# 遺伝子発現データベース/遺伝子発現解析ツール at AJACS番町3

情報・システム研究機構 データサイエンス共同利用基盤施設 ライフサイエンス統合データベースセンター 坊農 秀雅 bono@dbcls.rois.ac.jp

2019年8月7日

これは統合データベース講習会AJACS番町3「遺伝子発現データベース/遺伝子発現解析ツール」の資料です。

### 概要

本講習は、だれでも自由に使うことができる公共データベース(以下、DBと略します)やウェブツールを活用して、研究のさまざまな場面で調べることの多いオミックスDBを簡単に調べるための方法と基礎知識について学びます。とくに、需要の増している公共DBから遺伝子発現データを検索し取得してくる方法について詳しく説明、実習します。また、自ら行なった大規模発現解析の(あるいは公共DBから取得・解析した)結果として得られた数百~数千におよぶ遺伝子セットについて、生物学的な解釈をする方法とその結果の考察を実践します。

# 講習の流れ

今回の講習では、以下の内容について順番に説明します。 【実習】の部分はみんなでやる実習、 【応用】は早くできてしまった人用の応用課題です。

- 研究現場で頻繁に使われるDBやツールを知る: 統合TV
- 遺伝子発現DBとは
- 遺伝子発現DBの使い方
  - · AOE (All of gene expression)
    - 【実習1】AOEを使って、興味ある実験データセットを絞り込む
  - NCBI Gene Expression Omnibus (GEO)
  - EBI ArrayExpress
    - 【実習2】OmicsDIを使って、興味あるオミックスデータセットを検索する
- 個々の遺伝子の発現プロファイルなどを調べる
  - 。 【実習3】RefExを使って、組織特異的遺伝子を検索する
- 数十~数千の遺伝子群の生物学的解釈
  - 。 【実習4】 metascapeを用いて、発現データの結果を生物学的に解釈する
  - 。 【応用1】複数のエンリッチメント解析ツールを用いて、発現データの結果を生物学的に解釈する

# 講習に際しての注意とお願い

- みんなで同時にアクセスするとサイトにつながりにくくなることが予想されます。
  - 。 資料を見ながら自力で進められそうな方はどんどん先に、そうでない方は講師と一緒にすすめていきましょう。
  - サイトの反応が悪い時はタイミングをずらして実行してみてください。
  - 反応が無いからと言って何度もクリックするとますます繋がらなくなってしまいます。おおらかな気持ちで臨みましょう。
- わからないことがあったら挙手にてスタッフにお知らせください。
  - 。 遠慮は無用です(そのための講習会です!)。おいてけぼりは楽しくありません。

# 研究現場で頻繁に使われるデータベースやツールを知る

#### 統合TV

- 生命科学分野の有用なデータベースやツールの使い方を動画で紹介するウェブサイト
  - 。 本家ウェブサイト https://togotv.dbcls.jp/
  - 。 YouTube版もあります https://youtube.com/togotv/
  - ウェブサイトへのアクセスから結果の見方まで、操作の一挙手一投足がわかります。
    - 講義・講習などの参考資料や後輩指導の教材として利用できます。

- 本講習中、本家サイトが繋がらない時は、統合TVのYouTube版を見ればおおよその内容がわかるようになっています。
- 今回の講習に関連する内容の多くは、「発現解析」タグのついた動画にあります。
- 過去の講習会の内容はそのほとんどが統合TVに収録されており、いつでもどこでも繰り返し復習できるようになっています。
- 。 お探しの動画が見つからない or 統合TV未掲載の場合は、統合TV番組リクエストフォームへどうぞ!
- 。 統合TVを作ってくれる方、募集中!!

