ゲノム、遺伝子の比較

清水謙多郎

shimizuk@fc.jwu.ac.jp

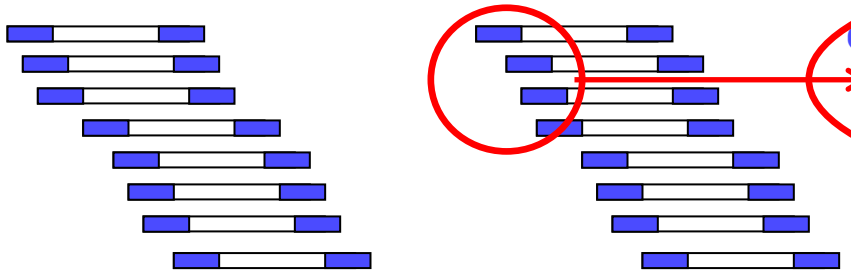
ゲノム配列をどう決めるか？

DNAシーケンサ

配列断片が生成される

現在のシーケンサでは、1回の実験で10^{12}塩基を越える解析が可能なものが存在

数百bp（ショートリード）
～数万bp（ロングリード）



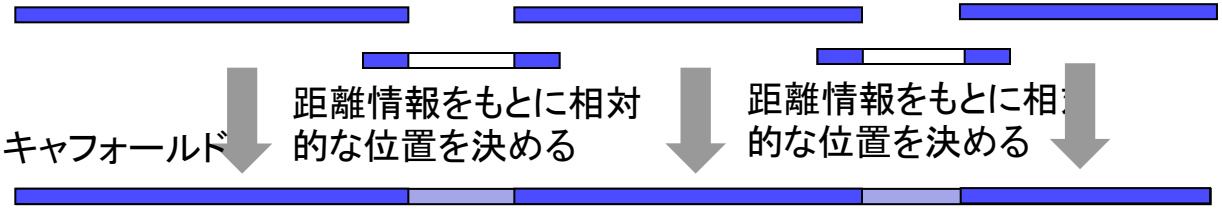
```
CGGAGTCAACTTACCTATA---------
        TTACCTATATTCTAATCG---
            CTATATTCTAATCGTAG--
                TATTCTAATCGTAGTA
```

アセンブリ

配列断片の重なりをもとにつなぐ

読み取りエラーも考慮

コンティグ配列

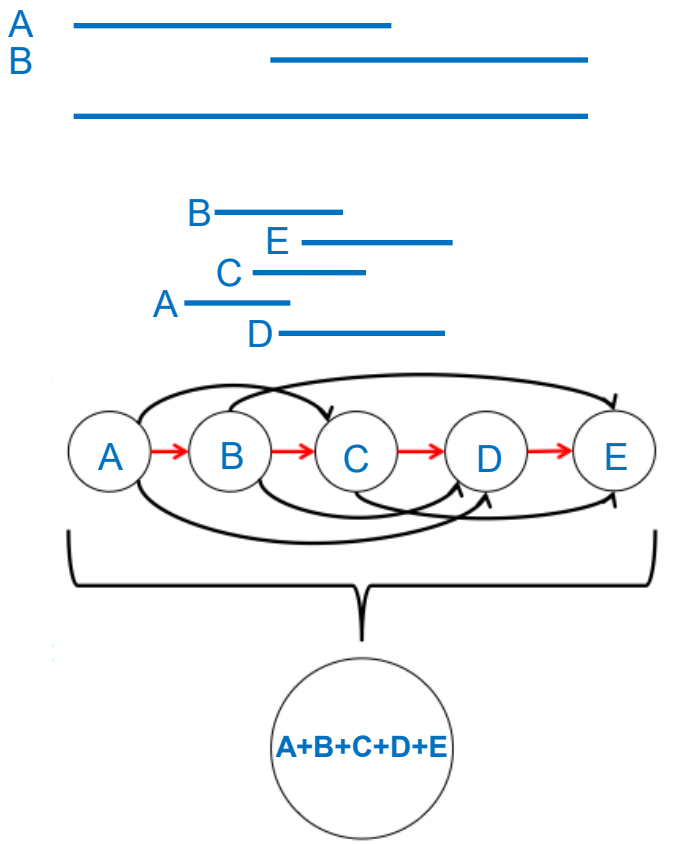


距離情報をもとに相対的な位置を決める

スキャフォールド

ゲノム配列の決定手法

- **Overlap-Layout-Consensus（OLC）**
  - ロングリードに対して用いられる
- Overlap
  - リードの重なりを検出し、リード間のつながりの関係をグラフの有向辺で表現
- Layout
  - リードを矛盾なく順序づけ、コンティグを形成する
  - リード間の包含関係、つながり関係を整理
- Consensus
  - リードのつながり関係をもとに配列を決定する



Consensusの例

ベースコールエラーに対する対処

```
          -ACCGT--
          ---CGTGC
          TACGGT--      塩基置換エラー
コンセンサス TACCGTGC

          -ACC-GT--
          ---CAGTGC
          TACC-GT--     挿入エラー
コンセンサス  TACCGTGC

          -ACCGT--
          ---C-TGC
          TACCGT--      欠失エラー
コンセンサス TACCGTGC
```

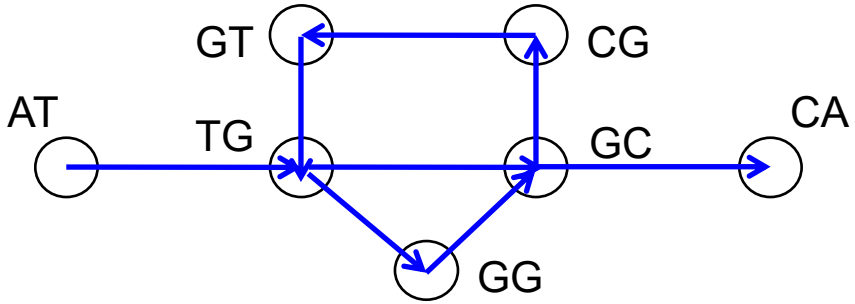
- 多数決
- 重み付き投票方式
  ベースコール（各塩基（A, T, G, C）の読み取り精度）のクオリティ情報（QV値）の利用が考えられる

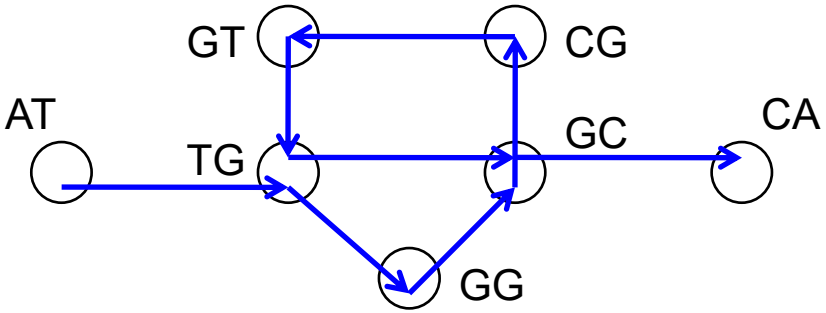
de Bruijn graph

- de Bruijn graph（デュブラングラフ）
  - ショートリードに対して用いられる
  - 配列の断片を k-mer（長さkの部分配列）に分割
  - $(k-1)$文字を頂点とする
  - 頂点をつないでk-merとなるよう有向辺を引く
- エッジをオイラー経路（一筆書き） のように全部たどると、元の配列が得られる
  - オイラー経路: すべての辺を1回ずつ訪れるパス
  - オイラー閉路: ある頂点から出発し、同じ頂点に戻るオイラー経路
- ゲノムシーケンシングでの利用
  - 短いリードを膨大な数（10^7～10^9）扱うとき、リードどうしを全部比較（OLC法）するのは重すぎる
  - k-merの共有で高速につなげることができ、現実に（基本的な手法として）よく用いられている

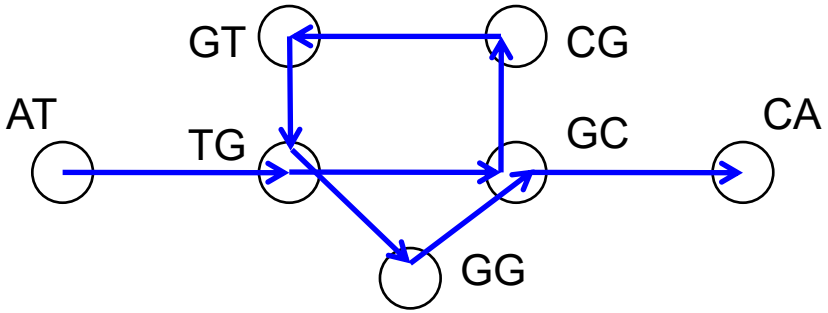
オイラー経路の探索例

S = { ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT }





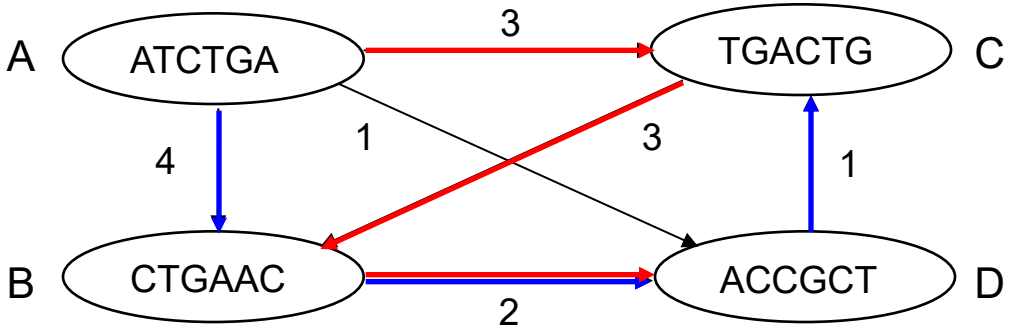
ATGGCGTGCA



ATGCGTGGCA

ハミルトン経路

- 最大重みハミルトン経路
  - 配列断片（リード）を頂点とし、頂点pの末尾k文字と頂点qの先頭k文字が一致しているとき、有向辺$p \rightarrow q$をもたせ、重みをkとする
  - ハミルトン経路: グラフ上の全ての頂点を1回ずつ通る経路
  - 重み付きハミルトン経路: 経路上の辺の重みを合計したもの
  - 最大重みのハミルトン経路を求め、その頂点を重ね合わせたものがShortest Common Superstring



貪欲法によるハミルトン経路: A → B → D → C　　重み: 7
　コンティグ: ATCTGAACCGCTGACTG　　長さ: 17

重みを最大にするハミルトン経路: A → C → B → D　　重み: 8
　コンティグ: ATCTGACTGAACCGCT　　長さ: 16

ハミルトン経路とオイラー経路

- ハミルトン閉路では、頂点の数をnとすると、$O(n!)$の時間を要する
  - 最大重みハミルトン経路は、ゲノムアセンブリの初期に用いられたモデル
  - 数百万～数千万リードを扱うゲノム解析には適用できない
- オイラー閉路では、辺の数をnとすると、$O(n)$の時間
  - 現在の主要なアセンブリー手法は、オイラー閉路を適用

Genome Data Viewer

- NCBI Genome Data Viewerのページを開いてください
  - Web検索で「Genome Data Viewer」、「NCBI GDV」などと入力すれば、恐らくトップでヒットします
  - この講義では、資料にリンクを埋め込んでいます

Web検索

ALDH2遺伝子の検索

An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

shimizu.jwu@gmail...

Genome Data Viewer

生物種（「Home Sapiens」（ヒト）がデフォルト）

Switch view

Search organisms

Homo sapiens (human)

GDV supports the exploration and analysis of NCBI-annotated and selected non-NCBI annotated eukaryotic genome assemblies. Currently, assemblies from over 3320 organisms are available.

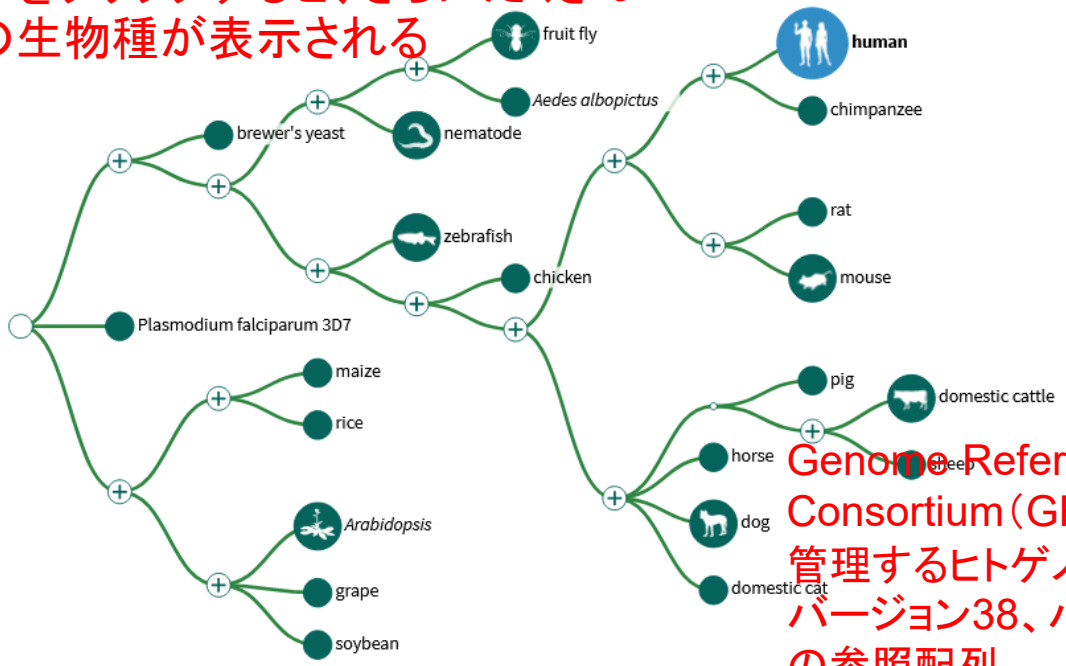
3320以上の生物を登録

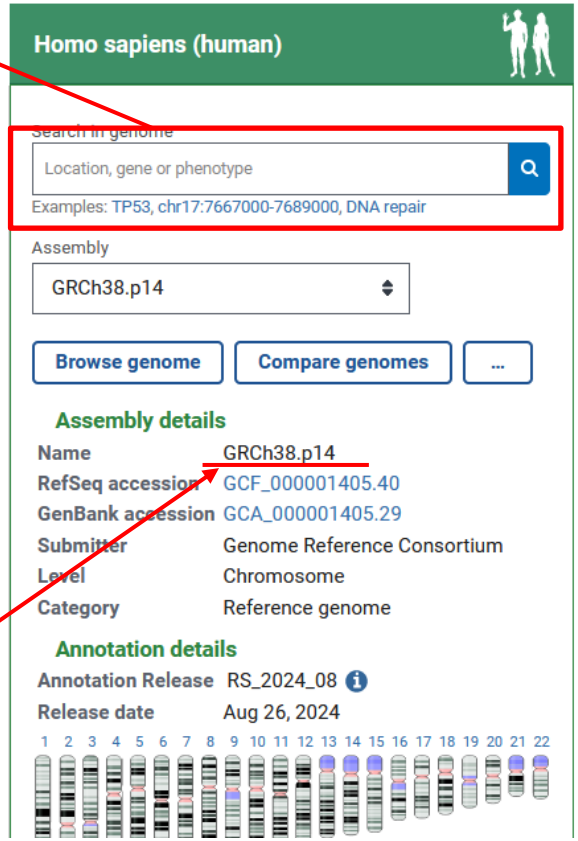
To view more organisms in the tree, click on nodes that have '+' signs. Press and hold the '+' to expand and reveal all the subgroups. Or, search for an organism using the search box above.

New! Click on Switch view at the top to see another way of navigating genomes.

