公共の遺伝子発現データに対するキュレーションの付加価値

概要

公共データベースに登録されている遺伝子発現データの数は急激に増えています。マイクロアレイのデータが現在登録されているデータの大部分を占めている一方、RNA-seq のデータが新たに公共データベースに登録されるデータの中心になりつつあります。これらの遺伝子発現データは様々なデータベースに登録され公開されています。それぞれのデータベースにはデータフォーマットや入力制約について固有の要件があります。しかしながら、最も重要な要素は発現解析を行ったサンプルについての生物学的な情報が入手できるかという点です。生物学的な情報はしばしば最小限の内容で登録されるため、情報の欠失や矛盾、登録者間での異なる語彙の使用の原因となります。公共データベース間の不均一性はさらに大きく、データベースや実験が異なる遺伝子発現データを統合することはデータを手作業でキュレーションする労力なしには不可能です。結果的に、大量の遺伝子発現データに存在する価値は十分に生かされていません。あるいは、実験が正しく準備されなかったり、エラーが訂正されなかったり、サンプル情報が十分に記述されなかったりした場合、間違った結論が出されることもあります。

いくつかの企業が自社のツールやデータベースに公共データを統合する機能を組み入れてきました。それらの機能のうち一部は単に公共データベースのデータを解析して自社の製品で読込み可能にする機能である一方、残りは一般的にキュレーションと呼ばれる処理、つまり、データの検証、標準化、さらにはメタデータの充実といった追加作業の必要性を認識しています。しかしながら、キュレーションの度合いやキュレーションを行うのに必要は専門的知識には大きなばらつきがあります。Nebion 社のキュレーションチームは 9 年間、公共データベースの遺伝子発現データを系統的にキュレーションして、品質管理を行い、データを統合してきました。キュレーションにより世界的にユニークな遺伝子発現データの抄録がつくられ、数千の実験に対する横断的な検索や完全に統合されたデータマイニングを可能にします。

本書では公共の遺伝子発現データのキュレーションとデータ統合について、その**目的**と**方法**および**付加価値**について述べています。特に、キュレーションの深さ、必要となる専門的知識、**最も品質の高いキュレーション**結果を得るために必要なチームの特徴について焦点を当てています。

公共の遺伝子発現データの情報源

公共データベース

公共の遺伝子発現データのほとんどは大学や公的研究機関により生成されています。ほとんどの論文誌は発現解析を含む論文を投稿する際に公共データベースへの生データの登録を課しています。遺伝子発現データの公共データベースはいくつか存在しますが、最もよく使われているのは GEO (Gene Expression Omnibus), ArrayExress, TCGA, GTEx, dbGAP などです。

公共データのデータ量

2005 年から 100 万以上のサンプルのトランスクリプトームのデータが生み出され、一般に公開されています。例えば、GEO では 130 万以上のサンプルを含む 60,000 以上の実験の遺伝子発現データが公開されています。これは科学的に高い価値を持つ大量の情報を意味します。これらの公共データベースの間にはかなりの重複がありますが、個々の実験の情報量や書式は大きく変化する可能性があります。

GEO や ArrayExpress のような大規模な公共データベースでは全ての技術プラットフォームやデータタイプが登録可能です。例えば、GEO には 14,000 以上の異なるプラットフォームが登録されていて、各プラットフォームは特定の仕様を満たす技術を意味します。図 1 は、GEO でサンプル数上位 100 プラットフォームを降順に並べたプラットフォームごとの発現解析されたサンプルの数です。上位 100 プラットフォームで GEO に登録されたサンプルの 85%以上を占めています。2015 年 6 月現在、Affymetrix Human U133 Plus2 プラットフォーム(図中の赤線)が圧倒的な首位を占めていて、続くどのプラットフォーム(図中のオレンジ線、図 1 下の上位 10 プラットフォームリスト参照)よりも著しく多いデータが登録されています。

