

Use of Public Resources for the Omics Studies

オミックス研究のための公共リソースの活用

粕川雄也 (Takeya KASUKAWA)

takeya.kasukawa@riken.jp

理化学研究所 生命医科学研究センター
(RIKEN IMS)

横浜市立大学 生命医科学研究科
(Grad. Sch. of of Medical Life Science, Yokohama City Univ.)

自己紹介 / Introduction

- 現在の所属（本務）：

- 理化学研究所 生命医科学研究センター
生命医科学大容量データ技術研究チーム チームディレクター
- 生命科学に関するデータ（主に遺伝子発現・転写制御）についてのデータに関する研究
 - 転写研究活用のためのデータリソースの開発
 - 転写データ処理・解析手法の開発
 - 大規模データ生産のためのデータ管理（データコーディネーション）

- 過去の所属

- (2013-2018) 理化学研究所 ライフサイエンス技術基盤研究センター ユニットリーダー
 - 上に同じ
- (2005-2013) 理化学研究所 発生・再生科学総合研究センター 研究員/専門職研究員
 - 遺伝子発現解析（マイクロアレイ）、概日時計、脳領域の発現プロファイリング
- (2000-2005) NTTソフトウェア株式会社
 - 完全長cDNAアノテーションのためのシステム開発
- (1998-2000) 大阪大学大学院基礎工学研究科 助手
 - XMLを用いた生命科学データの交換法についての研究

今回の内容 / Contents

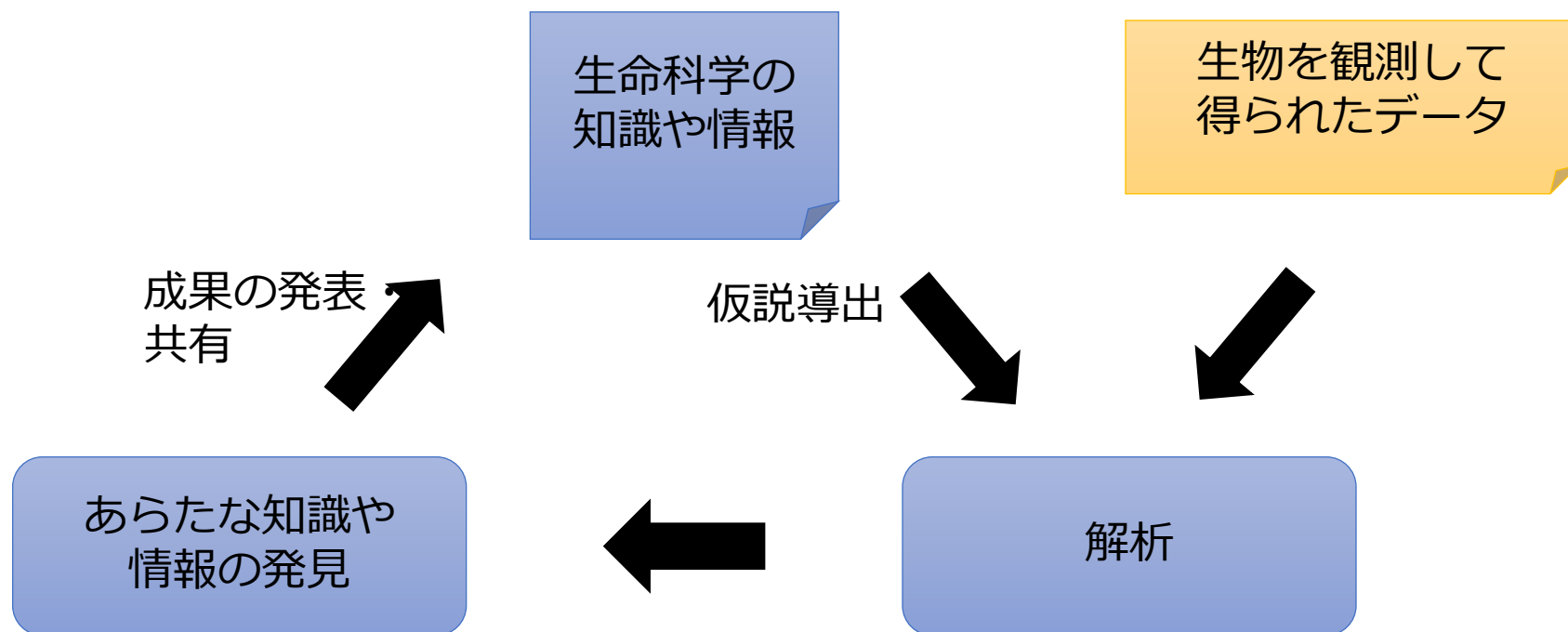
- 生命科学における「データリソース」「データベース」
“Database” and “Data resources” in life science
- 生命医学データ活用のためのデータリソース
Data resources for the use of biomedical data
- 我々が開発中のデータベース・データリソース
Databases/data resources developing in our team
- データベース研究の現在と今後の展開
Current and future of the database research

生命科学における「データリソース」「データベース」
“Database” and “Data resources” in life science

“Life-cycle” of the life-science studies (past)

生命医科学研究の進め方

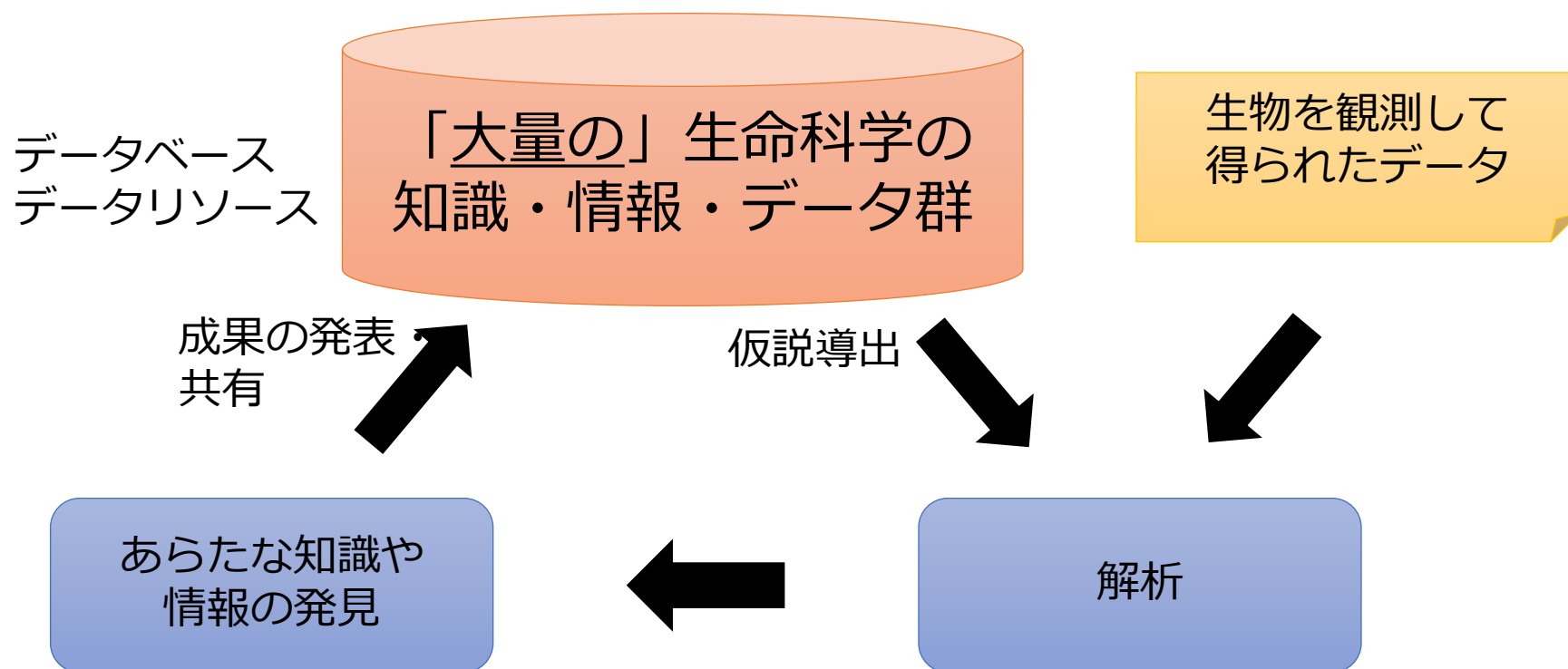
= データ取得・解析・蓄積のサイクル



“Life-cycle” of the life-science studies (current)

生命医科学研究の進め方

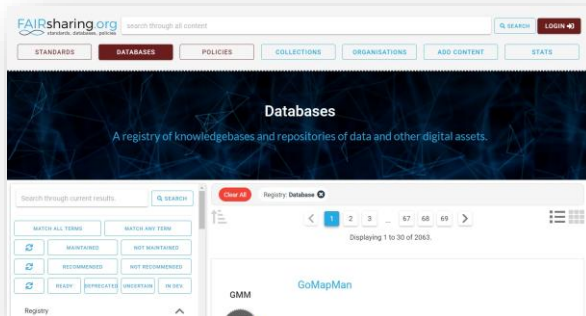
= データ取得・解析・蓄積のサイクル



生命科学研究・測定技術の発展により、利用可能な知識・情報・データが増大
→これらの活用が今後ますます重要に

Various life-science data resources/databases

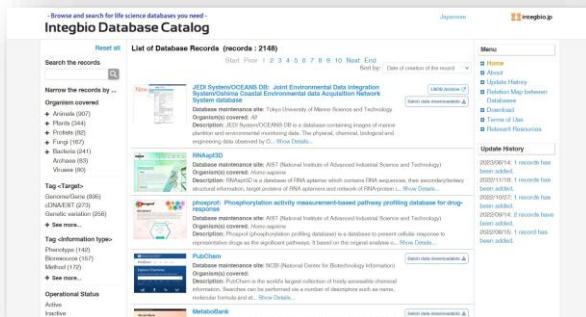
• 生命科学分野のデータベースの数



FAIRsharing

2,617 databases (as of Oct 25, 2025)

<https://fairsharing.org/databases/>



Integbio

2,582 databases (as of Oct 20, 2025)

<https://integbio.jp/dbcatalog/>



NAR Database summary papers

2,397 databases (as of Oct 25, 2024)

<https://www.oxfordjournals.org/nar/database/a/>

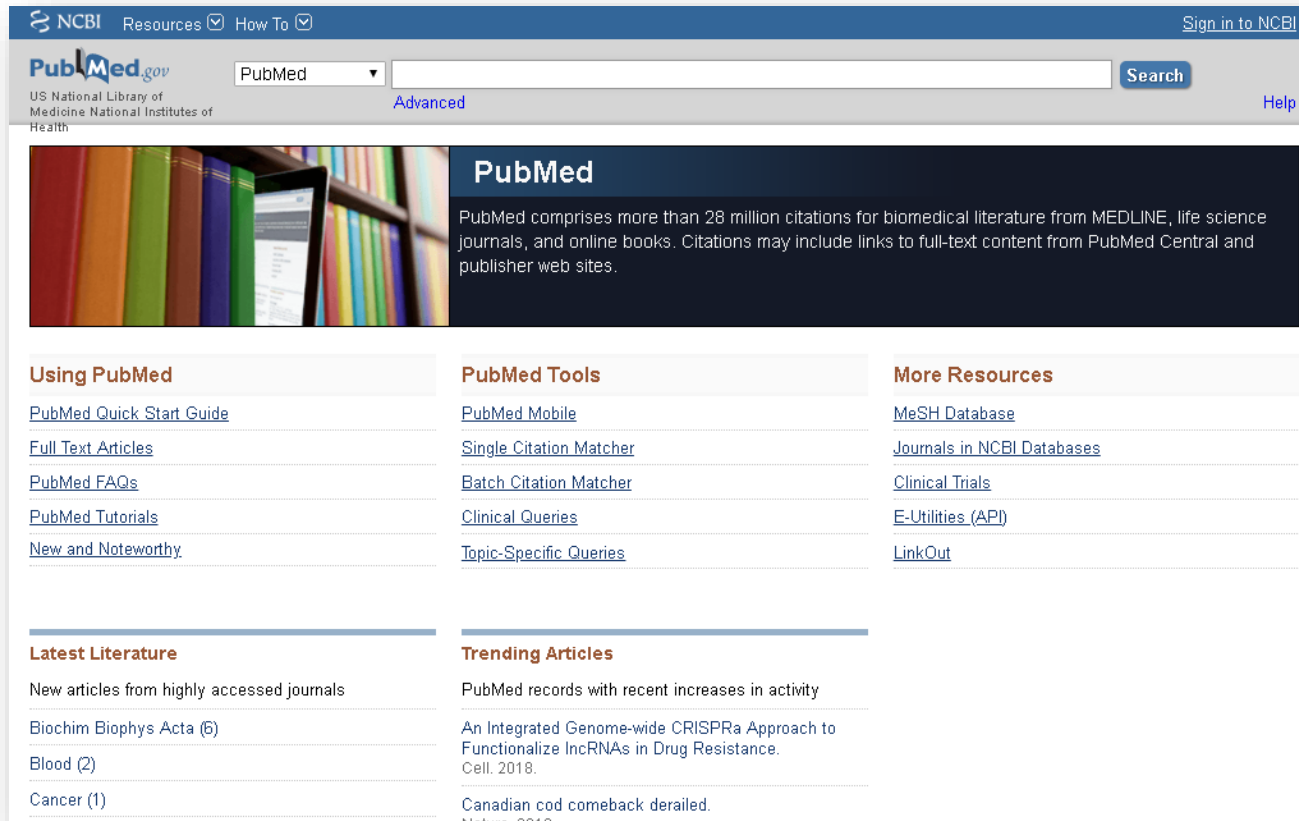
Various life-science data resources/databases

