txt
   htseq-countのページ
   hoge1.gtf
   sample_blekhman_36.txt
   Blekhman et al., Genome Res., 2010
2. 2019年07月08日 (PC使用)
   講義資料PDF(最終更新：2019.07.05)
   (Rで)塩基配列解析
   Blekhman et al., Genome Res., 2010
   TCC：Sun et al., BMC Bioinformatics, 2013
   Tang et al., BMC Bioinformatics, 2015
   Zhao et al., Biol. Proc. Online, 2018
   ReCount(website)：Frazee et al., BMC Bioinformatics, 2011
   recount2(website)：Collado-Torres et al., Nat Biotechnol., 2017
   recount(R package)：Collado-Torres et al., Nat Biotechnol., 2017
   rse_gene.Rdata(SRP001558)
   rse_gene.Rdata(ERP000546)
3. 2019年07月22日 (PC使用)
   講義資料PDF(最終更新：2019.07.17)
   kadai.txt
   (Rで)塩基配列解析
   TCC：Sun et al., BMC Bioinformatics, 2013
   edgeR：Robinson et al., Bioinformatics, 2010
   DESeq：Anders and Huber, Genome Biol, 2010
   DESeq2：Love et al., Genome Biol., 2014
   Blekhman et al., Genome Res., 2010
   Schurch et al., RNA, 2016
4. 2019年07月29日 (PC使用)
   講義資料PDF(最終更新：2019.07.22)
   (Rで)塩基配列解析
   rcode_ORA_basic.txt
   rcode_Pickrell.txt

①rcode_Pickrell.txt内で、Pickrellカウントデータファイル(SRP001540_23_13.txt)を入力として以下の三つを行っています：
・サンプル間クラスタリング（雄雌入り混じった状態）
・シルエットスコアの計算（非常に0に近い値）
・DEG検出（10%FDRで85個をDEGと判定）
このように一見hopelessな印象を受ける結果であったとしても、遺伝子セット解析によってカナリ説得力のある結果が得られることもあるというのが面白いですね

4. 2019年07月29日 (PC使用)
   講義資料PDF(最終更新：2019.07.22)
   (Rで)塩基配列解析
   rcode_ORA_basic.txt
   rcode_Pickrell.txt ①

最後に

講義日程（平成30年度）

1. 平成30年06月12日 (PC使用)
   講義資料PDF(約5MB; 2018.06.12版)
   (Rで)塩基配列解析
   QuasR : Gaidatzis et al., Bioinformatics, 2015
   HTSeq : Anders et al., Bioinformatics, 2015
   hoge10.txt
   htseq-countのページ
   hoge1.gtf
   sample_blekhman_36.txt
   Blekhman et al., Genome Res., 2010
2. 平成30年06月19日 (PC使用)
   講義資料PDF(約4MB; 2018.06.20版)
   (Rで)塩基配列解析
   Blekhman et al., Genome Res., 2010
   TCC : Sun et al., BMC Bioinformatics, 2013
   Tang et al., BMC Bioinformatics, 2015
   Zhao et al., Biol. Proc. Online, 2018
   ReCount(website) : Frazee et al., BMC Bioinformatics, 2011
   平成28年度NGSハンズオン講習会
   recount2(website) : Collado-Torres et al., Nat Biotechnol., 2017
   recount(R package) : Collado-Torres et al., Nat Biotechnol., 2017
   rse_gene.Rdata(SRP001558)
   rse_gene.Rdata(ERP000546)
3. 平成30年06月26日 (PC使用)
   講義資料PDF(約3MB; 2018.06.27版)
   kadai.txt
   (Rで)塩基配列解析
   TCC : Sun et al., BMC Bioinformatics, 2013
   edgeR : Robinson et al., Bioinformatics, 2010
   DESeq : Anders and Huber, Genome Biol, 2010
   DESeq2 : Love et al., Genome Biol., 2014
   Blekhman et al., Genome Res., 2010
   Schurch et al., RNA, 2016
4. 平成30年07月03日 (PC使用)
   講義資料PDF(約4MB; 2018.07.03版)
   (Rで)塩基配列解析
   rcode_ORA_basic.txt
   rcode_Pickrell.txt

①rcode_Pickrell.txt内で、Pickrellカウントデータファイル(SRP001540_23_13.txt)を入力として以下の三つを行っています：
・サンプル間クラスタリング（雄雌入り混じった状態）
・シルエットスコアの計算（非常に0に近い値）
・DEG検出（10%FDRで85個をDEGと判定）
このように一見hopelessな印象を受ける結果であったとしても、遺伝子セット解析によってカナリ説得力のある結果が得られることもあるというのが面白いですね

4. 平成30年07月03日 (PC使用)
   講義資料PDF(約4MB; 2018.07.03版)
   (Rで)塩基配列解析
   rcode_ORA_basic.txt
   rcode_Pickrell.txt ①