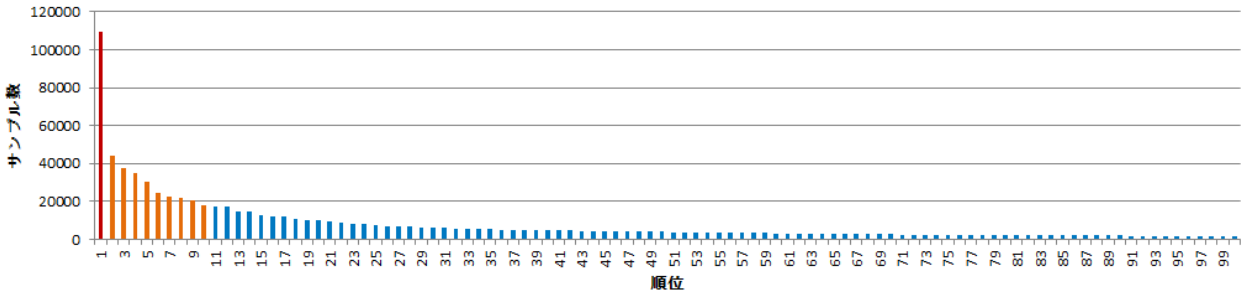

図 1 *GEO でサンプル数上位 100 プラットフォームを降順に並べたプラットフォームごとの遺伝子発現解析されたサンプル数。赤およびオレンジで図示された上位 10 プラットフォームは下の表を参照。サンプル数が最も多いのは 110,000 サンプルが登録されている Affymetrix Human U133 Plus2。*

GEO の上位 10 プラットフォーム(サンプル数が多い順にランキング、図 1 で赤およびオレンジで図示)

1. [GPL570] Affymetrix Human Genome U133 Plus 2.0 Array
2. [GPL1261] Affymetrix Mouse Genome 430 2.0 Array
3. [GPL96] Affymetrix Human Genome U133A Array
4. [GPL10558] Illumina HumanHT-12 V4.0 beadchip
5. [GPL13534] Illumina HumanMethylation450 BeadChip
6. [GPL6244] Affymetrix Human Gene 1.0 ST Array
7. [GPL13112] Illumina HiSeq 2000 (Mus musculus)
8. [GPL11154] Illumina HiSeq 2000 (Homo sapiens)
9. [GPL6947] Illumina HumanHT-12 V3.0 beadchip
10. [GPL6246] Affymetrix Mouse Gene 1.0 ST Array

データの種類

公共データベースは様々な種類のデータを保管しています。例えば、遺伝子発現データ(発現アレイ、RNA-seq、RT-qPCR)、変異情報(一塩基変異、コピー数変異)、メチル化、クロマチン免疫沈降、タンパク質の発現データなどです。図 2 に示すように、発現アレイは GEO に登録された情報の中で最も大きな割合を占めています。RNA-seq や ChIP-seq 等を含む NGS データが続きます。

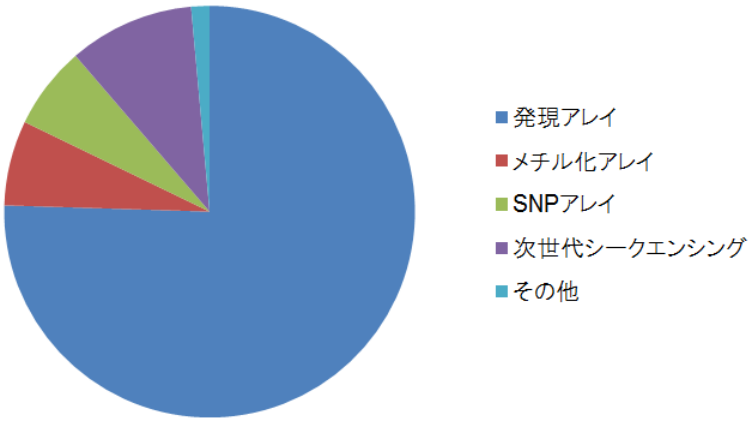

図 2 *GEO の上位 100 プラットフォームのデータの種類の割合。図中の全サンプルの数は 796,885。データの大部分である約 75%は発現アレイであり、GEO でいまだに最も重要なデータソース。*

