

2018.07.03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13-29あたり)」は確実に省略します。スライド125-151についても残り時間次第です

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/>

講義予定

- 第1回(2018年06月12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(2018年06月19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、クラスタリング結果の客観的な評価
- 第3回(2018年06月26日)
 - 発現変動解析(反復あり/なしの2群間比較)、M-A plot
 - 発現変動解析:3群間比較、デザイン行列
- 第4回(2018年07月03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

- 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

(Rで)塩基配列解析

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

このウェブページ
という前提で記述
(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

- 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/isoform | [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 | 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。

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

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

[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)

MSigDB

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)

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。

MSigDB

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、その後の

たものが多く濃子群がDEGのMootha et al. 20はGene Ontologyする遺伝子群の唱えています。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

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①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などが現実により得ます。また、exonやtranscriptを指し示すIDかもしれませんが、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の.entrez.gmt)とgene symbols(ファイル名の最後のほうの.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ぎりぎ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。

1. GOやKEGG
2. マイクロアレイ
3. それらを
4. エンドユーザに提供(

gmtファイルを手

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なくてはならない。

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、その後の

たものが多く濃子群がDEGのMootha et al. 20はGene Ontologyする遺伝子群の唱えています。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

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)です。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などが現実により得ます。また、exonやtranscriptを指し示すIDかもしれませんが、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に.entrez.gmt)とgene symbols(ファイル名の最後のほうに.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ぎりぎ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。

1. GOやKEGG
2. マイクロアレイ
3. それらを
4. エンドユーザを提供(

入力ファイルは2つ

解析 | 機能解析 | について **NEW**

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。

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)を意図した内容です。GSEAは、発現変動解析の核

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などが現実により得ます。また、exonやtranscriptを指し示すIDかもしれませんが、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に.entrez.gmt)とgene symbols(ファイル名の最後のほうに.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

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、その後の発展したものも多く、濃縮因子群がDEGのMootha et al., 2010はGene Ontologyを用いた遺伝子群の唱えています。Subramanian et al. (2005)は機能解析をトとしてKEGG以下に示すよう

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などが現実により得ます。また、exonやtranscriptを指し示すIDかもしれませんが、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に.entrez.gmt)とgene symbols(ファイル名の最後のほうに.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ぎゅ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。

1. GOやKEGG
2. マイクロアレイ
3. それらを
4. エンドユーザを提供(口)

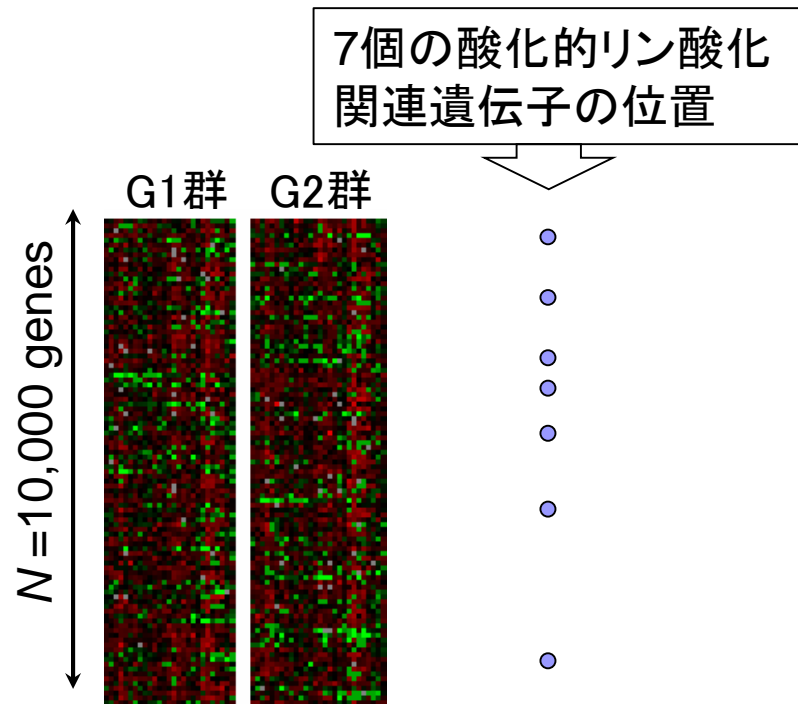
Contents

■ 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

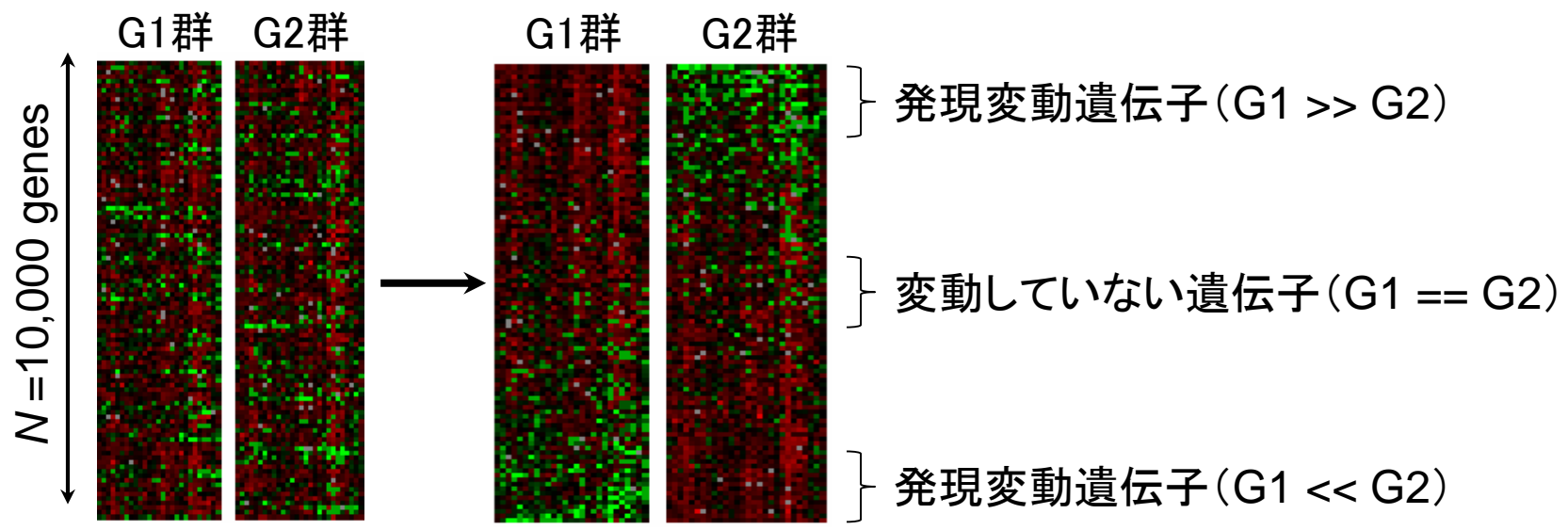
基本的な考え方

- 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: 酸化了的リン酸化 (=XXX) 関連遺伝子が $7 (=n)$ 個含まれている



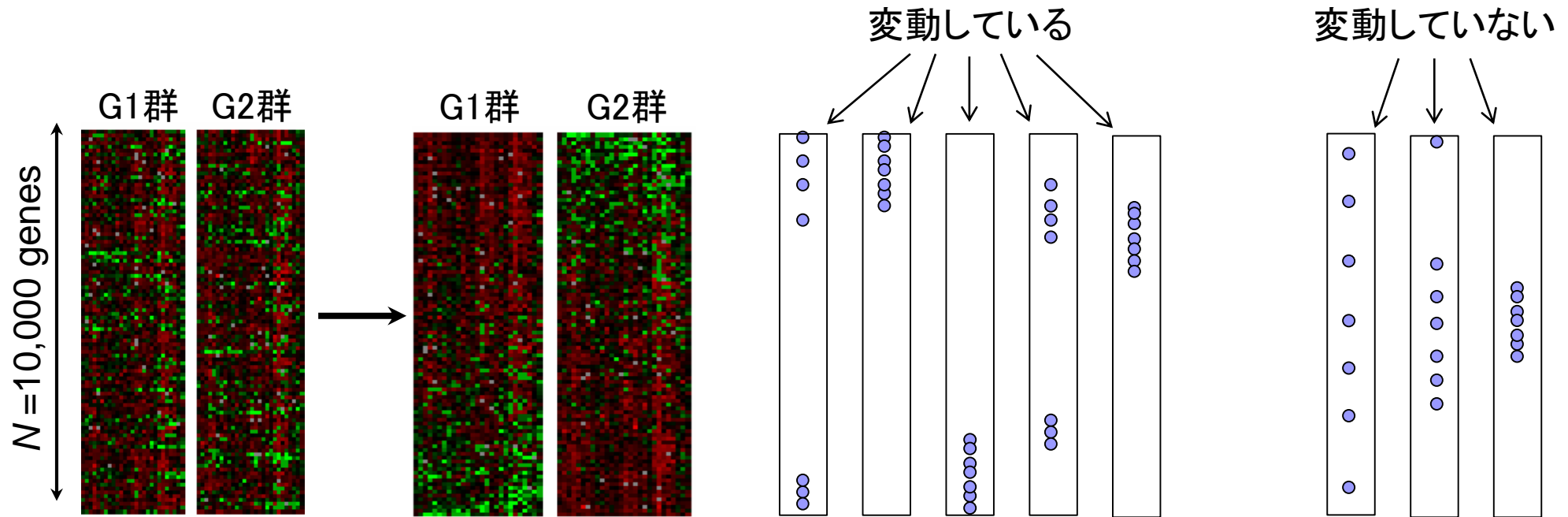
基本的な考え方

- 遺伝子ごとの発現変動の度合いを数値化
 - 例: t-統計量、 $\log_2(G2/G1)$ 、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



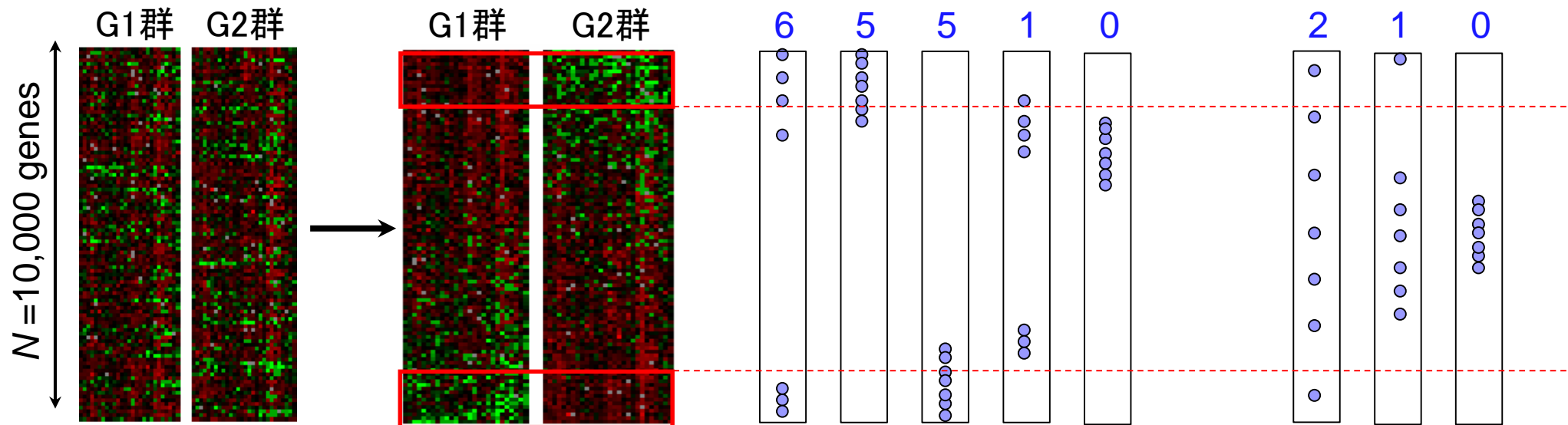
基本的な考え方

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 個が酸化リン酸化関連遺伝子であった

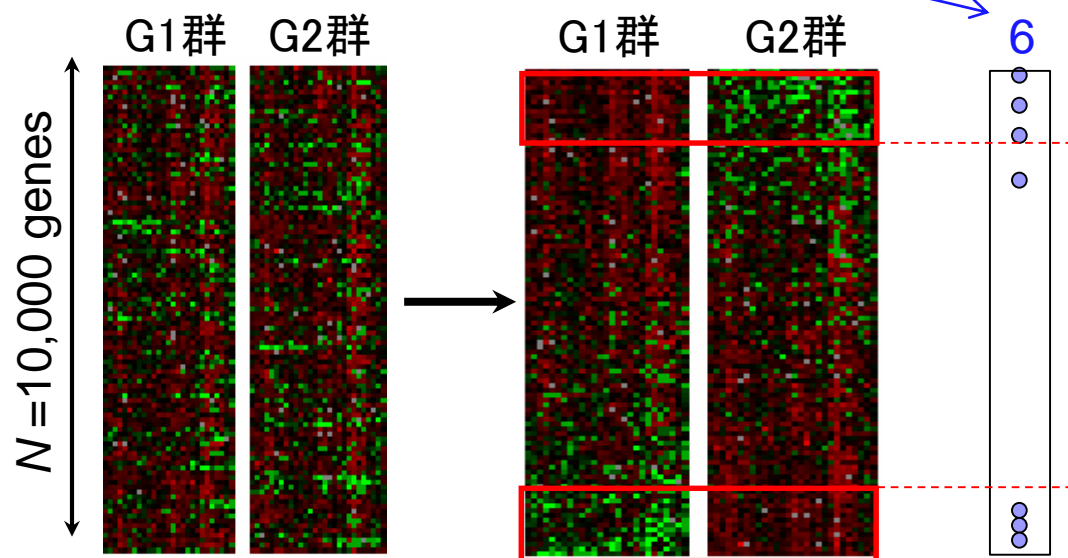
酸化リン酸化関連遺伝子セット ($n=7$) が変動していない場合: $x/n \doteq X/N (= 1500/10000)$
酸化リン酸化関連遺伝子セット ($n=7$) が変動している場合: $x/n \gg X/N (= 15\%)$



ORA

Over-Representation Analysis (ORA)

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



XXX=酸化了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	$N-n$	

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

```
R Console
> N <- 10000
> n <- 7
> X <- 1500
> x <- 6
> sum(dhyper(x=x:X, m=n, n=N-n, k=X))
[1] 6.892847e-05
> |
```

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

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

```
R Console
> ?dhyper
starting httpd help server ... done
> x <- 6
> m <- 7
> n <- 9993
> k <- 1500
> sum(dhyper(x=x:X, m=m, n=n, k=k))
[1] 6.892847e-05
> |
```

ORA (カイ二乗検定)

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
```

警告メッセージ:

```
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個の酸化リン酸化関連遺伝子が含まれていた
- ③ 帰無仮説:酸化リン酸化関連遺伝子の割合はDEGとnon-DEG間で差がない