# 遺伝子発現DBとは

#### mind the gap

遺伝子発現解析などのオミックス解析→DB(というか、公共アーカイブ)に登録

- 日本では論文投稿前が現在一般的
- 欧米では研究費の条件で多くの場合データを出したらわりとすぐ

#### その歴史

- マイクロアレイの発明→網羅的遺伝子発現定量が可能に→遺伝子発現DB
  - 。 アレイのイメージデータ(CELファイルなど)
  - 。 定量データ(Series Matrix File)
- マイクロアレイのデータベース
  - 。 DDBJ CIBEX (更新停止) → 2018年、Genomic Expression Archive (GEA)として仕切り直し
  - NCBI Gene Expression Omnibus (GEO)
  - EBI ArravExpress
    - ArrayExpressはGEOのデータを定期的に取り込み続けていた→こちらの使用を推奨していた→2017年に取り込み停止。
- 次世代シークエンサー(NGS: Next Generation Sequencer)の発明→個々のサンプルでのtranscript sequencing (RNA-seq)が現実的に
  - 。 データはNGS配列DB(SRA: Sequence Read Archive) and/or 遺伝子発現DBへ?
  - 。 NCBI,EBIでは遺伝子発現DB(それぞれGEO,ArrayExpress)に登録すれば、配列データ(FASTQファイル)がSRAにも登録される状況
    - DDBJ(日本)もその仕組みを提供中(上述のGEA)
  - 。 現状、どこかの公共DBに登録されていればOK

#### 発現定量のステップ

詳細は省くが、データを利用するので発現定量の概要を説明。

#### マイクロアレイ

- 1. ハイブリダイゼーション
- 2. 蛍光スキャナーで強度の読み取り=各プローブに対する発現量
- 3. 遺伝子ごとに発現量を正規化
- 4. 機能アノテーションを付加してデータ解析

### NGS(RNA-seq)

- 1. 配列データ(FASTQ形式)を得る
- 2. ゲノム配列にマッピング=ゲノムに対してアラインメント
- 3. マッピング結果をアッセンブル=転写単位ごとの発現量
- 4. 正規化
- 5. 機能アノテーションを付加してデータ解析



データ解析の詳細は、Dr.Bonoの生命科学データ解析 p174-p181などを参照。

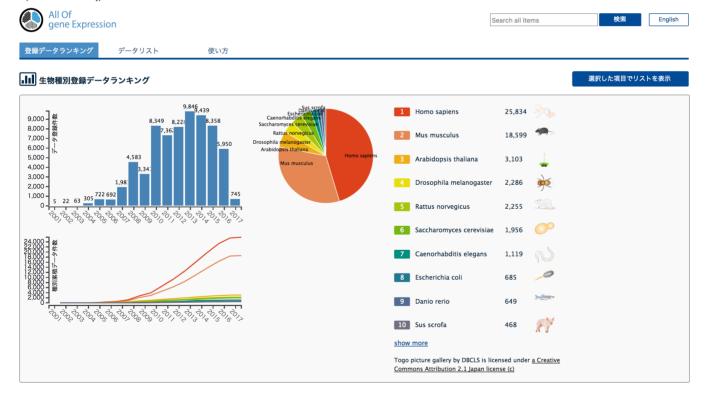
# 遺伝子発現DBの使い方

#### **AOE(All of gene Expression)**

- 遺伝子発現用のデータ目次 https://AOE.dbcls.jp/
  - 。 EBI ArrayExpress + NCBI GEO + DDBJ GEA をカバー
    - 年ごと、生物種ごとにヒストグラム表示
- マイクロアレイ(Affymetrix, Agilent, それ以外) + RNA-seq(Illumina, それ以外)でデータが分類
- キーワード検索結果もヒストグラム表示されてわかりやすい

