資料２－２

ライフサイエンス分野における、共通基盤としてのデータベース構築および基盤技術開発について

平成30年1月18日
国立研究開発法人科学技術振興機構
バイオサイエンスデータベースセンター

NBDC
National Bioscience Database Center

NBDCの概要

＜NBDC（National Bioscience Database Center）について＞

我が国におけるライフサイエンス研究の成果が、広く研究者コミュニティに共有かつ活用されることにより、基礎研究や産業応用研究につながる研究開発を含むライフサイエンス研究全体が活性化されることを目的として、**様々な研究機関によって作成されるライフサイエンス分野データベースの統合**に向けた、戦略の立案、ポータルサイトの構築・運用及び統合化に必要な研究開発を、平成２３年より推進している。

＜主な取り組み事項＞

※データベースの統合：RDF化等のセマンティック技術や、検索技術、オントロジー（高度な辞書）等を導入し、複数のデータベースを、縦横に検索・活用可能とすること。

※RDF： Resource Description Frameworkの略。セマンティック技術、LODを実現するための統一的な枠組み。

これまでの主な活動

◆ ライフサイエンス系データベースカタログ、データベース横断検索の運用

・厚生労働省、農林水産省、経済産業省各省のデータベース統合を推進している機関と連携
・カタログ（約1,600件）は国内の公開データベースはほぼ網羅、うち約600件が横断検索可能

◆ ライフサイエンス系データベースのアーカイブ化

・アーカイブ化（約130件）により、プロジェクト終了後も研究結果が継続的に公開・利用できるよう整備

◆ RDFポータルの開設

・分野横断的な研究の促進等に貢献するため、連携が容易で機械可読なＲＤＦ形式で統一したデータベースを集積したポータルサイトを平成27年11月に開設
・平成29年末現在、20件（約450億トリプル）のRDF形式の生命科学データベースを用意

◆ NBDCヒトデータベースの構築・運用

・ゲノム情報や画像情報等研究データを広く研究者間で共有するため、倫理面に配慮したガイドライン等を策定し、構築した国内初のプラットフォーム
・平成29年12月末時点で約130件の産学の研究プロジェクトから約30万検体分のデータ提供申請
・我が国で産出される人体由来データの収集と世界的な共有において中核的な拠点となっている

◆ 分野別データベースの統合支援

・ファンディングにより生物種やオミクス別のデータベース統合を推進（累計29課題）
・統合を支援したデータベースは、世界3極の一部を担うなど、それぞれ当該分野において枢要な存在となっている

◆ 公開前データのプロジェクト内での共有・活用支援

・AMEDと連携し、データの公開に先駆けて、プロジェクト内やグループ内におけるデータの共有を可能にする「AMEDゲノム制限共有データベース（AGD）」を 立ち上げ
・CREST「環境変動に対する植物の頑健性の解明と応用に向けた基盤技術の創出」領域において、トランスクリプトームデータの領域内共有を支援

http://www.biobank.amed.go.jp/itc/index.html

NBDCはFAIR原則※に先駆けて、同様の考えに沿った活動を展開

※データ共有を推進する研究者、出版社、ファンディング機関が参画する国際コミュニティ「FORCE11」が提唱した、データ公開に係る原則。Findable, Accessible, Interoperable, Re-usableの頭文字をとったもの。

分野別統合と分野を超えた統合のための技術開発

ファンディングによって分野別（生物種やオミクス別）のデータベース統合を推進するとともに、分野を超えた統合を実現するための基盤技術開発を実施。

分野別統合データベース群の構築支援

- 統合化推進プログラムによって、分野別に研究データの収集・標準化や、他のデータベースとの連携・統合化とそれに必要な技術開発、インターフェース設計、データ利活用のためのツール開発等を行っている。
- 開発したデータベースは、当該分野において枢要なものとなっている。