- どんなデータベースがあるのか？
 - ゲノム情報
 - DNA/RNA配列
 - 遺伝子情報
 - 遺伝子・タンパク質発現
 - アミノ酸配列
 - タンパク質ドメイン
 - タンパク質立体構造
 - 代謝パスウェイ・シグナリングネットワーク
 - 遺伝性疾患・ゲノム変異情報
 - 生物種情報
 - 文献・論文の情報
- などなど、多岐に渡る

Example of database (literature)

- データベースの例： PubMed
 - 出版された論文を検索できるデータベース
 - 生命科学、医科学系の論文をカバー

<https://pubmed.ncbi.nlm.nih.gov/>



Example of database (educational resource)

• TogoTV

<https://togotv.dbcls.jp/>

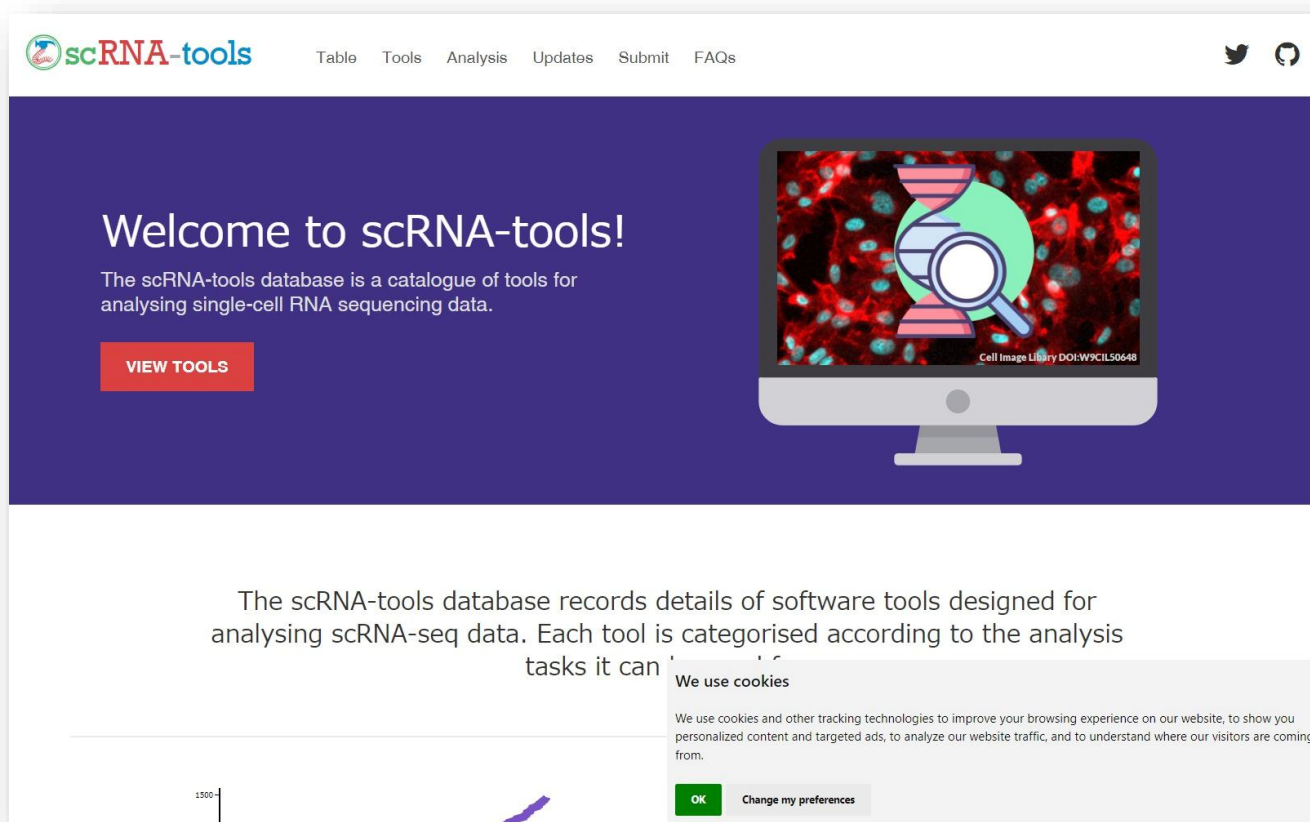
- データベースやツールの使い方を説明した動画のリポジトリ
- 狭義ではデータベースではないかもしれないが…

Example of database (bioinformatics tool)

- scRNA-tools

<https://www.scrna-tools.org/>

- 一細胞RNA-seqの解析に用いられるソフトウェアに関するデータベース



生命医科学データ活用のためのデータリソース
Data resources for the use of biomedical data

データリソースを活用するためには、生命科学データが
どのように表現されているかを知るのが非常に重要

オミックス研究で用いられる遺伝子・ゲノム
などのデータや情報がどのように表現されて
いるかを見ていきます

今回の講義で扱う情報・データ

- DNA配列データ
- ゲノム（領域）の情報
- 遺伝子の情報

用語について
Terms for life-science data

Terms in life-science data

• アノテーション / annotation

- 各種データや情報に、後から意味情報などを付与することや、その情報自身を指す
- 「注釈」「付加情報」とも呼ばれる。
- 元々は図書館学で、図書の分類や索引付けのことを指したもの
- 例えばゲノム配列について、どの部分が遺伝子といった情報や、遺伝子についてどういう機能を持っているかなど

• メタデータ / metadata

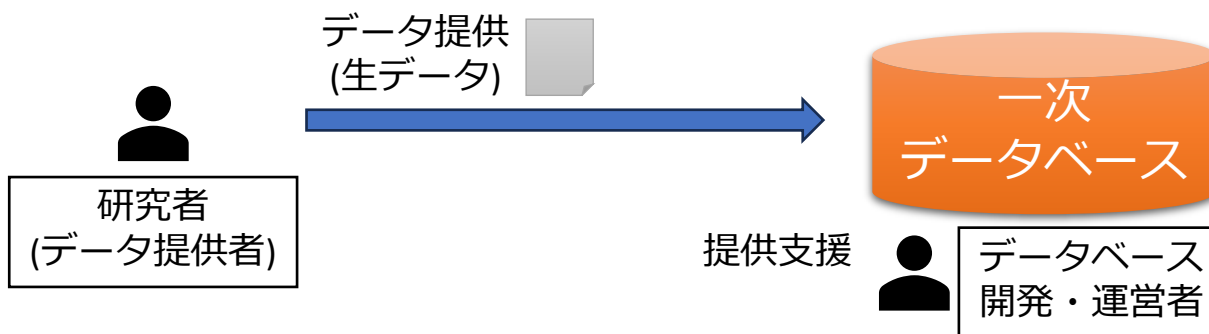
- ある「データ」について、そのデータの内容や属性などの情報/データを指す
- 「データ」を説明する「データ」ということで「メタデータ」と呼ばれる
- 例えば画像データについて、どういう対象を撮影したものか、サンプル準備条件、撮影情報など

• (マニユアル)キュレーション / (manual) curation

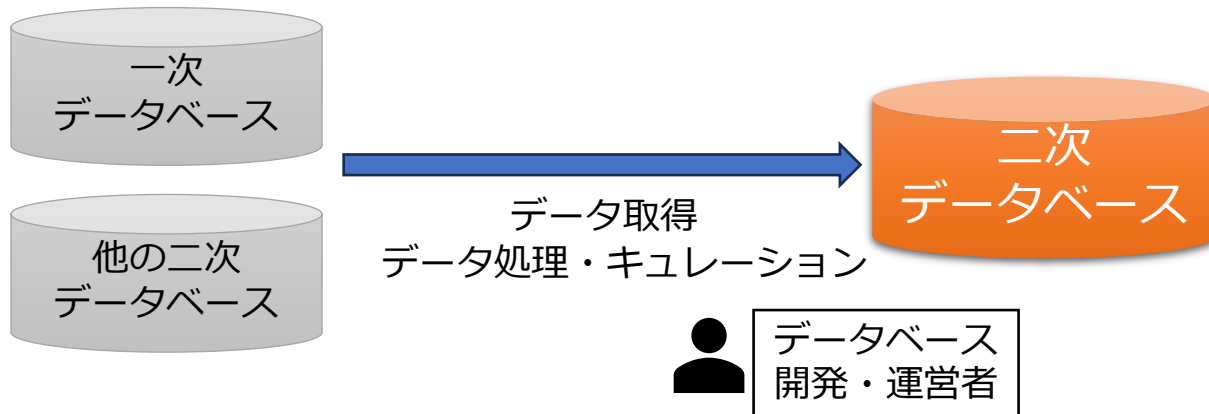
- 主として人手でアノテーション等を付与すること
- 元々は博物館学で、展示物にその情報を付けることを意味した

Classification of biomedical data resources/databases

- 一次データベース (primary databases)
リポジトリ型データベース (repository-type databases)



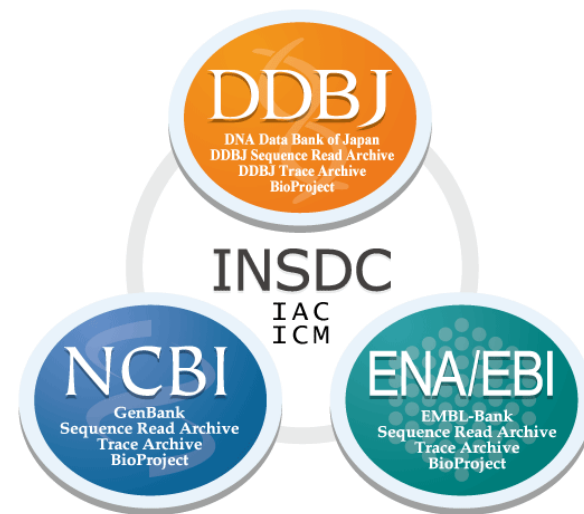
- 二次データベース (secondary databases)
知識ベース型データベース (knowledgebase-type databases)



DNA配列の情報
DNA sequence information

Treatment of DNA sequences

- DNA/RNA配列を読んで解析した論文を出すときは、配列データを公共リポジトリに登録することが事実上義務づけられている
- 登録するリポジトリ
 - INSDC (International Nucleotide Sequence Database Collection) のリポジトリのいずれか