いろいろな情報を使って検索できる

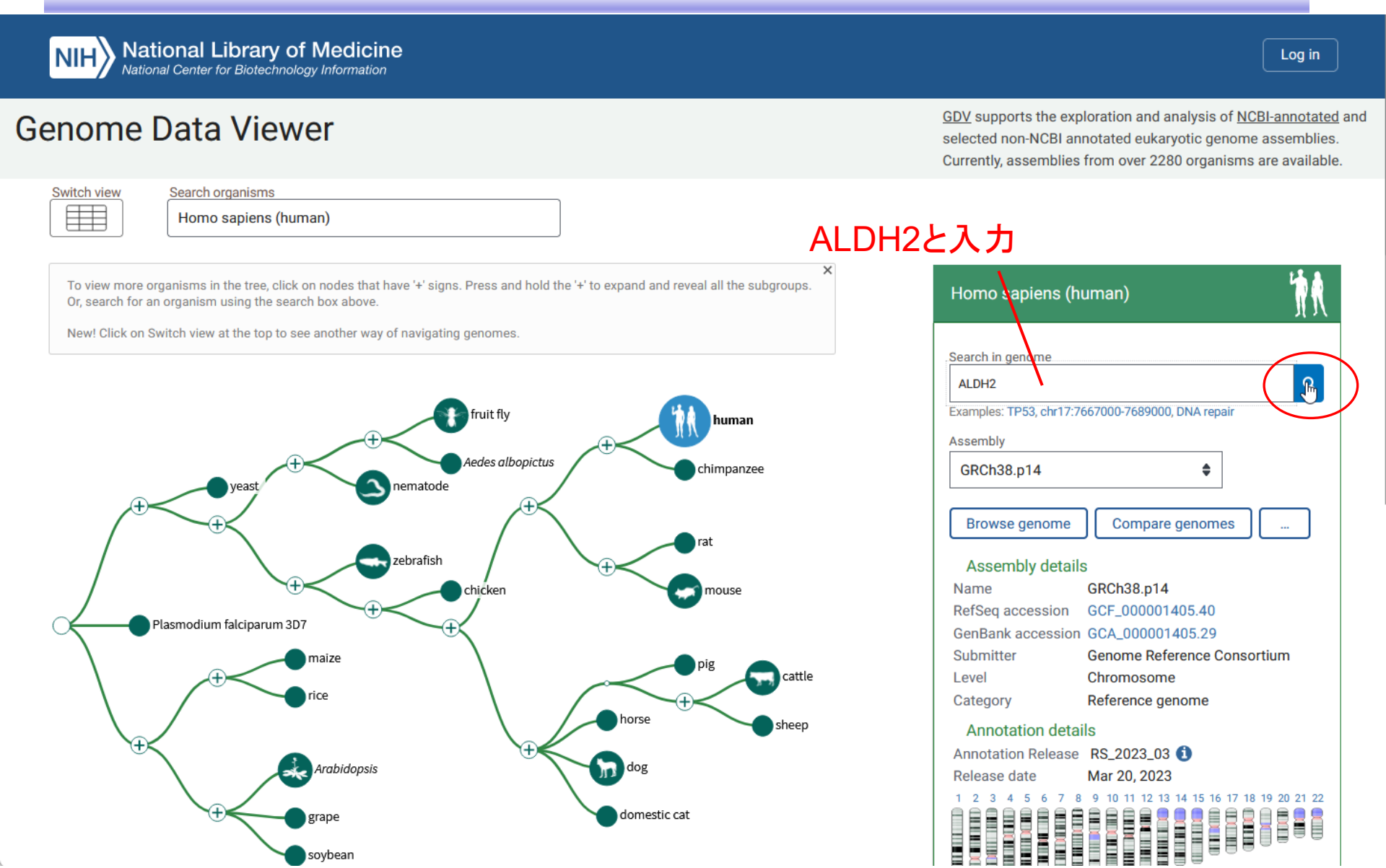
＋をクリックすると、さらにたくさんの生物種が表示される



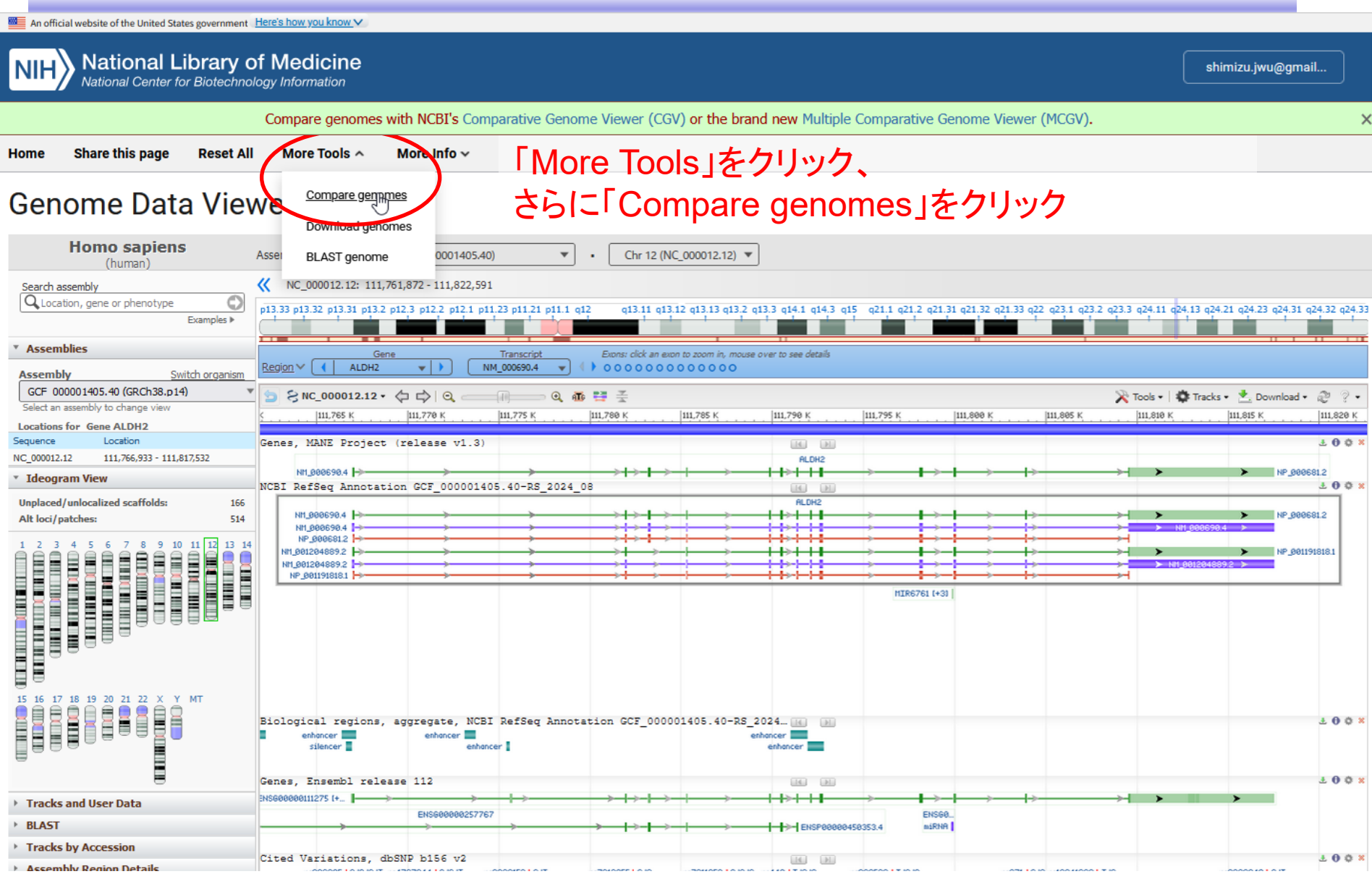


Genome Reference Consortium（GRC）が管理するヒトゲノムのバージョン38、パッチ14の参照配列

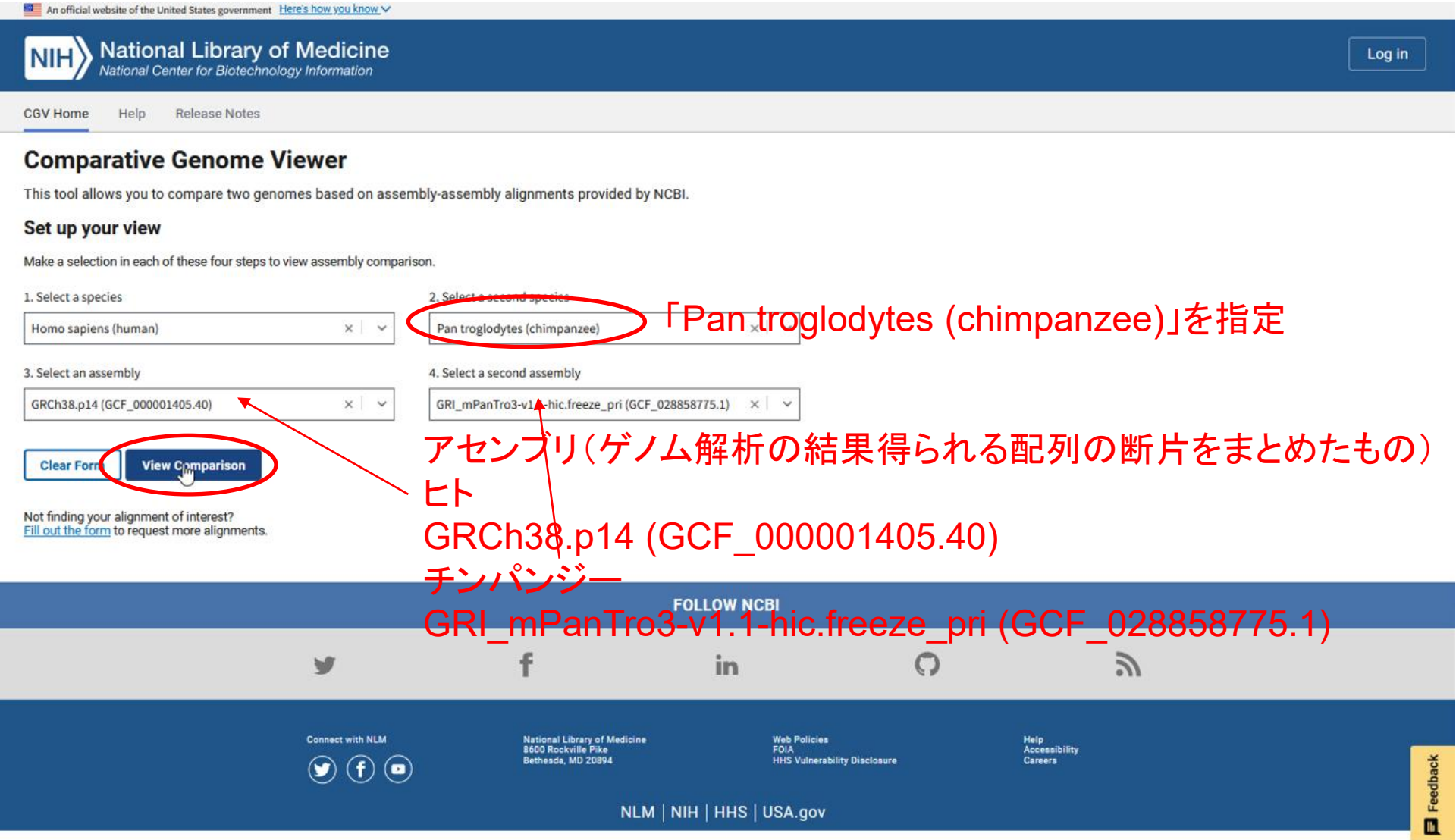
ゲノムブラウザで調べてみよう



他の生物種とゲノムを比較する



他の生物種とゲノムを比較する



他の生物種とゲノムを比較する

National Library of Medicine
National Center for Biotechnology Information

Log in

CGV Home Help Release Notes

Comparative Genome Viewer

You are ready to explore whole genome alignment between ***Pan troglodytes*** **NHGRI_mPanTro3-v1.1-hic.freeze_pri** (GCF_028858775.1) and ***Homo sapiens*** **GRCh38.p14** (GCF_000001405.40).

Find a gene in this alignment

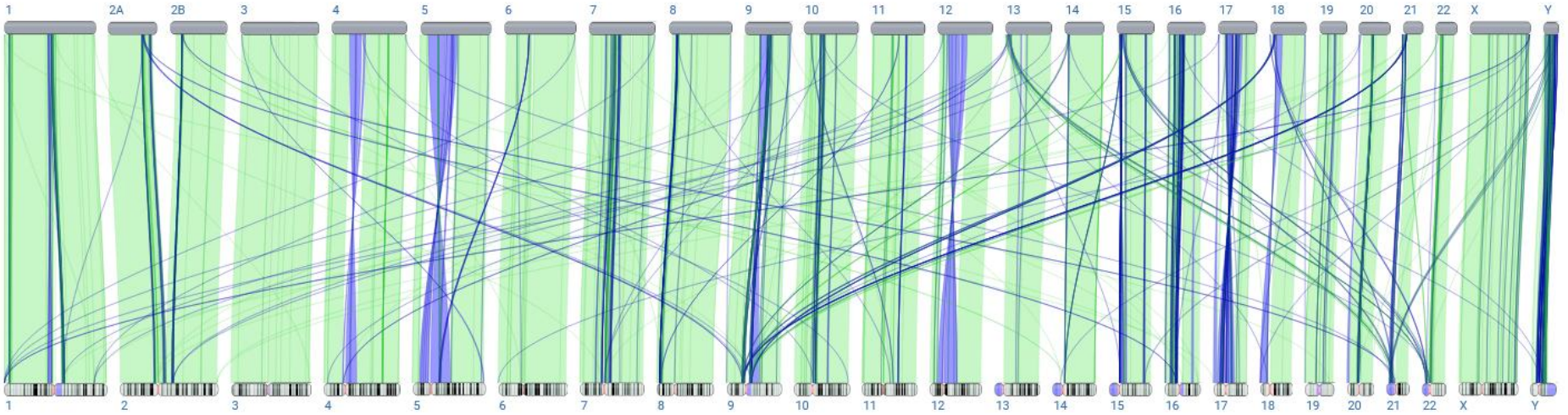
Type gene symbol or name, for example Ace2, or ... Search

緑は、順方向に対応している
青は、逆方向に対応している
領域としてねじれているような部分は逆位

Go to dotplot view

Reset to genome view

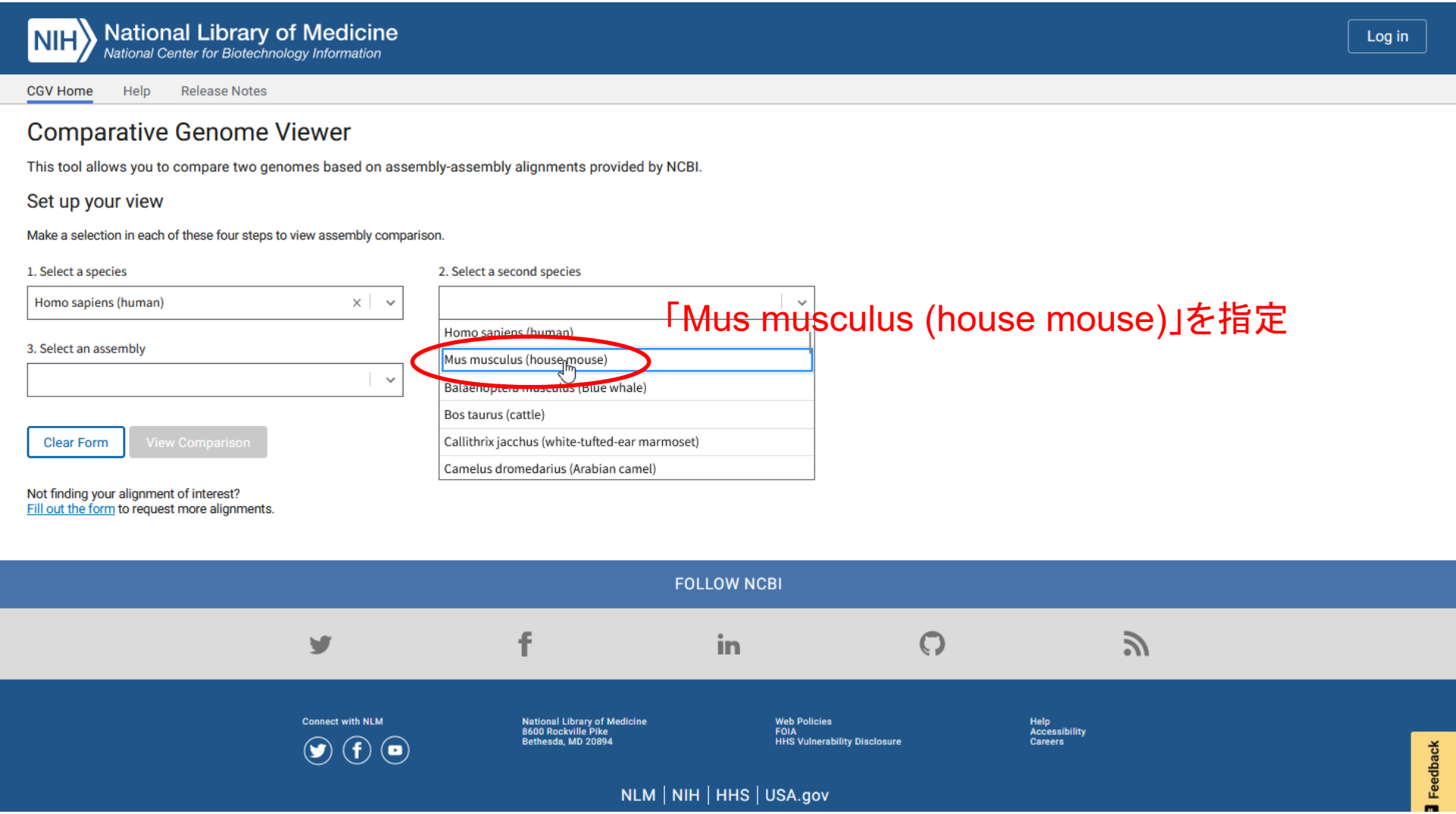
Pan troglodytes **NHGRI_mPanTro3-v1.1-hic.freeze_pri** (GCF_028858775.1)



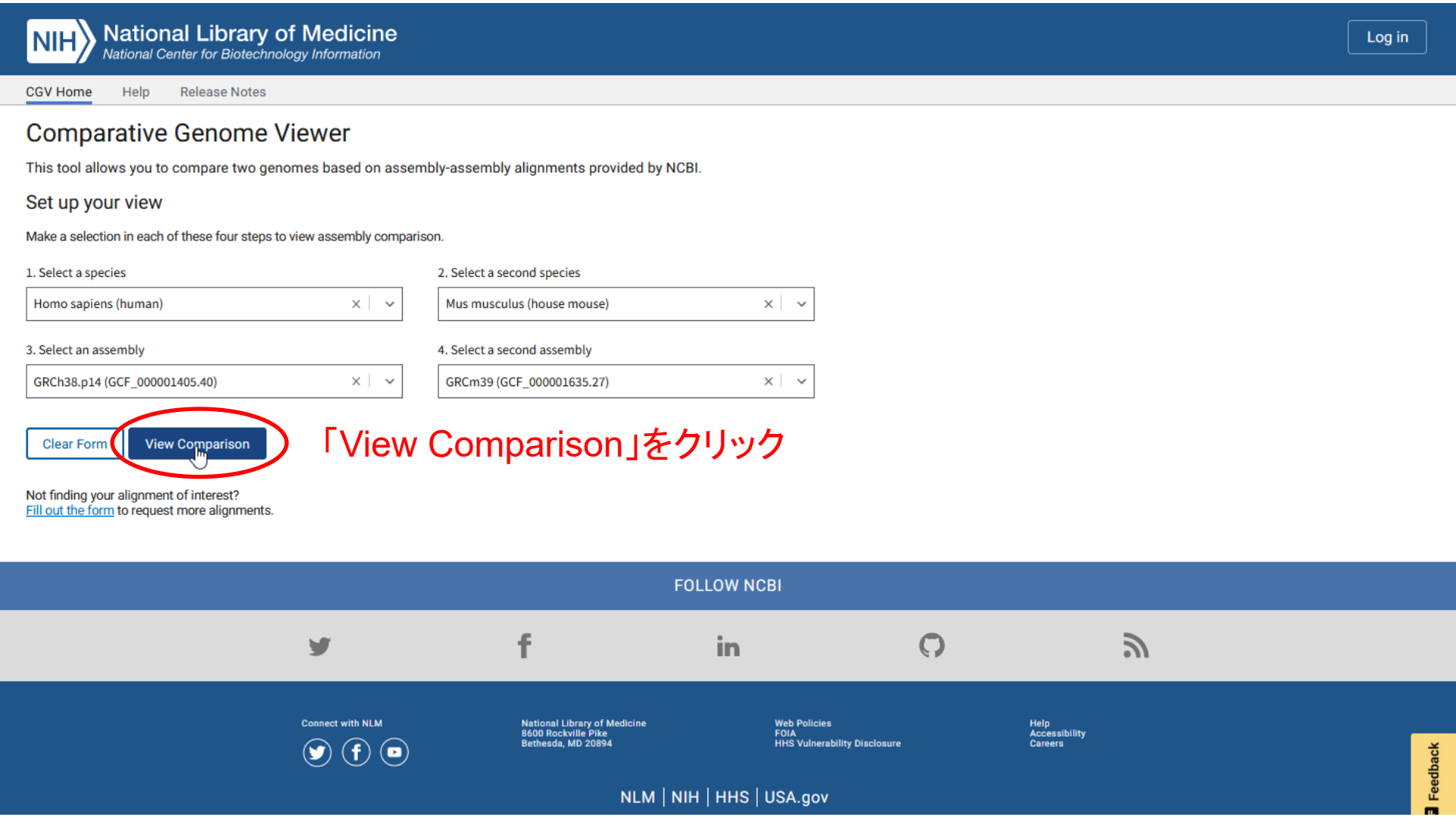
Homo sapiens **GRCh38.p14** (GCF_000001405.40)