統合的に活用

統合的に活用

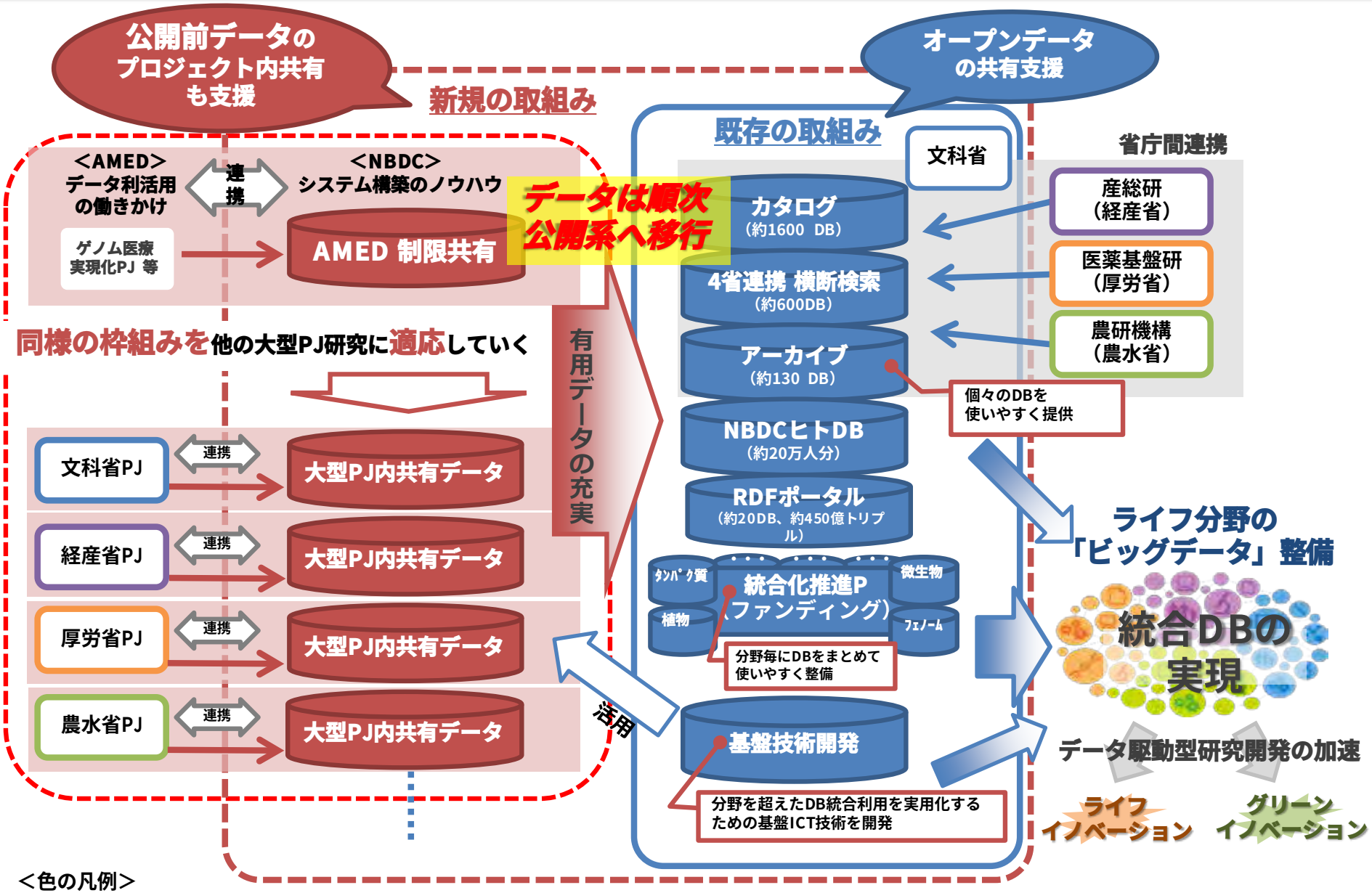
分野を超えた統合のための技術開発

- データ駆動型サイエンスに対応し、機械可読で統合可能なデータ整備（RDF化・オントロジー整備等）を行い、高度なデータ統合検索技術を開発
- データベースを最大限活用するための関連ツールの開発
- データベース統合化に向けた国際的標準化活動