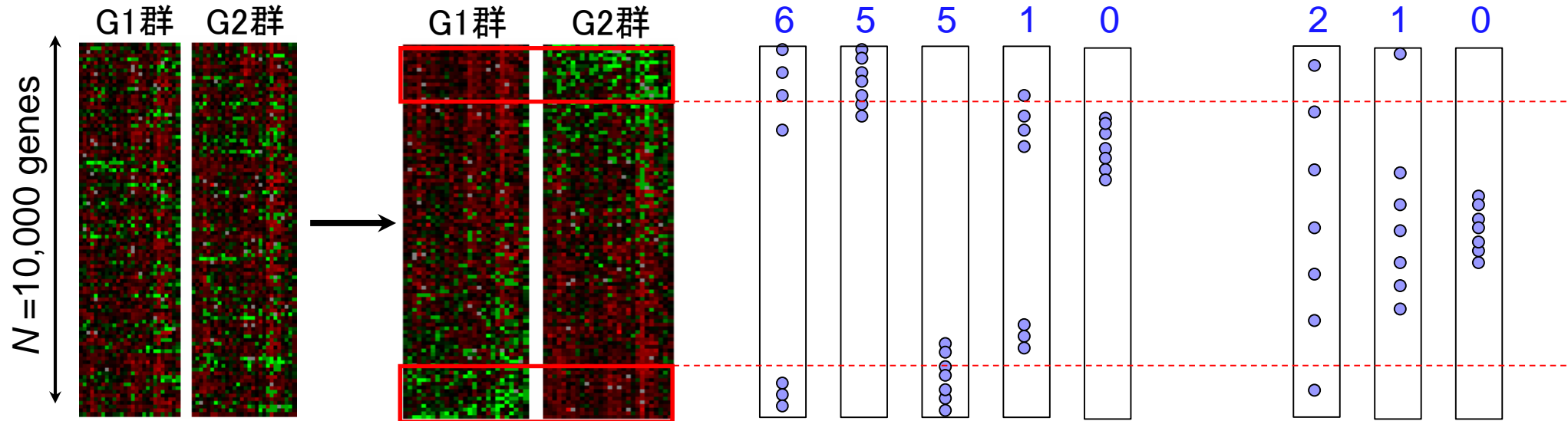
①の段階で、調べたい遺伝子セットは、 $7/10,000 = 0.07\%$ の割合だと考える。②で $6/1,500 = 0.4\%$ の割合に濃縮されていると考える。③今やっているのは、ある2群間比較。もし比較している群間で、この遺伝子セットが全体として発現変動していなかったとしたら…ランダムで1,500個とった時に、この遺伝子セット中の遺伝子が含まれる割合は0.07%なので、個数だと $1500 \times 0.07\% = 1.05$ 個程度しか含まれないはず。実際に得られたのは6個なので、偶然こんな結果が得られたとは考えにくい。起こるとしたら④くらい低い確率なんだね。だから「発現変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

```
R Console
> ?dhype
starting
> x <- 6
> m <- 7
> n <- 9993
> k <- 1500
> sum(dhyper(x=x:X, m=m, n=n, k=k))
④ [1] 6.892847e-05
> |
```

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の意味する

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

ORA

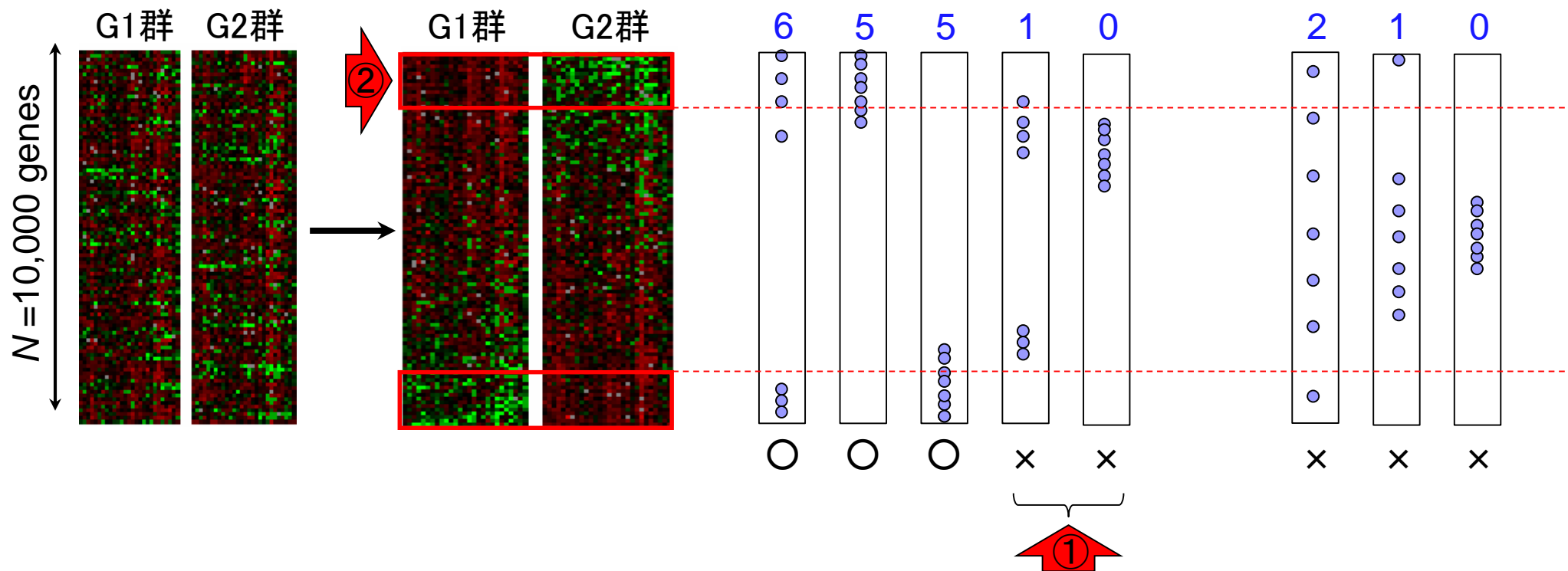
■ Over-Representation Analysis (ORA)

- GenMAPP (Dahlquist et al., *Nature Genet.*, **31**: 19–20, 2002) 
- 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)

	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)

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、 M 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)

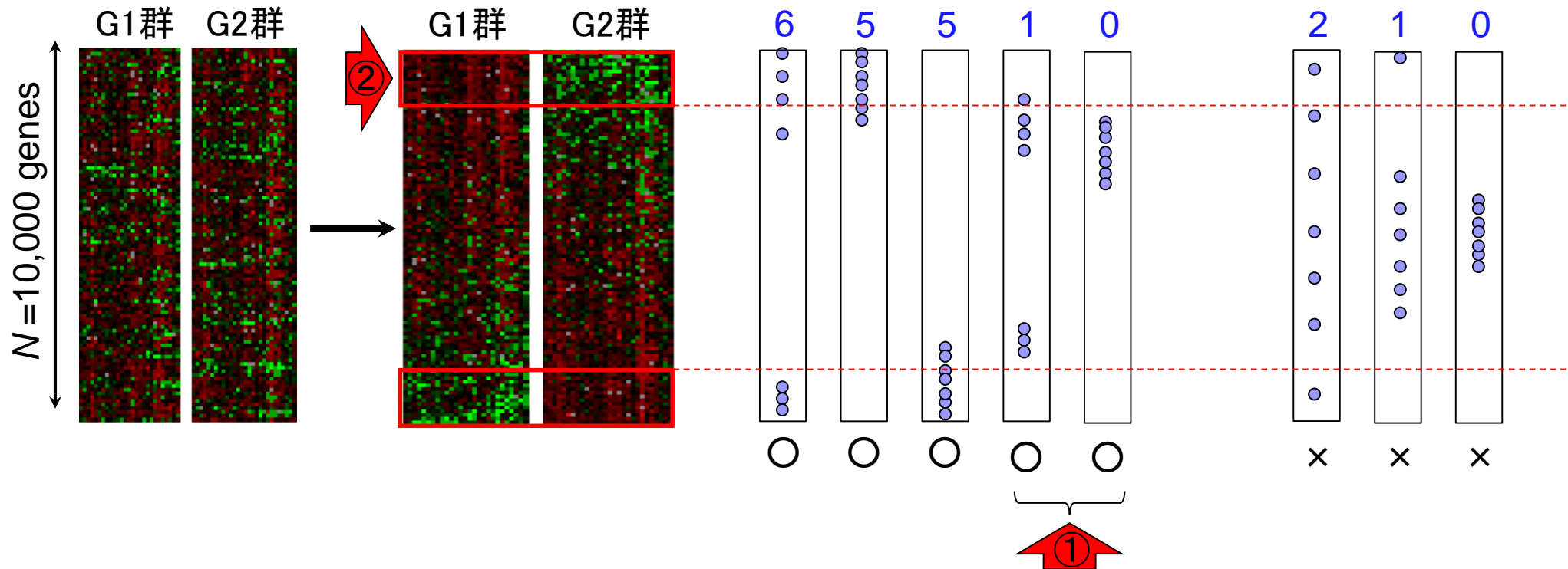
第2世代 (FCS)

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)

	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

①

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

遺伝子セット解析の課題

- (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, inferred 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%)

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

(Rで)塩基配列解析

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

このウェブページ
という前提で記述
(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

- 解析 | 発現変動 | exon/isoform | [DEXseq\(Anders 2012\)](#) (last modified 2014/06/23)
- 解析 | 機能解析 | [GO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) **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\(Hänzelmann 2012\)](#) (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](#) **①** (last modified 2018/06/27) **NEW**
- 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) **NEW**
- 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) **NEW**
- 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](#) (last modified 2018/06/27) **NEW**
- 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 | S](#)
- 解析 | 機能解析 | [パスウェイ\(Pathway\)解析 | C](#)
- 解析 | [分類 | について](#) (last modified 2018/06/2

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](#)
- [SeqGSEA\(各群5反復以上を要求\): Wang et al., Bioinformatics, 2014](#)

GO解析用

解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | につ

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
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
析も同じなので、そちらもチェックするといいかもかもしれません。

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
- [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
- [GSAR](#): Rahmatallah et al., BMC Bioinformatics, 2017
- [SeqGSA](#): Ren et al., BioData Min., 2017

R以外:

- [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
。2017.05.29の講義資料でも解説しています。「
2017.05.29」でページ内検索してください。



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が出ています。遺伝子セット解析(Gene Set Analysis; GSA)という枠組みではGO解析も同じなので、そちらもチェックするといいかもかもしれません。

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
- [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
- [GSAR](#): Rahmatallah et al., *BMC Bioinformatics*, 2017
- [SeqGSA](#): Ren et al., *BioData Min.*, 2017

R以外:

- [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.

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

Alhamdoosh M¹, Ng M¹, Wilson NJ¹, Sheridan JM^{2,3}, Huynh H¹, Wilson MJ¹, 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](http://www.bioconductor.org/packages/EGSEA/).

© The Author 2016. Published by Oxford University Press.

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

SeqGSEA

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 **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
- [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
- [GSAR](#): Rahmatallah et al., BMC Bioinformatics, 2017
- [SeqGSA](#): Ren et al., BioData Min., 2017



R以外:

- [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

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 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립ト](#): Gao et al., Bioinformatics, 2011
- [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, 2013
- [GSASeqSP](#): 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, 2017
- [EGSEA](#): Alhamdoosh et al., Bioinformatics, 2017 ②
- [AbsFilterGSEA](#)(small replicates用): Yoon et al., PLoS One, 2017
- [GSAR](#): Rahmatallah et al., BMC Bioinformatics, 2017
- [SeqGSA](#): Ren et al., BioData Min., 2017

R以外:

- [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 S¹, 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

FREE Full text

Save items

★ Add to Favorites

Similar articles

A flexible count data model to fit the wide [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 c [Biotechniques. 2003]

[See reviews...](#)

[See all...](#)

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

(Rで)塩基配列解析

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

このウェブページ
という前提で記述
(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](#)

- ・ [解析 | 発現変動 | exon/isoform | DEXseq\(Anders 2012\)](#) (last modified 2014/06/23)
- ・ [解析 | 機能解析 | について](#) (last modified 2018/06/24) **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 | 遺伝子オントロジー\(GO\)解析 | SeqGSEA\(Wang 2014\)](#)(last modified 2018/06/25) **NEW**
- ・ [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) **NEW**
- ・ [解析 | 機能解析 | 遺伝子オントロジー\(GO\)解析 | GOseq\(Li 2010\)](#) (last modified 2010/11/26)
- ・ [解析 | 機能解析 | パスウェイ\(Pathway\)解析 | について](#) (last modified 2018/06/27) **NEW**
- ・ [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 2018/06/25) **NEW**
- ・ [解析 | 機能解析 | パスウェイ\(Pathway\)解析 | GSVA\(Hänzelmann 2013\)](#)(last modified 2018/06/26) **NEW**
- ・ [解析 | 機能解析 | パスウェイ\(Pathway\)解析 | GOseq\(Li 2010\)](#) (last modified 2010/11/26)
- ・ [解析 | 分類 | について](#) (last modified 2018/06/27)

解析 | 機能解析 | パスウェイ(Pathway)解析 | について **NEW**

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。このページでは、[SeqGSEA](#)の解析が一通りできるようにしてあります(2014/03/30)。入力のカウントデータは、[Ensembl](#) (Zerbino et al., *Nucleic Acids Res.*, 2018) のアノテーション情報がおそらくデファクトスタンダードだと思います。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かもしれません。

R用:

- ・ [KEGGgraph: Zhang et al., Bioinformatics, 2009](#)
- ・ [GAGE: Luo et al., BMC Bioinformatics, 2009](#)
- ・ [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](#)

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 **NEW**

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かもしれません。

R用:

- [KEGGgraph: Zhang et al., Bioinformatics, 2009](#)
- [GAGE: Luo et al., BMC Bioinformatics, 2009](#)
- [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 2018/06/26, since 2010)

このウェブページ
という前提で記述
(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

- 解析 | 発現変動 | exon/isoform | [DEXseq\(Anders 2012\)](#) (last modified 2014/06/23)
- 解析 | 機能解析 | [| について](#) (last modified 2018/06/24) **NEW**
- 解析 | 機能解析 | [GMTファイル取得 | について](#) **①** (last modified 2018/06/27) **NEW**
- 解析 | 機能解析 | [GMTファイル取得 | EGSEAdat\(Alhamdoosh 2017\)](#) (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 | GSEA\(Hänzelmann 2013\)](#) (last modified 2018/06/26) **NEW**
- 解析 | 機能解析 | [遺伝子オンロジー\(GO\)解析 | について](#) (last modified 2018/06/27) **NEW**

解析 | 機能解析 | GMTファイル取得 | について **NEW**

[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)



- [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](#)

①最も有名なのはMSigDB。②2018年6月現在の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。

MSigDB ver. 6.1

(Rで)塩基配列解析

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

このウェブページ
という前提で記述
(2015/04/03)

- 解析 | 発現変動 | exon/isoform | [DEXseq\(Anders 2012\)](#) (last modified 2014/06/23)
- 解析 | 機能解析 | [| について](#) (last modified 2018/06/24) **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 | [遺伝子オンロジー\(GO\)解析 | | について](#) (last modified 2018/06/27) **NEW**
- 解析 | 機能解析 | 遺伝子オンロジー(GO)解析 | [SeqGSEA\(Wang 2014\)](#)(last modified 2018/06/25) **NEW**
- 解析 | 機能解析 | 遺伝子オンロジー(GO)解析 | [GSVA\(Hänzelmann 2013\)](#)(last modified 2018/06/26) **NEW**

What's new?