# 【実習1】AOEを使って、低酸素関連遺伝子発現データを検索する

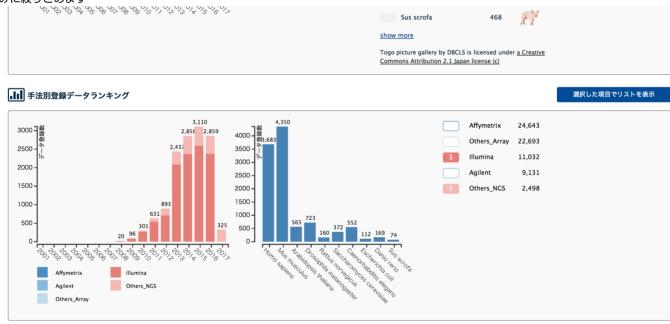
1. http://AOE.dbcls.jp/ にアクセスします



2. 「生物種別登録データランキング」で生物種のところ(8. Escherichia coli)をクリックすると、その生物種だけのデータ登録数に絞りこめます



3. 「手法別登録データランキング」で赤系の色(4. illuminaと5. Other\_NGS)だけクリックして残すと、NGSによる遺伝子発現測定のデータのみに絞りこめます



4. 上部の検索窓でキーワード(例えば、hypoxia)を入れて検索ボタンを押すと、そのキーワードを実験タイトルに含むエントリだけに絞りこめます



【参考】AOEを使って遺伝子発現データベースの統計を見ながら検索する

### **NCBI Gene Expression Omnibus (GEO)**

- NCBIが提供・維持管理している遺伝子発現情報(主にマイクロアレイ)のデータベース
  - https://www.ncbi.nlm.nih.gov/geo/

- 自分の興味のある発現データセットや遺伝子プロファイルを検索することができるだけでなく、それらの生データを自由にダウンロードすることが可能です。
  - 。 GEOのエントリについて(GEO ID番号の最初の3文字の意味)
    - GPL: Platform ー マイクロアレイチップの種類
    - GSM: Sample 1枚のマイクロアレイチップから得られたサンプルデータ
    - GSE: Series 1つの実験で得られたGSMのセット
    - GDS: DataSet NCBIのスタッフが解析可能なデータを集めて再編成したGSMのセット
- 【参考】NCBI GEOの使い方1~マイクロアレイデータの検索・取得~2017

#### **EBI ArrayExpress**

- EBIが提供・維持管理している遺伝子発現情報のデータベース
  - https://www.ebi.ac.uk/arrayexpress/
  - 。 NCBI GEOのデータも取り込んでいて、こちらのほうがデータ数が多い
    - が、2017年10月にそれを止めるとアナウンス
- EBI OmicsDIから検索可能
  - https://omicsDI.org/
  - 。 インターフェースが使いやすい

#### 【実習2】OmicsDIを使って、興味あるオミックスデータセットを検索する

- 1. OmicsDIのサイトにアクセス
- 2. 検索窓に'cancer'と入力、'search'ボタンをクリック
- 3. 左上の'Transcriptomics'をクリックして結果を絞り込む
- 【参考】OmicsDI を使って様々なデータベースからオミクスデータを縦断的に検索する

# 個々の遺伝子の発現プロファイルを調べる

# RefEx (Reference Expression dataset)

- ヒト、マウス、ラットの遺伝子発現情報リファレンスデータセット
  - http://refex.dbcls.ip/
- 4つの異なる実験手法(EST、GeneChip、CAGE、RNA-seq)によって得られた正常組織、初代培養細胞、細胞株における遺伝子発現データを検索、閲覧可能
  - 。 最近新たに、FANTOM5 CAGEデータが追加(ヒト556種、マウス286種)
  - 。 掲載しているデータやオリジナルデータなどの詳細については、RefExについて
- RefExで掲載されているデータはすべて再利用可能
  - 。 下部の「ダウンロード」からリンク
  - 。 figshare 【統合TV】というレポジトリにも doi:10.6084/m9.figshare.c.3812815
  - 。 「RefEx analysis」として論文に引用していただいたケースも
    - Aberrant IDH3a expression promotes malignant tumor growth by inducing HIF-1-mediated metabolic reprogramming and angiogenesis, Oncogene, (22 December 2014) I doi:10.1038/onc.2014.411
- 2017年に論文出ました doi:10.1038/sdata.2017.105
- このツールでできること
  - 。 正常組織における遺伝子発現データを調べる
  - 。 測定手法による遺伝子発現量の差異を比較する
  - 。 組織特異的遺伝子をワンタッチで検索可能
  - 。 遺伝子発現解析などで見出された不詳な遺伝子群の機能および関係性を調べる