など

ライフサイエンスにおけるビッグデータの活用基盤

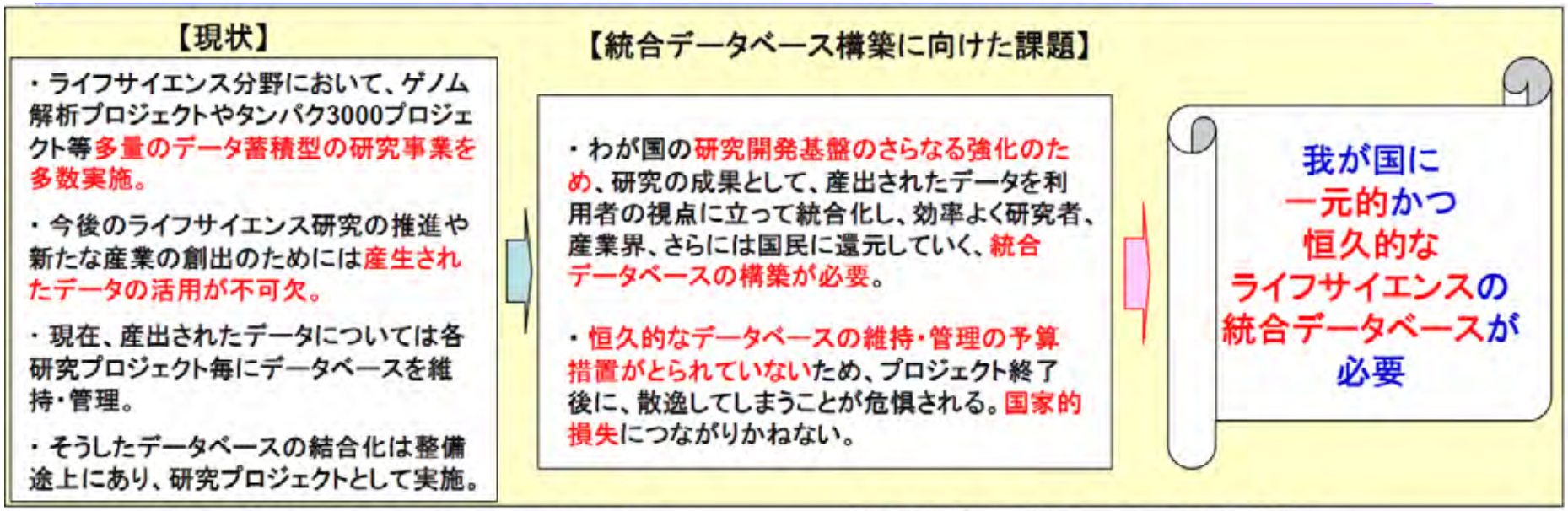
データベース活用に向けた新たな取組



＜色の凡例＞

大型PJ等と連携しつつ、NBDCが運営支援するもの	NBDCにおいて整備するもの	各機関において整備するもの

参考：NBDC発足について



「統合データベース タスクフォース報告書」（総合科学技術会議　ライフサイエンスPT）（H21年5月）より
⇒ライフサイエンス分野における我が国全体の恒久的且つ一元的な統合データベースの整備について方針をとりまとめた

JSTが実施していた「バイオインフォマティクス推進センター事業（BIRD）」と文部科学省が実施していた「統合データベースプロジェクト」が統合され、平成23年度にNBDCが発足

参考：NBDCの今後の取組み方針

○ 大型プロジェクト研究と連携し、未公開データのプロジェクト内での共有を支援する

公開前からの連携により、公開しやすく利活用性の高いデータの整備を支援するとともに、大規模かつ統一した様式で利活用しやすいデータの集積を目指す。

○ 応用につながる領域に焦点をあて、基礎研究データの統合を行う

AMEDやCRESTとの連携や、ファンディングで開発するデータベース、NBDCヒトデータベースを活かして、利用者からのフィードバックを得つつ、国内外の既存データや知見の統合を目指す。

○ 利用者との協業により、統合データの利活用に取り組む

大規模データ解析や人工知能の分野の研究者・機関との共同研究に取り組むほか、ファンディングによる研究開発においても利用者との協業を加速する。これらにより、データ解析・インフォマティクス分野と、生物科学分野の双方の利用者の観点をデータベース整備に反映し、データ利活用による仮説立案とデータ解析を支援する。

※ 研究データを保有する機関が、データ公開に取り組む際に、 NBDCはノウハウ・技術・オントロジーの提供等で支援する。なお、NBDCが当初からデータベース構築で連携するプロジェクトに関しては、データ公開（将来的なデータ公開も含む）を前提として、FAIR原則に則ったDB構築についてのノウハウ・技術・オントロジーの提供や開発支援等で連携・支援する。

参考：知識発見を支えるデータ基盤の構築

カタログ
データベースの所在情報を提供
国内の公開データベースは、ほぼ網羅。

横断検索
複数データベースを横断的に検索
Google検索での埋没リスクを回避。
国内の実施可能分は、ほぼ網羅。

アーカイブ
「統一フォーマット」でのダウンロードの実現
各省データベースをガイドラインに沿って
ライセンスを明確に整理しアーカイブ化。

データベース再構築
意味づけされた用語を整備
しより高度な検索を実現

Findable!
Accessible!
Interoperable!
Re-usable!