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

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ンロードした場合は、[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。

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

MSigDB ver. 6.1

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. 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)

①発現変動と関連するKEGGパスウェイを調べたいときは、3番目のc2というカテゴリーに属するCP: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。



MSigDB ver. 6.1

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. 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)

①発現変動と関連するGOのbiological processを調べたいときは、6番目のc5というカテゴリーに属するBP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

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
なので、これをダウンロードしておきます。

MSigDB ver. 6.1

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. 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

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ンロードした場合は、[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](#))

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。

③

②

c1.all.v6.1.entrez.gmt

① c1.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。

	A	B	C	D	E	F	G	H
1	chr5q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr5q23	5759	94033	51334	153163	133615	402229
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

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。

	A	B	C	D	E	F	G	H
1	chr5q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr5q23	5759	94033	51334	153163	133615	402229
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

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、1番目の遺伝子の遺伝子セット(chr5q23)よりも構成メンバー数が多いことがわかります。

自動保存 c1.all.v6.

ファイル ホーム 挿入 ページレイアウト 数式 テータ 校閲

CZ28

	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													

c1.all.v6.1.entrez

準備完了 - 100%



chr5q23

Excel screenshot showing a table of genomic data. The title bar indicates the file is 'c1.all.v6.1.entrez.gmt'. The ribbon shows 'ホーム' (Home) and '挿入' (Insert) tabs. The formula bar shows 'J18'. The table has columns A through H. A red arrow points to the URL in cell B1.

	A	B	C	D	E	F	G	H
1	chr5q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr5q23	5759	94033	51334	153163	133615	402229
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

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。

Excel spreadsheet showing a list of gene sets in columns A and B. The spreadsheet is titled "c1.all.v6.1.entrez.gmt".

	A	B	C	D	E	F	G	H	
1	chr5q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr5q23	chr5q23	5759	94033	51334	153163	133615	402229
2	chr16q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr16q24							
3	chr8q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr8q24							
4	chr13q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q11							
5	chr7p21	http://www.broadinstitute.org/gsea/msigdb/cards/chr7p21							
6	chr10q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q23							
7	chr13q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q13							
8	chr10q21	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q21							
9	chr1p13	http://www.broadinstitute.org/gsea/msigdb/cards/chr1p13							
10	chrxp21	http://www.broadinstitute.org/gsea/msigdb/cards/chrxp21							
11	chr4q12	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q12							
12	chr6q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr6q13							
13	chr2p22	http://www.broadinstitute.org/gsea/msigdb/cards/chr2p22							
14	chr18q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr18q23							
15	chr8p11	http://www.broadinstitute.org/gsea/msigdb/cards/chr8p11							

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	

Show membersで...

The screenshot shows the MSigDB website interface. On the left, a table lists various gene sets with columns for chromosome/region and a URL. On the right, a detailed view for the gene set 'chr5q23' is shown, including fields like Standard name, Systematic name, Brief description, Full description or abstract, Collection, Source publication, Exact source, Related gene sets, External links, Organism, Contributed by, Source platform, Dataset references, Download gene set, Compute overlaps, Compendia expression profiles, Advanced query, Gene families, Show members, and Version history. A red arrow points to the 'Show members' link in the 'Show members' field. Below the table, there is a row of numerical data: 138050, 6770, 79698, 25960, 793, 347028.

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

See [MSigDB license terms here](#). Please note that certain gene sets have special access terms.

chr8p11	138050	6770	79698	25960	793	347028
---------	--------	------	-------	-------	-----	--------

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

Show membersで...

自動保存 c1.all.v6.1.e

ファイル ホーム 挿入 ページレイアウト 数式 テータ 校閲 表

J18

	A	B
1	chr5q23	http://www.broadinstitute.org/gsea/msigdb/cards/
2	chr16q24	http://www.broadinstitute.org/gsea/msigdb/cards/
3	chr8q24	http://www.broadinstitute.org/gsea/msigdb/cards/
4	chr13q11	http://www.broadinstitute.org/gsea/msigdb/cards/
5	chr7p21	http://www.broadinstitute.org/gsea/msigdb/cards/
6	chr10q23	http://www.broadinstitute.org/gsea/msigdb/cards/
7	chr13q13	http://www.broadinstitute.org/gsea/msigdb/cards/
8	chr10q21	http://www.broadinstitute.org/gsea/msigdb/cards/
9	chr1p13	http://www.broadinstitute.org/gsea/msigdb/cards/
10	chrxp21	http://www.broadinstitute.org/gsea/msigdb/cards/
11	chr4q12	http://www.broadinstitute.org/gsea/msigdb/cards/
12	chr6q13	http://www.broadinstitute.org/gsea/msigdb/cards/
13	chr2p22	http://www.broadinstitute.org/gsea/msigdb/cards/
14	chr18q23	http://www.broadinstitute.org/gsea/msigdb/cards/
15	chr8p11	http://www.broadinstitute.org/gsea/msigdb/cards/

準備完了

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	(hide 86 members)

Original Member	Entrez Gene Id	Gene Symbol	Gene Description
ACTBP4	64	ACTBP4	actin, beta pse
ADAMTS2	9509	ADAMTS2	ADAM metallo
ADRA1B	147	ADRA1B	adrenergic, alp
ANKRD43	134548	ANKRD43	ankyrin repeat
ARGFXP1	503583	ARGFXP1	arginine-fifty h
CAMLG	819	CAMLG	calcium modul
CCDC100	153241	CEP120	centrosomal p

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。

Excel spreadsheet showing MSigDB data for Subramanian 2005. The spreadsheet has columns A through H. Row 1 is highlighted, and a red arrow points to the cell containing 'chr5q23' (labeled ①). Another red arrow points to the cell containing '5759' (labeled ②). A third red arrow points to the cell containing 'PTMAP2' (labeled ③). A pop-up window shows a list of genes with their Entrez IDs and symbols, with red arrows pointing to '5759' and 'PTMAP2'.

Gene	Entrez ID	Gene Symbol
NEUROG1	4762	NEUROG1
NME5	8382	NME5
NRG2	9542	NRG2
PACAP		
PGGT1B	5229	PGGT1B
PPIC	5480	PPIC
PPP2CA	5515	PPP2CA
PRR16	51334	PRR16
PRRC1	133619	PRRC1
PTMAP2	5759	PTMAP2
RAD50	10111	RAD50
RNF14	9604	RNF14
RNUXA	51808	PHAX

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。

自動保存 c1.all.v6.1.symbols.gmt. - 保存しました サインイン

ファイル ホーム 挿入 ページレイアウト 数式 テータ 校閲 表示 実行したい作業を入力してください 共有

	A	B	C	D	E	F	G	H																																																				
1	chr5q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr5q23	PTMAP2	IT	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-AS1																																																				
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	<table border="1"> <thead> <tr> <th>NEUROG1</th> <th>4762</th> <th>NEUROG1</th> <th>neurogenin 1</th> </tr> </thead> <tbody> <tr> <td>NME5</td> <td>8382</td> <td>NME5</td> <td>non-metastatic</td> </tr> <tr> <td>NRG2</td> <td>9542</td> <td>NRG2</td> <td>neuregulin 2</td> </tr> <tr> <td colspan="4">PACAP</td> </tr> <tr> <td>PGGT1B</td> <td>5229</td> <td>PGGT1B</td> <td>protein geranyl</td> </tr> <tr> <td>PPIC</td> <td>5480</td> <td>PPIC</td> <td>peptidylprolyl</td> </tr> <tr> <td>PPP2CA</td> <td>5515</td> <td>PPP2CA</td> <td>protein phosphatase 2A</td> </tr> <tr> <td>PRR16</td> <td>51334</td> <td>PRR16</td> <td>proline rich 16</td> </tr> <tr> <td>PRRC1</td> <td>133619</td> <td>PRRC1</td> <td>proline-rich coiled-coil 1</td> </tr> <tr> <td>PTMAP2</td> <td>5759</td> <td>PTMAP2</td> <td>prothymosin, alpha</td> </tr> <tr> <td>RAD50</td> <td>10111</td> <td>RAD50</td> <td>RAD50 homolog</td> </tr> <tr> <td>RNF14</td> <td>9604</td> <td>RNF14</td> <td>ring finger protein 14</td> </tr> <tr> <td>RNUXA</td> <td>51808</td> <td>PHAX</td> <td>phosphorylated</td> </tr> </tbody> </table>		NEUROG1	4762	NEUROG1	neurogenin 1	NME5	8382	NME5	non-metastatic	NRG2	9542	NRG2	neuregulin 2	PACAP				PGGT1B	5229	PGGT1B	protein geranyl	PPIC	5480	PPIC	peptidylprolyl	PPP2CA	5515	PPP2CA	protein phosphatase 2A	PRR16	51334	PRR16	proline rich 16	PRRC1	133619	PRRC1	proline-rich coiled-coil 1	PTMAP2	5759	PTMAP2	prothymosin, alpha	RAD50	10111	RAD50	RAD50 homolog	RNF14	9604	RNF14	ring finger protein 14	RNUXA	51808	PHAX	phosphorylated				
NEUROG1	4762	NEUROG1			neurogenin 1																																																							
NME5	8382	NME5			non-metastatic																																																							
NRG2	9542	NRG2			neuregulin 2																																																							
PACAP																																																												
PGGT1B	5229	PGGT1B			protein geranyl																																																							
PPIC	5480	PPIC			peptidylprolyl																																																							
PPP2CA	5515	PPP2CA			protein phosphatase 2A																																																							
PRR16	51334	PRR16			proline rich 16																																																							
PRRC1	133619	PRRC1			proline-rich coiled-coil 1																																																							
PTMAP2	5759	PTMAP2	prothymosin, alpha																																																									
RAD50	10111	RAD50	RAD50 homolog																																																									
RNF14	9604	RNF14	ring finger protein 14																																																									
RNUXA	51808	PHAX	phosphorylated																																																									
7	chr13q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr13q13																																																										
8	chr10q21	http://www.broadinstitute.org/gsea/msigdb/cards/chr10q21																																																										
9	chr1p13	http://www.broadinstitute.org/gsea/msigdb/cards/chr1p13																																																										
10	chrxp21	http://www.broadinstitute.org/gsea/msigdb/cards/chrxp21																																																										
11	chr4q12	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q12																																																										
12	chr6q13	http://www.broadinstitute.org/gsea/msigdb/cards/chr6q13																																																										
13	chr2p22	http://www.broadinstitute.org/gsea/msigdb/cards/chr2p22																																																										
14	chr18q23	http://www.broadinstitute.org/gsea/msigdb/cards/chr18q23																																																										
15	chr8p11	http://www.broadinstitute.org/gsea/msigdb/cards/chr8p11	HGSNAT	STAR	ZMAT4	GPR124	CAI B1	AFG3L2B																																																				

準備完了

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	NXP2	LOC72932	LOC44201	RAB1A
327								

326 gene setsなので...

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

	A	B	C	D	E	F	G	H
313	chr9p24	http://www.broadinstitute.org/gsea/msigdb/cards/chr9p24	729691	3641	414153	1E+08	11172	392285
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	727738
316	chr16q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr16q11	81831	79190	10265	441768	440375	388272
317	chr8p22	http://www.broadinstitute.org/gsea/msigdb/cards/chr8p22	137868	9788	2137	7991	11178	4481
318	chr14q24	http://www.broadinstitute.org/gsea/msigdb/cards/chr14q24	6547	122953	1112	317729	4053	729941
319	chr1q12	http://www.broadinstitute.org/gsea/msigdb/cards/chr1q12	6275	171419	54109	6279	728759	26861
320	chr15q21	http://www.broadinstitute.org/gsea/msigdb/cards/chr15q21	84419	6095	283665	8669	283659	145788
321	chr11p13	http://www.broadinstitute.org/gsea/msigdb/cards/chr11p13	119722	646077	4005	2001	25758	5080
322	chr17p12	http://www.broadinstitute.org/gsea/msigdb/cards/chr17p12	1770	9482	6143	136	10750	1352
323	chr4q32	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q32	57630	91351	152940	133332	201931	646865
324	chr4q11	http://www.broadinstitute.org/gsea/msigdb/cards/chr4q11	81608	6779	170825	3815	6767	3791
325	chr14q31	http://www.broadinstitute.org/gsea/msigdb/cards/chr14q31	730034	11099	54207	2957	730121	866
326	chr2p14	http://www.broadinstitute.org/gsea/msigdb/cards/chr2p14	57142	151516	11249	729324	442019	5861
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、本家もそうになっています。

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ンロードした場合は、[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

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、とりま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として、その後のものが多く濃子群がDEGのMootha et al. 2004はGene Ontology唱しています。Subramanian et al. 2005は機能解析トとしてKEGG以下に示すよう

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などが現実により得ます。また、exonやtranscriptを指し示すIDかもしれませんが、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ぎりぎ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。

1. GOやKEGG
2. マイクロアレイ
3. それらを
4. エンドユーザを提供(

機能解析の全体像

解析 | 機能解析 | について **NEW**

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。

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)を意図した内容です。GSEAは、発現変動解析の枠組みに属するものとして、その後のものが多く濃子群がDEGのMootha et al. 20はGene Ontologyする遺伝子群の唱えています。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

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などが現実により得ます。また、exonやtranscriptを指し示すIDかもしれませんが、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は.entrez.gmt)とgene symbols(ファイル名の最後のほうは.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。

1. GOや①
2. マイクロアレイ
3. それらを
4. エンドユーザを提供(

gene symbolsの場合

解析 | 機能解析 | について **NEW**

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。

多少間違えているかもしれませんが、とりあえず2018年6月現在の私の解析という項目ですが、実質的にはGene Set Enrichment Analysis (GSEA)2

組みに属するものとして、その後のものが多く濃子群がDEGのMootha et al. 20はGene Ontologyする遺伝子群の唱えています。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

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などが現実により得ます。また、exonやtranscriptを指し示すIDかもしれませんが、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は.entrez.gmt)とgene symbols(ファイル名の最後のほうは.symbols.gmt)の2種類が提供されています。①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。

1. GOやKEGG
2. マイクロアレイ
3. それらを
4. エンドユーザを提供(口

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 2018/06/29, since 2010)

このウェブページのR
います。初心者の方

- 解析 | 機能解析 | GMTファイル読込 | [GSEABase\(Morgan 2018\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | 遺伝子セット解析 | [GSVA\(Hänzelmann 2013\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | [GOseq\(Young 2010\)](#) (last modified 2018/06/27)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | [SeqGSEA\(Wang 2014\)](#) (last modified 2018/06/27)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | [GSVA\(Hänzelmann 2013\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | [GOseq\(Young 2010\)](#) (last modified 2018/06/27)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | [PathwayTools](#)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | [KEGG](#)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | [BioGRID](#)



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](#)
- [SeqGSEA\(各群5反復以上を要求\): Wang et al., Bioinformatics, 2014](#)
- [GSASeqSP: 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](#)



Pickrell data

Comparison of GSEA methods for RNA-seq data ②

The performances of GSEA methods were compared for published RNA-seq data from several aspects. First, two RNA-seq datasets denoted by Pickrell and Li data, respectively, were analyzed 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

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。

遺伝子オントロジー(GO)解析 | について NEW

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ありますが、Set Analysis; GSA)という枠組みではGO解析もパスウェイ解析もできないかもしれません。

- 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
- GSASeqSP: 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)

MSigDB C1

Comparison of GSEA methods for RNA-seq data

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

① 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, ② 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: ③ 'chryq11' (FDR = 0.0143) and 'chrp22' (FDR = 0.0514) both of which were sex-specific. These two gene-sets were ④ 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

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。

遺伝子オントロジー(GO)解析 | について NEW

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ありますが、Set Analysis; GSA)という枠組みではGO解析もパスウェイというよりも、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
- [GSASeqSP](#): 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

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

AbsFilterGSEA

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
- [SeqGSEA](#)(各群5反復以上を要求): Wang et al., Bioinformatics, 2014
- [GSASeqSP](#): 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

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

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説が難解です。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

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:

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.