Download data Download image

他の生物種とゲノムを比較する



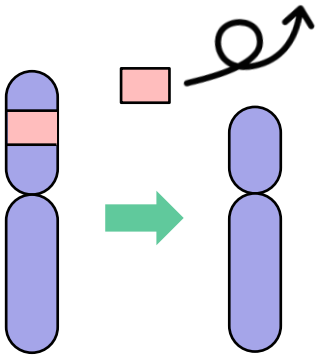
他の生物種とゲノムを比較する



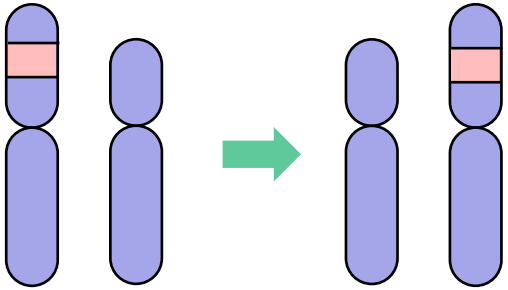
他の生物種とゲノムを比較する

染色体の再編成

疾患の原因と進化的多様化の原動力の両面をもつ重要なゲノム変化

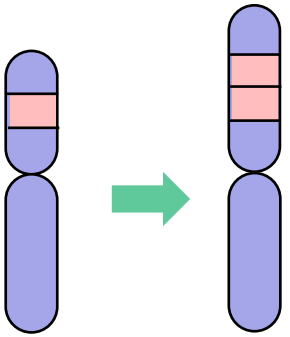


欠失

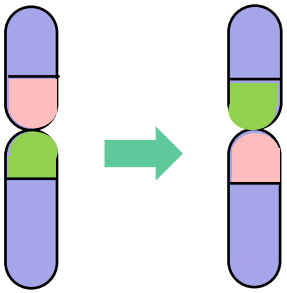


挿入

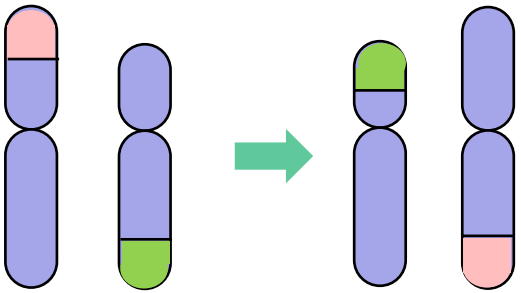
外来DNAのゲノムの一部
が取り込まれることもある



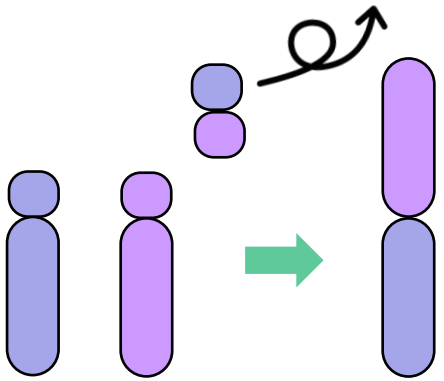
重複



逆位

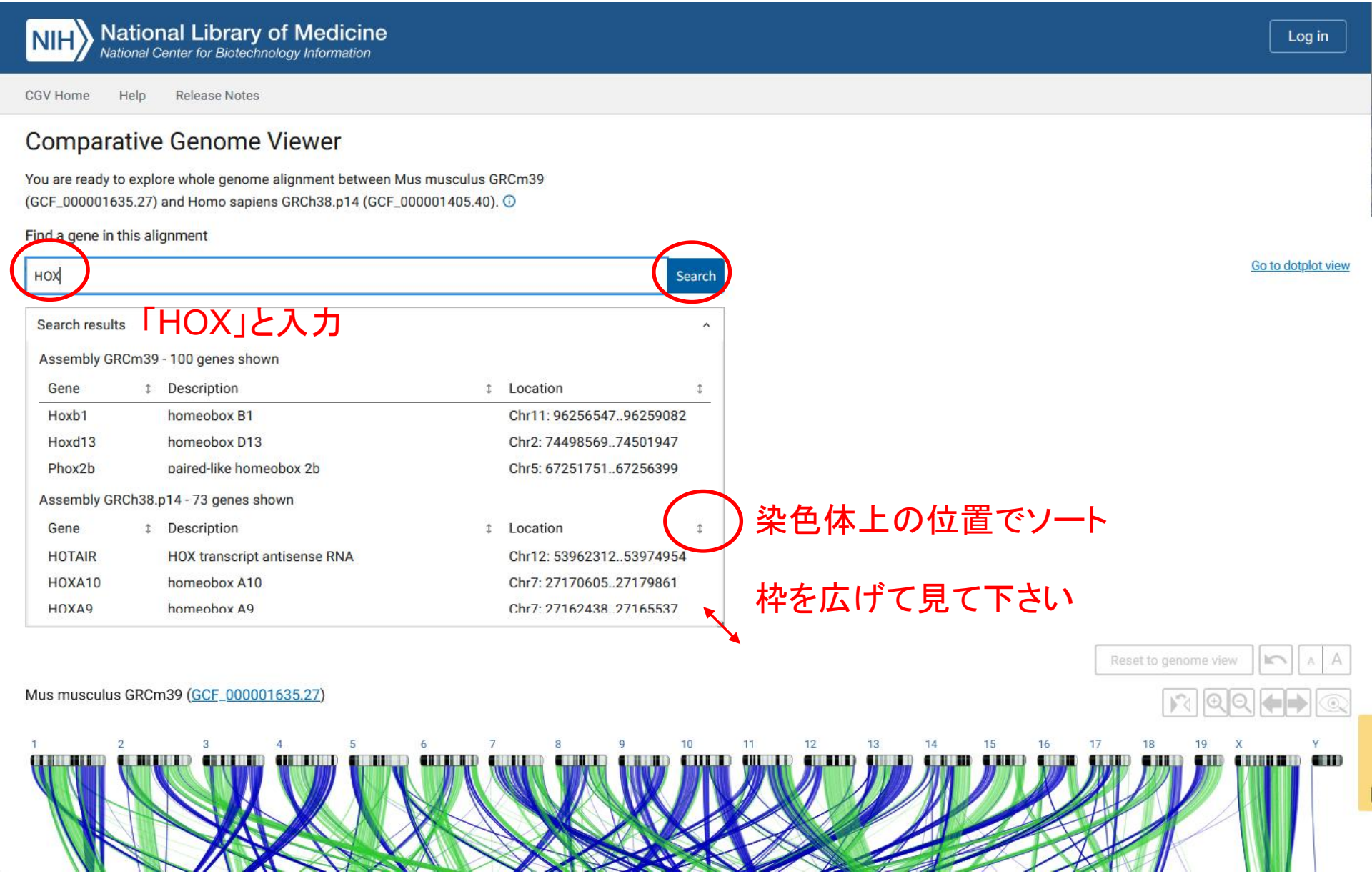


転座



Robertson型転座

遺伝子の位置の比較


National Library of Medicine
National Center for Biotechnology Information
Log in
CGV Home
Help
Release Notes
Comparative Genome Viewer
You are ready to explore whole genome alignment between Mus musculus GRCm39 (GCF_000001635.27) and Homo sapiens GRCh38.p14 (GCF_000001405.40).
Find a gene in this alignment
HOX
Search
Go to dotplot view
Search results
「HOX」と入力
Assembly GRCm39 - 100 genes shown
Gene
Description
Location
Hoxb1
homeobox B1
Chr11: 96256547..96259082
Hoxd13
homeobox D13
Chr2: 74498569..74501947
Phox2b
paired-like homeobox 2b
Chr5: 67251751..67256399
Assembly GRCh38.p14 - 73 genes shown
Gene
Description
Location
HOTAIR
HOX transcript antisense RNA
Chr12: 53962312..53974954
HOXA10
homeobox A10
Chr7: 27170605..27179861
HOXA9
homeobox A9
Chr7: 27162438..27165537
染色体上の位置でソート
枠を広げて見て下さい
Reset to genome view
Mus musculus GRCm39 (GCF_000001635.27)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y

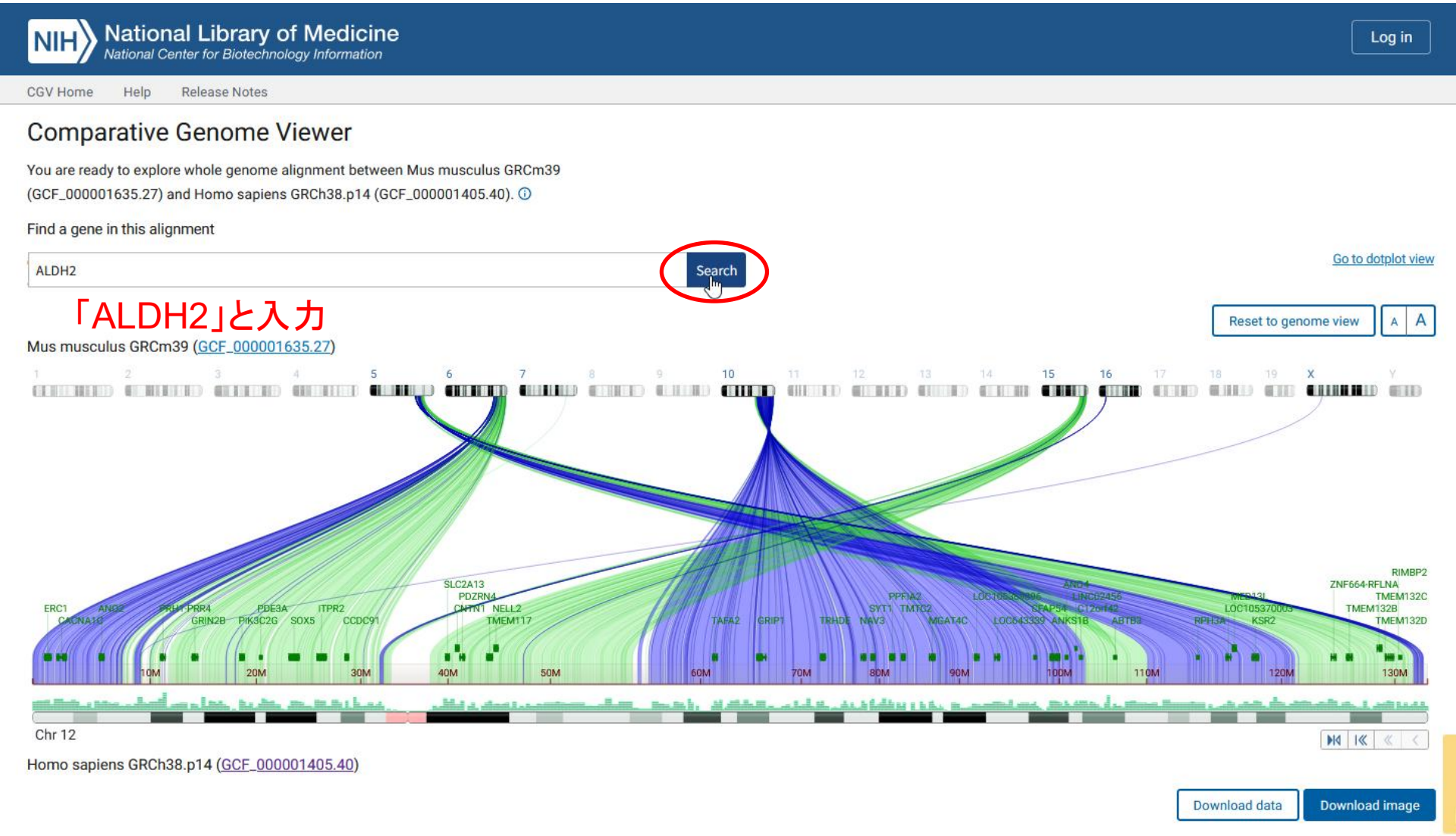
HOX遺伝子の比較

- ホメオボックス遺伝子群（Hox gene）：動物の体節形成や前後軸の形成、手足などのパターン形成を司る遺伝子群
- ヒトでは、染色体7番にHOXAクラスタが存在、マウスでは、染色体6番に存在
- エンハンサーHOTTIP（HOXA distal transcript antisense RNA）は染色体7番、ヒトではHOXAクラスタの近くに存在 → 手指の発達においてHOXAクラスターの遺伝子が一貫したパターンで発現し、手指の細かい運動や構造に関係している可能性がある

- **エンハンサー**: 特定の遺伝子の転写を活性化するDNA配列
- **遺伝子クラスター**: よく似た遺伝子が、染色体上で互いに近い位置にまとまって並んでいる状態

他の生物種とゲノムを比較する

他の生物種とゲノムを比較する



他の生物種とゲノムを比較する

Comparative Genome Viewer

You are ready to explore whole genome alignment between Mus musculus GRCm39 (GCF_000001635.27) and Homo sapiens GRCh38.p14 (GCF_000001405.40).

Find a gene in this alignment

ALDH2 | Search

Go to dotplot view

Search results

Assembly GRCm39 - 1 gene shown

Gene	Description	Location
Aldh2	aldehyde dehydrogenase 2, mitochondrial	Chr5: 121704090..121731887

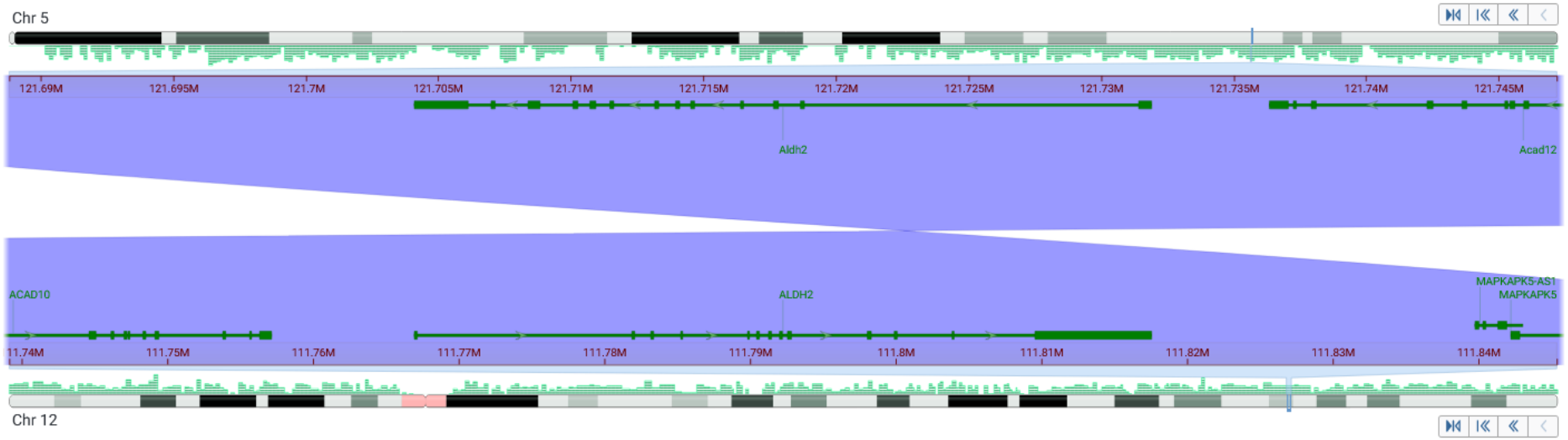
Assembly GRCh38.p14 - 1 gene shown

Gene	Description	Location
ALDH2	aldehyde dehydrogenase 2 family member	Chr12: 111766933..111817532

ヒト（下）とマウス（上）のそれぞれの染色体上の位置が表示される

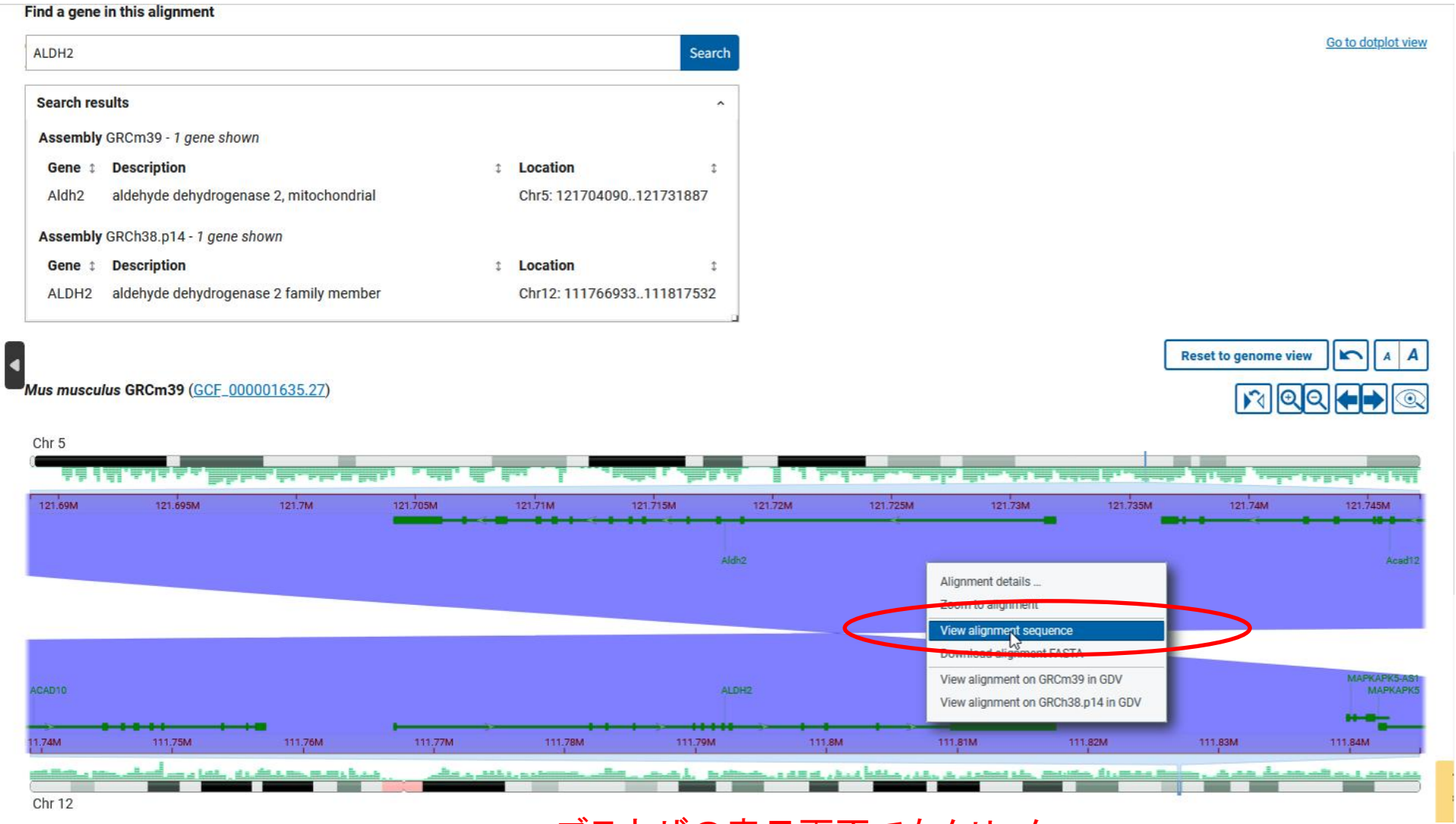
Reset to genome view | A | A

Mus musculus GRCm39 (GCF_000001635.27)



Homo sapiens GRCh38.p14 (GCF_000001405.40)

他の生物種とゲノムを比較する

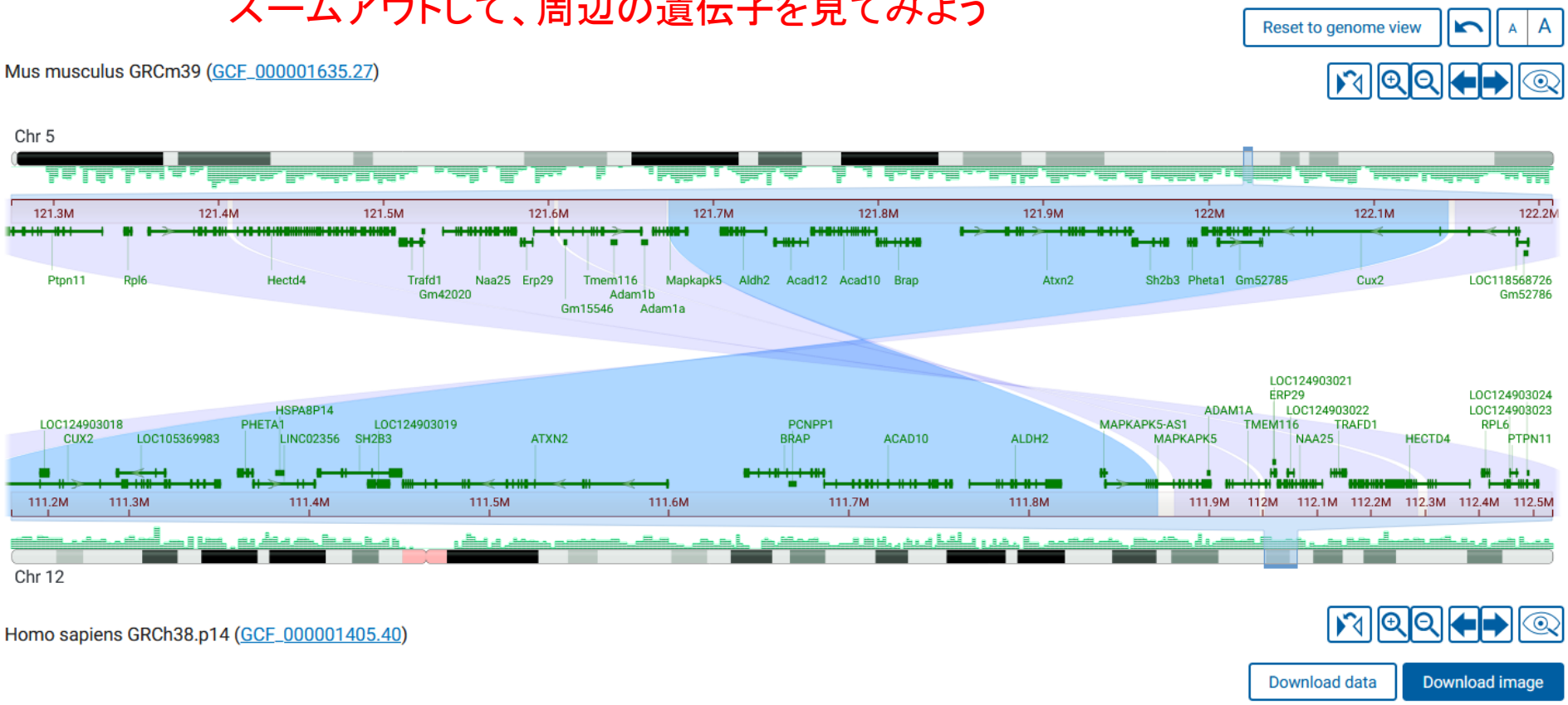


ブラウザの表示画面で右クリック
→「View alignment sequence」を選択

他の生物種とゲノムを比較する

他の生物種とゲノムを比較する

ズームアウトして、周辺の遺伝子を見てみよう



遺伝子配列の取得

National Library of Medicine
National Center for Biotechnology Information

Log in

CGV Home Help Release Notes

Comparative Genome Viewer

You are ready to explore whole genome alignment between ***Mus musculus*** **GRCm39** (GCF_000001635.27) and ***Homo sapiens*** **GRCh38.p14** (GCF_000001405.40).