FAIR principles
データ共有を推進する研究者、出版社、ファンディング機関が参画する国際コミュニティ「FORCE11」が提唱。

FAIR principlesと同様の考え方で、より早い時期からデータベース統合を実施
（4省連携の具体的方策として上記4ステップを提案、推進）

参考：各サービスの統合状況と利用状況

1．データベース統合数

	H24	H25	H26	H27	H28
データベースカタログ	1,258	1,362	1,421	1,544	1,597
データベース横断検索	355	418	504	568	612
データベースアーカイブ	60	80	99	113	129

		H25	H26	H27	H28
NBDC ヒトデータベース	公開研究プロジェクト数	4	15	35	52
	公開待機	2	10	12	27

2．NBDC関連サイト利用状況（統合化推進プログラム、基盤技術開発を除く）

	H24	H25	H26	H27	H28
アクセス数（年間合計）	2,895,000	4,088,000	4,047,000	4,247,000	4,547,000
ユニークIP数（月間平均）	15,000	41,000	53,000	35,000	40,000

※数値は各年度末時点のもの

参考：分野別統合データベース活用事例

ファンディングにより構築を支援している、分野別統合データベースを利活用した事例の一部を下記に示す。

＜事例１＞製薬企業での活用事例

新たな「痛み止め薬」を開発するにあたり、痛みの感覚を抑える伝達物質とその受容体による鎮痛システムに着目。未知であった受容体の立体構造推定にあたり、PDBjに登録されている類似受容体の立体構造を活用。候補となる化合物が推定した受容体に結合するか、コンピュータ上でシミュレーションを行い、新薬候補物質の選定に至った。

鎮痛に関わる伝達物質は分かっていたが、受容体の立体構造は未知。

未知の受容体の立体構造をPDBjに登録されている既知の受容体構造から推定、新薬候補の絞込みに活用。

（模式図はイメージ）

＜事例２＞食品企業での活用事例

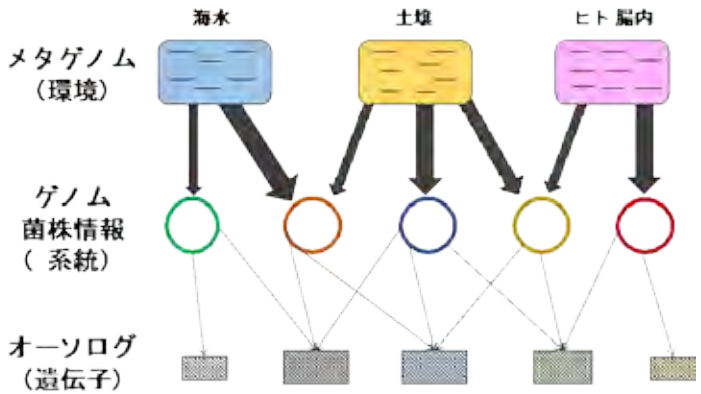
MicrobeDB.jpデータベースで提供している解析手法を用いて、乳児の腸内細菌がもつ遺伝子のメタゲノム解析を実施。母乳オリゴ糖の一種がビフィズス菌優勢の腸内フローラ（微生物叢）形成に重要であることが確認された。今後の製品開発への応用が期待される。

乳児の腸内細菌がもつ遺伝子を
MicrobeDB.jpで開発された手法で解析。

母乳オリゴ糖の一種フコシルラクトースがビフィズス菌優勢の腸内フローラ形成に重要であることを発見。

微生物統合データベース「MicrobeDB.jp」

微生物学の専門家のみならず非専門家も対象とし、微生物に関する膨大なゲノムやメタゲノム情報を容易に利活用できるDBを目指し開発。



微生物に関する膨大なデータを系統・遺伝子・環境の3つの軸に沿って整理し、ゲノムを核としてすべての情報を統合したフルRDFのDBを構築（約90億トリプル）

Data categories	Data sources	Ontologies
Genome	RefSeq Prokaryotes, Fungi, Algae	SO, FALDO, NCBITAX, INSDCO
Ortholog	MBGD	ORTHO
Culture collection	JCM, NBRC	MCCV, MPO
RNA-Seq	INSDC SRA	BAO
Genome & RNA-Seq Metadata	INSDC BioSample	MPO, MEO, MSV, PDO, CSSO
Metagenome	INSDC SRA	MEO, MSV