公共データ収集の目的

公共データベースで遺伝子発現データを収集、保管、公開する目的は全ての研究者が遺伝子発現データに容易にアクセスできるようにすることです。具体的には、研究者が特定の研究の成果を検証することを可能にし、他の研究者の研究から自らの研究成果を支持する証拠を見つける助けになります。

Nebion 社では、さらに一歩進んで大規模なデータの統合を通して公共データの価値を拡大することを目標にしています。これにより研究者は解析の最初の目的を超えて遺伝子発現データを活用することができます。実際、多くの遺伝子発現データを組み合わせて解析することで個々の実験結果を解析することでは見つけることができない新たな成果を見出すことができます。

公共データの可用性

多くの場合、公共データは無償で制限なく利用できます。最近の NGS の発展により被験者のプライバシーの問題が提起されています。そのため、データに制限がかかる可能性があります。

公共の遺伝子発現データを使う際の問題点

一般的に、論文を投稿する前に生データは公共データベースへ登録されます。登録作業は大まかに管理されているので、次のような問題が起きる場合があります。

- **不均一な語彙**：実験はたいてい登録者の好む単語を使って記述されます。オントロジーの利用を強制された場合、ほとんど全ての登録者はオントロジーの利用に不慣れでオントロジーの規則や落とし穴について知らないため、オントロジーの利用は体系的ではありません。

- **情報の不足**：ほとんどの場合、主な実験条件のみアノテーション情報が付加されています。実験、サンプル、被験者についてのより詳細な情報は頻繁に欠けているか、対応する論文や補足情報の中にのみ記載されています。

- **不正確な記述**：各サンプルの性質について正確に記述するより、多くの場合より一般的な単語が使われます。例えば、正確な抽出物の情報や細胞の種類よりも“血液”という単語が使われます。

- **重複**：同じサンプル群(例えば対照群)が複数の実験で使われることがあり、その際に複数のアクセッション番号が割り当てられます。ヒトのデータをキュレーションした際に今までのところ、10.1%のサンプルが他の実験と重複していることを見つけています。

- **グループ化されていないサンプル**：公共データベースでは一般にサンプルを実験群ごとにグループ化することは明確にモデル化されていません。

- **サンプルの誤分類**：サンプルはたまに間違って分類されます。例えば、実験群の代わりに対照群に分類されたり、実際は女性の被験者が男性と分類されたりします。

- **統計的に低品質なデータ**：全ての実験データはたいてい著者または受託業者によって品質管理が行われますが、品質管理の方法や閾値は大きく異なります。マイクロアレイのデータでは、公共データベースに登録されているサンプルのうち平均して 5%が NEBION 社の基準を満たさないことが分かっています。サンプル調製に標準的な手法を使わないなど様々な更なる問題により、RNA-seq ではこの割合が 20%を超えます。

- **マッピングの違い**：プローブやリードの参照ゲノム配列や遺伝子モデルへのマッピングは登録者間で大きく異なる場合があります。

- **標準化手法の違い**：標準化のための手法は数多く存在します。従って、標準化済のデータは非常に不均一であり、メタ分析のために複数の実験を矛盾なく直接統合することはできません。

データの内容や種類、品質、標準化手法の高い不均一性により複数の実験でゲノムデータや遺伝子発現データを直接統合することはできません。公共データベースからダウンロードした個別の実験の解析でさえ、上記の問題によりしばしば煩雑です。多くの場合、これらのデータを解析する研究者は実験についてのより詳細な情報を得るために頻繁に原著論文を読まなくてはなりません。その作業は面倒で、時には自らの専門を超えるレベルです。

一貫性のない記述や不十分なアノテーションの実験は科学的な知見の主要な障害です。科学者が本質的な発見を見落とす原因になるだけでなく、間違った結論に導くこともあります。従って、実験デザインやサンプルの属性の系統的で正確な記述は間違いなく必要不可欠なものです。

キュレーションの目的

- 管理された用語を使った実験情報の**体系化**
- サンプルのアノテーションと患者のメタデータの**検証**
- 一貫性のない情報の**修正**
- 公共データベースから入手できないサンプルの説明を追加するアノテーションの**充実**
- 専門家による疾患特異的な用語への言い換えによるアノテーションの**高度化**
- 解析に適した(階層が浅く重複がない)オントロジーの利用による**使い勝手の向上**
- キュレーター間の再確認によるアノテーションの**ピアレビュー**