<https://www.ddbj.nig.ac.jp/about/insdc.html>

INSDC repositories for DNA sequences

<https://www.insdc.org/>

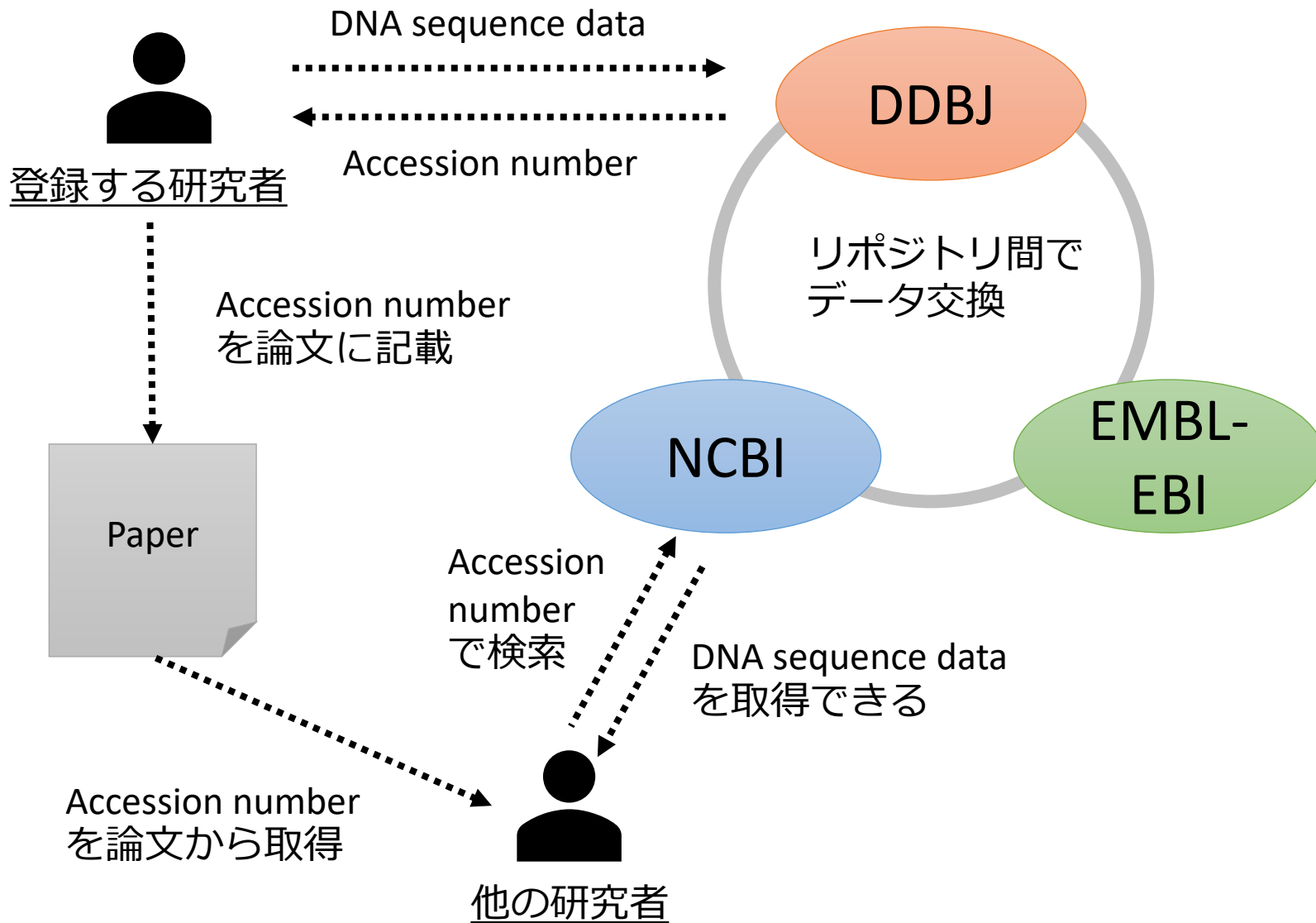
- DDBJ (DDBJ) / GenBank (NCBI) / EMBL (EBI)
 - DNA/RNA配列を蓄積した一次データベース（リポジトリ）
 - 基本的に、研究者がDNA配列を登録し、公開される
 - DNA配列だけでなく、そのメタデータも一緒に登録する
 - 3拠点で相互にデータ交換している（どこのデータベースに登録しても、他の拠点でも公開される）
- 運用開始年、DDBJ (日本・1987年)、GenBank (米国・1982年)、EMBL (英国・1981年)
- 大規模シーケンサー由来データ用に DRA (DDBJ), SRA (NCBI), ENA (EBI) というリポジトリが用意されている（最近はこちらが主）
- 4つのめのリポジトリとして中国の参加を準備中



Data type	DDBJ	EMBL-EBI	NCBI
Next Generation reads	Sequence Read Archive	European Nucleotide Archive	Sequence Read Archive
Assembled Sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

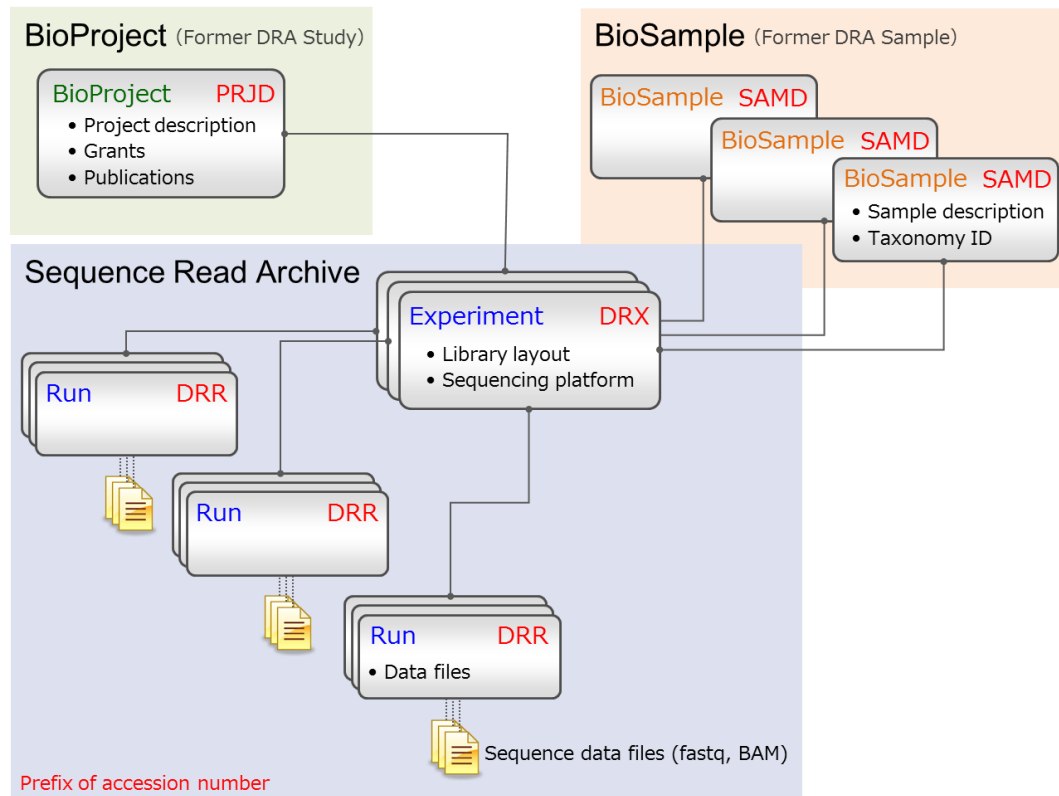
<https://www.insdc.org/>

Structure in INSDC



Biosamples and Bioprojects

- DNA配列データのうち以下はそれぞれ別のリポジトリに登録される
 - 採取元のサンプルについての情報→ BioSamplesへ登録
 - データを取得したプロジェクトについての情報→ BioProjectへ登録



Secondary data resources of sequences

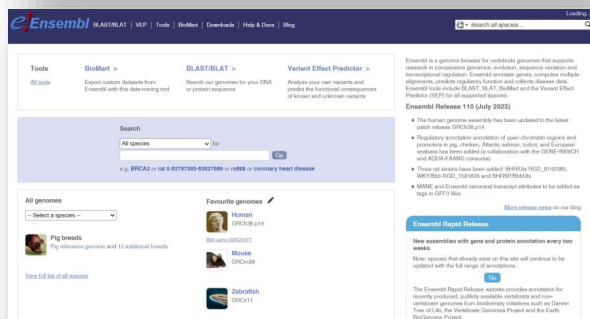


RefSeq

各遺伝子を代表するRNA/アミノ酸配列を決定
様々な生物種のものがある

NCBI (USA)により提供

<https://www.ncbi.nih.gov/refseq/>

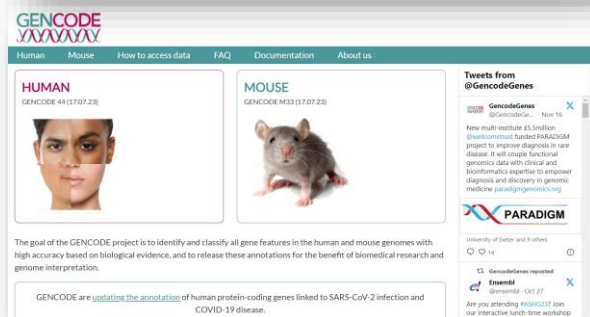


Ensembl Genes

ゲノムから予測した全遺伝子・RNA・アミノ酸配列
ゲノムの決定した様々な生物種のものがある

EBI (UK)により提供

<https://www.ensembl.org/>



GENCODE

ENSEMBL genesに加えて、いくつかのソースのデー
タを処理して作成された全遺伝子・RNAのセット
GENCODE Consortiumにより提供

<https://genencodegenes.org/>

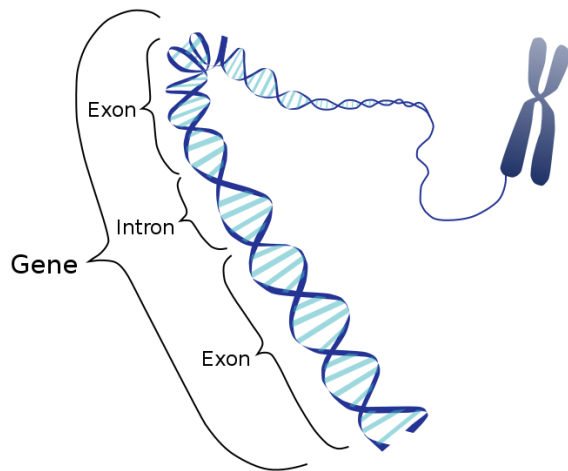
遺伝子発現解析等、さまざまなデータ解析に利用されている

ゲノム(領域)の情報

Genome and genomic region information

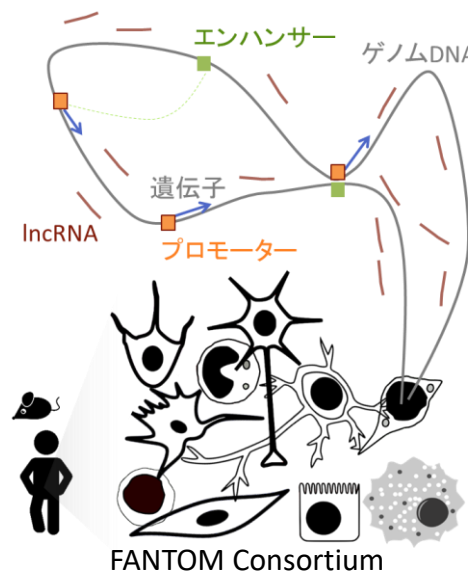
Various information in genomic regions

遺伝子領域



<https://www.biologyonline.com/dictionary/exon>

転写制御領域



ゲノム中の特定の領域に関するデータ・情報が蓄積されている

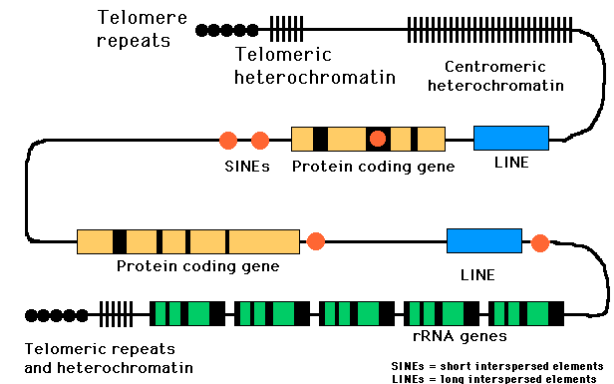
「ゲノム領域」の情報をどう扱っているか？

ゲノム変異領域

Structural Variation									
Reference	1	2	3						
Insertion	1	2	5	3					
Deletion	1	3							
Inversion	1	3	2						
Copy Number Variation	1	1	1	1	2	3			
Tandem Duplication	1	1	2	3					
Dispersed Duplication	1	2	1	3					
Mobile Element Insertion	1	2	Mobile Element	3					
Translocation	1								
	10	11	12	2	3				

Figure 1: Depiction of different types of structural variants compared to the reference genome. Each different number represents a different gene.