#### 【実習3】RefExを使って、組織特異的遺伝子を検索する

- 【復習用 統合TV】RefExの使い方
  - 1. http://refex.dbcls.jp/ を開きます。

2. 画面中央の「組織特異的に発現する遺伝子を見る」の臓器アイコンにカーソルを合わせると、更に詳細な部位のアイコンが出るので、調べたい臓器(例として肝臓)をクリックします。



3. 検索結果一覧が表示されます。検索結果一覧では、「ソート項目の切り替え」や「絞り込み検索」、「リストへの追加」ができます。 (手順11以降で解説します。) 4. 各遺伝子の青字の部分 (例 fibringen alpha chain)をクリックすると詳細情報を閲覧できます。



5. 「ヒートマップ on Bodyparts3D」では、表示する部位の切り替え(全身・体幹部・頭部)ができます。「皮膚・骨格筋を表示」もしくは「アニメーション表示」にチェックを入れるとどのように表示されるでしょうか。

6. 「組織40分類別データ」では、バーの上にマウスオーバーすると測定部位と発現値が表示されます。

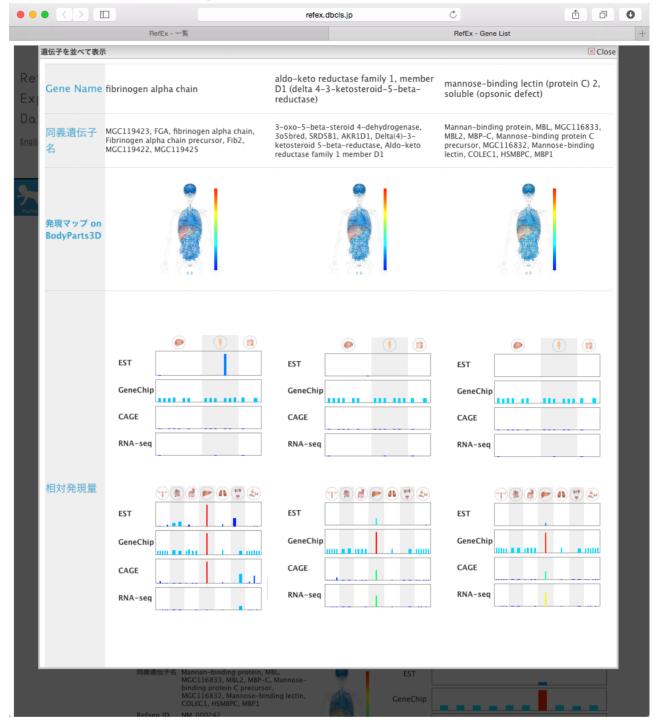


7. 「Download」をクリックすると、表示中の遺伝子の組織40分類別の発現データがタブ区切り形式でダウンロードできます。



- 9. 遺伝子オントロジー(GO ID)をクリックすると、そのGO termを持つ他の遺伝子を一括で検索できます。
  - 例として、GO:0007596 blood coagulation をクリックしてみましょう。
- 10. 右側のFANTOM5 CAGEのタブをクリックすると、FANTOM5 CAGEデータのビューアに切り替わります。
  - ビューアは上部が拡大図で、下部が全体表示になっています。
  - 検索窓にキーワードを入れるとサンプル名を検索できます。ヒットしたサンプルはオレンジ色で強調されます。
  - 右側に、サンプル名と発現値、サンプル分類が表示されます。
  - RefEx用に整理したサンプル情報一覧も閲覧可能です。
- 11. 検索結果一覧に戻ります。ソート項目を切り替えて、どのように結果が変わるでしょうか。
- 12. 様々な条件で検索結果を絞り込むことができます。絞り込み検索は左のバーから行えます。
  - 遺伝子名に「liver」を含むデータは何件あるでしょうか。