オントロジー(GO)解析 | について NEW

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あるが、Gene Set Analysis; GSA)という枠組みではGO解析もパスウェルといいかもしれません。

[informatics, 2009](#)

[ol, 2010](#)

[matics, 2010](#)

[matics, 2011](#)

[ormatics, 2013](#)

[Bioinformatics, 2013](#)

求): [Wang et al., Bioinformatics, 2014](#)

[ep., 2014](#)

[et al., Bioinformatics, 2015](#)

[BMC Bioinformatics, 2016](#)

(GSEA): [Hundt et al., BMC Bioinformatics, 2016](#)

[informatics, 2017](#)

用): [Yoon et al., PLoS One, 2016](#)

[Bioinformatics, 2017](#)

[n., 2017](#)

65919, 2016)

GSVA

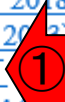
②GSVAは引用回数も多く、③Bioconductorのパッケージであったことが決め手。さらに解説PDFを読むと、Pickrellデータも例題として使われていることが判明。③をクリック

(Rで)塩基配列解析

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

このウェブページのR...
います。初心者の方

- 解析 | 機能解析 | GMTファイル読込 | [GSEABase\(Morgan 2018\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | 遺伝子セット解析 | [GSVA\(Hänzelmann 2013\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | [GSVA\(Hänzelmann 2013\)](#) (last modified 2018/06/27)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | [SeqGSEA\(Wang 2014\)](#) (last modified 2018/06/27)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | [GSVA\(Hänzelmann 2013\)](#) (last modified 2018/06/27)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | [GOseq\(Young 2010\)](#) (last modified 2018/06/27)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | [PathwayTools](#) (last modified 2018/06/27)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | [KEGG](#) (last modified 2018/06/27)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | [KEGG](#) (last modified 2018/06/27)



What's new?

- 「解析 | 一般 | Seq

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](#)
- [SeqGSEA\(各群5反復以上を要求\): Wang et al., Bioinformatics, 2014](#)
- [GSASeqSP: 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](#)



GSEA

The screenshot shows the Bioconductor website for the GSEA package. A red arrow labeled '2' points to the 'Bioconductor' logo in the top navigation bar. The page content includes the GSEA title, version information (3.7), a description of the method, authors, and installation instructions.

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](#)
- [GSEA: Hänzelmann et al., BMC Bioinformatics, 2013](#)
- [SeqGSEA\(各群5反復以上を要求\): Wang et al., Bioinformatics, 2014](#)
- [GSASeqSP: 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](#)

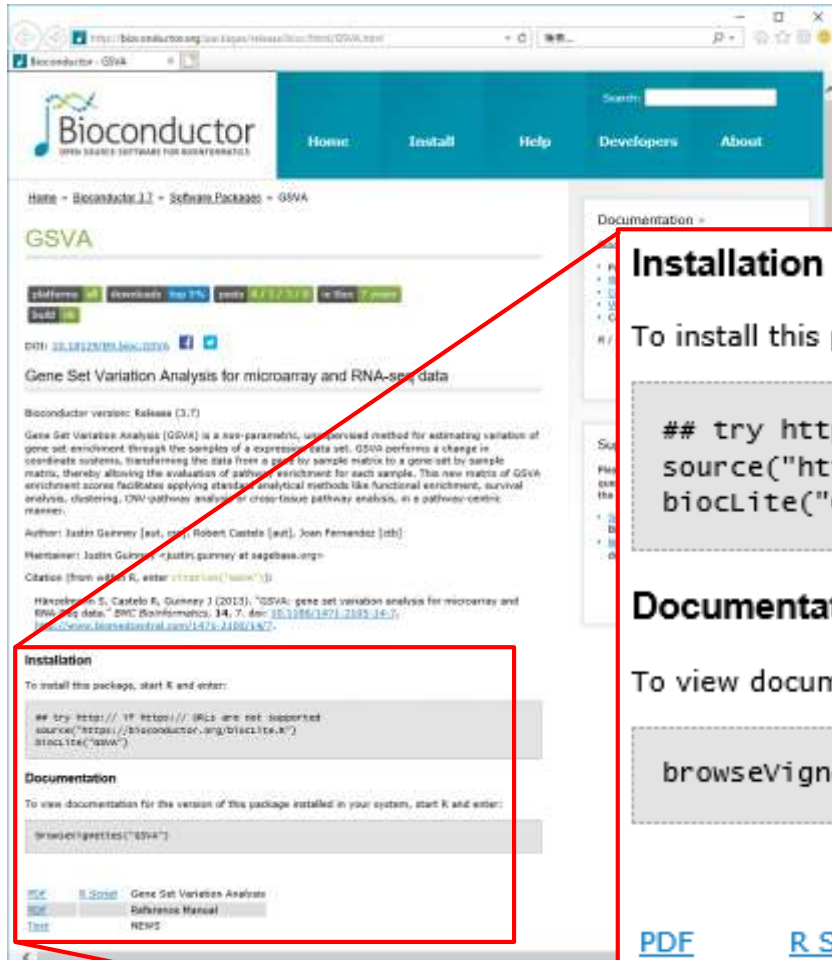


Contents

■ 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



Installation

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("Gsva")
```

Documentation

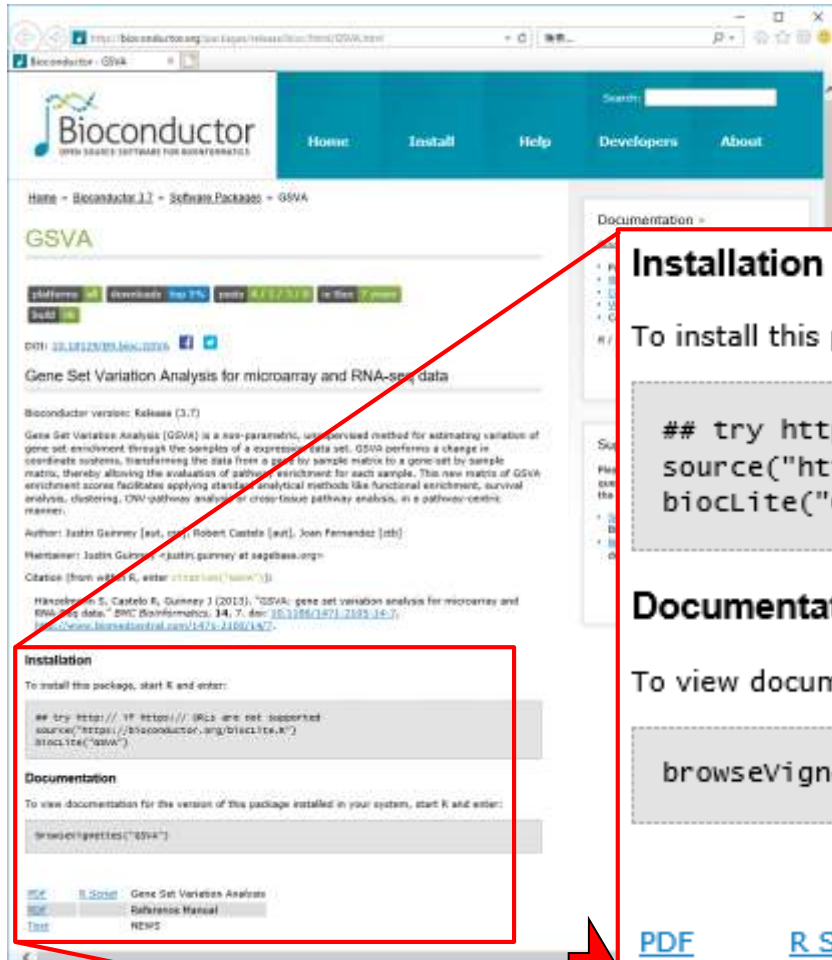
To view documentation for the version of this package installed in your system, start R and enter:

```
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 and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("GSVA")
```



Documentation

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

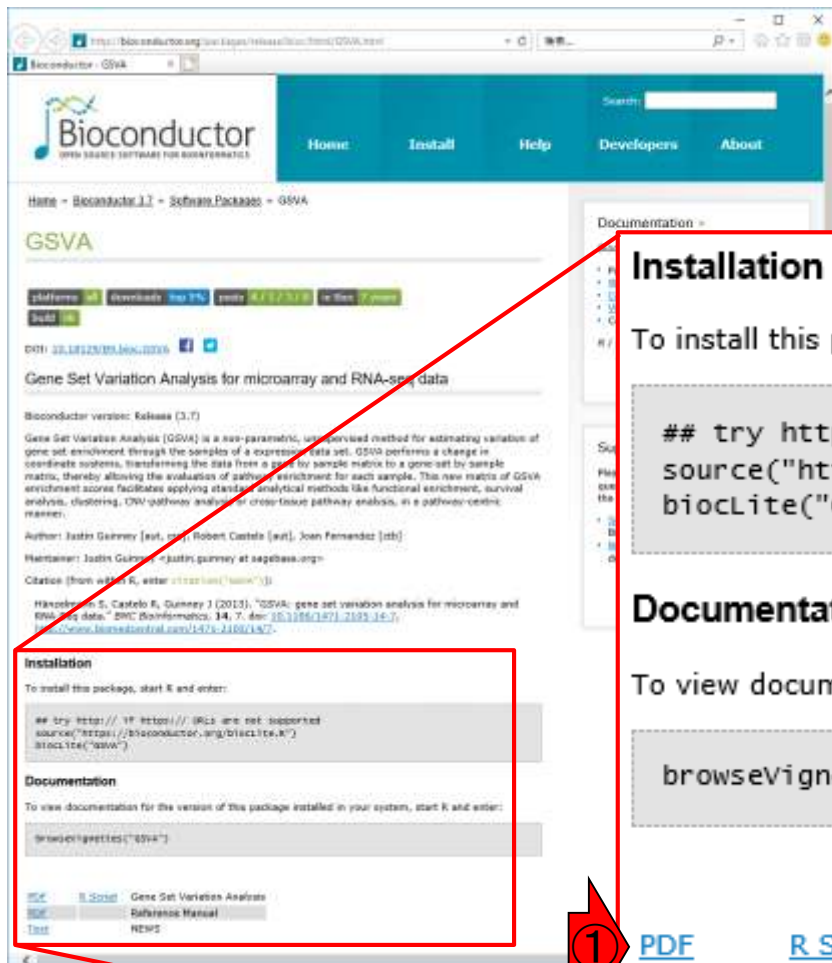
```
browseVignettes("GSVA")
```



- [PDF](#) [R Script](#) Gene Set Variation Analysis
- [PDF](#) Reference Manual
- [Text](#) NEWS

GSVAの解説PDF

①GSVAの解説PDF(bignettes; 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をコピペでも(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 and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("GSVA")
```

Documentation

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

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



[PDF](#)
[PDF](#)
[Text](#)

[R Script](#)

Gene Set Variation Analysis
Reference Manual
NEWS



GSVAの解説PDF

①GSVAの解説PDF(bignettes; 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をコピーでも(手元のPC内に存在する)解説PDFを開けます。③こんな感じ。

```
R Console  
> 'citation()' と入力してください。  
  
> 'demo()' と入力すればデモをみることができます$  
> 'help()' とすればオンラインヘルプが出ます。  
> 'help.start()' で HTML ブラウザによるヘルプが$  
> 'q()' と入力すれば R を終了します。  
  
> browseVignettes("GSVA")  
starting httpd help server ... done  
> |
```

```
enter:  
  
URLs are not supported  
(r.org/biocLite.R")
```

Citation (from within R, enter `citation("GSVA")`)

Härdge E, S, Castelo F, Guinney J (2013). "GSVA: gene set variation analysis for microarrays and RNA-seq data." *BMC Bioinformatics* 14, 7. doi: 10.1186/1471-2105-14-7. <https://doi.org/10.1186/1471-2105-14-7>.

Installation
To install this package, start R and enter:

```
# We try http:// if https:// URLs are not supported  
source("https://bioconductor.org/biocLite.R")  
biocLite("GSVA")
```

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

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の解説PDF

①GSVAの解説PDF(bignettes; ビニエツト)を起動して、利用法を学んでいく。例えば②のところを押すと、③のところ飛びます。基本的には解説PDF内を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る。

The screenshot shows a PDF viewer interface. At the top, the address bar contains the URL <http://127.0.0.1:16111/library/GSVA/doc/GSVA.pdf>. A red arrow labeled '1' points to this address bar. On the left sidebar, under the 'しおり' (Bookmarks) section, the 'Introduction' link is highlighted. A red arrow labeled '2' points to this link. The main content area displays the title 'GSVA: The Gene Set Variation Analysis package for microarray and RNA-seq data' and the authors 'Sonja Hänzelmann¹, Robert Castelo¹ and Justin Guinney²'. Below the title, the date 'October 30, 2017' is shown. The 'Abstract' section is visible, starting with 'The GSVA package implements a non-parametric unsupervised method, called Gene Set Variation Analysis (GSVA)'. A red arrow labeled '3' points to the '1 Introduction' section header at the bottom of the page.