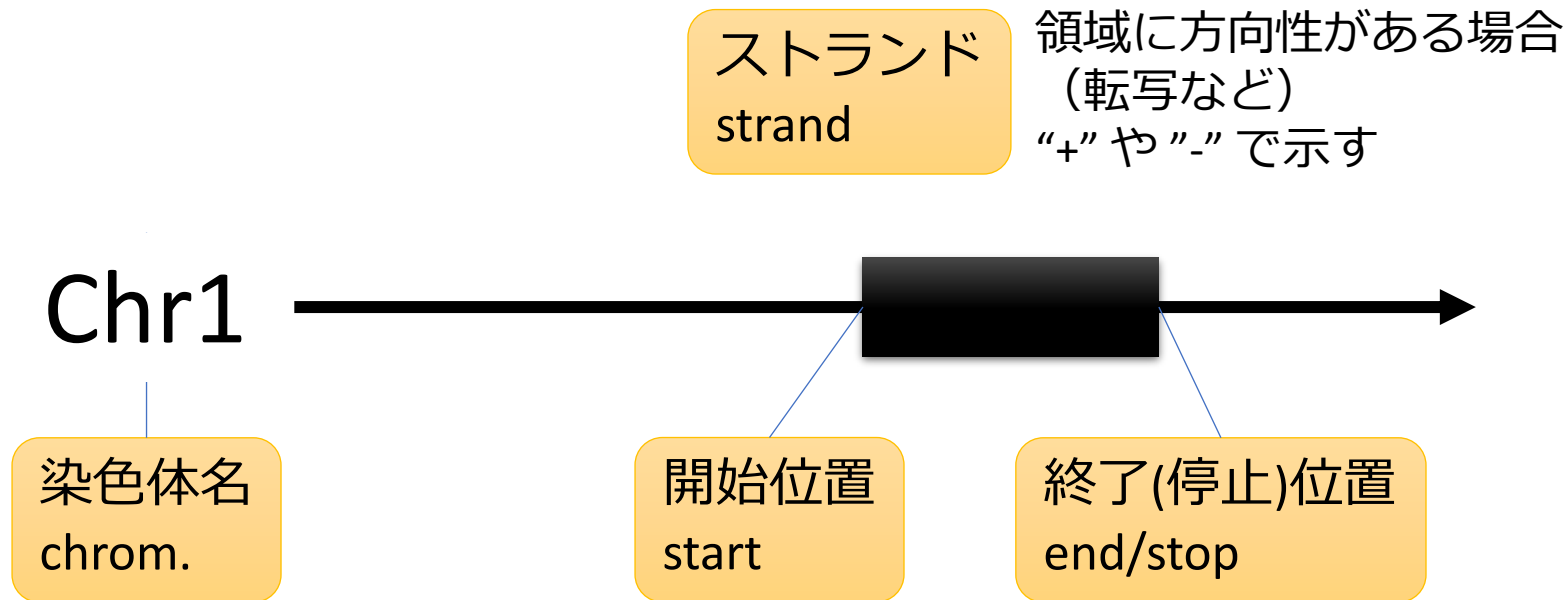
リピート領域



https://en.wikipedia.org/wiki/Structural_variation_in_the_human_genome

<https://web.pdx.edu/~newmanl/typesofdna2.html>

Representation of genomic regions



生物種や参照するゲノムにより
"chr1", "chr2", .., "chrX" と記載したり
"1", "2", ..., "X", "Y" と記載したりする
場合がある

付加情報として

- ・ スコア
- ・ 領域中の代表的な位置
- ・ 転写領域の場合、コード領域
などを示す場合もある

Notations of genomic regions

ゲノム領域の記載方法には2種類ある

0-based

Chr1: 3-6



1-based

Chr1: 4-6

Data format of genomics regions

「ゲノム領域」の情報を記述するためのデータ形式（例）

General formats

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- barChart and bigBarChart format
- bigBed format
- bigGenePred table format
- bigPsl table format
- bigMaf table format
- bigChain table format
- bigNarrowPeak table format
- bigLolly table format
- bigWig format
- Chain format
- CRAM format
- GenePred table format
- GFF format
- GTF format
- HAL format
- Hic format
- Interact and bigInteract format
- Longrange longTabix format
- MAF format
- Microarray format
- Net format
- Personal Genome SNP format
- PSL format
- VCF format
- WIG format

ENCODE-specific formats

- ENCODE broadPeak format
- ENCODE gappedPeak format
- ENCODE narrowPeak format
- ENCODE pairedTagAlign format
- ENCODE peptideMapping format
- ENCODE RNA elements format
- ENCODE tagAlign format

BED format

chr1	10609	10687	FCBS_1	0	.	10619	10687	0,0,255
chr1	29398	29439	FCBS_2	0	.	29398	29432	255,202,74
chr1	588413	588453	FCBS_3	0	.	588423	588453	255,202,74
chr1	605408	605588	FCBS_4	0	.	605433	605562	255,202,74
chr1	629081	629273	FCBS_5	0	.	629081	629229	255,202,74
chr1	629469	630126	FCBS_6	0	.	629469	629922	204,0,0
chr1	629938	630317	FCBS_7	0	.	629955	630144	153,0,0
chr1	630319	630749	FCBS_8	0	.	630330	630749	255,202,74
chr1	631483	631721	FCBS_9	0	.	631483	631535	255,202,74
chr1	631715	631857	FCBS_10	0	.	631715	631824	255,202,74
chr1	632051	632425	FCBS_11	0	.	632051	632269	255,202,74

1. 染色体名
2. 領域の開始位置 (0-based)
3. 領域の終了位置 (0-based)
4. 名前
5. スコア
6. スtrand (+ or - or .)
7. 太く示す領域の開始位置
8. 太く示す領域の終了位置
9. 表示色
10. 領域内のブロック数 (exonなど)
11. 各ブロックの長さ (, でつなぐ)
12. 各ブロックの開始位置 (, で繋ぐ)

元々 Genome Browserで表示させる領域を記述するためのデータ形式

その後、様々な領域データの記述に利用

1～3列めまでは必須

4列め以降は任意 (9列めまで使うBEDファイルは BED9形式とも記載)

その後に独自の列を追加することも可能 (例えば、BED9+ 形式とも記載)

GFF format (gff3)

```
##gff-version 3
chr1    fanta.bio    CRE      10609    10686    0        .        .        ID=FCHS_1;Name=FCHS_1
chr1    fanta.bio    CRE      29398    29438    0        .        .        ID=FCHS_2;Name=FCHS_2
chr1    fanta.bio    CRE      588413   588452    0        .        .        ID=FCHS_3;Name=FCHS_3
chr1    fanta.bio    CRE      605408   605587    0        .        .        ID=FCHS_4;Name=FCHS_4
chr1    fanta.bio    CRE      629081   629272    0        .        .        ID=FCHS_5;Name=FCHS_5
chr1    fanta.bio    CRE      629469   630125    0        .        .        ID=FCHS_6;Name=FCHS_6
chr1    fanta.bio    CRE      629938   630316    0        .        .        ID=FCHS_7;Name=FCHS_7
chr1    fanta.bio    CRE      630319   630748    0        .        .        ID=FCHS_8;Name=FCHS_8
chr1    fanta.bio    CRE      631483   631720    0        .        .        ID=FCHS_9;Name=FCHS_9
chr1    fanta.bio    CRE      631715   631856    0        .        .        ID=FCHS_10;Name=FCHS_10
```

1. 染色体名
2. データのソース
3. 領域の種類
4. 領域の開始位置 (1-based)
5. 領域の終了位置 (1-based)
6. スコア
7. スtrand (+ or - or .)
8. フェーズ (コドン中の位置)
9. 属性値

領域を記述するためのデータ形式（かなり昔からある）

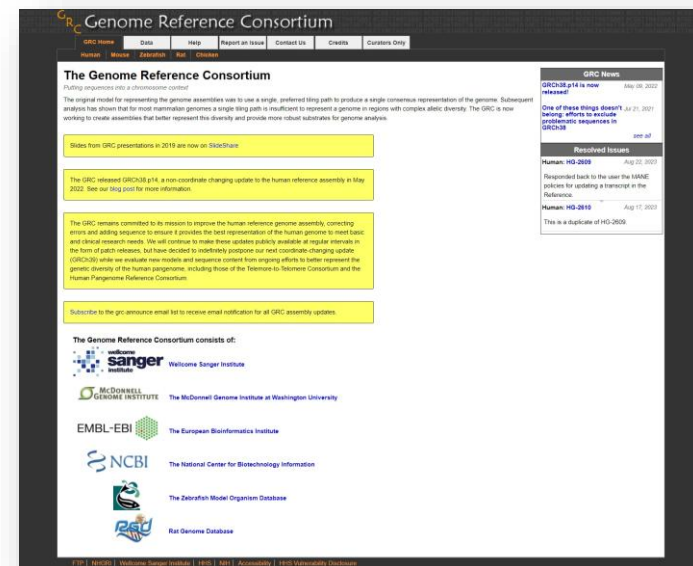
GFF version 1から始まって、改良されながら version 3が最新

GFF2 から派生した遺伝子領域の記述に特化した GTF形式というものもある
(9列めに gene_id と transcript_id の記述が必須となる)

Reference genomes

- 全員が違うゲノム配列を使ってしまうと、領域情報に互換性がなくなる
(実際に当初はデータ提供者ごとに領域の変換が必要だった)
- リファレンスゲノム：単一のゲノム配列を作成し、全員で共通に利用するようにしたもの
- human, mouse, rat, zebrafishi, chicken については Genome Reference Consortium (GRC)から提供
- 定期的な更新があり、それぞれにバージョンがつけられている
- 最新版は
 - ヒト: GRCh38 (GRCh38.p14)
(2022/2/3公開)
 - マウス : GRCm39 (2020/6/24公開)

Genome reference consortium
<https://www.ncbi.nlm.nih.gov/grc>



Data resources of genomic region information

UCSC Genome Browser

<https://genome.ucsc.edu/>

ENSEMBL Genome Browser

<https://www.ensembl.org/>

NCBI Genome Data Viewer

<https://www.ncbi.nlm.nih.gov/genome/gdv/>

これらは汎用的なものの例

他にも特定の生物種に特化したものや、特定の機能・分野に特化したものなど、数多く存在する

遺伝子情報
Gene information

これまでの生命科学では、「遺伝子」を中心とした研究が多く行われ、多くの知見が得られている。

「遺伝子」を単位にまとめられた情報・
データリソースが数多く存在する

Description of genes

- 遺伝子「自体」の記載方法
 - 体内時計遺伝子 BMAL1 の場合 (HGNCより)

HGNC data for BMAL1

Approved symbol ?	BMAL1
Approved name ?	basic helix-loop-helix ARNT like 1
Locus type ?	gene with protein product
HGNC ID ?	HGNC:701
Symbol status ?	Approved
Previous symbols ?	ARNTL
Previous names ?	" aryl hydrocarbon receptor nuclear translocator like "
Alias symbols ?	MOP3; JAP3; PASD3; bHLHe5; ARNTL1
Alias names ?	" brain and muscle ARNT-like 1 " " basic helix-loop-helix family member e5 " " PAS domain containing 3 " " member of PAS superfamily 3 "

BMAL1: 遺伝子シンボル (gene symbol)

アルファベット + 数字の短い名前。遺伝子名 (gene name) と呼ぶこともある
 basic helix-loop-helix ARNT like 1: 遺伝子名 (gene name)

短い文で記述した名前。遺伝子記述 (gene description) と呼ぶこともある
 ARNTL, MOP3, JAP3, PASD3, bHLHe5, ARNTL1: 別名 (synonym)

遺伝子につけられた別のシンボル

HGNC:701 : 遺伝子 ID (gene identifier)

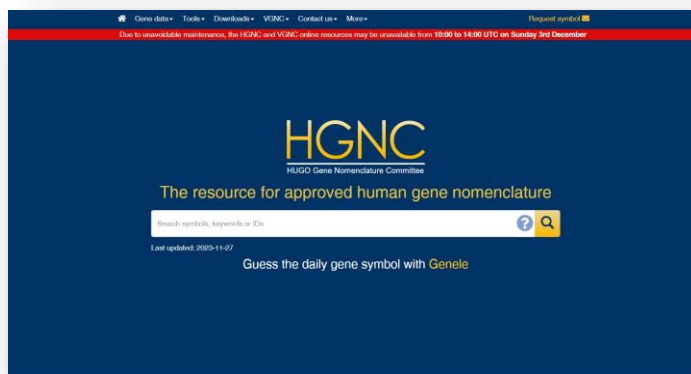
遺伝子に付与した識別子。各データリソースで独自に付与している

Official gene symbol

- 同じ遺伝子なのに遺伝子シンボルが複数付与されていることがある
 - 歴史的に遺伝子シンボルは遺伝子の発見者が（勝手に）つけていたため
 - 元々別の遺伝子だと考えられていたものが実は一緒だったケース
 - 機能未知な遺伝子につけられていたシンボルと機能判明後につけられたシンボルがあるケース
 - 研究コミュニティにより別々の名前が使われているケース
- 公式な遺伝子シンボルを決める活動

Human

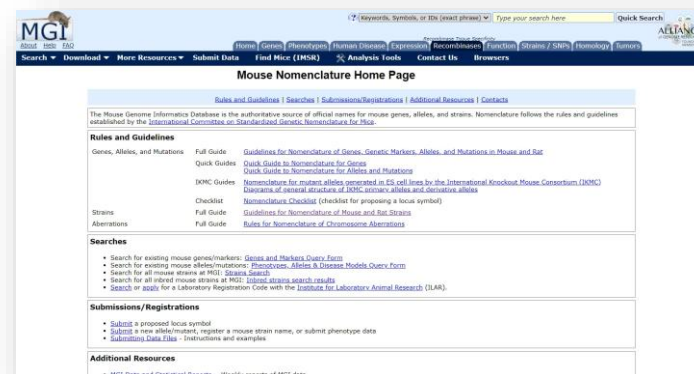
HUGO Gene Nomenclature Committee (HGNC)



<https://www.genenames.org/>

Mouse

Nomenclature Committee in IMGS/MGI



<https://www.informatics.jax.org/mgihome/nomen/>

Functions of genes

• 遺伝子「機能」の記載方法

• テキストによる記述

Summary The protein encoded by this gene is a basic helix-loop-helix protein that forms a heterodimer with CLOCK. This heterodimer binds E-box enhancer elements upstream of Period (PER1, PER2, PER3) and Cryptochrome (CRY1, CRY2) genes and activates transcription of these genes. PER and CRY proteins heterodimerize and repress their own transcription by interacting in a feedback loop with CLOCK/ARNTL complexes. Defects in this gene have been linked to infertility, problems with gluconeogenesis and lipogenesis, and altered sleep patterns. The protein regulates interferon-stimulated gene expression and is an important factor in viral infection, including COVID-19. [provided by RefSeq, Oct 2021]