キュレーションによるデータ解析の品質と範囲の改善

公共データベースに登録された実験データはピアレビューを受けておらず、しばしば実験を行った研究者以外の人間が登録します。さらに、論文が論文誌に掲載される前に生データは登録されるので、公開される情報の量はしばしば最小限になります。高品質なアノテーション情報を得て公開された実験データを系統的に統合するために、手作業でキュレーションする膨大な作業が必要になります。**公共データベースに登録された生データに関して唯一信頼できる情報源は、実験結果を報告している原著論文です。**実験データを適切にキュレーションする前に、論文の *Materials and Methods* と *Supplementary Materials* を詳細に確認する必要があります。実験、サンプル、被験者の情報は管理された語彙に変換されなくてはいけません。Nebion 社はキュレーションについて下記の基準で標準業務手順(SOP)を確立しています。

品質管理

- 品質管理に様々な無償あるいは有償の統計解析ツールを使用

標準化

- 生データから 2 段階の手法で全ての実験間の global normalization

アノテーション

Nebion 社でキュレーションされた各実験は、以下のアノテーション手順が必要不可欠です。

1. 実験ごとの説明：例えば、タイトル、著者、論文誌、公共データベースへのリンク、アクセッション番号、対応するオントロジーを使った治療領域と実験デザインによる分類。

2. サンプルごとの説明：例えば、
    a. 採取したサンプルの**組織と細胞の種類**(Nebion 社の *Anatomy* オントロジーを使用)
    b. **細胞株**の名前と原発組織(Nebion 社の *Cell Lines* オントロジーを使用)
    c. **癌種**(*ICD-10* と *ICD-O-3* のオントロジーを使用)
    d. 薬の投与、疾患、ストレス、遺伝子型、変異などを含むそれぞれ対応する対照群と比較した**刺激**(Nebion 社の *Perturbations* オントロジーを使用)
    e. 生物のライフサイクルの段階に対応する**発生段階**(Nebion 社の *Development* オントロジーを使用)

f. 例えば、年齢、性別、民族、健康状態、喫煙の有無、生存、病気の進行、マーカー、治療、疾患の段階や度合い、薬の投与量といったサンプルの状態に関連する全ての臨床的特徴や患者特性を含む追加のサンプルの変数(Nebion 社の *State Variables* オントロジーを使用)

上記 a.から f.のそれぞれのアノテーションで使用するオントロジー(管理された用語を含む)が作成され、継続的に維持されています。キュレーションの過程で使用される変数の数は実験ごとに大きく異なりますが、ほとんどの場合に一貫して上記 a.から e.のアノテーションが付けられます。f.のアノテーションの変数の数は実験デザインだけでなくキュレーターが実験を調査して得られた情報の量に依存します。追加の実験条件を見つけて、効率よく効果的に解析するのに必要な全ての変数を規定することは大きな労力を要します。

3. アノテーションの構造化：例えば、
    a. サンプルを biological replicates ごとにグループ化
    b. 比較する実験群の定義(例えば、患者と健常人、処理群と対照群)
    c. 同じ著者で重複するサンプルを含む複数の実験の統合

4. バージョン管理システムを使ったオントロジーの作成と更新

検証

- 各実験を一人目のキュレーターがキュレーションを行い、二人目のキュレーターが独立して**ピアレビューをします**。
- 性別の割当とマイコプラズマ感染はマーカーを使って検証します。
- 実験とサンプルの整合性はデータを可視化して検証します。

結果

このキュレーション作業により Nebion 社は**統計的に高品質でアノテーション情報が充実した**データの抄録を作ることができます。具体的には、Nebion 社のキュレーションの特長は下記の通りです。

- 数値データを標準化して、実験をまたいだ検索やメタ分析を可能にします。
- 管理された用語を使って実験やサンプルの記述(メタデータ)を標準化します。
- 実験デザインに基づいてサンプルを正しくグループ化します。
- 追加の因子を記述することでサンプルのアノテーション情報を充実させます。
- 公共データベースに存在するアノテーションエラーを取り除きます。
- 実験を統合して不必要な重複を見つけて取り除きます。