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

①Applicationのあたりまでページ下部に移動。このパッケージは、遺伝子セット情報として②MSigDBのC2コレクションを利用していることがわかる。

http://127.0.0.1:16111/library/GSVA/doc/GSVA.pdf

127.0.0.1

6 / 21

100%

サインイン

4 Applications

In this section we illustrate the following applications of GSVAdoc:

- 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 `GSVAdoc`, which stores these and other data employed in this vignette. These data can be loaded as follows:

```
> library(GSEABase)
> library(GSVAdoc)
> 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)
```

210 x 297 mm

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。

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 contains these and other data covered in this vignette. These data can be loaded as follows:

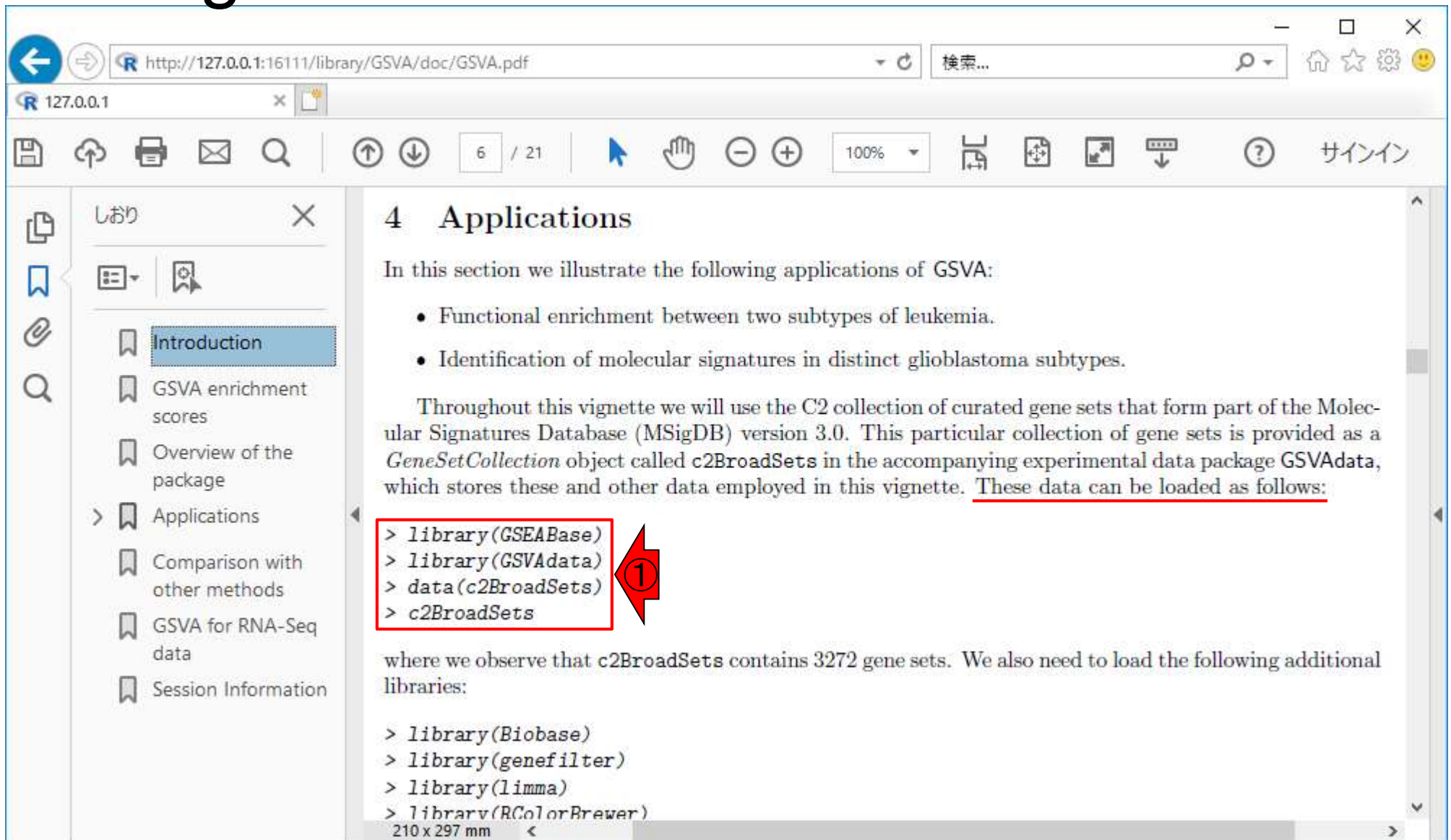
```
> library(GSEABase)
> library(GSVAdata)
> data(c2BroadSets)
> c2BroadSets
```

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



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
```

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

The image shows a web browser window displaying a PDF document titled "4 Applications". The document content includes a list of applications of GSVA: Functional enrichment, Identification of molecular signatures, and a vignette on using the GeneSetCollection object. Below this, there is a list of R libraries to load: Biobase, genefilter, limma, and RColorBrewer. A red box highlights the first four lines of the R console output: library(GSEABase), library(GSVAdata), data(c2BroadSets), and c2BroadSets. A red arrow with the number 1 points to this box.

Overlaid on the RGui window is a context menu with the following options and keyboard shortcuts:

コピー	Ctrl+C
ペースト	Ctrl+V
コマンドのみペースト	②
コピー & ペースト	Ctrl+X
ウインドウの消去	Ctrl+L
全て選択	
パッファに出力	Ctrl+W
ウインドウを常にトップに置く	

A red arrow with the number 2 points to the "コマンドのみペースト" option.

c2BroadSetsの確認

こんな感じになり、①GeneSetCollectionという形式のc2BroadSetsを無事読み込みました。

The screenshot shows the RGui interface with the R Console window open. The console displays the following commands and output:

```
> library(GSEAData)
1: パッケージ 'annotate' はバージョン 3.4.4 の R $
2: パッケージ 'XML' はバージョン 3.4.4 の R の下で$
> library(GSVAdat)
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, NAKAMUR$
unique identifiers: 5167, 100288400, ..., 57191 (29$
types in collection:
  geneIdType: EntrezIdentifier (1 total)
  collectionType: BroadCollection (1 total)
> |
```

Red arrows and a circled '1' highlight the `> library(GSEADat)` and `> data(c2BroadSets)` commands in the console, and the `> library(GSVAdat)` command in the background document.

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。

The screenshot shows the RStudio interface. On the left, a web browser displays a document titled "4 Applications". The document content includes a list of applications and a section titled "Throughout this...". A red box highlights the following commands in the browser's text area:

```
> library(GSEABase)
> library(GSVAdat
> data(c2BroadSe
> c2BroadSets
```

A red arrow with the number "1" points to the first command in the R Console window. The R Console shows the following output:

```
1: パッケージ 'annotate' はバージョン 3.4.4 の R $
2: パッケージ 'XML' はバージョン 3.4.4 の R の下で$
> library(GSVAdat)
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, NAKAMUR$
unique identifiers: 5167, 100288400, ..., 57191 (29$
types in collection:
  geneIdType: EntrezIdentifier (1 total)
  collectionType: BroadCollection (1 total)

> |
```


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	402229
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

?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。

The screenshot shows a web browser window displaying a PDF document titled "4 Applications". The document content includes a section titled "In this section we" with a list of bullet points: "Functional en" and "Identification". Below this, there is a paragraph starting with "Throughout thi" and "ular Signatures Da". A table of contents on the left lists sections like "Introduction", "GSVA enrichment scores", "Overview of the package", "Applications", "Comparison with other methods", "GSVA for RNA-Seq data", and "Session Information".

The R Console window shows the following commands and output:

```
1: パッケージ 'annotate' はバージョン 3.4.4 の R $
2: パッケージ 'XML' はバージョン 3.4.4 の R の下で$
> 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, NAKAMUR$
unique identifiers: 5167, 100288400, ..., 57191 (29$
types in collection:
  geneIdType: EntrezIdentifier (1 total)
  collectionType: BroadCollection (1 total)

> ?c2BroadSets|
```

?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。

http://127.0.0.1:16111/library/GSVAdata/html/c2BroadSets.html

検索...

R 127.0.0.1:16111/library/GSVAdata/html/c2BroadSets.html

R: C2 collection of canonica... x

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.

?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。

The screenshot shows the R documentation page for the `c2BroadSets` package. The page title is "c2BroadSets {GSVAdata}" and the subtitle is "C2 collection of canonical pathways from MSigDB 3.0". The "Description" section states "C2 Broad Sets." The "Usage" section shows the command `data(c2BroadSets)`. The "Details" section contains the text: "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." Five red arrows with numbers 1 through 5 point to specific parts of this text: 1 points to `c2BroadSets`, 2 points to `GeneSetCollection`, 3 points to `GSEABase`, 4 points to `getGmt()`, and 5 points to `c2.all.v3.0.entrez.gmt`. The "Source" section at the bottom lists "Subramanian, Tamayo, et al. *PNAS*, 102:15545-15550, 2005."

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。

gmtファイルの読込

(Rで)塩基配列解析

- 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) **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 2018/06/29) **NEW**

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ー。

② 1. [MSigDB](#)から得られた326 gene setsからなるgmtファイルの読込の基本形です。最後のgenesetオブジェクトの確認が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 ウィンドウ 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

解析 | 機能解析 | GMTファイル読込 | GSEABase

コピー実行結果。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。

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ー。

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

基本形です。最後のgenesetオブジェクトの確認で、geneIdType 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 ウィンドウ 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) ②
unique identifiers: TMAP2, FTMT, ..., GCA (30010 total)
types in collection:
  geneIdType: NullIdentifier (1 total)
  collectionType: NullCollection (1 total)
> |
```

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。

The screenshot shows an Excel spreadsheet with the following data:

	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-AS1
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	CTTNBP2M	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-AS1	ASTRN	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	CAI B1	AFG3L2B

c1.all.v6.1.symbols.gmt

これは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-AS1
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	CTTNBP2M	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	CYP11B1-AS1	ASTRN	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	CAI B1	AFG3L2B

c1.all.v6.1.symbols.gmt

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…

The screenshot shows a spreadsheet with the following data:

	A	B	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-AS1
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	CTTNBP2M	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-AS1	ASTRN	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	CAI B1	AFG3L2B

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。

例題1: 基本形

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ー。

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

基本形です。最後のgenesetオブジェクトの確認で、geneIdType 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 #パッケ  
library(GSEABase)  
#入力ファイルの読み込み #in_fで  
geneset <- getGmt(in_f) #確認し  
geneset
```

```
RGui (64-bit)  
ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ Vignettes  
R Console  
[4] "share"  
> 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  
GeneSetCollection  
names: chr5q23, chr16q11, ..., chr2p14 (326 t$)  
unique identifiers: PTMAP2, FTMT, ..., GCA (30010 t$)  
types in collection:  
  geneIdType: NullIdentifier (1 total)  
  collectionType: NullCollection (1 total)  
> |
```


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ではそうになっていなかった。

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ー。

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

基本形です。最後のgenesetオブジェクトの確認で、geneIdType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
#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$
```

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

R Console
[4] "share"
> in_f <- "c1.all.v6.1.symbols.gmt" #入力ファイル$
>
> #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 total)
types in collection:
  geneIdType: NullIdentifier (1 total)
  collectionType: NullCollection (1 total)
> c2BroadSets|
```

両者を比較

解析 | 機能解析 | GMTファイル読込 | GSEABase

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ー。

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

基本形です。最後のgenesetオブジェクトの確認で、geneIdType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
#必要なパッケージをロード
library(GSEABase) #パッケージをロード
#入力ファイルの読み込み
geneset <- getGmt(in_f) #in_fで指定したファイルを読み込み
geneset #確認
```

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、GSEAdata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。

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

R Console
> 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 total)
types in collection:
  geneIdType: NullIdentifier (1 total)
  collectionType: NullCollection (1 total)
> c2BroadSets #確認$
GeneSetCollection
names: NAKAMURA_CANCER_MICROENVIRONMENT_UP, NAKAMURA_CANCER_MICROENVIRONMENT_DOWN, ... (29 total)
unique identifiers: 5167, 100288400, ..., 57191 (29 total)
types in collection:
  geneIdType: EntrezIdentifier (1 total)
  collectionType: BroadCollection (1 total)
> |
```

例題3

(Rで)塩基配列解析

- (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) **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 2018/06/29) **NEW**

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ー。

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のところが無効Identifierに、そしてcollectionTypeのところが無効Collectionになっているのがわかります。

```
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
```

2. 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み
geneset #確認してるだけです
```

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:

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み
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C1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。

```
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み
geneset #確認してるだけです
```

① ↓
② ↓
③ ↑

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。これで見ただ目上は、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=EntrezIdentifier,
                  collectionType=BroadCollection)
geneset
```

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

R Console
> in_f <- "c1.all.v6.1.entrez.gmt" #入力ファイル$
>
> #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$
+                  #確認してる$
> geneset
GeneSetCollection
names: chr5q23, chr16q24, ..., chr2p14 (326 total)
unique identifiers: 5759, 94033, ..., 25801 (30012 $
types in collection:
  geneIdType: EntrezIdentifier (1 total)
  collectionType: BroadCollection (1 total)
> |
```

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のデータの様ですね。

4.1 Functional enrichment

In this section we illustrate how to identify functionally enriched gene sets between two phenotypes. As in most of the applications we start by calculating GSVA enrichment scores and afterwards, we will employ the linear modeling techniques implemented in the `limma` package to find the enriched gene sets.

The data set we use in this section corresponds to the microarray data from [Armstrong et al., 2002](#) which consists of 37 different individuals with human acute leukemia, where 20 of them have conventional childhood acute lymphoblastic leukemia (ALL) and the other 17 are affected with the MLL (mixed-lineage leukemia gene) translocation. This leukemia data set is stored as an `ExpressionSet` object called `leukemia` in the `GSVAdata` package and details on how the data was pre-processed can be found in the corresponding help page. Enclosed with the RMA expression values we provide some metadata including the main phenotype corresponding to the leukemia sample subtype.

```
> data(leukemia)
> leukemia_eset
```

ExpressionSet (storageMode: lockedEnvironment)
assayData: 12626 features, 37 samples
element names: exprs
protocolData
sampleNames: CL2001011101AA.CEL CL2001011102AA.CEL

ExpressionSet

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！

http://127.0.0.1:16111/library/GSVA/doc/GSVA.pdf

127.0.0.1 R: C2 collection of canonical p...

6 / 21 100%

しおり

- Introduction
- GSVA enrichment scores
- Overview of the package
- Applications
- Comparison with other methods
- GSVA for RNA-Seq data**
- Session Information

4.1 Functional enrichment

In this section we illustrate how to identify functionally enriched gene sets between two phenotypes. As in most of the applications we start by calculating GSVA enrichment scores and afterwards, we will employ the linear modeling techniques implemented in the `limma` package to find the enriched gene sets.

The data set we use in this section corresponds to the microarray data from [\(Armstrong et al., 2002\)](#) which consists of 37 different individuals with human acute leukemia, where 20 of them have conventional childhood acute lymphoblastic leukemia (ALL) and the other 17 are affected with the MLL (mixed-lineage leukemia gene) translocation. This leukemia data set is stored as an `ExpressionSet` object called `leukemia` in the `GSVAdata` package and details on how the data was pre-processed can be found in the corresponding help page. Enclosed with the RMA expression values we provide some metadata including the main phenotype corresponding to the leukemia sample subtype.

```
> data(leukemia)
> leukemia_eset
```

ExpressionSet (storageMode: lockedEnvironment)
assayData: 12626 features, 37 samples
 element names: exprs
protocolData
 sampleNames: CL2001011101AA.CEL CL2001011102AA.CEL

6

210 x 297 mm

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。

4.1 Functional enrichment

In this section we illustrate how to identify functionally enriched gene sets between two phenotypes. As in most of the applications we start by calculating GSVa enrichment scores and afterwards, we will employ the linear modeling techniques implemented in the limma package to find the enriched gene sets.