• 論文へのリンク

1. [Artificial induction of circadian rhythm by combining exogenous BMAL1 expression and polycomb repressive complex 2 inhibition in human induced pluripotent stem cells.](#)
Kaneko H, et al. Cell Mol Life Sci, 2023 Jul 8. PMID 37421441
2. [Protective effects of brain and muscle ARNT-like gene 1 on oxidized low-density lipoprotein-induced human brain microvascular endothelial cell injury by alleviating ferroptosis.](#)
Yu S, et al. Hum Exp Toxicol, 2023 Jan-Dec. PMID 3743012
3. [Bmal1 and Gut-lung axis in SARS-CoV-2 infection: New insight into the effects of melatonin on COVID-19 patients?](#)
Khezri MR, et al. Biomed Pharmacother, 2023 Aug. PMID 37257225, [Free PMC Article](#)

• 遺伝子オントロジー (Gene Ontology)

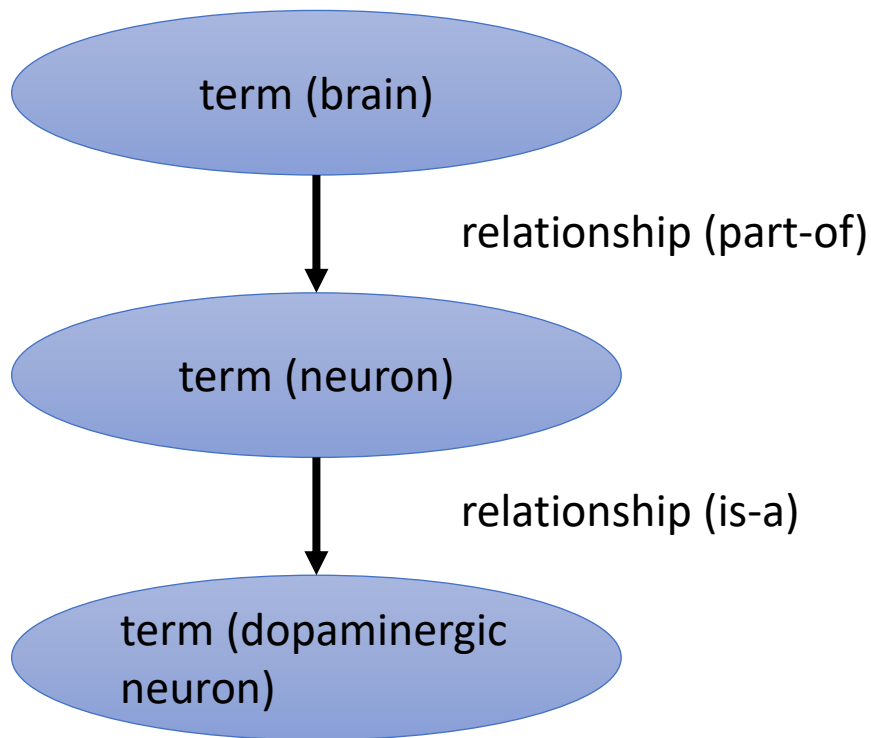
GO ID	Qualified GO term
GO:0000976 ^{3 5}	enables transcription cis-regulatory region binding
GO:0000978 ^{3 5}	enables RNA polymerase II cis-regulatory region sequence-specific DNA binding
GO:0000981 ^{3 5}	enables DNA-binding transcription factor activity, RNA polymerase II-specific
GO:0003677 ^{3 5}	enables DNA binding
GO:0003700 ⁵	DNA-binding transcription factor activity

Ontology

• オントロジー (Ontology)

- “An ontology is a formal representation of a body of knowledge within a given domain. Ontologies usually consist of a set of classes (or terms or concepts) with relations that operate between them.”
 - <http://geneontology.org/docs/ontology-documentation/>

特定の分野の知識体系を表現したもの
通常、クラス（ターム・概念）とクラス間の
関係性で構成する



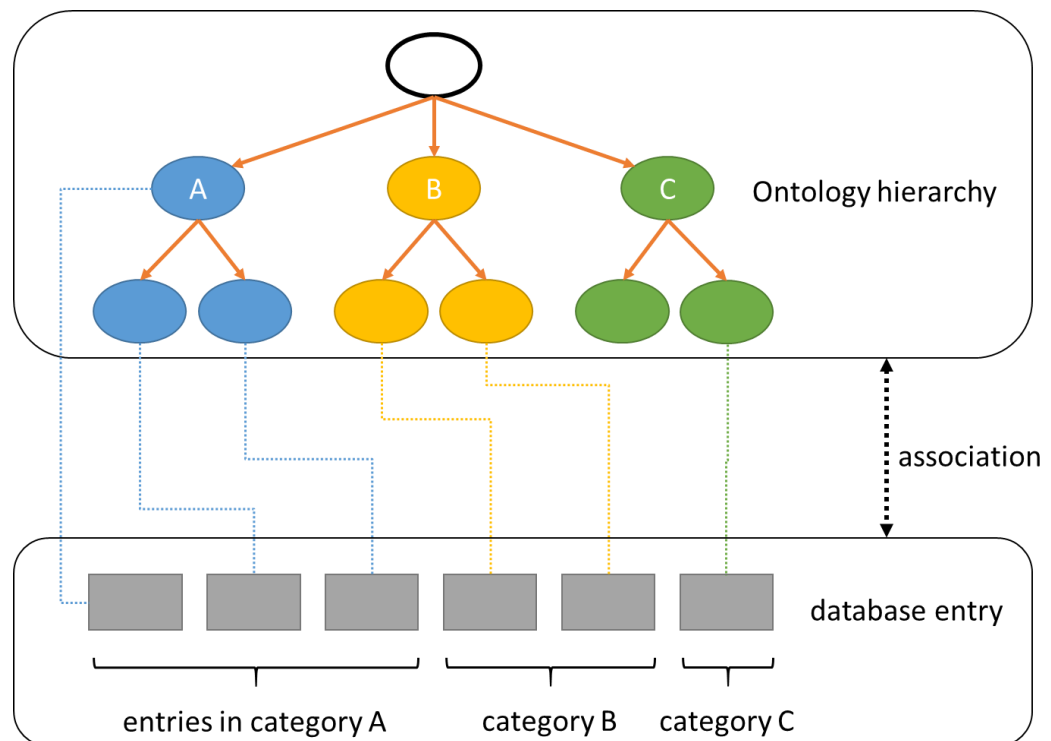
Application of ontology

- オントロジーを使うことで次のようなことが可能となる
 - オントロジーのタームを使うことで、特定の用語の統一化ができる
 - ターム間の関係性の情報を使うことで、分類することが可能となる

ID	sample
A	neural cell
B	neuron
C	neuronal cells
D	nerve cell sample



ID	sample by cell ontology
A	CL:0000540 (neuron)
B	CL:0000540 (neuron)
C	CL:0000540 (neuron)
D	CL:0000540 (neuron)

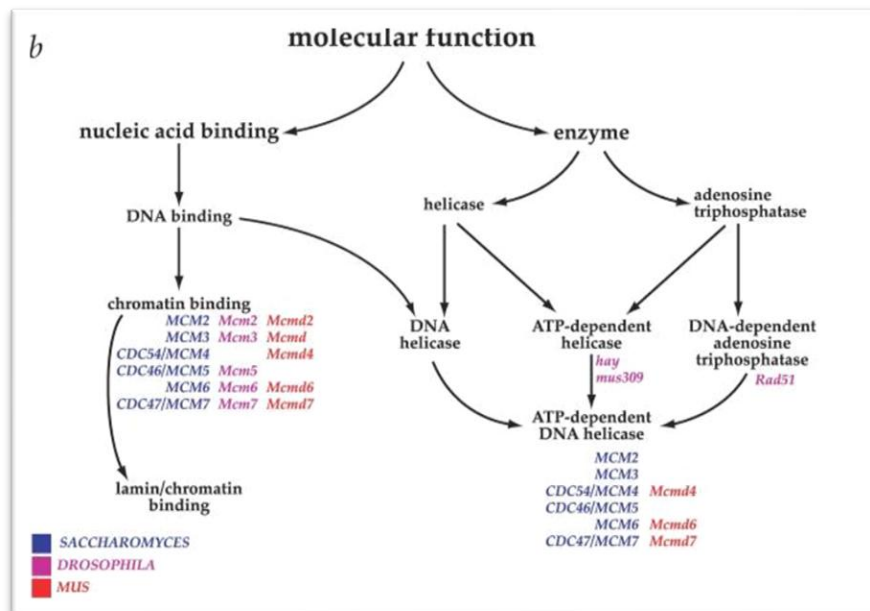


Gene ontology

• 遺伝子オントロジー (Gene Ontology) (<http://geneontology.org/>)

- The Gene Ontology knowledgebase provides a computational representation of our current scientific knowledge about the functions of genes (or, more properly, the protein and non-coding RNA molecules produced by genes) from many different organisms, from humans to bacteria.
- <https://geneontology.org/docs/introduction-to-go/>

Gene Ontologyによる知識ベースによって、遺伝子（正確には遺伝子から産出されるタンパク質やノンコーディングRNA）の機能に関する現在の科学的知見を計算機上で表現することができる



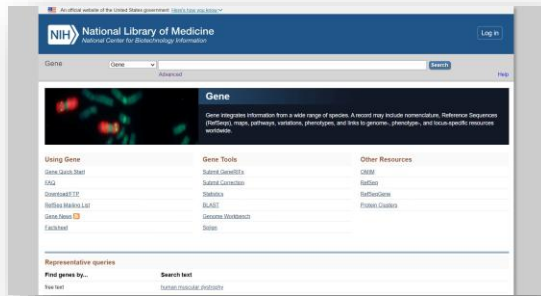
Example of Gene Ontology
 Ashburner M et al., Nat Genet, 2000,
 doi:10.1038/75556

Gene ontology

- 3種類のオントロジー

- Molecular function : 分子的機能
 - その遺伝子からの産物(タンパク質やノンコーディングRNA)が持つ分子的な機能
 - 例 : DNA結合、酵素、リン酸化など
- Biological process : 生物学的プロセス
 - その遺伝子が持つ生物学的な機能や関わる生体内プロセス
 - 例 : 転写制御、代謝、シグナル伝達など
- Cellular component
 - その遺伝子からの産物が機能を果たす細胞内の場所、器官
 - 例 : 核、細胞質、細胞膜など

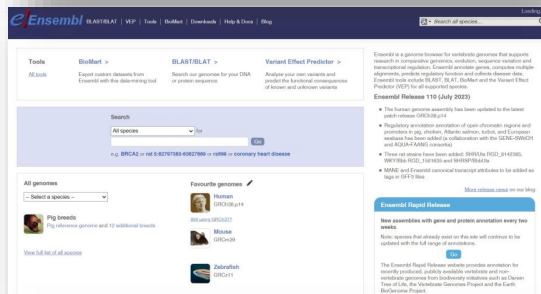
Gene information resources (common & human)



NCBI Genes (Entrez Genes)

様々な生物種をカバーしている

<https://www.ncbi.nih.gov/gene/>



Ensembl Genes

ゲノムの決定した様々な生物種のものがある

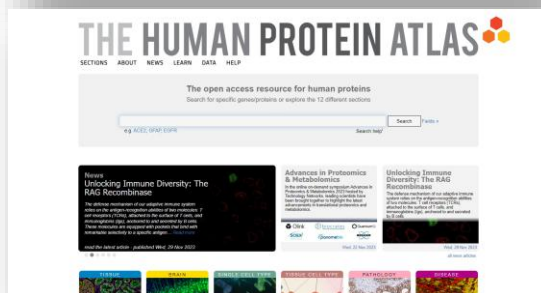
<https://www.ensembl.org/>



GeneCards

ヒトのみであるが、様々な情報をカバーしている

<https://encodegenes.org/>



Human Protein Atlas

ヒトのタンパク質のデータベースであったが、現在は遺伝子や転写情報など広くカバーされている

<https://www.proteinatlas.org/>

[illegible]

<https://www.informatics.jax.org/>

The screenshot displays the RGD (Rat Genome Database) homepage. At the top, a navigation bar includes links for Home, Tools, Analysis & Visualization, Diseases, Phenotypes & Models, Pathways, Community, and RSS. A search bar is located below the navigation bar. The main content area is divided into several sections: 'Search' with links to Genes, Strains, Ontology, etc.; 'Analysis and Visualization' with links to Browse, Variant Visualizer, VCMAP, OLGA, Disease Portals, and Phenotypes and Models; and 'Rat Genome Data' with links to Genome Browser, Map, etc. A row of animal icons (human, rat, mouse, etc.) is shown below the search bar. Social media links for Facebook, Twitter, and LinkedIn are visible in the top right corner.