キュレーターの専門的知識

信頼できるキュレーションは基本的な実験の十分な理解を前提としています。同様に、対**応する研究分野や疾患領域の専門的知識**を必要とします。Nebion 社のキュレーション部門は、神経、心臓、呼吸器、がん、皮膚、リウマチ、糖尿病を含むさまざまな治療領域の研究者から構成されています。長年にわたって各キュレーターは、腎臓、眼科、消化器、薬理、毒性など他の領域で 2 つ目の専門知識を習得しています。したがって、キュレーション部門はほとんどの研究領域の実験をキュレーションできます。

Nebion 社のキュレーターの大部分は、元はスイスの一流大学や製薬企業の実験中心の研究室で数年働いていた博士号を取得した研究者です。前職の経験からキュレーターは疾患特異的な詳細知識と同様に実験デザインや実験で使用された技術を完全に理解できます。

キュレーション部門の特徴

アノテーションの一貫性を最大にしてキュレーションの品質を最高にするため、Nebion 社では次の内容を保証しています。

- 情報と専門知識の連続性
- オントロジーおよびキュレーションでのその利用を含むキュレーションの全過程の十分な理解
- 管理された用語の一貫した使用

したがって、キュレーション部門は正社員の専門家から構成されています。キュレーションのために学生やアルバイトを雇用することはありません。

さらに、最大限の一貫性とキュレーション過程の全ての手順の標準化された処理を保証するため全てのキュレーターは**密接に交流するチーム**で仕事をしています。したがって、外部の企業やコンサルタントにキュレーションを外部委託することはありません。

キュレーションのメリット

公共の遺伝子発現データを専門家が手作業でキュレーションする主なメリットには次のようなものが挙げられます。

- **高度な標準化**
  - 制御された単語を使った一貫性のあるアノテーション情報を持つ情報の系統的な保存
  - 複数の公共データベースから集めた遺伝子発現データのフォーマットを統一
  - 遺伝子発現データの一括入出力
- **高品質なデータ**
  - ユーザーは各実験の品質を検証する必要がないため時間を節約できます。
  - ユーザーは論文で実験の詳細を検索する必要性を回避できます。
  - ユーザーは誤解を招く情報や情報の不足から間違った結論を出すことを避けることができます。
- **豊富な情報**
  - 単一のデータベースに集約された高度な多様性のある実験デザイン
  - サンプルや被験者の高度なアノテーション情報
  - いくつかの注目する分野の豊富な登録データ
- **より良い解析から得られるさらなる知見**
  - より正確な検索による詳細化解析と生物学的解釈
  - 数千の遺伝子発現データを同時に解析
  - 比較や検証の基準として参照データセットの利用
  - 高度な標準化の結果として複数の実験間の強固なメタ分析
  - サンプルの属性で測定値を集計可能
  - 他のアプリケーションへの出力用に関連する遺伝子発現データを直接かつ正確に集計