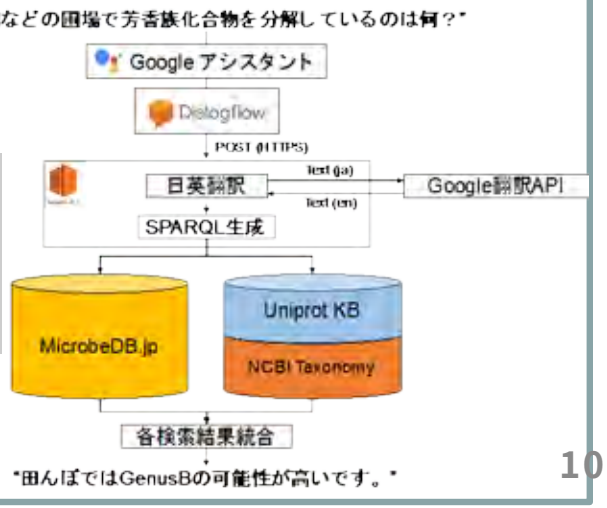
公開済みの約70万サンプルのメタゲノムデータ、約5万3千株のゲノム・ドラフトゲノムデータ、約1万7千株の菌株保存データを統合

RDFクエリ言語「SPARQL」により以下の問いにDB検索するだけで回答可能

- 肺炎に関与する細菌種はどのような環境に多く存在するのか？
- パーキンソン病の進行に伴い腸内細菌叢はどのように変化するのか？
- 活性汚泥が劣化する過程で顕著に優勢となる菌種および遺伝子は何か？
- ある汚染物質の分解に関わる遺伝子はどのような環境に多く存在するのか？
- 河川環境の汚染度の指標となり得る細菌群集パターンは？

Google Homeを利用した応答システム

MicrobeDBと国外のRDF-DBとの統合解析も可能



遺伝子発現統合データベース「RefEx」

RefEx

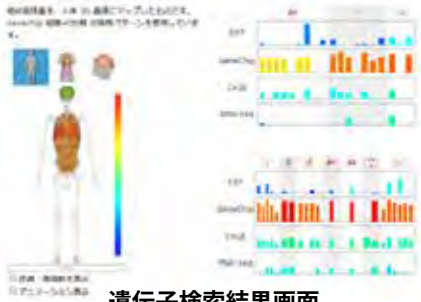
公共データベースにある再利用価値の高い遺伝子発現データを、測定サンプルや手法による発現量の違いをひと目で比較できるDBを目指し開発。

公共DBにある遺伝子発現データを収集・再解析・整理

RefExで採用している4つの実験手法	
EST	INSD（国際塩基配列データベース)のEST部門から得られたもので、各ESTエントリーのcDNA注釈をもとに組織別に分類し、カウントした発現データ
GeneChip	NCBI GEOから取得したAffymetrix社が作製したDNAマイクロアレイ「GeneChip」 によって測定された発現データ
CAGE	RIKEN FANTOM5 プロジェクトで集められたCAGEデータ
RNA-Seq	NCBI Sequence Read Archive、またはEuropean Nucleotide Archive より取得したIllumina Genome Analyzerで測定されたRNA sequencingの発現データ

活用事例

がん治療の標的となる遺伝子の候補を数十個得ていた研究者が、RefExを用いてそれらの発現状況を検索。
正常組織における発現量が低い遺伝子は治療標的として有用では、と仮説。
追加検証実験の対象を効率的に優先順位付け、絞り込むことができ、新規発見につながった。

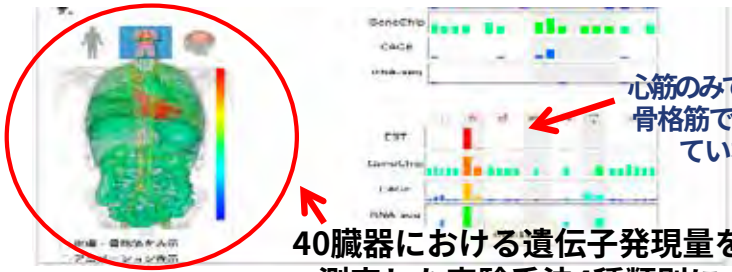


遺伝子検索結果画面

Aberrant IDH3α expression promotes malignant tumor growth by inducing HIF-1-mediated metabolic reprogramming and angiogenesis. Oncogene. 2015 Sep 3;34(36):4758-66. doi: 10.1038/onc.2014.411. Epub 2014 Dec 22.