The data set we use in this section corresponds to the microarray data from [Armstrong et al., 2002](#) which consists of 37 different individuals with human acute leukemia, where 20 of them have conventional childhood acute lymphoblastic leukemia (ALL) and the other 17 are affected with the MLL (mixed-lineage leukemia gene) translocation. This leukemia data set is stored as an ExpressionSet object called leukemia in the GSVa data package and details on how the data was pre-processed can be found in the corresponding help page. Enclosed with the RMA expression values we provide some metadata including the main phenotype corresponding to the leukemia sample subtype.

```
> data(leukemia)
> leukemia_eset
```

① ExpressionSet (storageMode: lockedEnvironment)
assayData: 12626 features, 37 samples
element names: exprs
protocolData
sampleNames: CL2001011101AA.CEL CL2001011102AA.CEL

6 **③**

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。

The screenshot shows a PDF viewer displaying a document page. The page content includes a paragraph of text and a block of R code. Annotations are present:

- A red arrow labeled ① points to the number '7' in the text.
- A red arrow labeled ② points to the variable `leukemia_eset` in the R code.
- A red arrow labeled ③ points to the function `nsFilter` in the R code.

The R code block is as follows:

```
> filtered_eset <- nsFilter(leukemia_eset, require.entrez=TRUE, remove.dupEntrez=TRUE,
+                           var.func=IQR, var.filter=TRUE, var.cutoff=0.5, filterByQuant:
+                           feature.exclude="^AFFX")
> filtered_eset
```

The output of the R code is:

```
$eset
210 x 297 mm
```

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。

The screenshot shows a PDF viewer displaying a document page. The page content includes the following text:

We carry out a non-specific filter... ability, probesets without Entrez ID... the annotation, and Affymetrix quality control probes.

7

```
> filtered_eset <- nsFilter(leukemia_eset, require.entrez=TRUE, remove.dupEntrez=TRUE,  
+ var.func=IQR, var.filter=TRUE, var.cutoff=0.5, filterByQuant:  
+ feature.exclude="^AFFX")  
> filtered_eset
```

\$eset
210 x 297 mm

Red arrows point to the following parts of the code:

- ② points to `leukemia_eset`
- ③ points to `require.entrez=TRUE`
- ④ points to `remove.dupEntrez=TRUE`
- ① points to `feature.exclude="^AFFX"`

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

GSVAの入力

http://127.0.0.1:16111/library/GSVA/doc/GSVA.pdf

検索...

127.0.0.1 R: C2 collection of canonical p...

8 / 21 100%

サインイン

しおり

- Introduction
- GSVA enrichment scores
- Overview of the package
- Applications
- Comparison with other methods
- GSVA for RNA-Seq data
- Session Information

We carry out a non-specific filtering step by discarding the 50% of the probesets with smaller variability, probesets without Entrez ID annotation, probesets whose associated Entrez ID is duplicated in the annotation, and Affymetrix quality control probes:

7

```
> filtered_eset <- nsFilter(leukemia_eset, require.entrez=TRUE, remove.dupEntrez=TRUE,
+                           var.func=IQR, var.filter=TRUE, var.cutoff=0.5, filterByQuant:
+                           feature.exclude="^AFFX")
> filtered_eset
```

\$eset
210 x 297 mm

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

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

GSVAの入力

The screenshot shows a PDF viewer displaying the GSVA documentation. The browser address bar shows the URL: `http://127.0.0.1:16111/library/GSVA/doc/GSVA.pdf`. The document title is "R: C2 collection of canonical p...". The viewer interface includes a toolbar with navigation icons and a search bar. The left sidebar shows a table of contents with "Applications" selected. The main content area shows the following R code and text:

```
> 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:

8

ExpressionSet形式

重複除去の実行結果は、①filtered_eset。① filtered_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入力のようにです。③ leukemia_filtered_esetは、ExpressionSet形式です

The screenshot shows a web browser displaying a PDF document. The browser's address bar shows the URL: `http://127.0.0.1:16111/library/GSVA/doc/GSVA.pdf`. The browser's title bar indicates the page is from "127.0.0.1" and the document is "R: C2 collection of canonical p...". The browser's toolbar includes navigation icons and a search bar. The PDF content is displayed in a white area with a dark border. On the left side, there is a sidebar with a search icon and a list of document sections: "しおり", "Introduction", "GSVA enrichment scores", "Overview of the package", "Applications", "Comparison with other methods", "GSVA for RNA-Seq data", and "Session Information". The main content area shows R code and text. The R code is:

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

 The text below the code reads: "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." Below this text is another R code block:

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

 The text below the code reads: "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:" The page number "8" is visible at the bottom of the PDF content. Red arrows with circled numbers 1 through 5 point to specific elements: 1 points to `filtered_eset`, 2 points to `$eset`, 3 points to the page number "8", 4 points to `leukemia_filtered_eset` in the text, and 5 points to `leukemia_filtered_eset` in the code.

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。

> leukemia_filtered_eset <- f

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:

8

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。

The screenshot shows a web browser displaying a PDF document titled "GSVA.pdf". The browser's address bar shows the URL `http://127.0.0.1:16111/library/GSVA/doc/GSVA.pdf`. The PDF content includes the following text and code:

```
> 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 using a simple linear model and moderated t-statistics computed by the `limma` Bayes shrinkage method (see Smyth, 2004). We are going to examine both changes at pathway level and since, as we shall see below, there are plenty employ the following stringent cut-offs to attain a high level of statistical an

Red arrows point to the following elements in the document:

- ① Points to the text "normally distributed ES."
- ② Points to the argument `min.sz=10` in the code block.
- ③ Points to the argument `max.sz=500` in the code block.
- ④ Points to the command `?gsva` in the R Console inset window.

?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 {GSVA} R Documentation

Gene Set Variation Analysis

Description

Estimates GSVA enrichment scores.

Usage

```
## 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),  
      ...)
```

?gsva

何を書いているのか (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...

```
## 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,
```


?gsva

何を書いているのか(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。**

```
## S4 method for signature 'ExpressionSet,list' ④
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' ③
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)
```



```
## 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,
```

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。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

```
R: Gene Set Variation Analysis x
  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,
```

?gsva

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

R: Gene Set Variation Analysis

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

Arguments

expr
Gene expression data which can be given either as an ExpressionSet object or as a matrix of expression values where rows correspond to genes and columns correspond to samples.

gset.idx.list
Gene sets provided either as a list object or as a GeneSetCollection object.

annotation
In the case of calling `gsva()` with expression data in a matrix and gene sets as a `GeneSetCollection` object, the `annotation` argument can be used to supply the name of the Bioconductor package that contains annotations for the class of gene identifiers occurring in the row names of the expression data matrix. By default `gsva()` will try to match the identifiers in `expr` to the identifiers in `gset.idx.list` just as they are, unless the `annotation` argument is set.

method

?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。

divided by the square-root of the size of the gene set, while in the case of plaque they are used to calculate the singular value decomposition (SVD) over the genes in the gene set and use the coefficients of the first right-singular vector as pathway activity profile.

kcdf

Character string denoting the kernel to use during the non-parametric estimation of the cumulative distribution function of expression levels across samples when `method="gsva"`. By default, `kcdf="Gaussian"` which is suitable when input expression values are continuous, such as microarray fluorescent units in logarithmic scale, RNA-seq log-CPMs, log-RPKMs or log-TPMs. When input expression values are integer counts, such as those derived from RNA-seq experiments, then this argument should be set to `kcdf="Poisson"`. This argument supersedes arguments `rnaseq` and `kernel`, which are deprecated and will be removed in the next release.

rnaseq

This argument has been deprecated and will be removed in the next release. Please use the argument `kcdf` instead.

abs.ranking

Flag used only when `mx.diff=TRUE`. When `abs.ranking=FALSE` (default) a



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

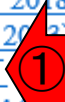
②GSVAの解説PDFには、Pickrellデータも例題として使われています。実際にPickrellデータを眺めます。

(Rで)塩基配列解析

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

このウェブページのR
います。初心者の方

- 解析 | 機能解析 | GMTファイル読込 | [GSEABase\(Morgan 2018\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | 遺伝子セット解析 | [GSVA\(Hänzelmann 2013\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | [GO解析 | について](#) (last modified 2018/06/27)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | [SeqGSEA\(Wang 2014\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | [GSVA\(Hänzelmann 2013\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | 遺伝子オントロジー(GO)解析 | [GOseq\(Young 2010\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | [Pathway解析 | について](#) (last modified 2018/06/29)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | [Pathway解析 | について](#) (last modified 2018/06/29)
- 解析 | 機能解析 | パスウェイ(Pathway)解析 | [Pathway解析 | について](#) (last modified 2018/06/29)



What's new?

- 「解析 | 一般 | Seq

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](#)
- [SeqGSEA\(各群5反復以上を要求\): Wang et al., Bioinformatics, 2014](#)
- [GSASeqSP: 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](#)



GSVA for RNA-seq data

①をクリックすると、②GSVA for RNA-seq dataのところに飛びます。

127.0.0.1 R: C2 collection of canonical p...

14 / 21 100%

サインイン

しおり

- Introduction
- GSVA enrichment scores
- Overview of the package
- > Applications
- Comparison with other methods
- 1** GSVA for RNA-Seq data
- Session Information

6 GSVA for RNA-Seq data **2**

In this section we illustrate how to use GSVA with RNA-seq data and, more importantly, how the method provides pathway activity profiles analogous to the ones obtained from microarray data by using samples of lymphoblastoid cell lines (LCL) from HapMap individuals which have been profiled using both technologies [Huang et al. \(2007\)](#); [Pickrell et al. \(2010\)](#). These data form part of the experimental package `GSVAdata` and the corresponding help pages contain details on how the data were processed. We start loading these data and verifying that they indeed contain expression data for the same genes and samples, as follows:

```
> data(commonPickrellHuang)
> stopifnot(identical(featureNames(huangArrayRMAnoBatchCommon_eset),
```

14

210 x 297 mm

この画面内で実際に行うのは、①
GSVAdataパッケージをロードした後、②を
コピー実行することのみ。③は無視でよい。

6 GSVA for RNA-Seq data

In this section we illustrate how to use GSVA with RNA-seq data and, more importantly, how the method provides pathway activity profiles analogous to the ones obtained from microarray data by using samples of lymphoblastoid cell lines (LCL) from HapMap individuals which have been profiled using both technologies [Huang et al. \(2007\)](#); [Pickrell et al. \(2010\)](#). These data form part of the experimental package `GSVAdata` and the corresponding help pages contain details on how the data were processed. We start loading the data and verifying that they indeed contain expression data for the same genes and samples, as follows:

```
> data(commonPickrellHuang)
> stopifnot(identical(featureNames(huangArrayRMA),
```

14

210 x 297 mm

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、featureNames関数でgene ID情報を取り出した後、identical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が、このあたりは深入りする価値はありません

The screenshot shows a web browser window with a document page. The address bar shows the URL: `http://127.0.0.1:16111/library/GSVA/doc/GSVA`. The page content includes a table of contents on the left and a main text area. The main text area contains the following text:

6 GSVA

In this section we describe the method for processing samples of lymphoblastoid cell lines (LCL) from HapMap individuals which have been profiled using both technologies [Huang et al. \(2007\)](#); [Pickrell et al. \(2010\)](#). These data form part of the experimental package GSVA and the corresponding help pages contain details on how the data were processed. We start loading these data and verifying that they indeed contain expression data for the same genes and samples, as follows:

```
> data(commonPickrellHuang)
> stopifnot(identical(featureNames(huangArrayRMAnoBatchCommon_eset),
```

14

Red arrows point to the following elements:

- ①: `Huang et al. (2007)`
- ②: `Pickrell et al. (2010)`
- ③: `stopifnot`
- ④: `data`

p15~16

genes in the DE gene set; (G, H) strong signal-to-noise ratio, 80% of DE gene in the DE gene set.

15

```

+           featureNames(pickrellCountsArgonneCQNcommon_eset)))
> stopifnot(identical(sampleNames(huangArrayRMABatchCommon_eset),
+           sampleNames(pickrellCountsArgonneCQNcommon_eset)))

```

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)))]

```

210 x 297 mm

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。

p16の上のほう

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)))]
```

①

```
> canonicalC2BroadSets
```

GeneSetCollection

```
names: KEGG_GLYCOLYSIS_GLUconeogenesis, KEGG_CITRATE_CYCLE_TCA_CYCLE, ..., BIOCARTA_ACT
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
 geneIds: 266, 84663, ..., 353513 (total: 34)
 geneIdType: EntrezId
 collectionType: Broad

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chryq1であった。

C2ではなくC1

http://127.0.0.1:16111/library/GSVA/doc/GSVA.pdf

R 127.0.0.1 x R: C2 collection of canonical p...

16 / 21 100%

サインイン

しおり

- Introduction
- GSVA enrichment scores
- Overview of the package
- Applications
- Comparison with other methods
- GSVA for RNA-Seq data
- Session Information

```

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_ACT
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
geneIds: 266, 84663, ..., 353513 (total: 34)
geneIdType: EntrezId
collectionType: Broad
  
```

①

②

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

性特異的セットを追加

① canonicalC2BroadSets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。

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)))]
```

①

```
GeneSetCollection
names: KEGG_GLYCOLYSIS_GLUconeogenesis, KEGG_CITRATE_CYCLE_TCA_CYCLE, ..., BIOCARTA_ACT
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:

②

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

```
setName: MSY
geneIds: 266, 84663, ..., 353513 (total: 34)
geneIdType: EntrezId
collectionType: Broad
```


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。

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_ACT
 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
 geneIds: 266, 84663, ..., 353513 (total: 34)
 geneIdType: EntrezId
 collectionType: Broad

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。

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
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
```

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

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

http://127.0.0.1:16111/library/GSVA/doc/GSVA.pdf

R 127.0.0.1 R: C2 collection of canonical p...