<https://rgd.mcw.edu/>

[illegible]

<https://wormbase.org/>


Tools • Downloads • Links • Community • Species • About • Help • Archives •

FLYBASE, Inc. (renewed September 26, 2009)
A Database of *Drosophila* Genes & Genomes

[JGI](#)
[Help](#)
[Contact Us](#)


Information


Genetics


ReBase


JBrowse


New to FlyBase


How To FlyBase


FlyBase 4.0


Validation


ID Validation


Batch Download


FlyTrack


FlyBase 4.0


Latest Technical Abstracts

Coming FlyBase

External Resources

- News and Outreach

- Drosophila Journals and Publications

- Drosophila Resources [Wiki]

- Meetings/Conferences

- Drosophila in Clinical Research

- Inter-species


Quick Search

[Human Diseases](#)
[Protein Domains](#)
[Gene Groups](#)
[Pathways](#)
[GO](#)
[Data Class](#)

[Search Flies](#)
[Interlog](#)
[GAT 4.0](#)
[Expression](#)
[Phenotype](#)
[References](#)

[Click here to submit multiple Drosophila IDs.](#)

Commentary
[view all Commentaries](#)

[New FlyBase user guide online access kit!](#)
 Please visit our FlyBase Resource Center... at \$10.00 PER person / PER year.
 (includes only the FlyBase 4.0 database access guide)

Note: 100 users or less for additional your search kit.

<https://flybase.org/>

開発中のデータベース・データリソース
Databases/data resources developing in our
team

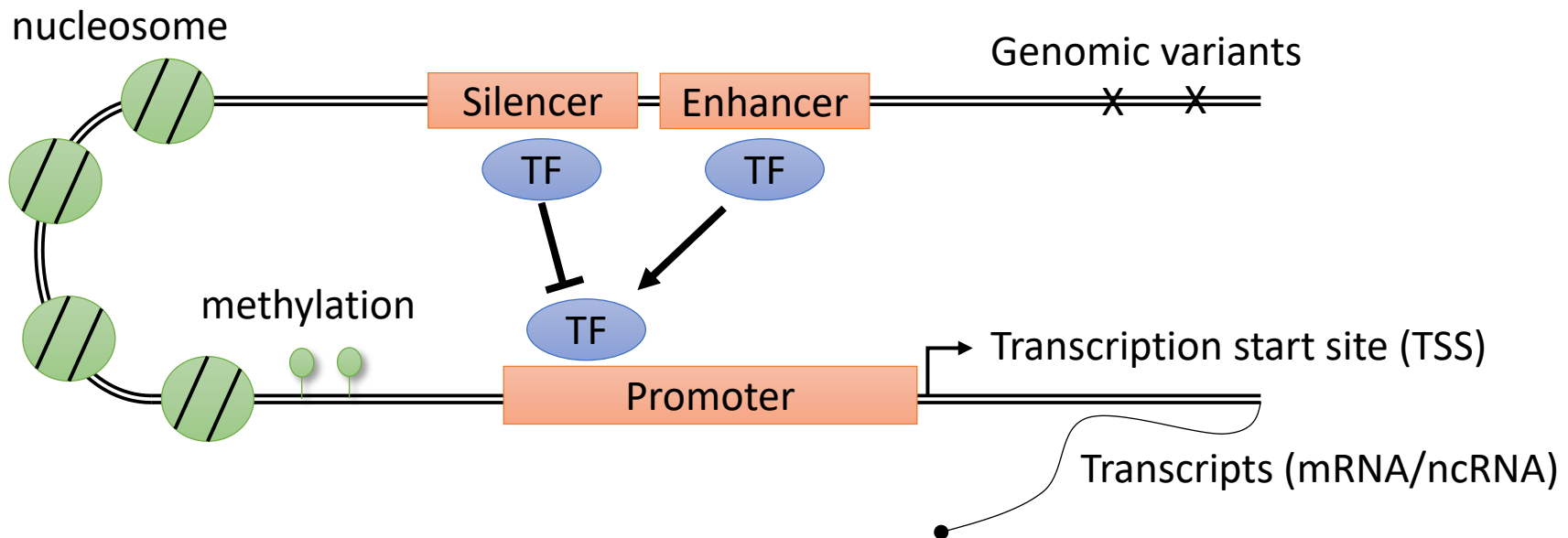
Various sequence data related to transcriptional regulations

Epigenome

ChIP-seq (18,214)	ATAC-seq (7,765)
Hi-C (3,516)	Mnase-seq (688)
Bisulfite-seq (3,816)	DNase-seq (864)
MeDIP-seq (639)	FAIRE-seq (298)

Genomic Variation

Whole genome (44,015)
Whole exome (3,094)



転写制御の理解

= 多様で大規模なデータの
効率的な統合利用が必須

Transcriptome

RNA-seq (90,935)	miRNA-seq (2,605)
CAGE (1,234)	Long-read RNA (905)

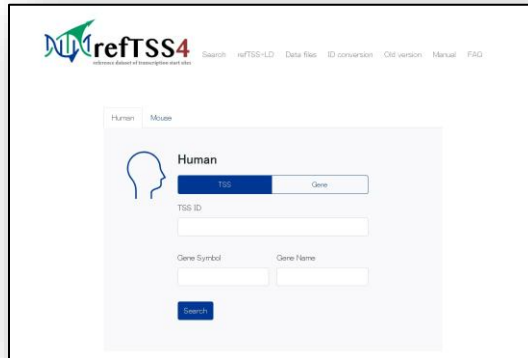
(xxx) : the number of records in BioProject (as of 2025-10-20 in DDBJ Search)

Database/resource activities in my team

大規模データ生産プロジェクトに参加し、
データのコーディネーションを行う
(primary databases/resources)

公共データを活用して、転写や転写制御に関
するデータリソースを構築し、公開する
(secondary databases/resources)

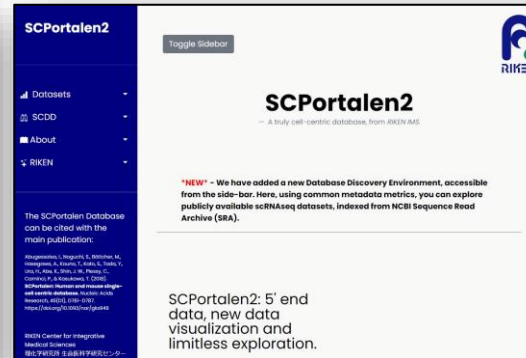
Our data resources for transcription and their regulations



refTSS 4

<https://refTSS.riken.jp/>

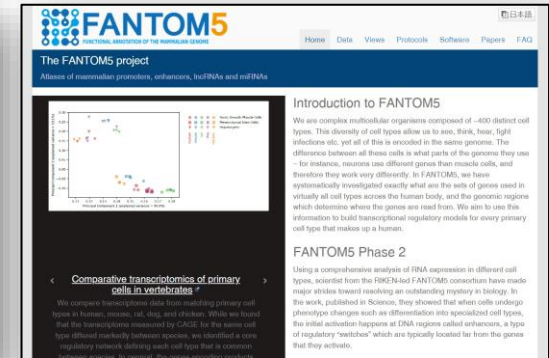
転写開始点のリファレンスセット



SCPortalen 2

<https://single-cell.riken.jp/>

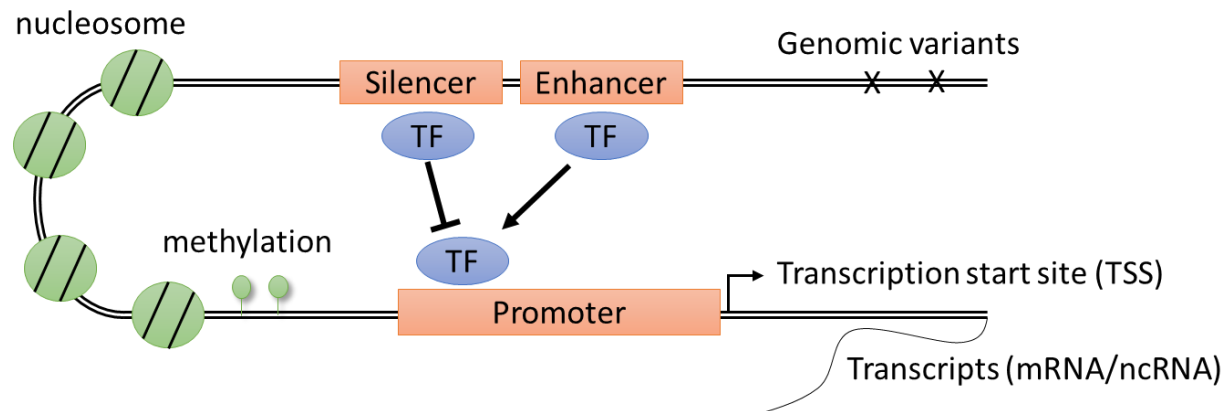
公共の一細胞RNA-seq
データ再利用のための
データベース



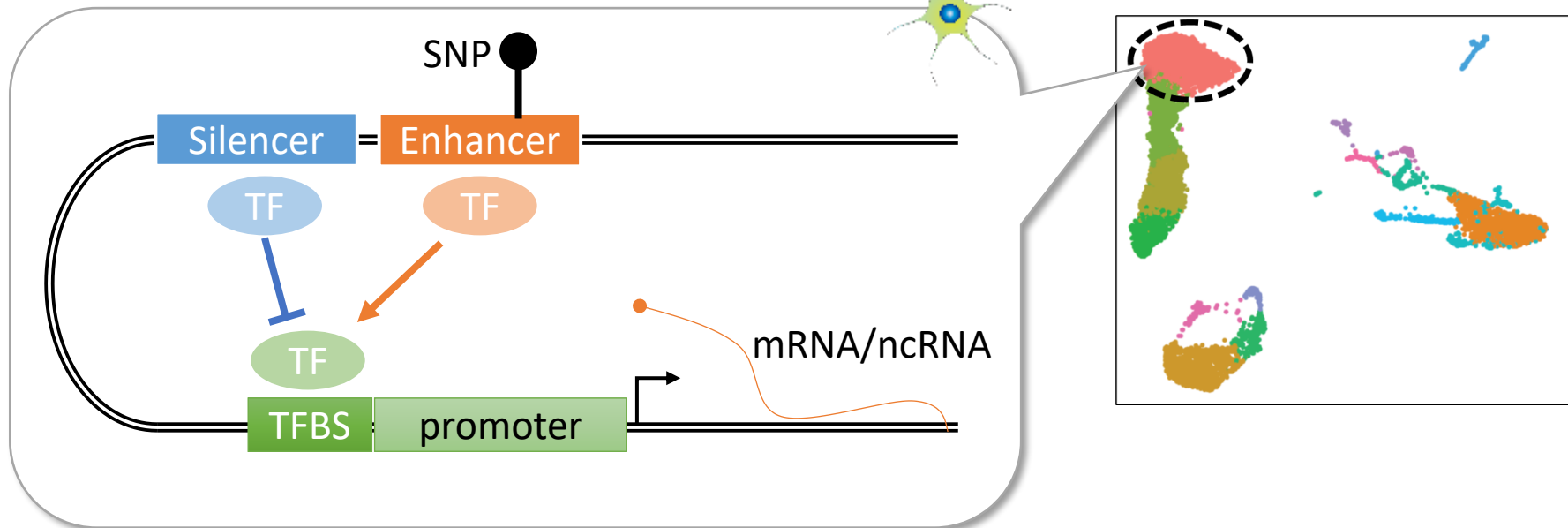
FANTOM5 web resource

<https://fantom.gsc.riken.jp/5/>

ヒト・マウスを含む哺乳
類3,000サンプルを対象と
したプロモーター・エン
ハンサーアトラス

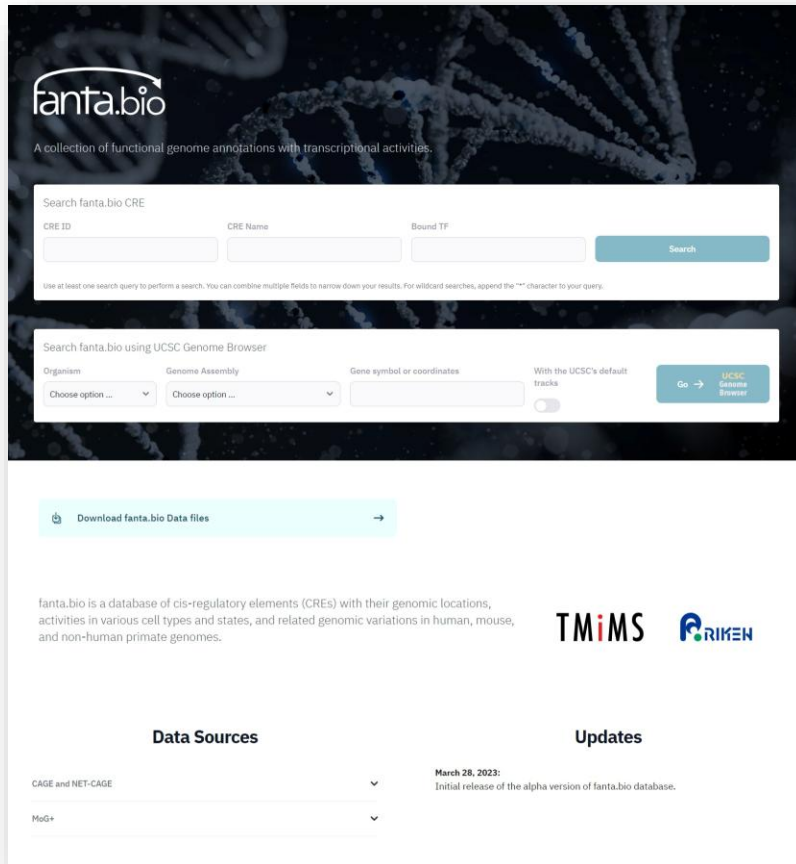


cis-regulatory elements (CREs)