臓器・組織特異的な発現パターンを示す遺伝子をワンタッチで検索

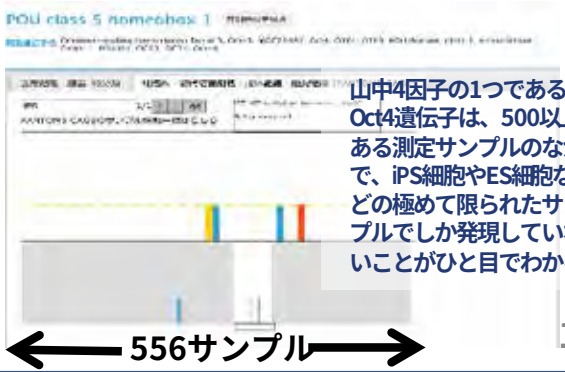
キーワード検索のほか、臓器や遺伝子ファミリーなどで検索可



心筋のみで発現し、骨格筋では発現していない

40臓器における遺伝子発現量を測定した実験手法4種類別に一覧することができる

臓器以外の**細胞株や初代培養細胞についても閲覧可能**



山中4因子の1つであるOct4遺伝子は、500以上ある測定サンプルのなかで、iPS細胞やES細胞などの極めて限られたサンプルでしか発現していないことがひと目でわかる

補足資料：NBDCの各サービスについて

カタログ
・国内外の公開データベース（約1,600 DB）の情報（名称、URL、概要説明、運用機関など）を収録。
・国内のデータベースについてはほぼ網羅、国外は主要なデータベースを収録。

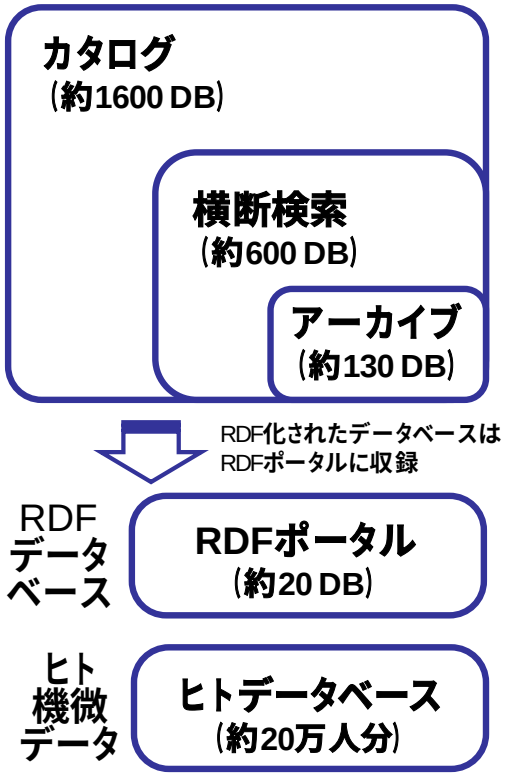
横断検索
・カタログに収録したデータベースのうち、技術的に可能なもの（ログイン不要など）、キーワード検索が可能なもの（属性情報や説明文が記述されている）約600 DBを検索対象としている。
・データベース内の個別データを横串で検索可能。

アーカイブ
・カタログに収録したデータベースのうち、大規模プロジェクトから産生されたもの、オリジナルサイトが運用を停止したもの、一括ダウンロード機能がない約130のデータベースを収録。
・データベースを丸ごとダウンロードして利用できるように、利用ライセンスを明確に整理し統一フォーマットで提供。

RDFポータル
・統合利用がしやすいよう、統合化推進プログラムや、DBCLS・DDBJでRDF化したデータベースを約20件収録。

ヒトデータベース
・国立遺伝学研究所DDBJと連携し、ヒトゲノム等のデータ、約20万人分を収録。
・収録データのうち、機微なものはアクセス制限付きで提供。

○ 関係イメージ図



※生物種やオミクス別のデータベース統合を推進する、統合化推進プログラムで開発したデータベースは、研究終了後、アーカイブ・RDFポータル・ヒトデータベースのいずれかに寄託。