マイクロアレイや RNA-SEQ の公共データのキュレーションの例

実験例	*問題点*	*解決策*
例 1 GSE25640: 野生型マウスと FIZZ2 ノックアウトマウスの肺の発現データ	GEO での実験条件はブレオマイシン処理を 21 日間行ったと記載されている。一方、論文には処理を 7 日間行ったと記載されている。	キュレーターは著者に 7 日間が正しい期間であることを確認して、GEO の情報が間違っていることを裏付けた。
例 2 GSE12385: 身体活動により誘導された末梢血単核球の発現変動	GEO のマトリックスファイルの一部で、24 日間の運動の前後のサンプルに同じ値がアノテーションされていた。	キュレーターは対応する論文に記載された内容に基づいてアノテーション情報を修正した。
例 3 GSE32512: 高血糖と GCKR 遺伝子の一般的な変異	GEO ではほとんど何も情報が得られない。	キュレーターは対応する 2 つの論文からサンプルの変数を 14 個見つけて、それらに適切なアノテーションを行った。
例 4 GSE41177: 心房細動の患者の左心房での領域特異的な発現	GEO ではほとんど何も情報が得られない。	キュレーターは論文からサンプルの変数を 18 個見つけてアノテーションを行った。
例 5 GEO に登録された 150 個のシロイヌナズナの RNA-seq の実験	詳細なキュレーションを行うために必要ないくつかの問題点がありました。	キュレーターは次のような結論に至りました。 • 実験のうち 37%は replicates がなかったり 4 サンプル未満だったりする理由でキュレーションする価値がありませんでした。 • 残りのサンプルのうち 35%は生データの欠失、低品質、3’末端のみのシークエンス、ダウンロード不可などさまざまな問題がありました。
例 6 Nebion 社で 2015 年 5 月にキュレーションを行った 23 個のヒトのマイクロアレイ実験から構成される代表的な実験群	この実験群は、キュレーションして GENEVESTIGATOR に統合することを顧客から依頼された GEO の典型的な実験群です。	キュレーターは次のような結論に至りました。 • 23 個の実験のうち 10 個で、少なくとも 1 つのサンプルが低品質の理由から除外する必要がありました。いくつかの実験では、最大 7%のサンプルが十分な品質を持っていませんでした。 • 23 個の実験のうち 18 個で、GEO には存在しない新しい実験条件(例えば、年齢、性別、治療計画、生存、遺伝子型)のアノテーションを付けました。アノテーションされた実験条件の数の平均は、キュレーション前が 5.5、キュレーション後が 9.2 でした。 • 1 つの実験の対照群のサンプルが他の 2 つの実験で重複していることが分かりました。3 つの実験全てを修正しました。 • 生データがないため 1 つのサンプルを削除しました。

NEBION 社のキュレーションの優先順位

公共データベースに登録された遺伝子発現データは、大学や公的研究機関に大規模な予算が付く治療領域、例えば癌や神経疾患、に強く偏っています。希少疾患や一般的には興味をもたれない実験条件はほとんどデータが収集されていません。

しかしながら、データ探索や標的遺伝子／バイオマーカー発見の視点から多種多様な生物学的情報の利用できることは重要です。例えば、特定の化合物の兆候を発見するには、できるだけ多くの疾患由来のデータが存在する必要があります。また、バイオマーカーや標的遺伝子の候補を発見あるいは検証するために、実験条件の大規模な一覧を利用できることは必要不可欠です。したがって、Nebion 社は次のようなキュレーションの優先順位の基準を使ってキュレーションされた遺伝子発現データベースを作っています。