Find a gene in this alignment

ALDH2 Search

Go to dotplot view

Search results

Assembly GRCm39 - *1 gene shown*

Gene	Description	Location
Aldh2	aldehyde dehydrogenase 2, mitochondrial	Chr5: 121704090..121731887

Assembly GRCh38.p14 - *1 gene shown*

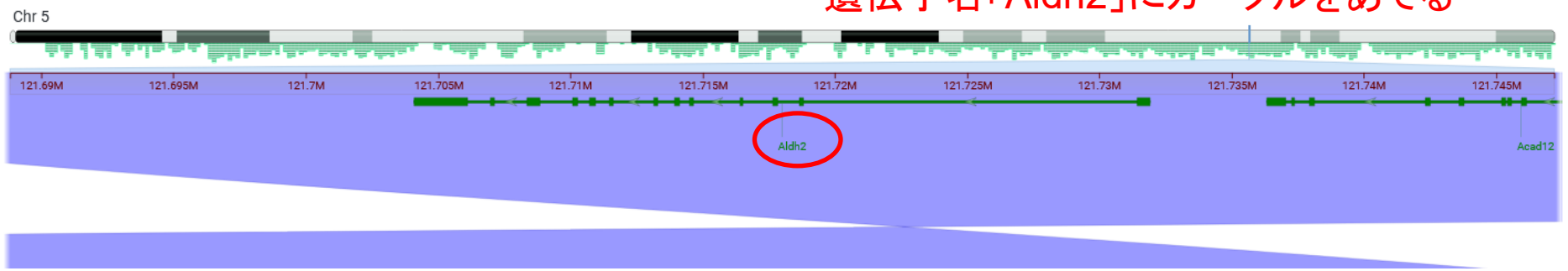
Gene	Description	Location
ALDH2	aldehyde dehydrogenase 2 family member	Chr12: 111766933..111817532

他の画面でもよいが、遺伝子名の表示された状態にあることが必要

Reset to genome view A A

Mus musculus **GRCm39** (GCF_000001635.27)

遺伝子名「Aldh2」にカーソルをあてる



遺伝子配列の取得

「Aldh2」をクリック

遺伝子配列の取得

National Library of Medicine
National Center for Biotechnology Information

Log in

Gene | Gene | Search | Advanced | Help

Full Report | Send to: | Hide sidebar >>

遺伝子Aldh2（マウス）のページ

Aldh2 aldehyde dehydrogenase 2, mitochondrial [*Mus musculus* (house mouse)]

Download Datasets

Gene ID: 11669, updated on 6-Oct-2023

配列をダウンロードするには
「Download Datasets」をクリック

Summary

Official Symbol	Aldh2 provided by MGI
Official Full Name	aldehyde dehydrogenase 2, mitochondrial provided by MGI
Primary source	MGI:MGI:99600
See related	Ensembl:ENSMUSG00000029455 AllianceGenome:MGI:99600
Gene type	protein coding
RefSeq status	VALIDATED
Organism	Mus musculus
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae; Murinae; Mus; Mus
Also known as	Ahd5; ALDHI; Ahd-5; AHD-M1; ALDH-E2
Summary	Predicted to enable NADH binding activity; aldehyde dehydrogenase (NAD+) activity; and identical protein binding activity. Predicted to be involved in several processes, including behavioral response to ethanol; cellular detoxification of aldehyde; and intrinsic apoptotic signaling pathway in response to oxidative stress. Located in mitochondrion. Is expressed in several structures, including alimentary system; genitourinary system; nervous system; respiratory system; and sensory organ. Human ortholog(s) of this gene implicated in several diseases, including alcohol use disorder; artery disease (multiple); diabetes mellitus (multiple); diabetic neuropathy; and liver disease (multiple). Orthologous to human ALDH2 (aldehyde dehydrogenase 2 family member). [provided by Alliance of Genome Resources, Apr 2022]
Expression	Broad expression in adrenal adult (RPKM 574.9), liver adult (RPKM 432.0) and 18 other tissues See more
Orthologs	human all

NEW Try the new Gene table
Try the new Transcript table

Genomic context

Location: 5 F; 5 61.86 cM — See Aldh2 in Genome Data Viewer

Exon count: 13

Annotation release	Status	Assembly	Chr	Location
RS_2023_04	current	GRCm39 (GCF_000001635.27)	5	NC_000071.7 (121704090..121731887, complement)
108.20200622	previous assembly	GRCm38.p6 (GCF_000001635.26)	5	NC_000071.6 (121566027..121593824, complement)

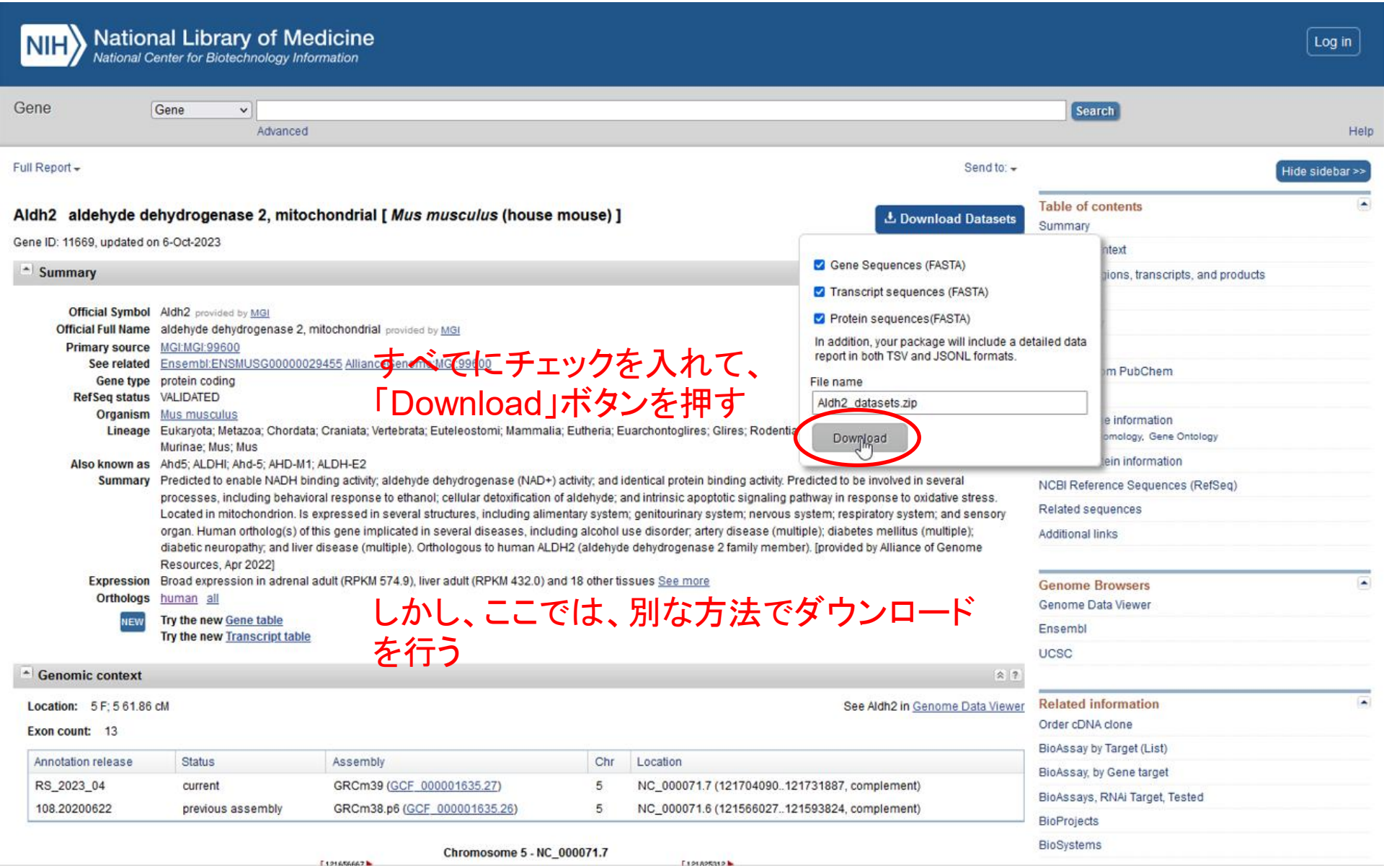
Chromosome 5 - NC_000071.7

Table of contents
- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Expression
- Bibliography
- Variation
- Pathways from PubChem
- Interactions
- General gene information
  - Markers, Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

Genome Browsers
- Genome Data Viewer
- Ensembl
- UCSC

Related information
- Order cDNA clone
- BioAssay by Target (List)
- BioAssay, by Gene target
- BioAssays, RNAi Target, Tested
- BioProjects
- BioSystems

遺伝子配列の取得


National Library of Medicine
National Center for Biotechnology Information
Log in
Gene
Search
Advanced
Help
Full Report
Send to:
Hide sidebar >>
Aldh2 aldehyde dehydrogenase 2, mitochondrial [Mus musculus (house mouse)]
Gene ID: 11669, updated on 6-Oct-2023
Download Datasets
Gene Sequences (FASTA)
Transcript sequences (FASTA)
Protein sequences(FASTA)
In addition, your package will include a detailed data report in both TSV and JSONL formats.
File name
Aldh2_datasets.zip
Download
すべてにチェックを入れて、
「Download」ボタンを押す
Summary
Official Symbol Aldh2 provided by MGI
Official Full Name aldehyde dehydrogenase 2, mitochondrial provided by MGI
Primary source MGI:MGI:99600
See related Ensembl:ENSMUSG00000029455 Alliance Genome:MGI:99600
Gene type protein coding
RefSeq status VALIDATED
Organism Mus musculus
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Murinae; Mus; Mus
Also known as Ahd5; ALDHI; Ahd-5; AHD-M1; ALDH-E2
Summary Predicted to enable NADH binding activity; aldehyde dehydrogenase (NAD+) activity; and identical protein binding activity. Predicted to be involved in several processes, including behavioral response to ethanol; cellular detoxification of aldehyde; and intrinsic apoptotic signaling pathway in response to oxidative stress. Located in mitochondrion. Is expressed in several structures, including alimentary system; genitourinary system; nervous system; respiratory system; and sensory organ. Human ortholog(s) of this gene implicated in several diseases, including alcohol use disorder; artery disease (multiple); diabetes mellitus (multiple); diabetic neuropathy; and liver disease (multiple). Orthologous to human ALDH2 (aldehyde dehydrogenase 2 family member). [provided by Alliance of Genome Resources, Apr 2022]
Expression Broad expression in adrenal adult (RPKM 574.9), liver adult (RPKM 432.0) and 18 other tissues See more
Orthologs human all
NEW Try the new Gene table
Try the new Transcript table
しかし、ここでは、別な方法でダウンロードを行う
Genomic context
Location: 5 F; 5 61.86 cM
See Aldh2 in Genome Data Viewer
Exon count: 13
Annotation release | Status | Assembly | Chr | Location
RS_2023_04 | current | GRCm39 (GCF_000001635.27) | 5 | NC_000071.7 (121704090..121731887, complement)
108.20200622 | previous assembly | GRCm38.p6 (GCF_000001635.26) | 5 | NC_000071.6 (121566027..121593824, complement)
Chromosome 5 - NC_000071.7
Table of contents
Summary
NCBI Reference Sequences (RefSeq)
Related sequences
Additional links
Genome Browsers
Genome Data Viewer
Ensembl
UCSC
Related information
Order cDNA clone
BioAssay by Target (List)
BioAssay, by Gene target
BioAssays, RNAi Target, Tested
BioProjects
BioSystems

遺伝子配列の取得

National Library of Medicine
National Center for Biotechnology Information

Log in

Gene | Gene | Search | Advanced | Help

Full Report | Send to: | Hide sidebar >>

遺伝子Aldh2（マウス）のページ

Aldh2 aldehyde dehydrogenase 2, mitochondrial [*Mus musculus* (house mouse)]

Download Datasets

Gene ID: 11669, updated on 6-Oct-2023

Summary

Official Symbol	Aldh2 provided by MGI
Official Full Name	aldehyde dehydrogenase 2, mitochondrial provided by MGI
Primary source	MGI:MGI:99600
See related	Ensembl:ENSMUSG00000029455 AllianceGenome:MGI:99600
Gene type	protein coding
RefSeq status	VALIDATED
Organism	Mus musculus
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae; Murinae; Mus; Mus
Also known as	Ahd5; ALDHI; Ahd-5; AHD-M1; ALDH-E2
Summary	Predicted to enable NADH binding activity; aldehyde dehydrogenase (NAD+) activity; and identical protein binding activity. Predicted to be involved in several processes, including behavioral response to ethanol; cellular detoxification of aldehyde; and intrinsic apoptotic signaling pathway in response to oxidative stress. Located in mitochondrion. Is expressed in several structures, including alimentary system; genitourinary system; nervous system; respiratory system; and sensory organ. Human ortholog(s) of this gene implicated in several diseases, including alcohol use disorder; artery disease (multiple); diabetes mellitus (multiple); diabetic neuropathy; and liver disease (multiple). Orthologous to human ALDH2 (aldehyde dehydrogenase 2 family member). [provided by Alliance of Genome Resources, Apr 2022]
Expression	Broad expression in adrenal adult (RPKM 574.9), liver adult (RPKM 432.0) and 18 other tissues See more
Orthologs	human all

NEW **Try the new Gene table**
Try the new Transcript table

「Transcript table」をクリック

Genomic context

Location: 5 F; 5 61.86 cM

See Aldh2 in Genome Data Viewer

Exon count: 13

Annotation release	Status	Assembly	Chr	Location
RS_2023_04	current	GRCm39 (GCF_000001635.27)	5	NC_000071.7 (121704090..121731887, complement)
108.20200622	previous assembly	GRCm38.p6 (GCF_000001635.26)	5	NC_000071.6 (121566027..121593824, complement)

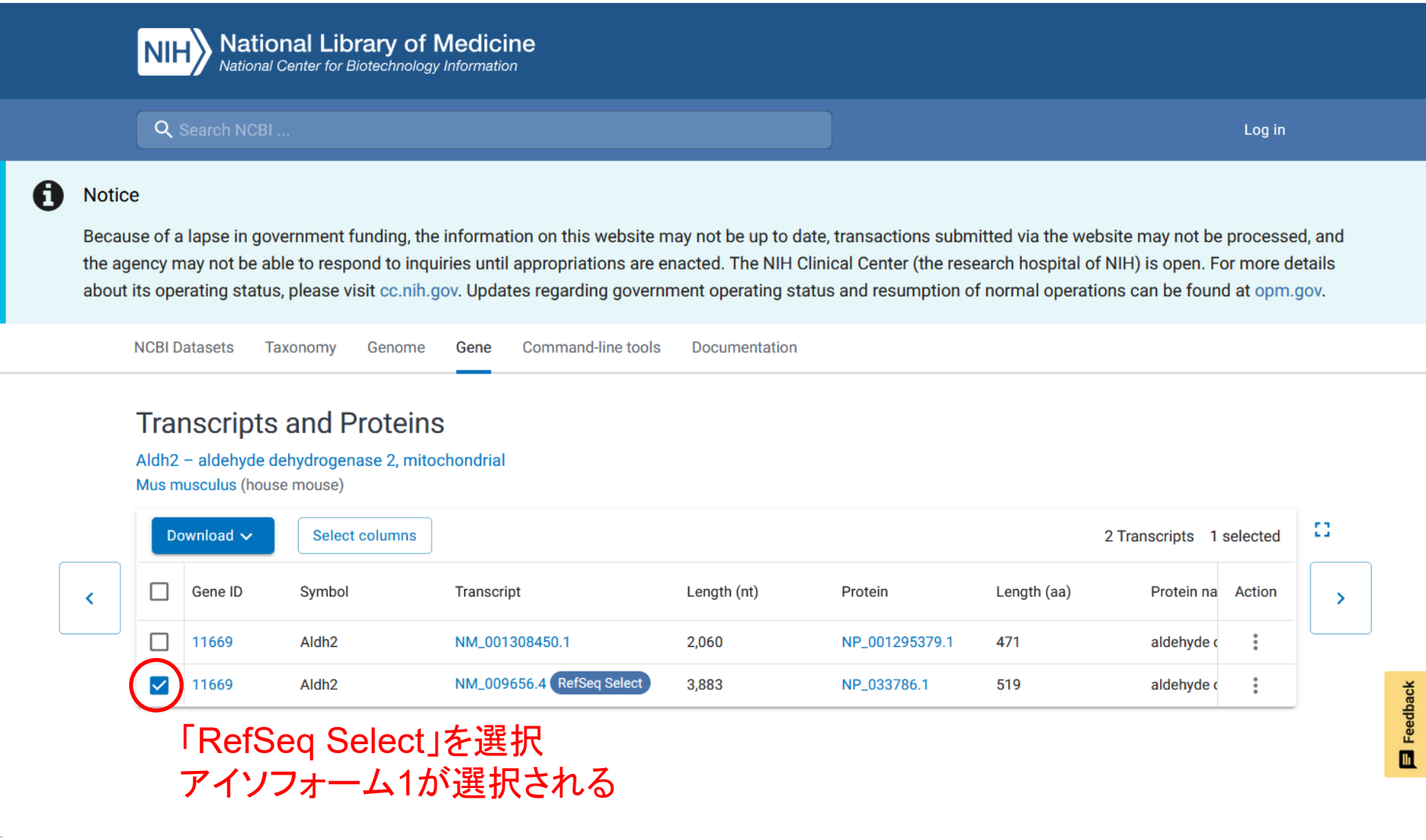
Chromosome 5 - NC_000071.7

Table of contents
- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Expression
- Bibliography
- Variation
- Pathways from PubChem
- Interactions
- General gene information
  Markers, Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

Genome Browsers
- Genome Data Viewer
- Ensembl
- UCSC

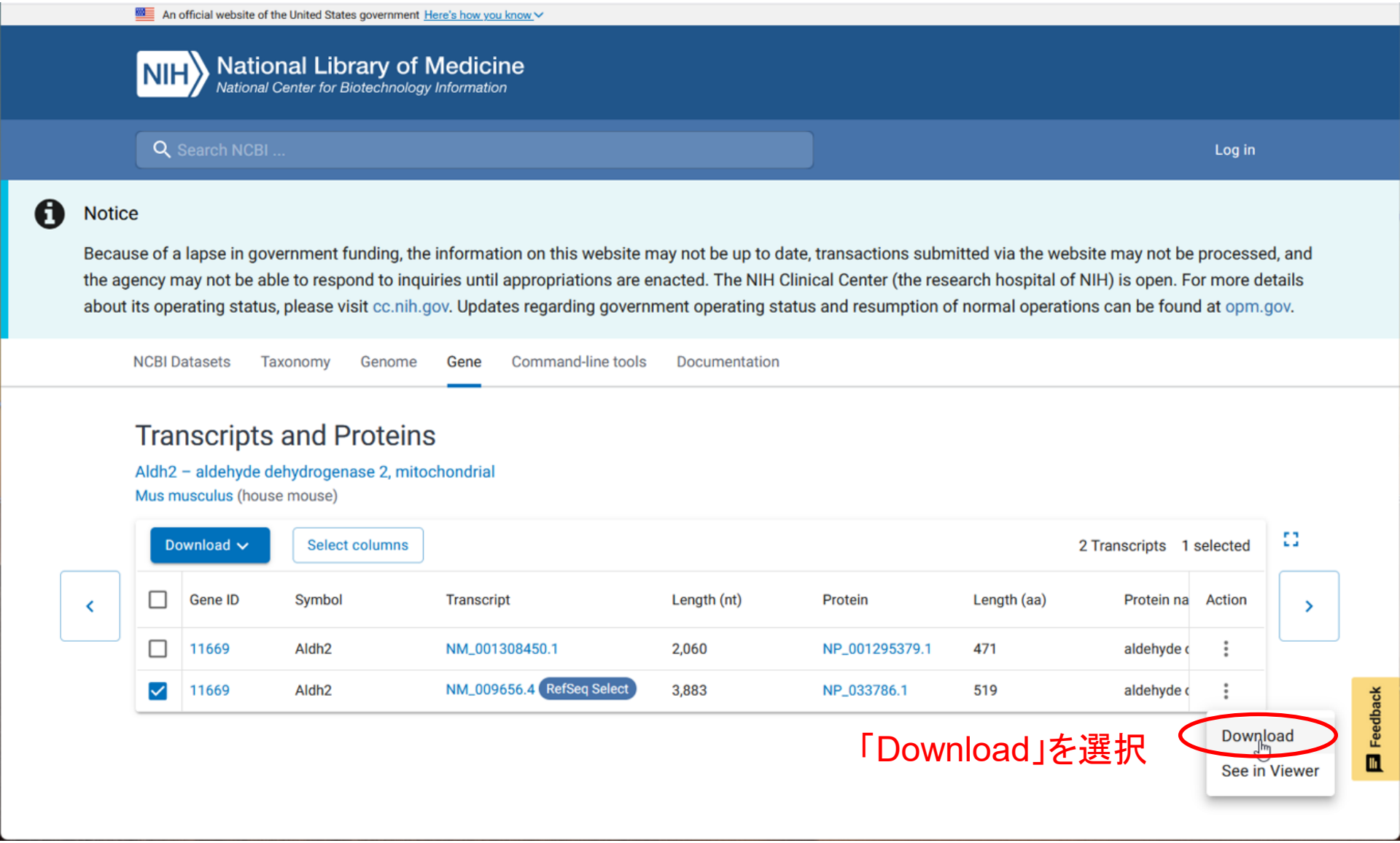
Related information
- Order cDNA clone
- BioAssay by Target (List)
- BioAssay, by Gene target
- BioAssays, RNAi Target, Tested
- BioProjects
- BioSystems