16 / 21 100% サインイン

```

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

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

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

GeneSetCollection
names: KEGG_GLYCOLYSIS_GLUconeogenesis, KEGG_CITRATE_CYCLE_TCA_CYCLE, ..., XiE (835 tot
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
the RNA-seq integer count data. Note that the only requirement to do the latter is to set the argument
kcdf="Poisson" which is "Gaussian" by default. Note, however, that if our RNA-seq derived expression
  
```

②

⑦

⑥

③

④

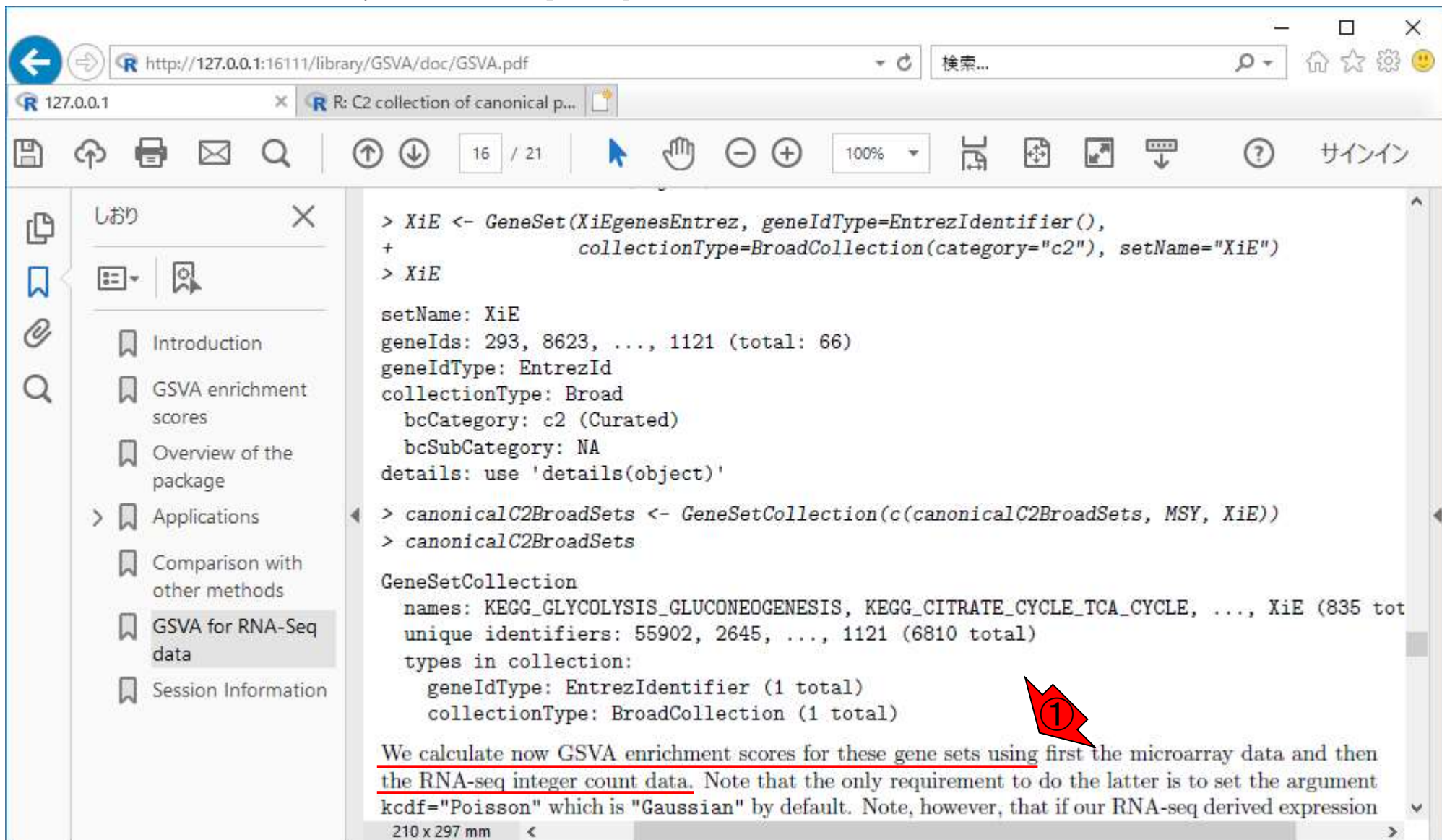
⑤

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

GSVAの実行準備完了



The screenshot shows a web browser window displaying the GSVA package documentation. The address bar shows the URL: `http://127.0.0.1:16111/library/GSVA/doc/GSVA.pdf`. The browser tabs include "127.0.0.1" and "R: C2 collection of canonical p...". The page content shows R code and its output:

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

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

The next command and its output are:

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

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

A red arrow with the number "1" points to the output of the `canonicalC2BroadSets` command. Below the code, there is a paragraph of text:

We calculate now GSVA enrichment scores for these gene sets using first the microarray data and then the RNA-seq integer count data. Note that the only requirement to do the latter is to set the argument `kcdf="Poisson"` which is "Gaussian" by default. Note, however, that if our RNA-seq derived expression

p17の上のほう

The screenshot shows a PDF viewer displaying a document with R code. The code is as follows:

```
> esmicro <- gsva(huangArrayRMAnoBatchCommon_eset, canonicalC2BroadSets, min.sz=5, max.sz=
+ mx.diff=TRUE, verbose=FALSE, parallel.sz=1)
> dim(esmicro)
Features Samples
      806      36

> esrnaseq <- gsva(pickrellCountsArgonneCQNcommon_eset, canonicalC2BroadSets, min.sz=5, n
+ kcdf="Poisson", mx.diff=TRUE, verbose=FALSE, parallel.sz=1)
> dim(esrnaseq)
Features Samples
      806      36
```

The line `> esrnaseq <- gsva(pickrellCountsArgonneCQNcommon_eset, canonicalC2BroadSets, min.sz=5, n + kcdf="Poisson", mx.diff=TRUE, verbose=FALSE, parallel.sz=1)` is highlighted with a red box and a red arrow labeled '1' pointing to it from the right.

Below the code, the text reads: "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:"

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。

http://127.0.0.1:16111/library/GSVA/doc/GSVA.pdf

R 127.0.0.1 x R: C2 collection of canonical p...

17 / 21 100%

しおり

- Introduction
- GSVA enrichment scores
- Overview of the package
- Applications
- Comparison with other methods
- GSVA for RNA-Seq data
- Session Information

```
> esmicro <- gsva(huangArrayRMAnoBatchCommon_eset, canonicalC2BroadSets, min.sz=5, max.sz +
+               mx.diff=TRUE, verbose=FALSE, parallel.sz=1)
> dim(esmicro)
Features Samples
      806     36

> esrnaseq <- gsva(pickrellCountsArgonneCQNcommon_eset, canonicalC2BroadSets, min.sz=5, n
+               kcdf="Poisson", mx.diff=TRUE, verbose=FALSE, parallel.sz=1)
> dim(esrnaseq)
Features Samples
      806     36
```

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:

210 x 297 mm

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。

The screenshot shows a PDF viewer displaying a document page with R code and its output. The browser address bar shows the URL: `http://127.0.0.1:16111/library/GSVA/doc/GSVA.pdf`. The document title is "R: C2 collection of canonical". The page number is 17 / 21.

The left sidebar shows a table of contents with the following items:

- Introduction
- GSVA enrichment scores
- Overview of the package
- Applications
- Comparison with other methods
- GSVA for RNA-Seq data
- Session Information

The main content area shows R code and its output:

```
> esmicro <- gsva(huangArrayRMAnoBatchCommon_eset, canonicalC2BroadSets, min.sz=5, max.sz +
+                 mx.diff=TRUE, verbose=FALSE, parallel.sz=1)
> dim(esmicro)
Features Samples
      806      36

> esrnaseq <- gsva(pickrellCountsArgonneCQNcommon_eset, canonicalC2BroadSets, min.sz=5, n
+                 kcdf="Poisson", mx.diff=TRUE, verbose=FALSE, parallel.sz=1)
> dim(esrnaseq)
Features Samples
      806      36
```

Red arrows point to the following elements:

- Arrow ③ points to the `esmicro` code block.
- Arrow ① points to the `dim(esrnaseq)` output.
- Arrow ② points to the text below the `dim(esrnaseq)` output.

The text below the `dim(esrnaseq)` output reads: "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:"

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

The screenshot shows an RStudio terminal window with the following R code and output:

```
> library(edgeR)
> data(annotEntrez220212)
> head(annotEntrez220212)
```

	Length	GCcontent
1	2301	0.6292916
10	1344	0.3816964
100	2612	0.5153139
1000	4380	0.4502283
10000	7091	0.3989564
1008586	606	0.4339934

```
> cpm <- cpm(exprs(pickrellCountsArgonneCQNcommon_eset))
> dim(cpm)
```

```
[1] 11508 36
```

```
> common <- intersect(rownames(cpm), rownames(annotEntrez220212))
> length(common)
```

Red arrows and numbers 1-4 indicate the following steps:

- ①: Points to the `head(annotEntrez220212)` command and its output table.
- ②: Points to the `cpm <- cpm(exprs(pickrellCountsArgonneCQNcommon_eset))` command.
- ③: Points to the `common <- intersect(rownames(cpm), rownames(annotEntrez220212))` command.
- ④: Points to the `cpm` object.

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

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)

このウェブページは、初

What's new

• 「解析」

- マップ後 | カウント 情報取得 | paired-end | トランスクリプトーム | [QuasR\(Gaidatzis 2015\)](#) (last modified 2016/02/12)
- 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 | |](#) [について](#) (last modified 2018/06/22) **NEW**
- カウント 情報取得 | リアルデータ | SRP012167 | [recount\(Collado-Torres 2017\)](#) (last modified 2018/06/26) **NEW**
- カウント 情報取得 | リアルデータ | SRP012167 | [parathyroidSE\(Haglund 2012\)](#) (last modified 2018/06/26) **NEW**
- カウント 情報取得 | リアルデータ | SRP001558 | [recount\(Collado-Torres 2017\)](#) (last modified 2018/06/15) **NEW**
- カウント 情報取得 | リアルデータ | SRP001540 | [recount\(Collado-Torres 2017\)](#) (last modified 2018/06/25) **NEW**
- カウント 情報取得 | リアルデータ | SRP001540 | [GSVadata\(Hänzelmann 2013\)](#) (last modified 2018/07/03) **NEW**
- カウント 情報取得 | リアルデータ | ERP000546 | [recount\(Collado-Torres 2017\)](#) (last modified 2018/06/10) **NEW**
- カウント 情報取得 | [シミュレーションデータ](#) (last modified 2018/06/09) **NEW**

カウント情報取得 | リアルデータ | SRP001540 | [GSVadata\(Hänzelmann_2013\)](#) **NEW**

• [配列長とカ](#)

• [正規化 | |](#)

• [正規化 | 基](#)

• [正規化 | 基](#)

• [正規化 | 基](#)

[GSVadata](#) パッケージを用いて、[SRP001540\(Pickrell et al., Nature, 2010\)](#); ブラウザは正以外を推奨) 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

1. [pickrellCountsArgonneCQNcommon_eset](#) の場合:
原著論文(Pickrell et al., Nature, 2010) の ExpressionSet オブジェクトです。

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

```

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Desktopに変更してから、①ls()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。単に全員的环境を揃えているだけなので、意味が分かるヒトはやらなくてもよい。

7. pickrellCountsYaleCQNcommon_esetの場合:

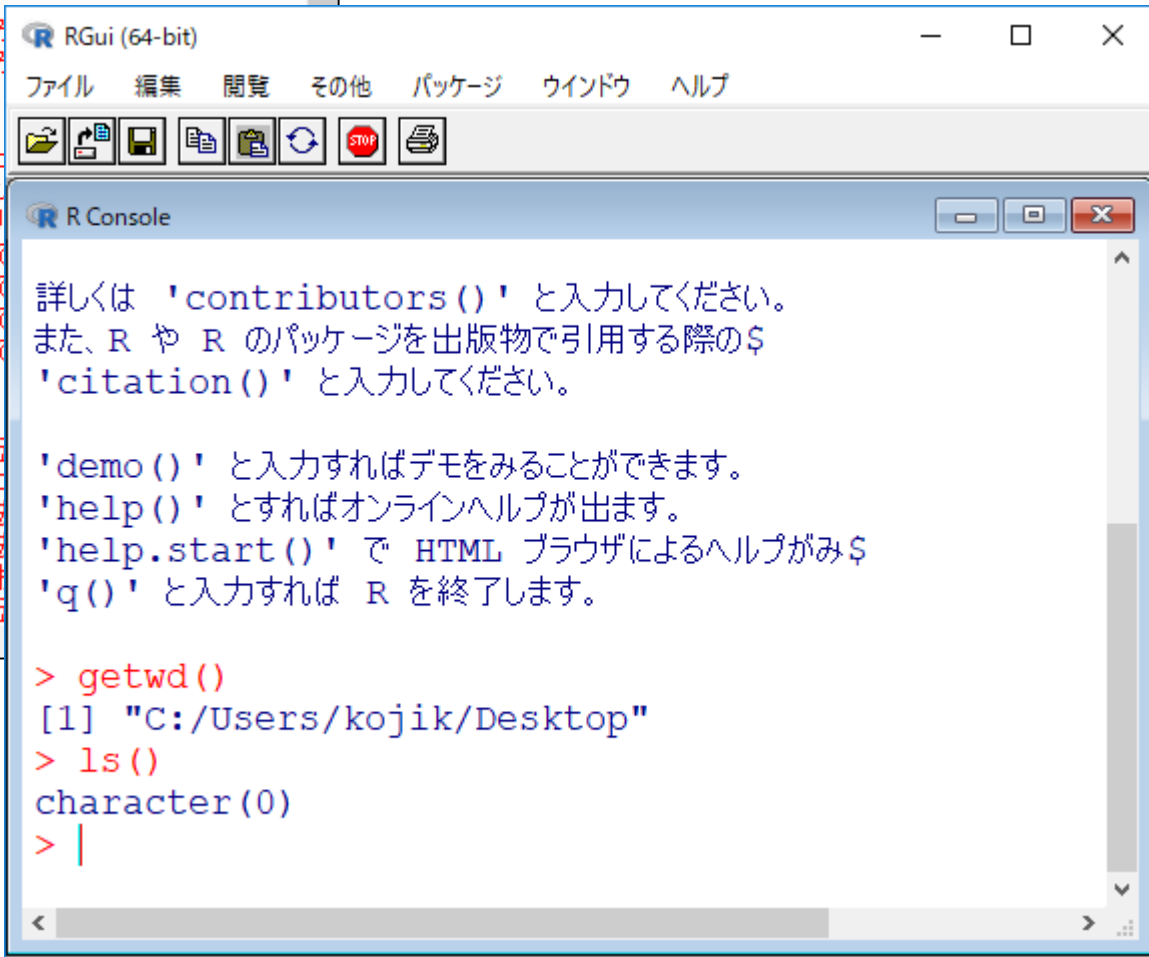
例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニークにしています。11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルは[hoge7.txt](#)です。

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

#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を読み込み
ls() #利用可能なオブジェクトを確認
eset <- pickrellCountsYaleCQNcommon_eset #esetとして取り出し
dim(pData(eset)) #確認してるだけ
head(pData(eset)) #確認してるだけ
pData(eset)$Gender #確認してるだけ
table(pData(eset)$Gender) #確認してるだけ

#本番(重複除去)
length(rownames(exprs(eset))) #重複除去前の遺伝子数
length(unique(rownames(exprs(eset)))) #重複除去前のユニークな遺伝子数
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として取り出し
length(rownames(exprs(eset))) #重複除去前の遺伝子数
```



例題7

①赤枠内をコピー。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11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイル

```

out_f <- "hoge7.txt" #出力ファイル名

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

#読み込み(目的のデータセットをロード)
data(commonPickrellHuang)
ls()

eset <- pickrellCountsYaleCQNcommon_eset
dim(pData(eset))
head(pData(eset))
pData(eset)$Gender
table(pData(eset)$Gender)

#本番(重複除去)
length(rownames(exprs(eset)))
length(unique(rownames(exprs(eset))))
hoge <- nsFilter(eset, var.filter=F,
                 remove.dupEntrez=T)
eset <- hoge$eset
length(rownames(exprs(eset)))

```



RGui (64-bit) 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\$
 >
 > #読み込み(目的のデータセットをロード)
 > data(commonPickrellHuang) #param\$
 > ls() #利用\$
 [1] "huangArrayRMAanoBatchCommon_eset"
 [2] "out_f"
 [3] "pickrellCountsArgonneCQNcommon_eset"
 [4] "pickrellCountsYaleCQNcommon_eset"
 >