  - 「遺伝子名」の下の「条件なし」をクリックして表示されるウインドウに「liver」と入力し、「Include」をクリックし、「この条 件で絞り込み」を押します。
  - 「遺伝子名」の項目で「Exclude」に「solute」を加えると、検索結果はどう変わるでしょうか。
  - 「組織」の項目で、データ元をRNA-seqに変更したり、臓器の指定を追加すると検索結果はどう変わるでしょうか。
  - 「必ず含むデータセット」の「ALL」にチェックを入れると、検索結果はどう変わるでしょうか。
- 13. 個々の遺伝子の詳細情報は、リストに追加することで並列に比較することができます。
  - 肝臓特異的遺伝子の検索結果一覧に移動して、3つの遺伝子を「リストに追加」してみましょう。
  - 追加した件数は「リストを見る」の横に表示されます。
  - 「リストを見る」をクリックするとリストに移動します。
  - 『並べて表示する』にチェックを入れて、「遺伝子を並べて表示」をクリックします。

■ 並列に比較することで見えてくる「違い」はなんでしょうか。



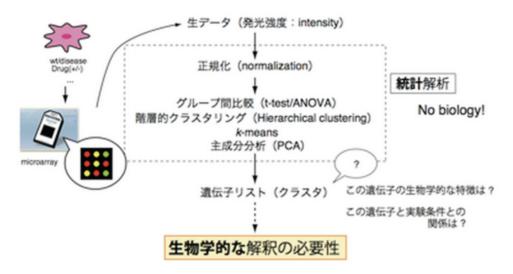


14. 自分の研究テーマに関連する、また興味のある遺伝子について検索してみましょう。

# 数十~数千の遺伝子群の生物学的解釈

### metascape

- マイクロアレイやRNA-seqデータの生物学的な解釈
- マイクロアレイやRNA-seq、すなわち遺伝子発現解析の一般的な目的は、実験条件によって得られた数十〜数千の遺伝子群の発現が生物学的にどういう意味を持つかを考えることです。



• 今回は、その方法の一つとして、遺伝子リストにある遺伝子群に機能アノテーション(Gene Ontologyなど)を付与し、生物学的な解釈を行います。

#### 遺伝子リストの準備

- サンプルデータとして、公共遺伝子発現データベースからメタ解析してえた遺伝子リストを用います。この遺伝子リストaffy10.txtは、ある刺激前後の2群間で発現増加した実験が10回以上あった遺伝子群のリストです。
  - → affy10.txt (右クリックして「新しいタブで開く」もしくは「名前を付けてリンク先を保存」してください。)
- このデータは、どのような実験から得られたデータなのか、どのように解釈できるのかをDAVIDを使って考察してみましょう!

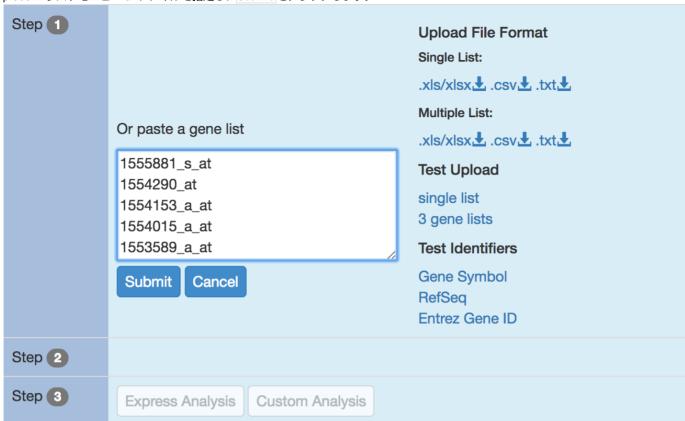
#### 【実習4】metascapeを用いて、発現データの結果を生物学的に解釈する