遺伝子配列の取得

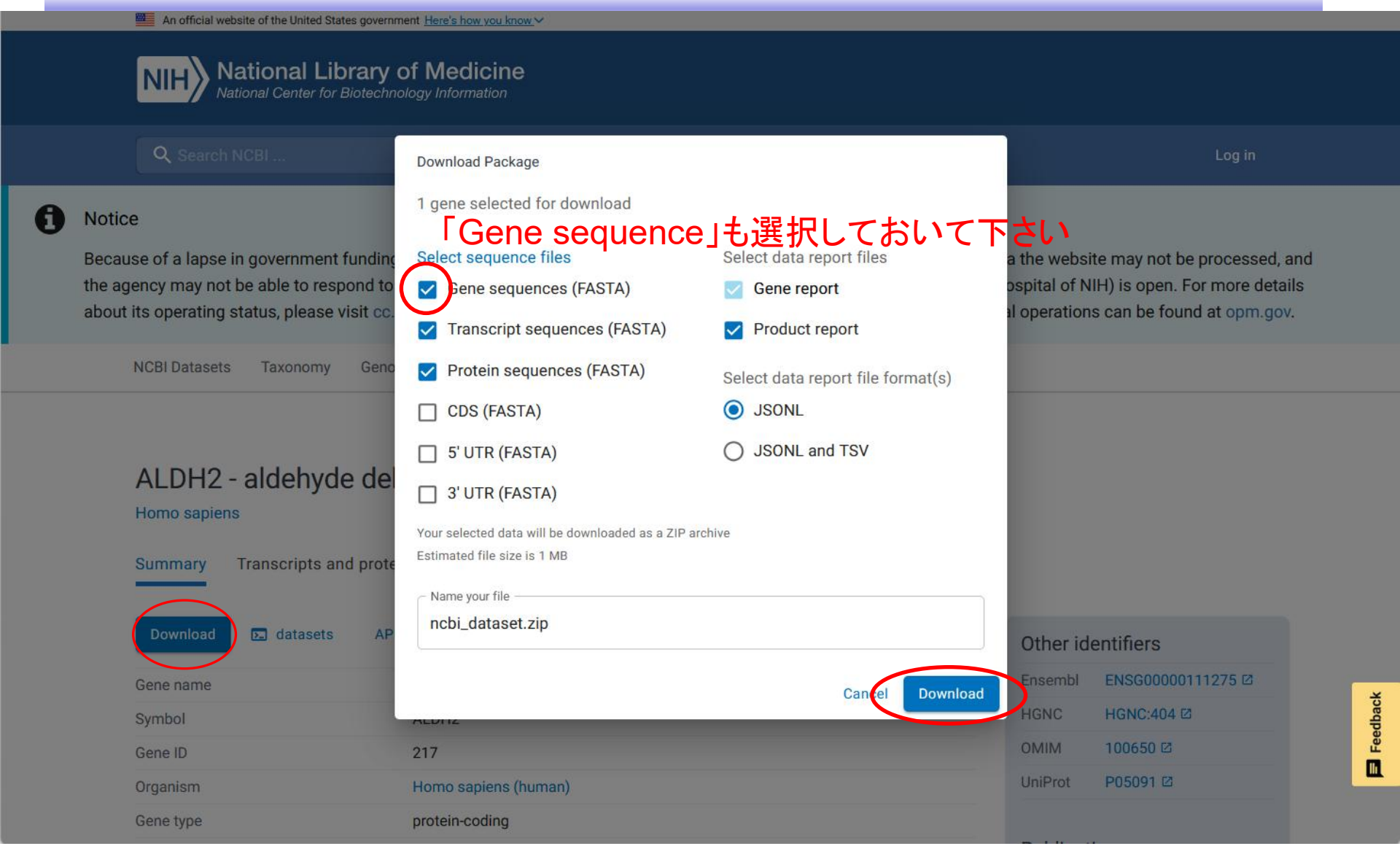


「RefSeq Select」を選択
アイソフォーム1が選択される

遺伝子配列の取得


An official website of the United States government Here's how you know
National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
Log in
Notice
Because of a lapse in government funding, the information on this website may not be up to date, transactions submitted via the website may not be processed, and the agency may not be able to respond to inquiries until appropriations are enacted. The NIH Clinical Center (the research hospital of NIH) is open. For more details about its operating status, please visit cc.nih.gov. Updates regarding government operating status and resumption of normal operations can be found at opm.gov.
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
Transcripts and Proteins
Aldh2 – aldehyde dehydrogenase 2, mitochondrial
Mus musculus (house mouse)
Download
Select columns
2 Transcripts 1 selected
Gene ID
Symbol
Transcript
Length (nt)
Protein
Length (aa)
Protein na
Action
11669
Aldh2
NM_001308450.1
2,060
NP_001295379.1
471
aldehyde
11669
Aldh2
NM_009656.4
RefSeq Select
3,883
NP_033786.1
519
aldehyde
Download
See in Viewer
「Download」を選択
Feedback

遺伝子配列の取得


An official website of the United States government Here's how you know
National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
Log in
Notice
Because of a lapse in government funding
the website may not be processed, and
the agency may not be able to respond to
ospital of NIH) is open. For more details
about its operating status, please visit cc.
l operations can be found at opm.gov.
NCBI Datasets
Taxonomy
Geno
ALDH2 - aldehyde de
Homo sapiens
Summary
Transcripts and prote
Download
datasets
AP
Gene name
Symbol
Gene ID
217
Organism
Homo sapiens (human)
Gene type
protein-coding
Other identifiers
Ensembl
ENSG00000111275
HGNC
HGNC:404
OMIM
100650
UniProt
P05091
Feedback
Download Package
1 gene selected for download
「Gene sequence」も選択しておいて下さい
Select sequence files
Gene sequences (FASTA)
Transcript sequences (FASTA)
Protein sequences (FASTA)
CDS (FASTA)
5' UTR (FASTA)
3' UTR (FASTA)
Select data report files
Gene report
Product report
Select data report file format(s)
JSONL
JSONL and TSV
Your selected data will be downloaded as a ZIP archive
Estimated file size is 1 MB
Name your file
ncbi_dataset.zip
Cancel
Download

遺伝子配列の取得

- ncbi_dataset.zipという圧縮ファイルとして、ダウンロードされる
  - 内容によらず、同じ名前のファイルになるので注意
- これを解凍する

遺伝子配列の取得

- 圧縮ファイルの中の`ncbi_dataset/data/`の下に配列が存在
  - `gene.fna`
  - `rna.fna`
  - `protein.faa`

拡張子	意味
`.fasta, .fas, .fa, .fsa, .seq`	とくに内容を指定しないfasta ファイル
`.fna`	遺伝子/RNAの塩基配列のfastaファイル
`.ffn`	ゲノムの翻訳領域（coding region）を含む fastaファイル
`.faa`	アミノ酸配列のfastaファイル
`.mpna`	複数のアミノ酸配列を含むfastaファイル
`.frn`	Non-coding RNAのfastaファイル

遺伝子配列の取得

- Gene Sequences (FASTA): 遺伝子の塩基配列
  - .fnaファイル
- Transcript Sequences (FASTA): mRNAの塩基配列
  - .rnaファイル
  - cDNA（complementary DNA）の配列
  - cDNA: mRNAから逆転写酵素を用いた逆転写反応によって合成された二本鎖DNA ↔ データベース上はエキソン領域（coding + UTR）を正しい順序でつなげた「成熟mRNA」に対応する塩基（ACGT）の配列
- Protein Sequences (FASTA): タンパク質のアミノ酸配列
  - .proteinファイル

遺伝子配列の取得

- マウスのgene.fnaの内容

```
>NC_000071.7:c121731887-121704090 Aldh2 [organism=Mus musculus]
[GeneID=11669] [chromosome=5]
ATTCTCTTCGCCGCCATATCTGCACAGATGTGAGCCTTAGGCGCCAGCCACCCTGCTAGGAGCGCACACC
ACTCTGGCTAGGCTTTCTCAGGGTTCTGCAAACTCCATCTCTGACTTGGCTTTGGGAGCCAGGGGTCGCG
CCCCTTAGGCCGTGAGGGGCTGGGACTCCCTGACCACGCCCCCGTGTCTCCGCCTCCCATTGGCGGCTGC
AGGGGGCGGAGGCGAGGACTTGTTCTTCAACGCTGCAGTCGCCCTCCGATCGGCAAGGCTTCTCTCGGCT
CCGTTCGGCTCGGCTCGCCCATTTCAGTTCAGTTCGGGTCAGTTAAGCTCCGCTCAGTTCAGCATGCTGC
GCGCCGCACTCACCACTGTCCGCCGCGGACCGCGCCTGAGCCGCCTGTTGTCCGCCGCCGCCACCAGCGC
GGTGCCAGCCCCCAACCATCAGCCTGAGGTCTTCTGCAACCAGGTGAGACTCACTTCGCCTCCTCGGGCC
CTAGGGCGGGACGGAGCGGGCGACGCGGCCTCAGTTTCCCCTCCGCGCCCAGCGATCTTTGCGCGCGTCT
TCCCTTGCCCATGTGGCCGCAGTGTGTCGCCTCTCGGCCTGGGACAGTCCTGGGCCTCTTTACCCCGCAT
CGATGCAGCCTGCTCCTTCCCCGCGCGGTCCGGGTTCACCCAGGAAGGCCCAGCCGTCCCAGCTCCTCTC
TGCGAGACCTCCCATACACCCCTACCCGATTCACGCAGACCTGAGCCAGAGTTCAGAAGCAAAAGCTGGG
TCAGGACCCCTCTGTGGCCGCTCTTACCCAACCGGGACTTTGGAATGTTGTCTTAGCTTTTTGAGTTTTG
GTAAAGTCTGGAATCACATGAAAGAACGTGGTTGTGAAGTATTCTTTTTGTTTGTGCAACAAGTGGGACT
TGGCTGCCCTCTTCTGTAAAATGGGAAAGCAGTCAGGGCTGATATCTTGATGTGTTATTTTGAAAGAAGA
TCTCATTTTGTAGCCAAGGCTGGCCCGGAATTCTGTGTTCTTCCGTCAGCCACTCCCACCCCACTTTGGA
TTTGTGCAGGCCAGACACGCAGCTACCACCTCCGGTCTTCAGCTTTCTCTCTACTTTTTCTTTGAAACAG
.....
ACTCACCTAGTGAGTGAACAAGGCGTGGAGAGCAAGCTGCCATCACAGGCACAAGAAACGGACGGTGAGC
TTAGCTTTAGAACTAGCCAGTCAGAGGCAGAGCTGAGGGTAGAAGGCTGATGAAGCCCTGAAGTTGTCCT
TCGACCTCCATATACACATCCCTGTATGTGCATGCGCACTCAATGAAATAAATAAGTAAATACAATTTTT
AAAGATCA
```

遺伝子配列の取得

- マウスのrna.fnaの内容

```
>NM_009656.4 Aldh2 [organism=Mus musculus] [GeneID=11669]
[transcript=1]
ATTCTCTTCGCCGCCATATCTGCACAGATGTGAGCCTTAGGCGCCAGCCACCCTGCTAGGAGCGCACACC
ACTCTGGCTAGGCTTTCTCAGGGTTCTGCAAACTCCATCTCTGACTTGGCTTTGGGAGCCAGGGGTCGCG
CCCCTTAGGCCGTGAGGGGCTGGGACTCCCTGACCACGCCCCCGTGTCTCCGCCTCCCATTGGCGGCTGC
AGGGGGCGGAGGCGAGGACTTGTTCTTCAACGCTGCAGTCGCCCTCCGATCGGCAAGGCTTCTCTCGGCT
CCGTTCGGCTCGGCTCGCCCATTTCAGTTCAGTTCGGGTCAGTTAAGCTCCGCTCAGTTCAGCATGCTGC
GCGCCGCACTCACCACTGTCCGCCGCGGACCGCGCCTGAGCCGCCTGTTGTCCGCCGCCGCCACCAGCGC
GGTGCCAGCCCCCAACCATCAGCCTGAGGTCTTCTGCAACCAGATCTTCATTAACAATGAGTGGCACGAC
.....
ACGGTGAGCTTAGCTTTAGAACTAGCCAGTCAGAGGCAGAGCTGAGGGTAGAAGGCTGATGAAGCCCTGA
AGTTGTCCTTCGACCTCCATATACACATCCCTGTATGTGCATGCGCACTCAATGAAATAAATAAGTAAAT
ACAATTTTTAAAGATCAAAAAAAAAAAAAAAA
```

```
>NM_001308450.1 Aldh2 [organism=Mus musculus] [GeneID=11669]
[transcript=2]
ATTCTCTTCGCCGCCATATCTGCACAGATGTGAGCCTTAGGCGCCAGCCACCCTGCTAGGAGCGCACACC
ACTCTGGCTAGGCTTTCTCAGGGTTCTGCAAACTCCATCTCTGACTTGGCTTTGGGAGCCAGGGGTCGCG
CCCCTTAGGCCGTGAGGGGCTGGGACTCCCTGACCACGCCCCCGTGTCTCCGCCTCCCATTGGCGGCTGC
AGGGGGCGGAGGCGAGGACTTGTTCTTCAACGCTGCAGTCGCCCTCCGATCGGCAAGGCTTCTCTCGGCT
CCGTTCGGCTCGGCTCGCCCATTTCAGTTCAGTTCGGGTCAGTTAAGCTCCGCTCAGTTCAGCATGCTGC
GCGCCGCACTCACCACTGTCCGCCGCGGACCGCGCCTGAGCCGCCTGTTGTCCGCCGCCGCCACCAGCGC
GGTGCCAGCCCCCAACCATCAGCCTGAGGTCTTCTGCAACCAGATCTTCATTAACAATGAGTGGCACGAC
.....
TGTGACGGTGAGAGCCCAGGCTGAGCATAAACAACTGCTCAGCGCCTGCTTCATCACCGTTAGGATTAAA
GAAAGAGCAAACGTGGATGGGGCTGCATAGAGATCAGTACCCTCTGCTTGACTGGTTGGAATAAAATAGT
CAGCTACTGTGGAAAAAAAAAAAAAAAA
```

アイソフォームを指定しないと、2つのアイソフォームの配列が取得される

遺伝子配列の取得

- マウスのprotein.faaの内容

```
>NP_001295379.1 Aldh2 [organism=Mus musculus] [GeneID=11669] [isoform=2
precursor]
MLRAALTTVRRGPRLSRLLSAAATSAVPAPNHQPEVFCNQIFINNEWHDAVSRKTFPTVNPSTGEVICQV
AEGNKEDVDKAVKAARAAFQLGSPWRRMDASDRGRLLYRLADLIERDRTYLAALETLDNGKPYVISYLVD
LDMVLKCLRYYAGWADKYHGKTIPIDGDFFSYTRHEPVGVCGQIIPWNFPLLMQAWKLGPALATGNVVVM
KVAEQTPLTALYVANLIKEAGFPPGVVNIVPGFGPTAGAAIASHEGVDKVAFTGSTEVGHLIQVAAGSSN
LKRVTLELGGKSPNIIMSDADMDWAVEQAHFALFFNQGQCCCAGSRTFVQENVYDEFVERSVARAKSRVV
GNPFDSRTEQGPQVDETQFKKILGYIKSGQQEGAKLLCGGGAAADRGYFIQPTVFGDVKDGMTIAKEEIF
GPVMQILKFKTIEEVVGRANDSKYGLAAAVFTKDLDKANYLSQALQAGTVW
```

```
>NP_033786.1 Aldh2 [organism=Mus musculus] [GeneID=11669] [isoform=1
precursor]
MLRAALTTVRRGPRLSRLLSAAATSAVPAPNHQPEVFCNQIFINNEWHDAVSRKTFPTVNPSTGEVICQV
AEGNKEDVDKAVKAARAAFQLGSPWRRMDASDRGRLLYRLADLIERDRTYLAALETLDNGKPYVISYLVD
LDMVLKCLRYYAGWADKYHGKTIPIDGDFFSYTRHEPVGVCGQIIPWNFPLLMQAWKLGPALATGNVVVM
KVAEQTPLTALYVANLIKEAGFPPGVVNIVPGFGPTAGAAIASHEGVDKVAFTGSTEVGHLIQVAAGSSN
LKRVTLELGGKSPNIIMSDADMDWAVEQAHFALFFNQGQCCCAGSRTFVQENVYDEFVERSVARAKSRVV
GNPFDSRTEQGPQVDETQFKKILGYIKSGQQEGAKLLCGGGAAADRGYFIQPTVFGDVKDGMTIAKEEIF
GPVMQILKFKTIEEVVGRANDSKYGLAAAVFTKDLDKANYLSQALQAGTVWINCYDVFGAQSPFGGYKMS
GSGRELGEYGLQAYTEVKTVTVKVPQKNS
```

アイソフォームを指定しないと、2つのアイソフォームの配列が取得される

FASTA形式

- 1行の説明とその後に続く配列データから構成される
  - ヘッダ行: 先頭の1行。‘>’ で始まる。配列の説明。
  - 配列データ: 配列を表す文字列。複数行にわたってよい。
    - IUB/IUPACで規定されているアミノ酸または核酸コード
    - ただし、
      - 小文字は大文字に変換される
      - ‘-’ でギャップを表す
      - 数字は受け付けない

A	adenosine	M	A C (amino)
C	cytidine	S	G C (strong)
G	guanine	W	A T (weak)
T	thymidine	B	G T C
U	uridine	D	G A T
R	G A (purine)	H	A C T
Y	T C (pyrimidine)	V	G C A
K	G T (keto)	N	A G C T (any)
		-	gap of indeterminate length

A	alanine	P	proline
B	aspartate or asparagine	Q	glutamine
C	cystine	R	arginine
D	aspartate	S	serine
E	glutamate	T	threonine
F	phenylalanine	U	selenocysteine
G	glycine	V	valine
H	histidine	W	tryptophan
I	isoleucine	Y	tyrosine
K	lysine	Z	glutamate or glutamine
L	leucine	X	any
M	methionine	*	translation stop
N	asparagine	-	gap of indeterminate

```
>sp|P61626|LYSC_HUMAN Lysozyme C OS=Homo sapiens GN=LYZ PE=1 SV=1
MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATNYNAGDRST
DYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRDPQGIRAWVAWRNRCQNRDVR
QYVQGCGV

>ref|NC_000012.11|:69742134-69748013 Homo sapiens chromosome 12, GRCh37
AAATACTGGGGCCAGCTCACCCTGGTCAGCCTAGCACTCTGACCTAGCAGTCAACATGAAGGCTCTCATT
GTTCTGGGGCTTGTCCTCCTTTCTGTTACGGTCCAGGGCAAGGTCTTTGAAAGGTGTGAGTTGGCCAGAA
CTCTGAAAAGATTGGGAATGGATGGCTACAGGGGAATCAGCCTAGCAAACTGTAAGTCTACTCTCCATAA
TTCCAGAGAATTAGCTACGTATGGAACAGACACTAGGAGAGAAGGAAGAAGAAGAAGGGGCTTTGAGTGA
```