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を読み込み
ls() #利用可能なオブジェクトを確認

eset <- pickrellCountsYaleCQNcommon_eset #esetとして取り出す
dim(pData(eset)) #確認してるだけ
head(pData(eset)) #確認してるだけ
pData(eset)$Gender #確認してるだけ
table(pData(eset)$Gender) #確認してるだけ

#本番(重複除去)
length(rownames(exprs(eset))) #重複除去前の遺伝子数
length(unique(rownames(exprs(eset)))) #重複除去前のユニークな遺伝子数
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として取り出す
length(rownames(exprs(eset))) #重複除去前の遺伝子数
    
```

RGui (64-bit) window showing the R Console output:

```

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という名前を取り扱い、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、ユニークな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)
ls()
eset <- pickrellCountsYaleCQNcommon_eset #paramで指定した利用可能なオブジェクトとして取り出す
dim(pData(eset)) #確認してるだけ
head(pData(eset)) #確認してるだけ
pData(eset)$Gender #確認してるだけ
table(pData(eset)$Gender) #確認してるだけ

#本番(重複除去)
length(rownames(exprs(eset))) #重複除去前のユニークなエンタrez IDの数
length(unique(rownames(exprs(eset)))) #重複除去前のユニークなエンタrez IDの数
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として取り出す
length(rownames(exprs(eset))) #重複除去前のユニークなエンタrez IDの数

```

R Console Output:

```

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"
> |

```

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。

女23人、男13人

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)
ls()
eset <- pickrellCountsYaleCQNcommon_eset #paramで指定したオブジェクト
dim(pData(eset)) #利用可能なオブジェクト
head(pData(eset)) #確認してるだけ
pData(eset)$Gender #確認してるだけ
table(pData(eset)$Gender) #確認してるだけ

#本番(重複除去)
length(rownames(exprs(eset))) #重複除去前の遺伝子数
length(unique(rownames(exprs(eset)))) #重複除去前のユニークな遺伝子数
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として取り替える
length(rownames(exprs(eset))) #重複除去後の遺伝子数

```



コード下部に移動

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tmp1に格納
write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F) #tmpの中
```



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、ユニークな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

```

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

#本番(重複除去)
length(rownames(exprs(eset)))
length(unique(rownames(exprs(eset))))
hoge <- nsFilter(eset, var.filter=F,
                 remove.dupEntrez=T)
eset <- hoge$eset
length(rownames(exprs(eset)))

#後処理(列名を変更し、列をソート)
data <- exprs(eset)
colnames(data) <- pData(eset)$Gender
head(data[1:2,])
data <- data[, order(colnames(data))]
head(data[1:2,])

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

```

① 重複除去前の遺伝子IDのユニーク化
#同一IDの重複除去
#esetとして取り扱ったデータ
#重複除去前の遺伝子ID

#dataとして取り扱ったデータ
#列名を変更
#確認してるだけです
#列名でソート
#確認してるだけです

#保存したい情報を追加

```

R R Console
[25] "Female" "Female" "Female" "Male"
[29] "Female" "Male" "Female" "Male"
[33] "Female" "Female" "Female" "Male"
> table(pData(eset)$Gender)
Female Male
      23  13
> dim(exprs(eset))
[1] 11508 36
> length(rownames(exprs(eset)))
[1] 11508
> length(unique(rownames(exprs(eset))))
[1] 11482
> |

```

①

nsFilter関数の実行

①重複除去を行うnsFilter関数を実行。② var.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(平均しているのかどうかなど)を行っているか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、ユニークな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)$Gender
table(pData(eset)$Gender)
```

#確認してるだけです
#確認してるだけです

#本番(重複除去)

```
length(rownames(exprs(eset)))
length(unique(rownames(exprs(eset))))
hoge <- nsFilter(eset, var.filter=F,
                 remove.dupEntrez=T)
eset <- hoge$eset
length(rownames(exprs(eset)))
```

#重複除去前の遺伝子数
#重複除去前のユニークな遺伝子数
#同一IDの重複除去
#同一IDの重複除去
#esetとして取り扱
#重複除去前の遺伝子数

#後処理(列名を変更し、列をソート)

```
data <- exprs(eset)
colnames(data) <- pData(eset)$Gender
head(data[1:2,])
data <- data[, order(colnames(data))]
head(data[1:2,])
```

#dataとして取り扱
#列名を変更
#確認してるだけ
#列名でソート
#確認してるだけ

#ファイルに保存

```
tmp <- cbind(rownames(data), data)
write.table(tmp, out_f, sep="\t", append=F, quote=F, r
```

#保存したい情報を

RGui (64-bit) window showing the R Console output:

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

Annotations in the screenshot:

- Red arrow ① points to the `nsFilter` function call.
- Red arrow ② points to the `length(unique(rownames(exprs(eset))))` output.
- Red arrow ③ points to the final `length(rownames(exprs(eset)))` output.

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ので、列名はFemale or Maleで充分

列名変更の必要性

7. pickrellCountsYaleCQNcommon_esetの場合:

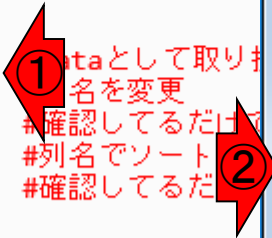
例題6をベースとして、重複したEntrez gene IDのものが存在するので、ユニークにしていきます。
11,482 Entrez gene IDs×36 samplesのカウントデータです。出力ファイルはhoge7.txtです。

```
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
length(unique(rownames(exprs(eset)))) #重複除去前のユニークな遺伝子数
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として取り出す
length(rownames(exprs(eset))) #重複除去前の遺伝子数

#後処理(列名を変更し、列をソート)
data <- exprs(eset) #dataとして取り出す
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) #保存したい情報を追加
write.table(tmp, out_f, sep="\t", append=F, quote=F, r
```



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

R Console
> 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"
> |
```

①赤枠内コピー実行後。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。

列名変更後

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、ユニーク化しています。
 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。

```

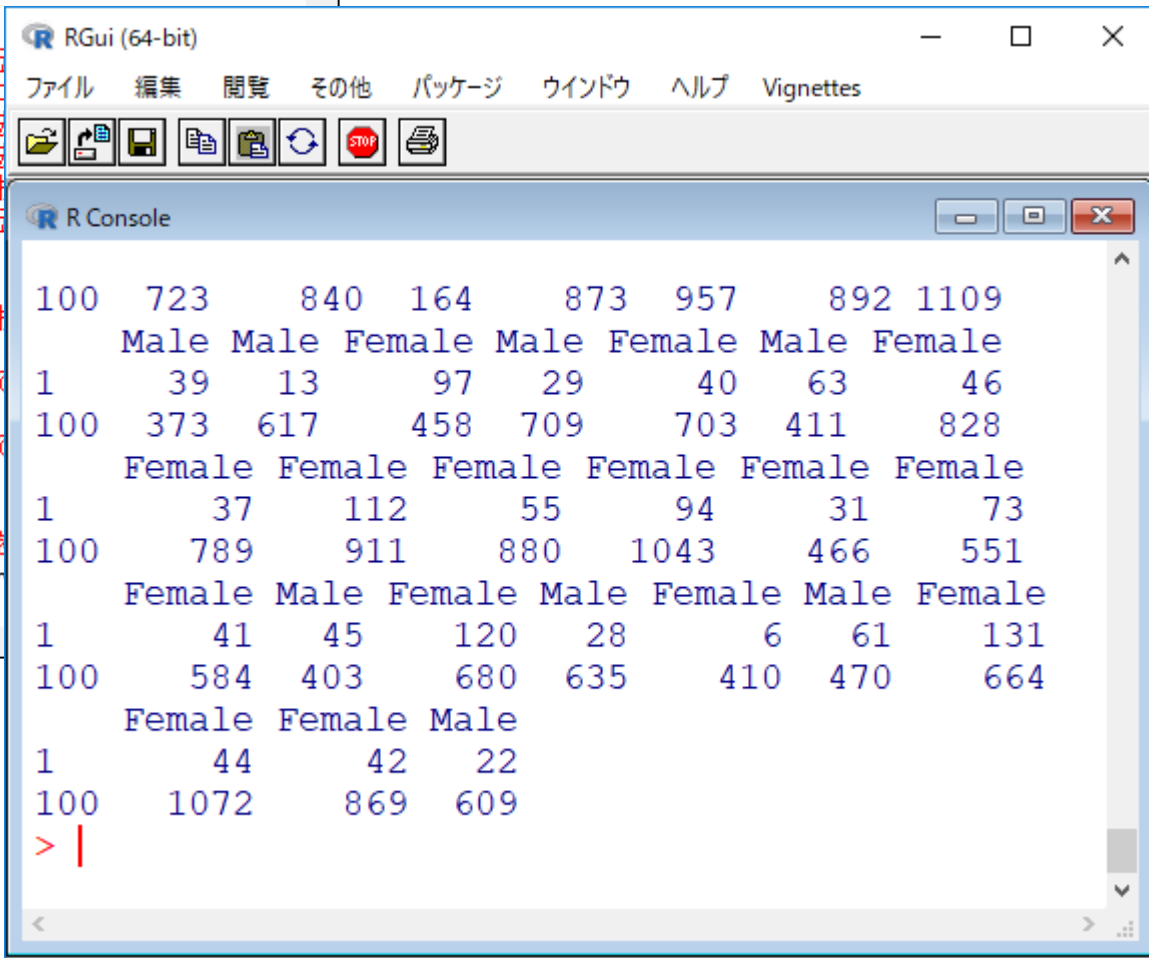
pData(eset)$Gender
table(pData(eset)$Gender)
#確認してるだけです
#確認してるだけです

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

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

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

① dataとして取り扱った列名を変更
 #確認してるだけです
 #列名でソート
 #確認してるだけです



列名でソート

①赤枠内コピー実行後。確かに列名でソートできていることがわかります。order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)$Gender
table(pData(eset)$Gender) #確認してるだけです
                             #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
length(unique(rownames(exprs(eset)))) #重複除去前のユニーク遺伝子数
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として取り出す
length(rownames(exprs(eset))) #重複除去前の遺伝子数

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

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

```



```

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

R Console
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 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
> |

```

例題7を最後までコピー後

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
    
```


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

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。

C1コレクションでGSVA

(Rで)塩基配列解析

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

このウェブページのR...
います。初心者の方

What's new?

「解析 | 一般 | Seq

- 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)
- 解析 | 機能解析 | [GMTファイル取得 |MSigDB\(Subramanian 2005\)](#) (last modified 2018/06/27)
- 解析 | 機能解析 | [GMTファイル読み込み |GSEABase\(Morgan 2018\)](#) (last modified 2018/06/27)
- 解析 | 機能解析 | [遺伝子セット解析 |GSVA\(Hänzelmann 2013\)](#) (last modified 2018/06/29) **1**
- 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)
- 解析 | 機能解析 | [遺伝子オントロジー\(GO\)解析 |SeqGSEA\(Wang 2014\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | [遺伝子オントロジー\(GO\)解析 |GSVA\(Hänzelmann 2013\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | [遺伝子オントロジー\(GO\)解析 |GOseq\(Young 2010\)](#) (last modified 2018/06/29)
- 解析 | 機能解析 | [パスウェイ\(Pathway\)解析 |について](#) (last modified 2018/06/27) **NEW**
- 解析 | 機能解析 | [パスウェイ\(Pathway\)解析 |Pathway\(Chen 2013\)](#) (last modified 2018/06/27)
- 解析 | 機能解析 | [パスウェイ\(Pathway\)解析 |Pathway\(Chen 2013\)](#) (last modified 2018/06/27)
- 解析 | 機能解析 | [パスウェイ\(Pathway\)解析 |Pathway\(Chen 2013\)](#) (last modified 2018/06/27)
- 解析 | [分類 |について](#) (last modified 2018/06/29)
- 解析 | [菌叢解析 |について](#) (last modified 2018/06/29)

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chrp22)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(Enrichment Score)をサンプルごとに算出し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別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ー。

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)
```

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.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で指定したファイルの読み込み
geneset <- getGmt(in_f2, geneIdType=EntrezIdentifier(), #in_f2で指定したファイルの読み込み
                 collectionType=BroadCollection(category="c1")) #in_f2で指定したファイルの読み込み
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以上)
           mx.diff=T, verbose=F, parallel.sz=1) #GSVAの実行
dim(out) #確認してるだけです
    
```


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、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"
in_f2 <- "c1.all.v6.1.entrez.gmt"
out_f <- "hoge1.txt"
param_G1 <- 23
param_G2 <- 13

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

#入力ファイルの読み込み
data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t")
geneset <- getGmt(in_f2, geneIdType=EntrezIdentifier(), collectionType=BroadCollection(category="Chromosome"))

#本番(GSVAの実行)
data <- as.matrix(data)
out <- gsva(data, geneset, min.sz=5, max.sz=500, kcdf="Poisson", mx.diff=T, verbose=F, parallel.sz=1)

dim(out)
    
```



#入力ファイル名を指定してin_f1に格納(発現ファイル)
 #入力ファイル名を指定してin_f2に格納(gmtファイル)
 #出力ファイル名を指定してout_fに格納
 #G1群のサンプル数を指定してparam_G1に格納
 #G2群のサンプル数を指定してparam_G2に格納
 #パッケージの読み込み
 #パッケージの読み込み
 #データの型をmatrixにする
 #GSVAの実行
 #確認してるだけ
 #確認してるだけ

```

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

R Console
[1] 298 36
>
> #後処理(wilcox.testを実行)
> data.cl <- c(rep(1, param_G1), rep(2, param_G2))
> pvalue <- NULL
> for(i in 1:nrow(out)){
+   pvalue <- c(pvalue, wilcox.test(out[i, data.cl]))
+ }
>
> #ファイルに保存
> tmp <- cbind(row.names(out), out, pvalue) #保存$
> tmp <- tmp[order(pvalue),] #pvalue$
> write.table(tmp, out_f, sep="\t", append=F, q$)
> |
    
```

実行結果ファイル

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 | [GSVAdat\(Hänzelmann 2013\)](#)」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_trez.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で指定したファイル
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

```


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…

hoge1.txt

自動保存

同じ結果が得られた

① AbsFilterGSEAの論文の結果と同じですね。② chryq11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 RNA-seq data from several 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 positive rate as well, these two gene-sets were regarded as true positives. Then, five samples were randomly selected from



遺伝子オントロジー(GO)解析 | について NEW

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ありますが、Set Analysis; GSA)という枠組みではGO解析もパスウェイ分析も難しいかもしれません。

- 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](#)
- GSASeqSP: [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](#)



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が。

hoge1.txt

Excel screenshot showing a data table with columns A through AL. The table contains numerical values for various gene sets (Female.1 to Female.22, Male.1 to Male.12) and a pvalue column. A red box highlights the row for 'chryq11' in columns Y through AK, and a red arrow points to the value '1' in column AE.

	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	-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	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	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	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	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	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	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	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	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	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	-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	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 (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かなり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