- 遺伝子発現の調節に関わるゲノム領域
- 「プロモーター」・「エンハンサー」など非コード領域にあるものが一般的
- 細胞種ごとの特異的な発現を調節
- 表現型に係する多型・変異が頻出する
- 全貌解明はまだ途上である

fanta.bio database for CRE information



The screenshot shows the fanta.bio website interface. At the top, the logo and tagline "A collection of functional genome annotations with transcriptional activities." are visible. Below this is a search section titled "Search fanta.bio CRE" with input fields for "CRE ID", "CRE Name", and "Bound TF", and a "Search" button. A note below the search fields states: "Use at least one search query to perform a search. You can combine multiple fields to narrow down your results. For wildcard searches, append the '*' character to your query." Below the search section is another section titled "Search fanta.bio using UCSC Genome Browser" with dropdown menus for "Organism" and "Genome Assembly", a text input for "Gene symbol or coordinates", a "With the UCSC's default tracks" toggle, and a "Go" button with a "UCSC Genome Browser" link. At the bottom left, there is a "Download fanta.bio Data files" button. Below this, a paragraph describes the database: "fanta.bio is a database of cis-regulatory elements (CREs) with their genomic locations, activities in various cell types and states, and related genomic variations in human, mouse, and non-human primate genomes." To the right of this paragraph are logos for "TMiMS" and "RIKEN". At the bottom, there are two sections: "Data Sources" with a dropdown menu showing "CAGE and NET-CAGE" and "Motif+", and "Updates" with a date "March 28, 2023:" and the text "Initial release of the alpha version of fanta.bio database."

公共データを再処理し、ゲノム中の転写制御に関わる領域（シスエレメント/CRE）について

- ゲノム中のCREの位置
 - 細胞種や細胞状態ごとのCREの活性
 - CREと関係するゲノム変異
- の情報をまとめたデータベース

対象とする生物種

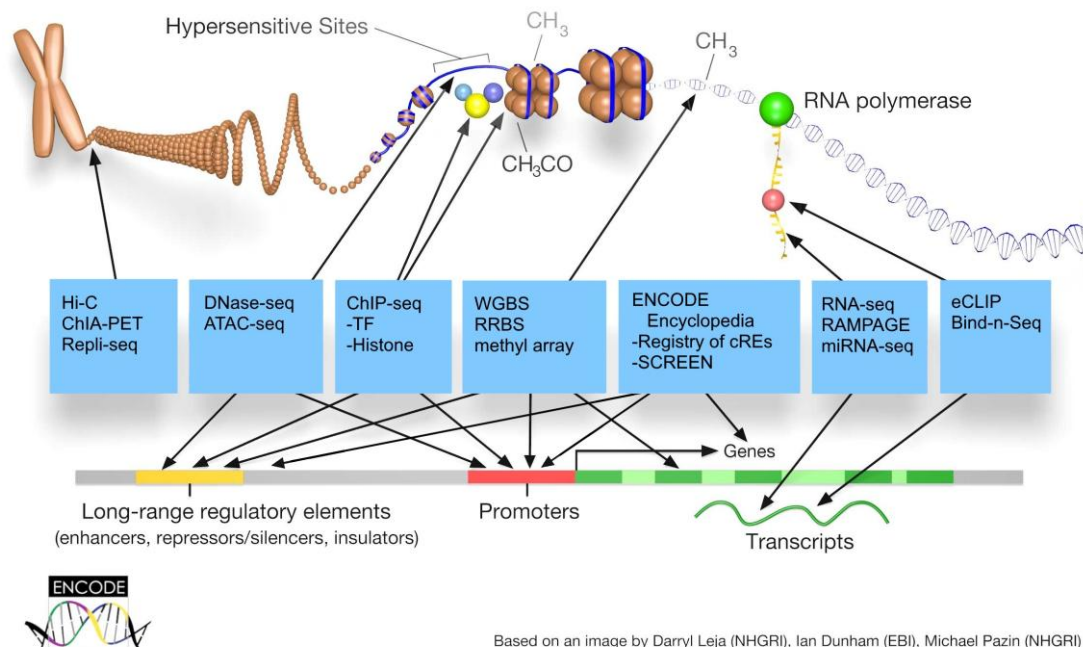
- ヒト、マウス、非ヒト霊長類

特徴：シスエレメント領域を転写情報から同定する

<https://fanta.bio/>

Chromatin-based method to identify CREs

- シスエレメント領域ではヒストンのメチル化パターンやオープンクロマチン状態などが特徴的な変化を示すことを利用

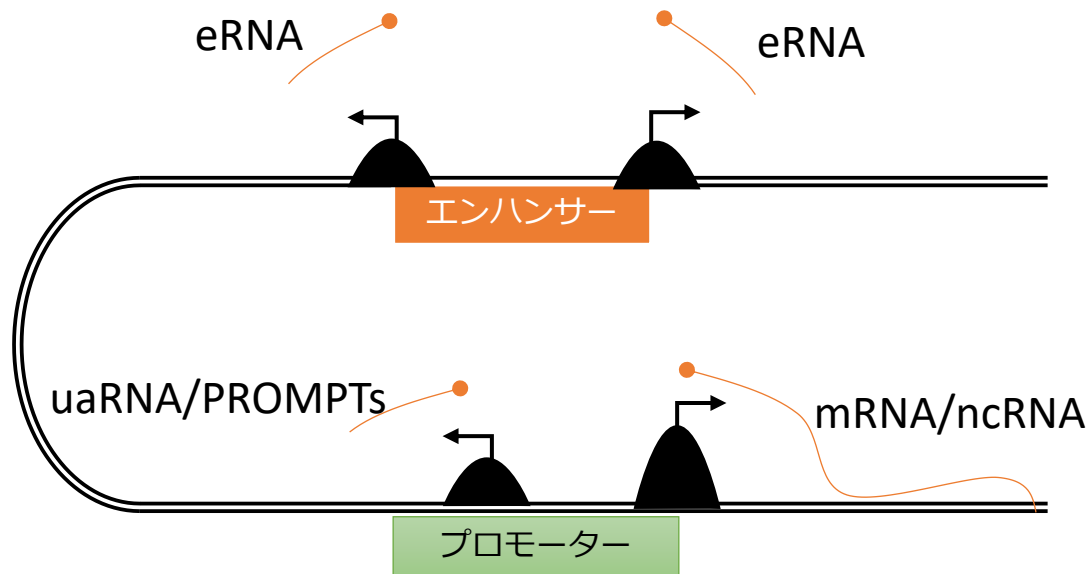


<https://screen.encodeproject.org/>

異なるメチル化ヒストンを対象とするChIP-seq結果の組み合わせで同定

Transcriptome-based method to identify CREs

- 「転写情報」(RNAの5'端による転写開始点)を用いたシスエレメントの同定



転写開始点の場所・方向や
活性量を元に同定

CRE identification comparison (Pros & Cons)

クロマチン情報ベースの方法

- 同定の感度が高い（発見できる個数が多い）
- × 複数のシーケンシング実験が必要（Histone ChIP-seq, ATAC-seq, FAIRE-seq）

転写情報ベース

- 1種類のシーケンシング実験(5'端 RNAのシーケンシング) のみで同定できる
- 転写も見られる、一細胞レベルでも同定可能
- × 感度はクロマチンベースのものほどではない

感度の問題が解決できれば、転写情報の方が情報が収集しやすい

「転写情報」による方法を主として用い、
クロマチン情報はサポート的に用いる（エビデンスとして等）

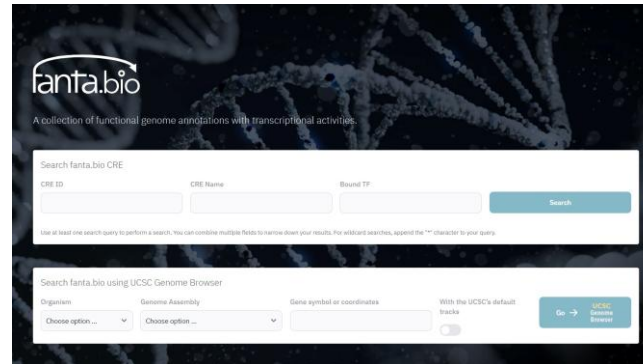
CRE annotations in fanta.bio (tentative)

Items
CRE names
Overlapped FANTOM5 CAGE Peaks (promoters)
Overlapped FANTOM5 Enhancers
Overlapped ENCODE SCREEN cCRE
Overlapped refTSS TSSs
Overlapped ChIP-Atlas TF and Others binding sites

<https://fanta.bio/>

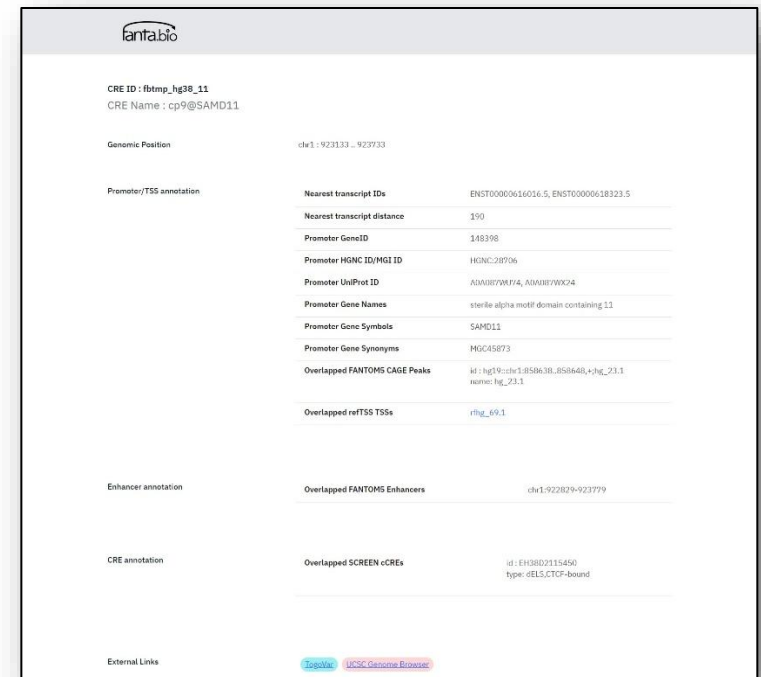
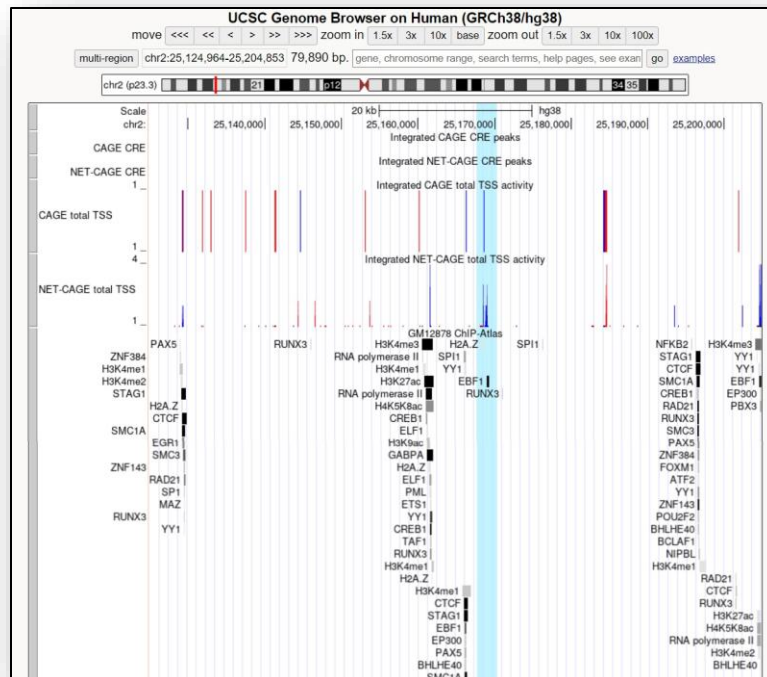
UCSC Genome Browser