（途中まで）

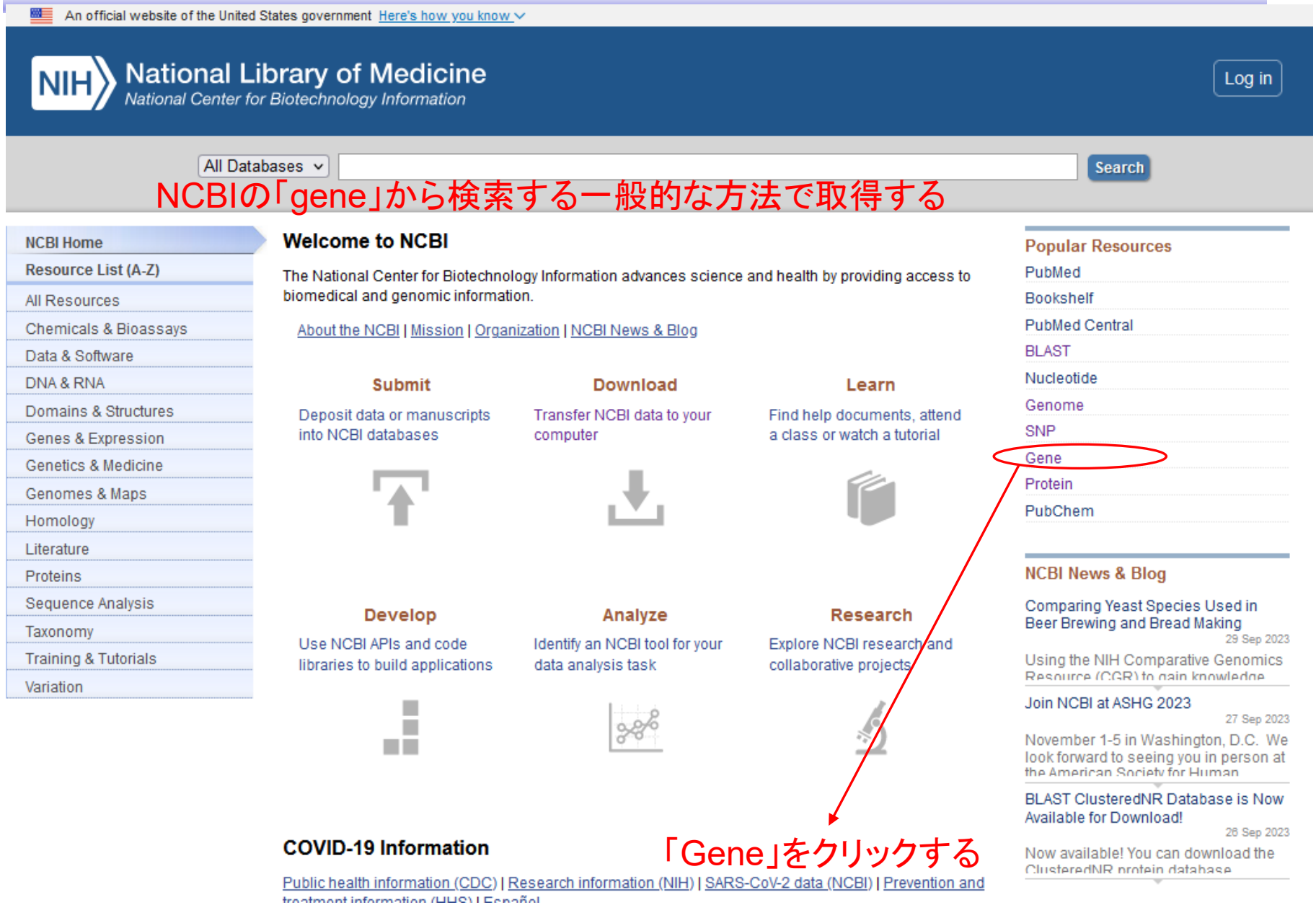
IUPAC: International Union of Pure and Applied Chemistry（国際純正応用化学連合）
IUB: International Union of Biochemistry（国際生化学連合）

配列の準備

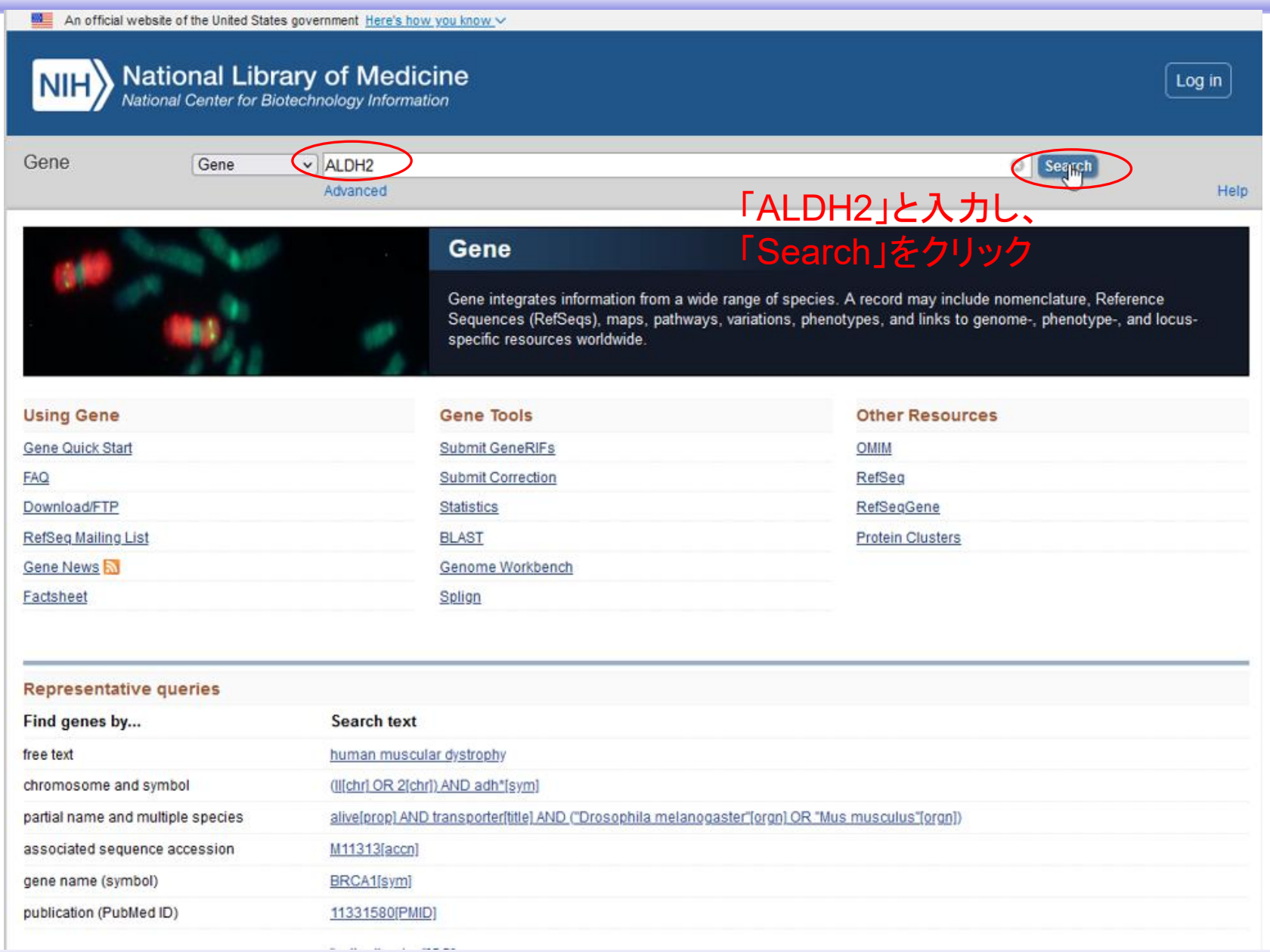
- 取得したファイルは以下のように名前を変更する
  - `rna.fna → mouse-rna.fna`
  - `protein.faa → mouse-protein.faa`
- 同様に、ヒトの配列をダウンロードして下さい
- NCBIの「gene」から検索する一般的な方法で取得する
- 同様に、取得したファイルの名前を変更する
  - `rna.fna → human-rna.fna`
  - `protein.faa → human-protein.faa`

遺伝子配列の取得

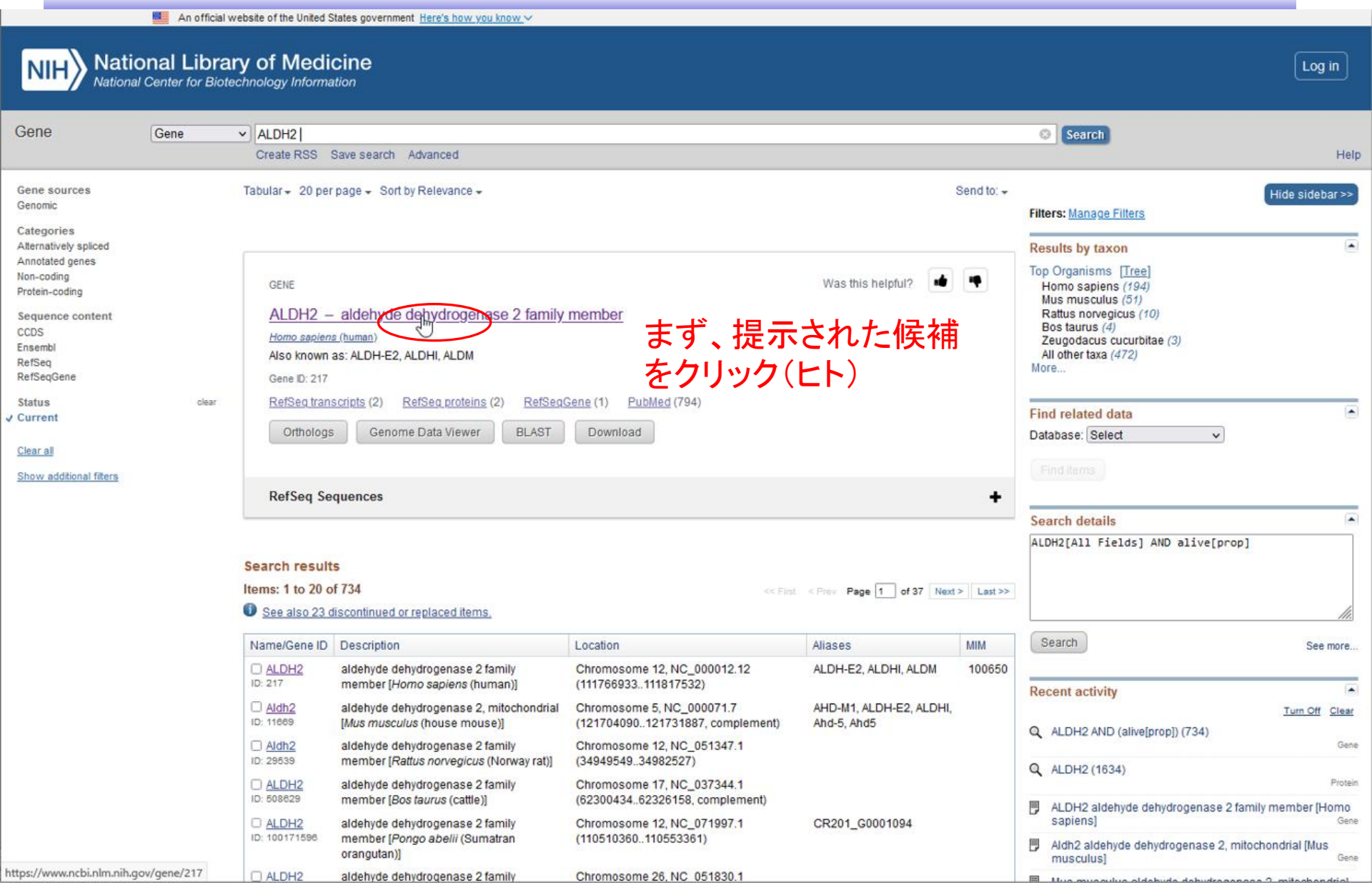
以下、作業を飛ばして結構です



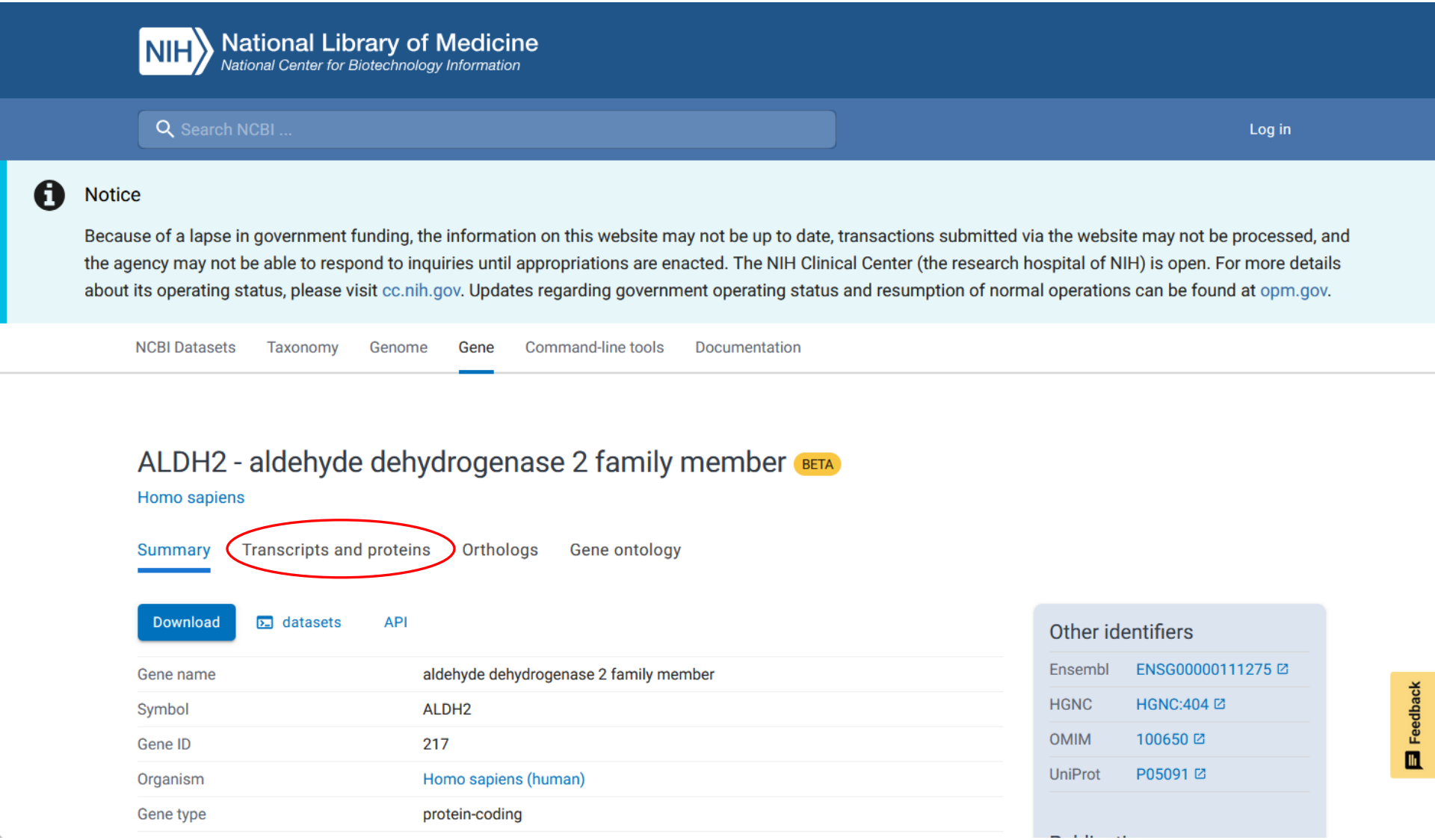
遺伝子配列の取得


An official website of the United States government Here's how you know
National Library of Medicine
National Center for Biotechnology Information
Log in
Gene
Gene
ALDH2
Search
Advanced
Help
「ALDH2」と入力し、
「Search」をクリック
Gene
Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.
Using Gene
Gene Quick Start
FAQ
Download/FTP
RefSeq Mailing List
Gene News
Factsheet
Gene Tools
Submit GeneRIFs
Submit Correction
Statistics
BLAST
Genome Workbench
Splign
Other Resources
OMIM
RefSeq
RefSeqGene
Protein Clusters
Representative queries
Find genes by...
Search text
free text
human muscular dystrophy
chromosome and symbol
(II[chr] OR 2[chr]) AND adh*[sym]
partial name and multiple species
alive[prop] AND transporter[title] AND ("Drosophila melanogaster"[orgn] OR "Mus musculus"[orgn])
associated sequence accession
M11313[accn]
gene name (symbol)
BRCA1[sym]
publication (PubMed ID)
11331580[PMID]

遺伝子配列の取得


An official website of the United States government Here's how you know
National Library of Medicine
National Center for Biotechnology Information
Log in
Gene
Gene
ALDH2
Search
Create RSS Save search Advanced
Help
Gene sources
Genomic
Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Sequence content
CCDS
Ensembl
RefSeq
RefSeqGene
Status
clear
Current
Clear all
Show additional filters
Tabular 20 per page Sort by Relevance
Send to:
Hide sidebar >>
Filters: Manage Filters
GENE
Was this helpful?
ALDH2 – aldehyde dehydrogenase 2 family member
Homo sapiens (human)
Also known as: ALDH-E2, ALDHI, ALDM
Gene ID: 217
RefSeq transcripts (2) RefSeq proteins (2) RefSeqGene (1) PubMed (794)
Orthologs Genome Data Viewer BLAST Download
RefSeq Sequences
まず、提示された候補をクリック（ヒト）
Search results
Items: 1 to 20 of 734
<< First < Prev Page 1 of 37 Next > Last >>
See also 23 discontinued or replaced items.
Name/Gene ID Description Location Aliases MIM
ALDH2 ID: 217 aldehyde dehydrogenase 2 family member [Homo sapiens (human)] Chromosome 12, NC_000012.12 (111766933..111817532) ALDH-E2, ALDHI, ALDM 100650
Aldh2 ID: 11669 aldehyde dehydrogenase 2, mitochondrial [Mus musculus (house mouse)] Chromosome 5, NC_000071.7 (121704090..121731887, complement) AHD-M1, ALDH-E2, ALDHI, Ahd-5, Ahd5
Aldh2 ID: 29539 aldehyde dehydrogenase 2 family member [Rattus norvegicus (Norway rat)] Chromosome 12, NC_051347.1 (34949549..34982527)
ALDH2 ID: 508629 aldehyde dehydrogenase 2 family member [Bos taurus (cattle)] Chromosome 17, NC_037344.1 (62300434..62326158, complement)
ALDH2 ID: 100171596 aldehyde dehydrogenase 2 family member [Pongo abelii (Sumatran orangutan)] Chromosome 12, NC_071997.1 (110510360..110553361) CR201_G0001094
ALDH2 aldehyde dehydrogenase 2 family Chromosome 26, NC_051830.1
Results by taxon
Top Organisms [Tree]
Homo sapiens (194)
Mus musculus (51)
Rattus norvegicus (10)
Bos taurus (4)
Zeugodacus cucurbitae (3)
All other taxa (472)
More...
Find related data
Database: Select
Find items
Search details
ALDH2[All Fields] AND alive[prop]
Search
See more...
Recent activity
Turn Off Clear
ALDH2 AND (alive[prop]) (734) Gene
ALDH2 (1634) Protein
ALDH2 aldehyde dehydrogenase 2 family member [Homo sapiens] Gene
Aldh2 aldehyde dehydrogenase 2, mitochondrial [Mus musculus] Gene
https://www.ncbi.nlm.nih.gov/gene/217