1 番目：**顧客の要望**に基づいた実験のキュレーション（これらの実験は最も高い優先度です）

2 番目：疾患、化合物、遺伝子型、その他の実験条件の数が最大になるような**基本的な遺伝子発現データの一覧**の作成と拡張

3 番目：少数の**詳細な治療領域**に着目したキュレーション

- 癌
- 神経変性疾患
- 呼吸器疾患
- 自己免疫疾患
- 糖尿病と代謝性疾患
- 循環器疾患

治療領域ごとの登録内容

治療領域ごとに分割した GENEVESITGATOR のキュレーションされた遺伝子発現データベースの現在の登録内容は図 3 の通りです。

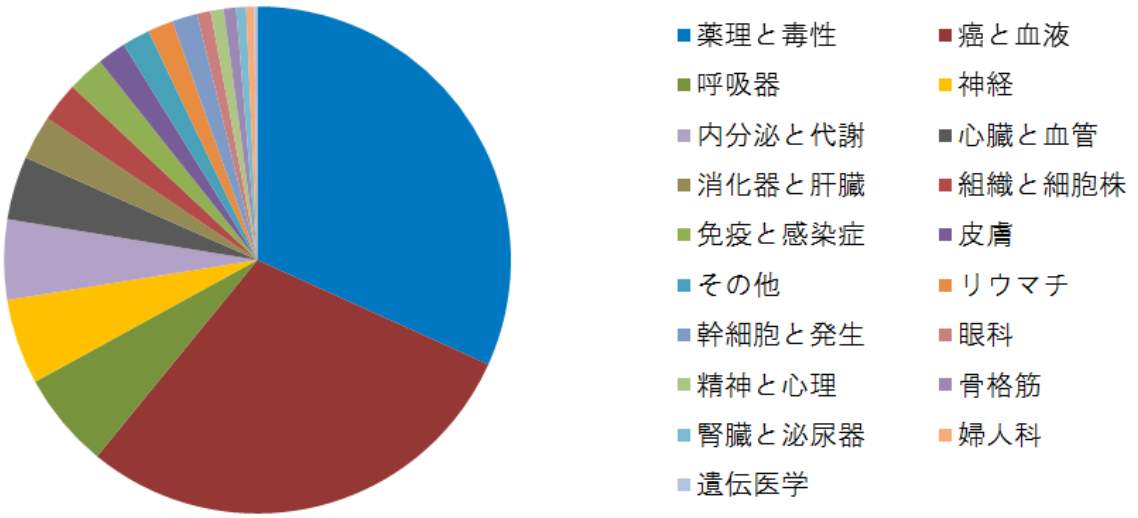

図 3 *2015 年 3 月時点の GENEVESTIGATOR(製薬企業向け)の全てのキュレーションされた遺伝子発現データ。全ての実験が複数の領域に帰属させることができるかもしれませんが、一つの治療領域に帰属されています。毒性が最も大きな区分ですが、この領域はほとんど全ての治療領域に関連があります。治療領域の中では、癌(28%)、呼吸器疾患(6%)、神経(5%)、内分泌(5%)、循環器疾患(4%)、消化器(3%)が最も多い領域です。このグラフの全サンプル数は 110,000 です。キュレーションされた遺伝子発現データ全体と比べた神経、呼吸器、循環器、皮膚、リウマチ、血液の割合は公共データベースにおける割合より高いです。すなわち、これらの領域は Nebion 社のデータベースで濃縮されています。癌はこのデータベースで最も大きな区分ですが、その割合でさえ公共データベースにおける割合よりも高いです。*

薬理と毒性に関連する登録内容

図 3 で最も大きな区分である毒性は 2015 年 6 月時点で次のような登録内容です。

毒性に関連した実験数(サンプル数)

- ヒト: 110 (13,118)
- マウス: 55 (1,572)
- ラット: 58 (26,465)
- 全体: **223 実験 (41,155 サンプル)**

評価された重複のない化合物の数

- ヒトのサンプルで 1,521
- ラットのサンプルで 520
- マウスのサンプルで 70
- 全体: **1,761 化合物**

生物種

生物学的な情報の多様性の観点から、製薬業界や食品業界、化粧品業界向けの GENEVESTIGATOR のデータベースは 2015 年 6 月時点で次のような登録内容を含みます。

生物種	**実験数**	**サンプル数**	*実験条件*	*遺伝子型*	*組織*	*ガン*
ヒト	**1,086**	**73,471**	5,748	1,534	377	649
マウス	**383**	**9,411**	814	339	284	6
ラット	**163**	**28,893**	7,269	28	120	0
ブタ	**43**	**1,261**	175	24	55	---
ショウジョウバエ	**123**	**2,345**	192	258	48	---
酵母	**63**	**1,771**	168	156	4	---
大腸菌	**21**	**617**	118	90	0	---

- 「実験条件」は異なる外部または内部の刺激(外的な要因または遺伝子組み換え)の総数を意味します。
- 「遺伝子型」はヒトでは異なる細胞株を意味する一方、マウスでは疾患モデルマウスあるいはノックアウト／ノックダウンマウスを意味します。
- 「組織」は異なる種類の器官、組織、初代培養細胞の数を意味します。
- 「ガン」は ICD-10 と ICD-O-3 で分類された異なる種類の癌を意味します。

お問い合わせ先

キュレーションのメリットは GENEVESTIGATOR をお使いいただくとはっきりと実感できます。GENEVESTIGATOR は統合された遺伝子発現データ全体に対して他にない検索を提供します。キュレーションや登録内容についての詳細は下記までお問い合わせ下さい。

メール：sales@molsis.co.jp

ウェブサイト：https://www.molsis.co.jp/