- 【復習用 統合TV】Metascapeを使って、遺伝子リストの生物学的解釈をする
- 1. metascapeのウェブサイトにアクセスします。

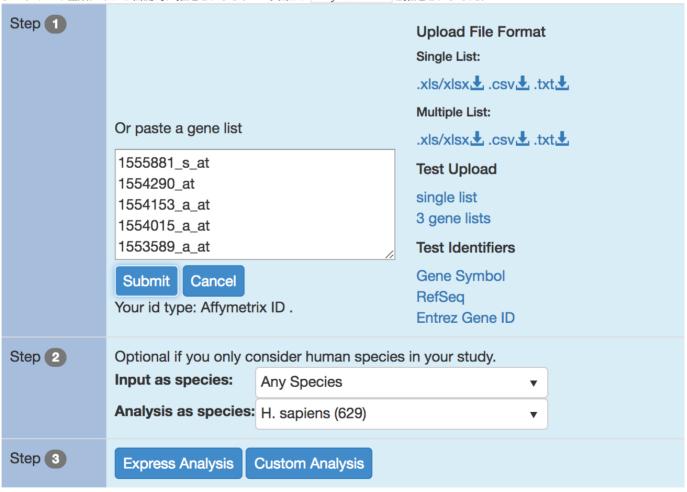




2. probe IDリストをコピペ or ファイルを指定し、Submit をクリックします。

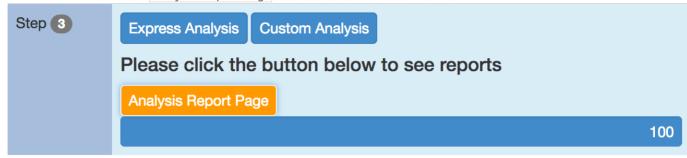


3. リストのIDの種類タイプが自動的に推定されます。... 今回は、Affymetrix IDと推定されました。



4. Express Analysis をクリックするとデータ解析が実行されます。

5. しばらく待つと解析が終了し Analysis Report Page が現れます。それをクリックします。



- 6. 解析を続けます。真ん中の「Functional Annotation Tool」をクリックします。
- 7. Enrichment解析の結果が表示されます。

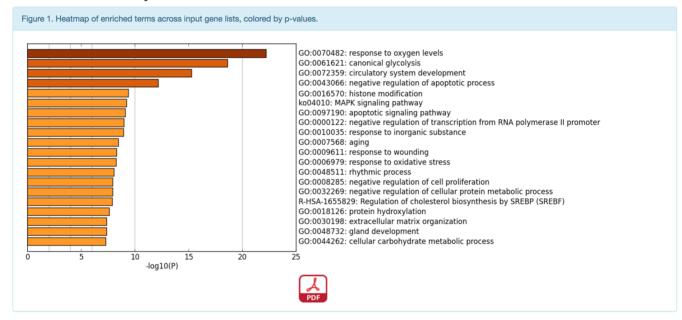
Gene List Report Excel Sheets Gene List Report PPT file All in One Zip File

Warning: this report page can be accessed within the next 72 hours; all data will be deleted afterwards.

# Metascape Gene List Analysis Report

metascape.org1

#### **Enrichment Summary**



- 8. 下部のPDFアイコンをクリックすると、Enrichment解析のヒストグラムがPDF形式でダウンロードできます。
- 9. Gene List Report Excel Sheets や Gene List Report PPT file をクリックするとそれぞれExcelとPowerPointのファイル形式で結果が得られます。
- 10. このページの下部には、Enrichment解析の詳細な結果などが表示されていますので見てみましょう。

#### 【応用1】複数のツールが統合されたサイトを用いて、発現データの結果を生物学的に解釈する

これまでの課題が早くできてしまった向けに。

iDEPというRNA-seqデータ解析でよく使われる発現差解析やパスウェイ解析のためのツールが統合されたウェブアプリケーションを試してみましょう。