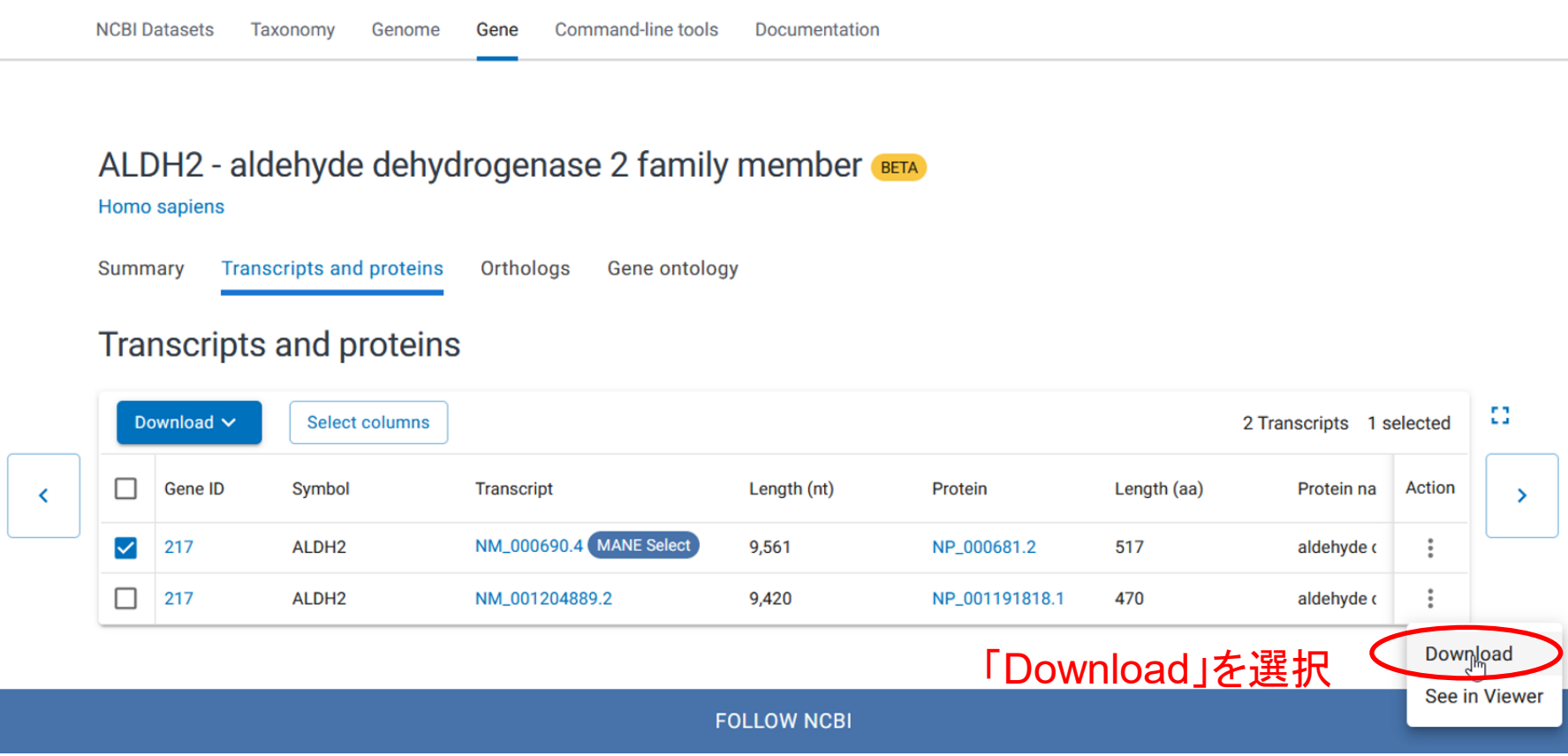
遺伝子配列の取得


National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
Log in
Notice
Because of a lapse in government funding, the information on this website may not be up to date, transactions submitted via the website may not be processed, and the agency may not be able to respond to inquiries until appropriations are enacted. The NIH Clinical Center (the research hospital of NIH) is open. For more details about its operating status, please visit cc.nih.gov. Updates regarding government operating status and resumption of normal operations can be found at opm.gov.
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
ALDH2 - aldehyde dehydrogenase 2 family member
BETA
Homo sapiens
Summary
Transcripts and proteins
Orthologs
Gene ontology
Download
datasets
API
Gene name
aldehyde dehydrogenase 2 family member
Symbol
ALDH2
Gene ID
217
Organism
Homo sapiens (human)
Gene type
protein-coding
Other identifiers
Ensembl
ENSG00000111275
HGNC
HGNC:404
OMIM
100650
UniProt
P05091
Feedback

遺伝子配列の取得


An official website of the United States government Here's how you know
National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
Log in
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
ALDH2 - aldehyde dehydrogenase 2 family member
BETA
Homo sapiens
Summary
Transcripts and proteins
Orthologs
Gene ontology
Transcripts and proteins
Download
Select columns
2 Transcripts 1 selected
Gene ID
Symbol
Transcript
Length (nt)
Protein
Length (aa)
Protein na
Action
217
ALDH2
NM_000690.4
MANE Select
9,561
NP_000681.2
517
aldehyde
217
ALDH2
NM_001204889.2
9,420
NP_001191818.1
470
aldehyde
「MANE Select」を選択
アイソフォーム1が選択される
FOLLOW NCBI

遺伝子配列の取得


NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
ALDH2 - aldehyde dehydrogenase 2 family member
BETA
Homo sapiens
Summary
Transcripts and proteins
Orthologs
Gene ontology
Transcripts and proteins
Download
Select columns
2 Transcripts 1 selected
Gene ID
Symbol
Transcript
Length (nt)
Protein
Length (aa)
Protein na
Action
217
ALDH2
NM_000690.4
MANE Select
9,561
NP_000681.2
517
aldehyde
217
ALDH2
NM_001204889.2
9,420
NP_001191818.1
470
aldehyde
Download
See in Viewer
「Download」を選択
FOLLOW NCBI

遺伝子配列の取得

- ヒトのgene.fnaの内容

```
>NC_000012.12:111766933-111817532 ALDH2 [organism=Homo sapiens]
[GeneID=217] [chromosome=12]
GAGACCCTAGCTCTGCTCTCGGTCCGCTCGCTGTCCGCTAGCCCGCTGCGATGTTGCGCGCTGCCGCCCG
CTTCGGGCCCCGCCTGGGCCGCCGCCTCTTGTCAGCCGCCGCCACCCAGGCCGTGCCTGCCCCCAACCAG
CAGCCCGAGGTCTTCTGCAACCAGGTGAGCCCACCGGCCGGGCTCGCGCTTTGTTTTCCGGCCCGAGTCC
CCCGCAGGCCCCTAGGAAGGCCCCGCGCCGCCGTGGGCCTTAGTGTACTCATCTGGGGCTCGAGGGGTTT
GCAGAGGCTGACCTGGAAGCACATCTGCCCCCTCCTCTCCTGCAAGCATTTTAGCCCCTTCGCCGCCTCT
GACAGTCCCCGTCCCATTTCCCCGTGACTTGGGCCCCCTCTCCGTCTTCGCCCAGCCTCGGTCTGTTTTC
CAGGGCCAACTCTCGGGGTTCCGTTCTCCCCATGGTCCTTGCTTTCGGGTCTCCGCAGGGTCCCCACCCT
.....
GGGACAGATTAACAGGAATCCATAGCCCAGGGATGCGACCCAAAATTATCAAAGTAGAGACATTATGTGT
TTGCAATGTGCTATGATTAATGCAGTGACATAACTGGCAAGATTTACAGGTCAATTGGATATTGTTTACC
TGGAGCTGGTCCTTCTTAGTTGCCAAAAAAGACATAAGGATTAAAAAGACAAACCATAAA
>NC_060936.1:111743846-111794443 ALDH2 [organism=Homo sapiens]
[GeneID=217] [chromosome=12]
GAGACCCTAGCTCTGCTCTCGGTCCGCTCGCTGTCCGCTAGCCCGCTGCGATGTTGCGCGCTGCCGCCCG
CTTCGGGCCCCGCCTGGGCCGCCGCCTCTTGTCAGCCGCCGCCACCCAGGCCGTGCCTGCCCCCAACCAG
CAGCCCGAGGTCTTCTGCAACCAGGTGAGCCCACCGGCCGGGCTCGCGCTTTGTTTTCCGGCCCGAGTCC
CCCGCAGGCCCCTAGGAAGGCCCCGCGCCGCCGTGGGCCTTAGTGTACTCATCTGGGGCTCGAGGGGTTT
GCAGAGGCTGACCTGGAAGCACATCTGCCCCCTCCTCTCCTGCAAGCATTTTAGCCCCTTCGCCGCCTCT
GACAGTCCCCGTCCCATTTCCCCGTGACTTGGGCCCCCTCTCCGTCTTCGCCCAGCCTCGGTCTGTTTTC
CAGGGCCAACTCTCGGGGTTCCGTTCTCCCCATGGTCCTTGCTTTCGGGTCTCCGCAGGGTCCCCACCCT
.....
GACAGATTAACAGGAATCCATAGCCCAGGGATGCGACCCAAAATTATCAAAGTAGAGACATTATGTGTTT
GCAATGTGCTATGATTAATGCAGTGACATAACTGGCAAGATTTACAGGTCAATTGGATATTGTTTACCTG
GAGCTGGTCCTTCTTAGTTGCCAAAAAAGACATAAGGATTAAAAAGACAAACCATAAA
```

遺伝子配列の取得

- ヒトのrna.fnaの内容

```
>NM_000690.4 ALDH2 [organism=Homo sapiens] [GeneID=217] [transcript=1]
GAGACCCTAGCTCTGCTCTCGGTCCGCTCGCTGTCCGCTAGCCCGCTGCGATGTTGCGCGCTGCCGCCCG
CTTCGGGCCCCGCCTGGGCCGCCGCCTCTTGTCAGCCGCCGCCACCCAGGCCGTGCCTGCCCCCAACCAG
CAGCCCGAGGTCTTCTGCAACCAGATTTTCATAAACAATGAATGGCACGATGCCGTCAGCAGGAAAACAT
TCCCCACCGTCAATCCGTCCACTGGAGAGGTCATCTGTCAGGTAGCTGAAGGGGACAAGGAAGATGTGGA
CAAGGCAGTGAAGGCCGCCCGGGCCGCCTTCCAGCTGGGCTCACCTTGGCGCCGCATGGACGCATCACAC
AGGGGCCGGCTGCTGAACCGCCTGGCCGATCTGATCGAGCGGGACCGGACCTACCTGGCGGCCTTGGAGA
CCCTGGACAATGGCAAGCCCTATGTCATCTCCTACCTGGTGGATTTGGACATGGTCCTCAAATGTCTCCG··
···
CCATAGCCCAGGGATGCGACCCAAAATTATCAAAGTAGAGACATTATGTGTTTGCAATGTGCTATGATTA
ATGCAGTGACATAACTGGCAAGATTTACAGGTCAATTGGATATTGTTTACCTGGAGCTGGTCCTTCTTAG
TTGCCAAAAAAGACATAAGGATTAAAAAGACAAACCATAAA
>NM_001204889.2 ALDH2 [organism=Homo sapiens] [GeneID=217]
[transcript=2]
GAGACCCTAGCTCTGCTCTCGGTCCGCTCGCTGTCCGCTAGCCCGCTGCGATGTTGCGCGCTGCCGCCCG
CTTCGGGCCCCGCCTGGGCCGCCGCCTCTTGTCAGCCGCCGCCACCCAGGCCGTGCCTGCCCCCAACCAG
CAGCCCGAGGTCTTCTGCAACCAGATTTTCATAAACAATGAATGGCACGATGCCGTCAGCAGGAAAACAT
TCCCCACCGTCAATCCGTCCACTGGAGAGGTCATCTGTCAGGTAGCTGAAGGGGACAAGGCCTTGGAGAC
CCTGGACAATGGCAAGCCCTATGTCATCTCCTACCTGGTGGATTTGGACATGGTCCTCAAATGTCTCCGG
TATTATGCCGGCTGGGCTGATAAGTACCACGGGAAAACCATCCCCATTGACGGAGACTTCTTCAGCTACA
CACGCCATGAACCTGTGGGGGTGTGCGGGCAGATCATTCCGTGGAATTTCCCGCTCCTGATGCAAGCATG
·····
CATAGCCCAGGGATGCGACCCAAAATTATCAAAGTAGAGACATTATGTGTTTGCAATGTGCTATGATTAA
TGCAGTGACATAACTGGCAAGATTTACAGGTCAATTGGATATTGTTTACCTGGAGCTGGTCCTTCTTAGT
TGCCAAAAAAGACATAAGGATTAAAAAGACAAACCATAAA
```

アイソフォームを指定しないと、2つのアイソフォームの配列が取得される

遺伝子配列の取得

- ヒトのprotein.faaの内容

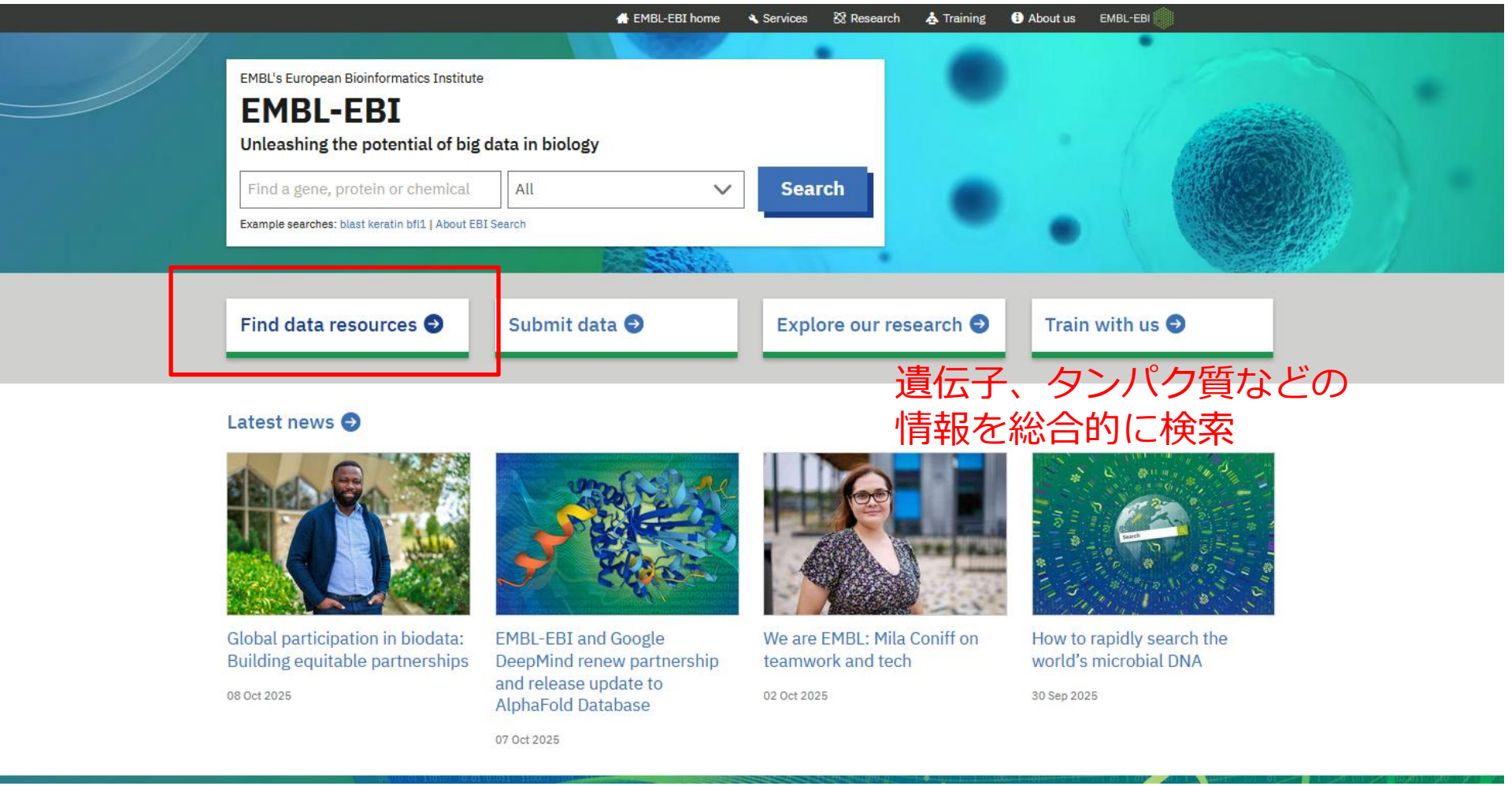
```
>NP_000681.2 ALDH2 [organism=Homo sapiens] [GeneID=217] [isoform=1
precursor]
MLRAAARFGPRLGRRLLSAAATQAVPAPNQQPEVFCNQIFINNEWHDAVSRKTFPTVNPSTGEVICQVAE
GDKEDVDKAVKAARAAFQLGSPWRRMDASHRGRLLNRLADLIERDRTYLAALETLDNGKPYVISYLVDLD
MVLKCLRYYAGWADKYHGKTIPIDGDFFSYTRHEPVGVCGQIIPWNFPLLMQAWKLGPALATGNVVVMKV
AEQTPLTALYVANLIKEAGFPPGVVNIVPGFGPTAGAAIASHEDVDKVAFTGSTEIGRVIQVAAGSSNLK
RVTLELGGKSPNIIMSDADMDWAVEQAHFALFFNQGQCCCAGSRTFVQEDIYDEFVERSVARAKSRVVGN
PFDSKTEQGPQVDETQFKKILGYINTGKQEGAKLLCGGGIAADRGYFIQPTVFGDVQDGMTIAKEEIFGP
VMQILKFKTIEEVVGRANNSTYGLAAAVFTKDLDKANYLSQALQAGTVWVNCYDVFGAQSPFGGYKMSGS
GRELGEYGLQAYTEVKTVTVKVPQKNS
```

```
>NP_001191818.1 ALDH2 [organism=Homo sapiens] [GeneID=217] [isoform=2
precursor]
MLRAAARFGPRLGRRLLSAAATQAVPAPNQQPEVFCNQIFINNEWHDAVSRKTFPTVNPSTGEVICQVAE
GDKALETLDNGKPYVISYLVDLDMVLKCLRYYAGWADKYHGKTIPIDGDFFSYTRHEPVGVCGQIIPWNF
PLLMQAWKLGPALATGNVVVMKVAEQTPLTALYVANLIKEAGFPPGVVNIVPGFGPTAGAAIASHEDVDK
VAFTGSTEIGRVIQVAAGSSNLKRVTLELGGKSPNIIMSDADMDWAVEQAHFALFFNQGQCCCAGSRTFV
QEDIYDEFVERSVARAKSRVVGNPFDSKTEQGPQVDETQFKKILGYINTGKQEGAKLLCGGGIAADRGYF
IQPTVFGDVQDGMTIAKEEIFGPVMQILKFKTIEEVVGRANNSTYGLAAAVFTKDLDKANYLSQALQAGT
VWVNCYDVFGAQSPFGGYKMSGSGRELGEYGLQAYTEVKTVTVKVPQKNS
```