CREのゲノム上でのビュー



CRE annotation view

CREについての各種
アノテーション
情報のビュー



データベース研究の現在と今後の展開

Current and future of the database research

Current of the biomedical database research

ここ最近の生命科学データベース研究は激動の時代

様々なAI-like algorithmを
用いたシステムの実用化

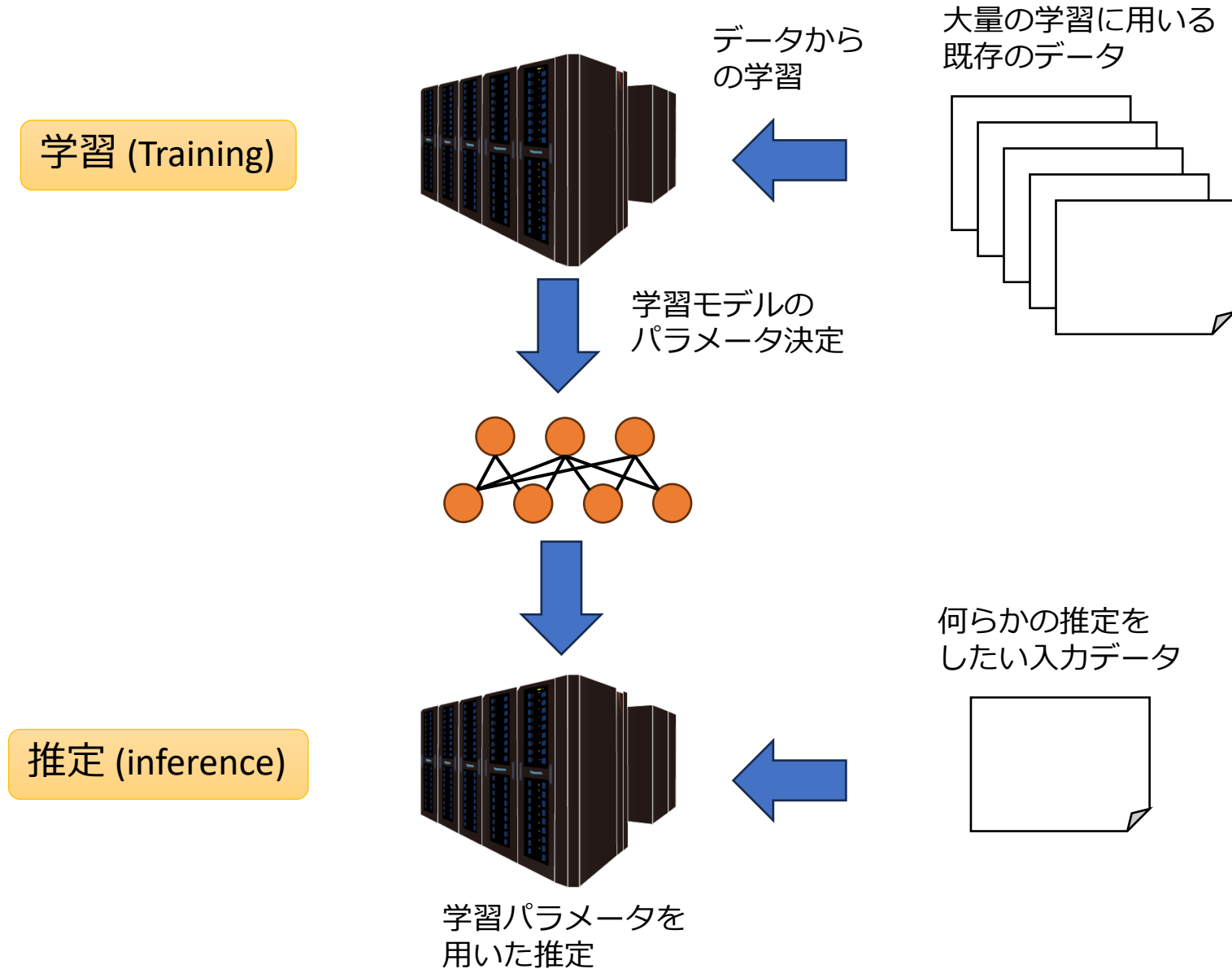
AI and database

- AIとデータリソースの2種類の関わり方
 - AI-based システムを開発するためのデータリソースの活用
 - データリソースを開発するための AI-based システムの活用

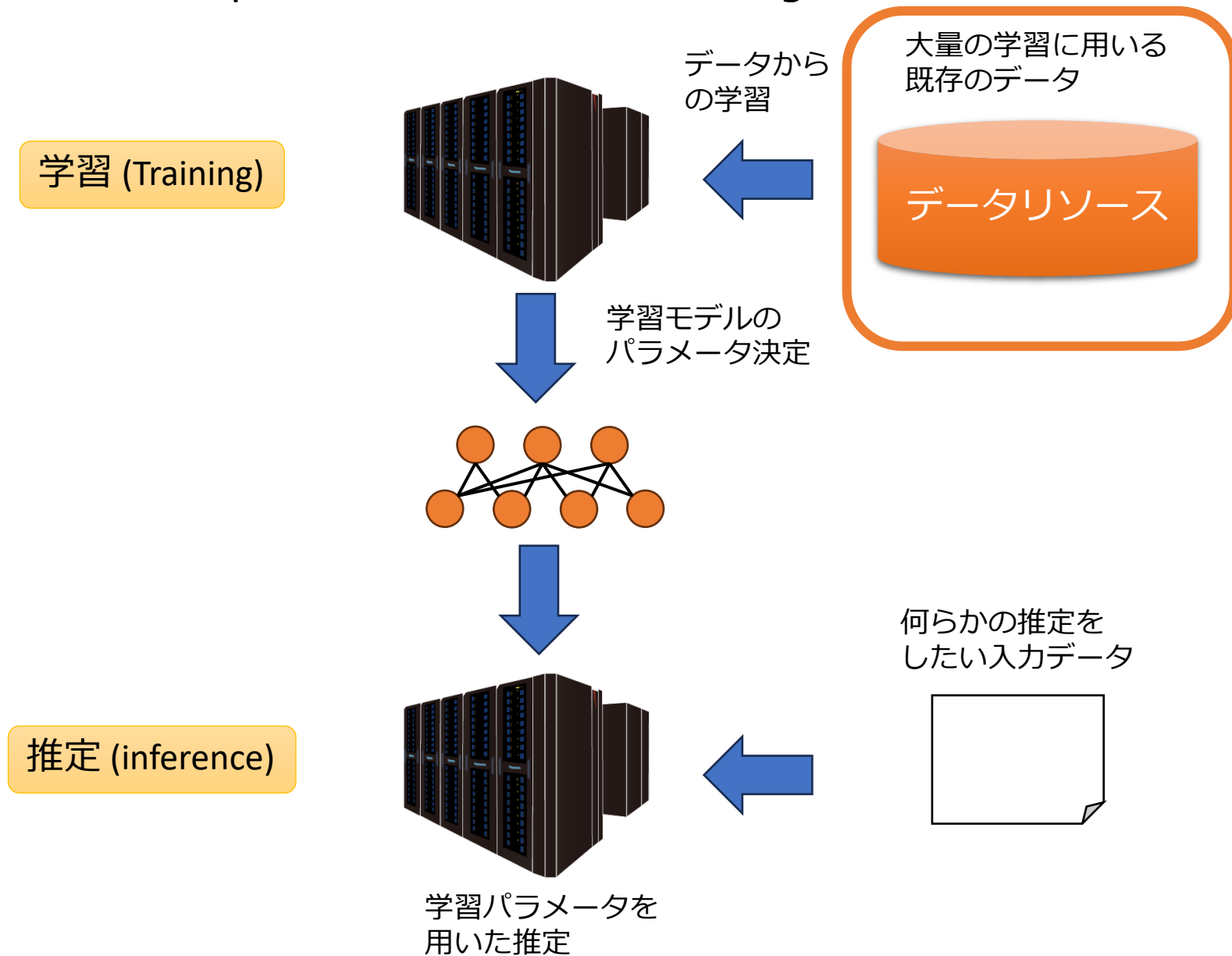
AI and database

- AIとデータリソースの2種類の関わり方
 - AI-based システムを開発するためのデータリソースの活用
 - データリソースを開発するための AI-based システムの活用

Basic concept of AI / machine learning



Basic concept of AI / machine learning



AI and database

- AIとデータベースの2種類の関わり方
 - AI-based システムを開発するためのデータベースの活用
 - データベースを開発するための AI-based システムの活用

Flow of life-science database development

- 入れるデータを作る

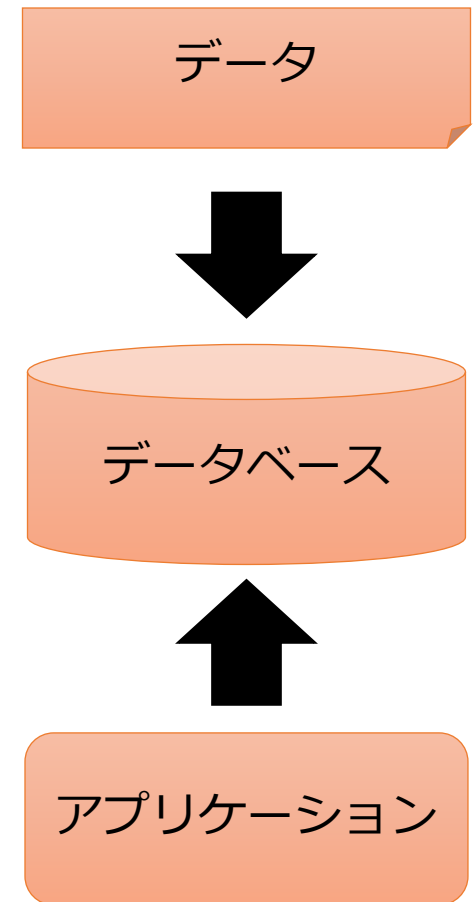
- 入手可能なデータを加工・処理して、データベースに格納できる形に変換する
- データの curation を行う

- データを格納するデータベースを作る

- データモデルを設計
- データベース管理システムの選択

- インタフェース（アプリケーション）を作る

- データベースにアクセスするためのインタフェースを開発する
- データベースの説明書を用意する



AI-based systems applicable to database developments

- 大規模言語モデル（LLM）を用いた論文のサマリー作成・抽出
 - chatGPT, SciSummary, Scholarcy, PaperBrain AI, ...
- コンピューターコードの自動生成・支援
 - GitHub Copilot, Google Cloud, IBM watsonx Code Assistant, TabNine
- ユーザーインターフェースの自動生成・支援
 - Khroma, designs.ai, uizard, GeniusUI, Midjourney

Bioinformatics workflow generation with ChatGPT



You

FASTQファイルから発現量のexpression tableを生成するnextflow workflowを書いてください



ChatGPT

Nextflowを使用したFASTQファイルから発現量のexpression tableを生成するワークフローを以下に示します。この例では、Trinityを用いてトランスクリプトームアセンブリを行い、その後、Salmonを使用して発現量の推定を行います。このワークフローは簡略化されており、実際のプロジェクトに合わせて変更する必要があります。

nextflow

Copy code

```
// main.nf

// インプットディレクトリ
params.reads = "./data/*_R{1,2}.fastq.gz"

// アウトプットディレクトリ
outdir "./results"

// トランスクリプトームアセンブリ
process trinity {
    input:
        file reads

    output:
        file "trinity_out/Trinity.fasta" into trinity_assembly

    script:
```



AI-supports for life-science database development

• 入れるデータを作る

- 入手可能なデータを加工・処理して、データベースに格納できる形に変換する
- データの curation を行う

LLM

Code generation

• データを格納するデータベースを作る

- データモデルを設計
- データベース管理システムの選択

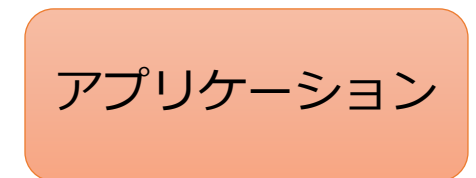
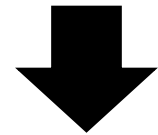
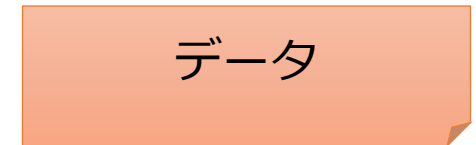
Code generation

• インタフェース（アプリケーション）を作る

- データベースにアクセスするためのインタフェースを開発する
- データベースの説明書を用意する

LLM

UI generation



Discussion

AI-like algorithmを用いた様々なシステムが実用化されてくるとデータリソース、データベースはどうなっていくのか？

Negative comments

- 公共データリソースはAI学習に使われるだけになってしまう？
- AIがすべて探索してくれるから、特に二次データベースはもう必要ない？
- データベースだけ作っても評価されなくなる？

Positive comments

- 今までの公共データベースにAI chatbot 検索がついて情報探索機能が向上
- AIシステムを使えば、二次データベースの構築・維持管理が効率化できる
- データベース+AIの統合が今後ますます重要となる

Summary (Take-home messages)

- 現在の生命科学研究において、公共リソースの活用は必要不可欠
- 様々な情報・データ・知識がインターネット上に蓄積されており、利用可能になっている
- 公共リソースの活用のためには、今までの知識がどのように蓄積されてきているのかを知るのが重要
- 公共リソースの提供も重要な研究・開発項目の1つ（シスエレメントデータベースの例）
- 今後はAIアルゴリズムをどのように用いるかを考えることが重要（データベース分野に限らず）
 - AIアルゴリズム・システムを敵視せず、味方にする方法を考えるべき