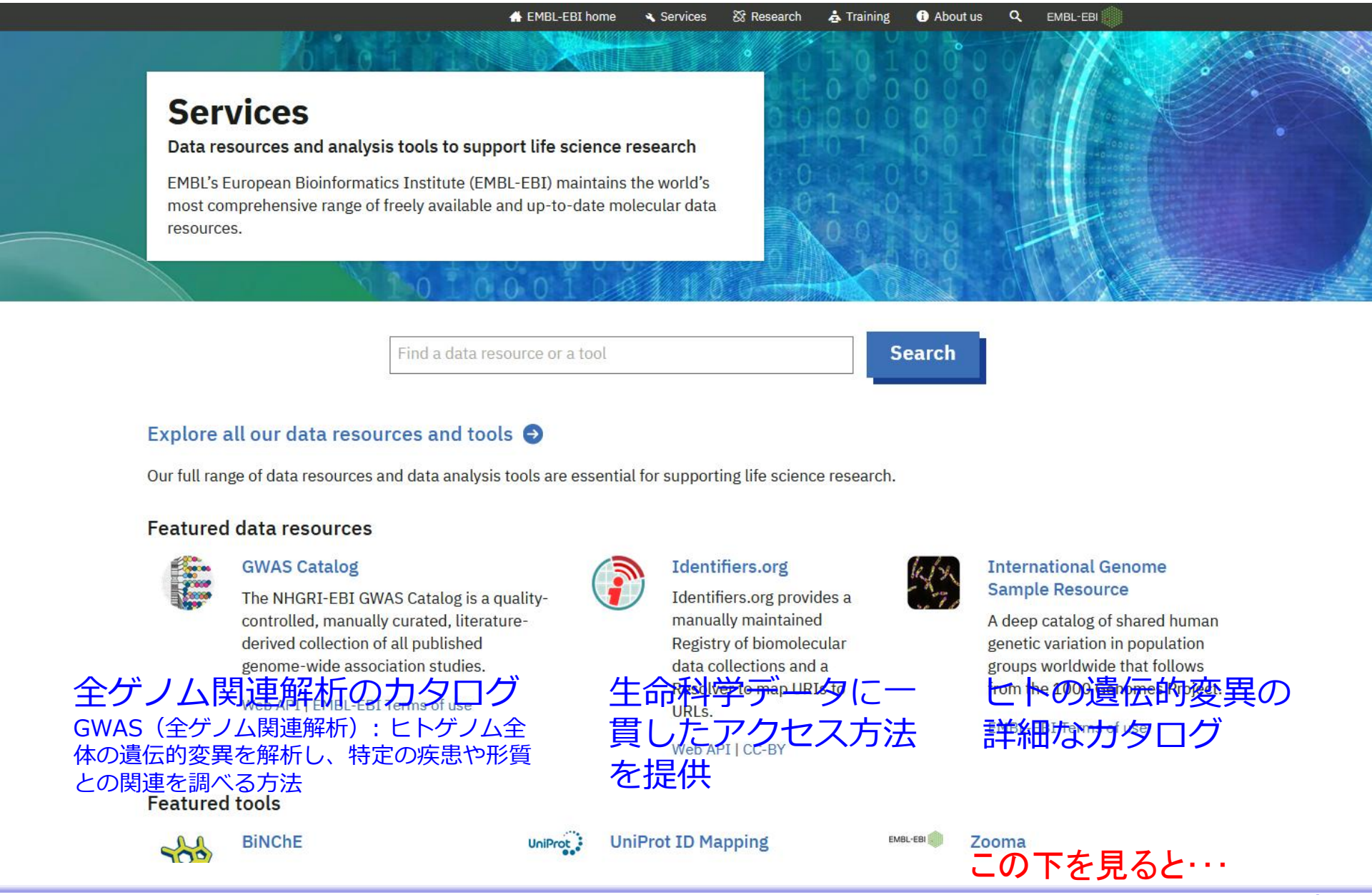
アイソフォームを指定しないと、2つのアイソフォームの配列が取得される

EMBL-EBI

- European Bioinformatics Institute（欧州バイオインフォマティクス研究所）
- 欧州分子生物学研究所（EMBL）の下部組織、生命科学の研究所で、バイオインフォマティクスのデータおよびツールのリソースを提供
- https://www.ebi.ac.uk/



EMBL-EBI



全ゲノム関連解析のカタログ

GWAS（全ゲノム関連解析）：ヒトゲノム全体の遺伝的変異を解析し、特定の疾患や形質との関連を調べる方法

生命科学データに一貫したアクセス方法を提供

ヒトの遺伝的変異の詳細なカタログ

この下を見ると･･･

EMBL-EBI

Data submission | Sequence analysis | Support and feedback | Service teams

配列解析

EMBL-EBI's Data Stewardship Principles

Open

EMBL's Open Science policy makes databases, code and software freely available to scientists whenever possible. Our Terms of Use impose no additional restrictions on data use, beyond those imposed by the data owner. Our Licensing policy aims to implement machine-readable cc0 licenses for data resources wherever possible. Our teams work open-source and many of our software systems can be downloaded and installed locally.

Long-Term Resilience

Delivery of our mission, to help scientists realise the potential of big data in biology, depends on the long-term resilience of our data resources. This is ensured through resource life cycle management, staffing and infrastructure continuity measures, data backup and delivery through international partners. Our approach to data preservation and service continuity.

Expert Added Value

Our communities of expert biologists enhance the data we steward via curation and annotation. This includes integration with data from other sources such as the scientific literature. Increasingly, curators use emerging technologies such as text mining, machine learning, large-language models and other AI tools to automate and scale, subject to rigorous evaluation and quality control. Our work integrating the latest advances in AI

Covid-19のデータポータル

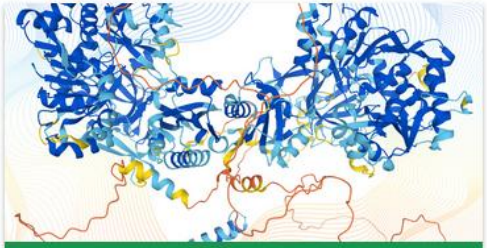
タンパク質モデル構造データベース

ヒトの細胞の遺伝学に基づく参照マップ

Highlight projects

COVID-19 Data Portal

The European COVID-19 Data Platform facilitates data sharing and analysis in order to accelerate coronavirus research.



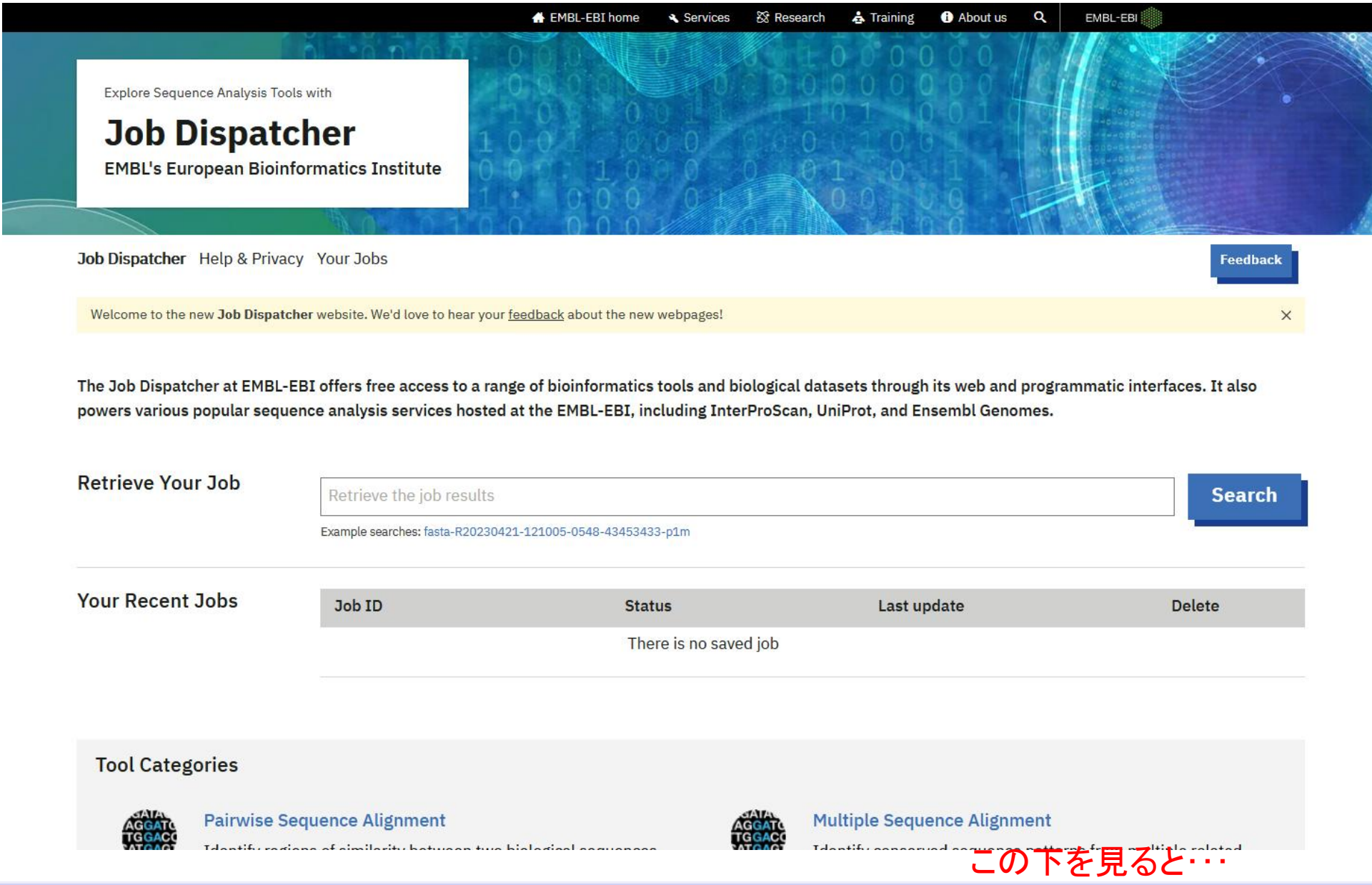
AlphaFold DB

Enjoy open access to over 990,000 protein structure predictions for the human proteome and other key proteins.

Human Cell Atlas

Aiming to map every cell type in the human body, the Human Cell Atlas will transform our understanding of biology and disease.

配列の比較



この下を見ると・・・

配列の比較

Tool Categories

Pairwise Sequence Alignment
Identify regions of similarity between two biological sequences.
Needle | Stretcher | GGEARCH2SEQ | Water | Matcher | More...

「Pairwise Sequence Alignment」を選ぶ

Multiple Sequence Alignment
Identify conserved sequence patterns from multiple related sequences.
Clustal Omega | Cons | Kalign | MAFFT | MUSCLE | T-Coffee | More...

Sequence Similarity Search
Find sequences in databases based on similarity.
NCBI BLAST | PSI-BLAST | FASTA | SSEARCH | PSI-Search | More...

Sequence Translation
Emboss sequence translation and back translation tools.
Transeq | Sixpack | Backtranseq | Backtranambig

Sequence Statistics
Set of tools for performing sequence statistics.
Pepinfo | Pepstats | Pepwindow | SAPS | Cpgplot | More...

Phylogeny
Phylogenetic tree generation using the ClustalW2 program.
Simple Phylogeny

Protein Functional Analysis
Assign biological or biochemical roles to proteins.
InterProScan | PfamScan | HMMER3 hmmscan | More...

Sequence Operations
Generate checksums for protein and nucleotide sequences.
SeqCksum

RNA Analysis
Set of tools for performing various RNA analysis.
Infernal cmscan | R2DT

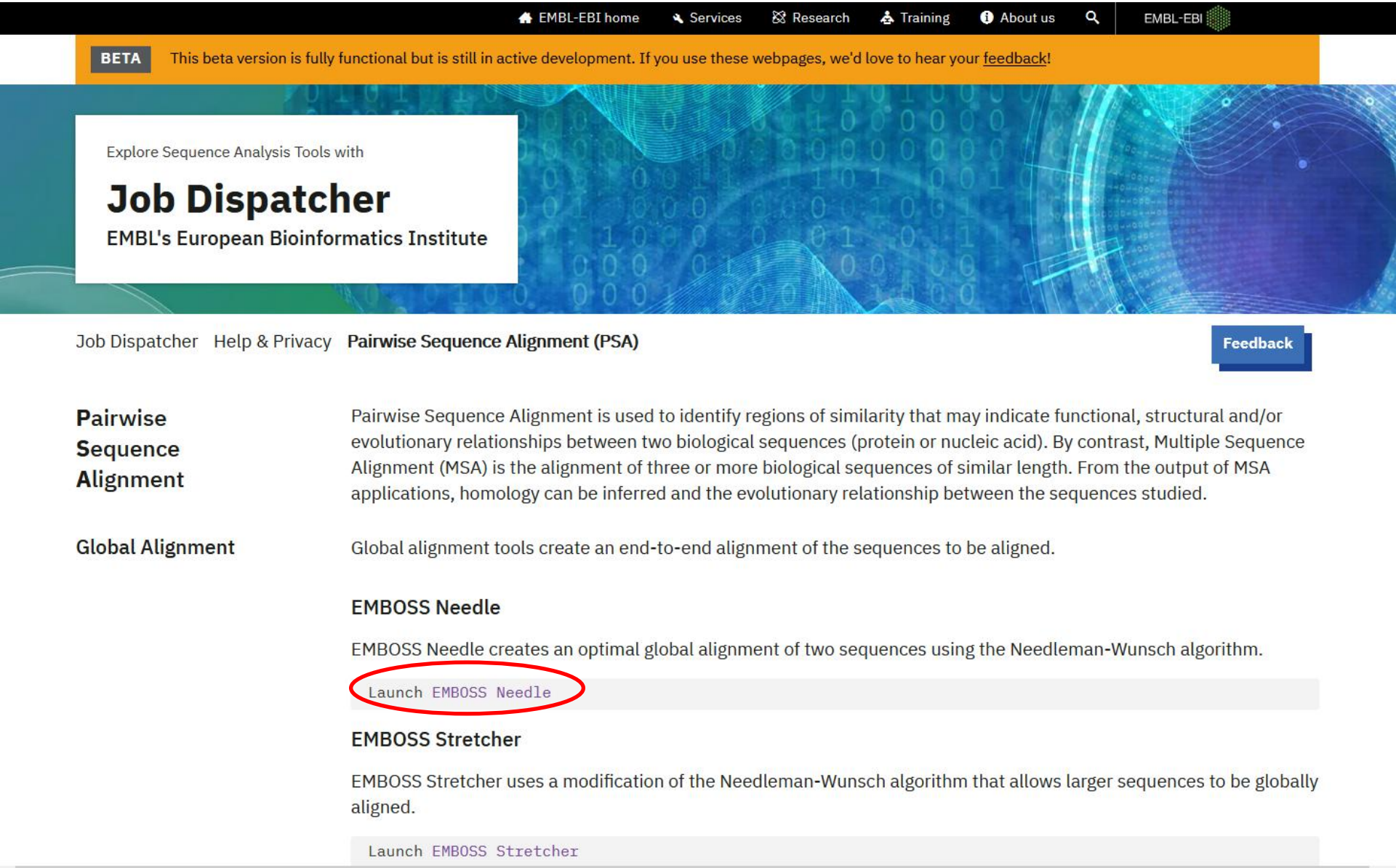
Sequence Format Conversion
Read different biological sequence formats and convert them to other formats.
Seqret | MView

Dbfetch
Single point of access to retrieve entries from various EMBL-EBI databases.

EMBOSS Tools
Set of selected EMBOSS tools for sequence analysis.
Cons | Needle | Stretcher | More...

配列の比較

手順としては3つ先のスライドに飛びます



配列を比較する方法

遺伝子やタンパク質の配列を並べて比較する
アラインメントという方法が使われる

ACGAAGCTCTA

ACCAGAGTCA

ギャップ

ACGA-AGCTCTA

| | | | | | | | | | | |

ACCAGAG-TC-A

配列を、類似した部分が並ぶよう、配列要素（文字）を対応づける操作

長さをそろえて、
文字と文字の「最適な」
対応関係を調べる

必要に応じてギャップ文字「-」を入れる

- 対応する文字がないということ
- 進化の過程で文字が挿入されたり失われたりすることがあるため

グローバルアラインメントとローカルアラインメント

- **グローバルアラインメント**: 配列全体にわたり類似性を考慮してアラインメント
  - 全体的に類似した配列の異なる部位を調べる
    - 同じタンパク質の生物種間の違いを調べる
    - 進化の解析を行う
  - 最適なペアワイズアラインメントのアルゴリズムとして**Needleman-Wunsch**のアルゴリズムがある
- **ローカルアラインメント**: 局所的な類似部分をアラインメント
  - 機能に関わる配列パターンを考慮したアラインメント
  - 長さが大きく異なる配列、配列類似性が低い配列の比較に用いられる
  - 最適なペアワイズアラインメントのアルゴリズムとして**Smith-Waterman**のアルゴリズムがある

アラインメントの例

グローバルアラインメント（ヒトのヘモグロビンとLupinレグヘモグロビン）

```
HBA_HUMAN     1 -MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSF----PTTKTYFP     45
                ..|:.:....||::|.:..|:..::    ..|.|:..    |..|..|.
LGB2_LUPLU    1 MGALTESQAALVKSSWEEFNANIPKH----THRFFILVLEIAPAAKDLFS     46

HBA_HUMAN    46 HF----DLSHGSAQVKGHGKKVADALTNA-----VAHVDDMPNALSALSD     86
                ..    ::...:.:::.|..||...:..|     |..|.....|..|..
LGB2_LUPLU   47 FLKGTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGS     96

HBA_HUMAN    87 LHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTV    136
                :|..| .|...:|.:::...:|.|:...:.|:::.:::::.....::.|
LGB2_LUPLU   97 VHVSK-GVADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIV    145

HBA_HUMAN   137 LTSKYR---    142
                :..:..
LGB2_LUPLU  146 IKKEMNDAA    154
```

```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 159
# Identity:      26/159 (16.4%)
# Similarity:    60/159 (37.7%)
# Gaps:          22/159 (13.8%)
# Score: 40.5
```

ローカルアラインメント（ヒトのヘモグロビンとLupinレグヘモグロビン）

```
HBA_HUMAN     3 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSF----PTTKTYFPHF-     47
                |:.:....||::|.:..|:..::    ..|.|:..    |..|..|...
LGB2_LUPLU    4 LTESQAALVKSSWEEFNANIPKH----THRFFILVLEIAPAAKDLFSFLK     49

HBA_HUMAN    48 ---DLSHGSAQVKGHGKKVADALTNA-----VAHVDDMPNALSALSDLHA     89
                   ::...:.:::.|..||...:..|     |..|.....|..|..:|.
LGB2_LUPLU   50 GTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHV     99

HBA_HUMAN    90 HKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHAS    125
                .| .|...:|.:::...:|.|:...:.|:::.:::::
LGB2_LUPLU  100 SK-GVADAHFPVVKEAILKTIKEVVGAKWSEELNSA    134
```

```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 136
# Identity:      25/136 (18.4%)
# Similarity:    55/136 (40.4%)
# Gaps:          18/136 (13.2%)
# Score: 48.5
```

EMBOSS（The European Molecular Biology Open Software Suite）
Pairwise Sequence Alignment

配列の比較


EMBL-EBI home
Services
Research
Training
About us
EMBL-EBI
BETA
This beta version is fully functional but is still in active development. If you use these webpages, we'd love to hear your feedback!
EMBOSS Needle
Pairwise Sequence Alignment (PSA)
Job Dispatcher
Help & Privacy
Input form
Feedback
ヒトとマウスのmRNAの配列の比較
EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.
Input Sequence
PROTEIN
DNA
Paste in your sequence or use the example
参照...
ファイルが選択されていません。
aldh2-human-rna.fna
Paste in your sequence or use the example
参照...
ファイルが選択されていません。
aldh2-mouse-rna.fna
Use the example
Clear Sequence
Find more example inputs
Parameters
OUTPUT FORMAT
pair
More Options

配列の比較

Input sequence

Sequence type

○ Protein ● DNA

Paste your first sequence here - or use the example sequence

```
>NM_001204889.2 ALDH2 [organism=Homo sapiens] [GeneID=217] [transcript=2]
GAGACCCTAGCTCTGCTCTCGGTCCGCTCGCTGTCCGCTAGCCCGCTGCGATGTTGCGCG
CTGCCGCCCGCTTCGGGCCCCGCCTGGGCCGCCGCCTCTTGTCAGCCGCCGCCACCCAGG
CCGTGCCTGCCCCCAACCAGCAGCCCGAGGTCTTCTGCAACCAGATTTTCATAAACAATG
AATGGCACGATGCCGTCAGCAGGAAAACATTCCCCACCGTCAATCCGTCCACTGGAGAGG
TCATCTGTCAGGTAGCTGAAGGGGACAAGGCCTTGGAGACCCTGGACAATGGCAAGCCCT
ATGTCATCTCCTACCTGGTGGATTTGGACATGGTCCTCAAATGTCTCCGGTATTATGCCG
```

参照... ファイルが選択...ていません。

Paste your second sequence here - or use the example sequence

```
>NM_009656.4 Aldh2 [organism=Mus musculus] [GeneID=11669] [transcript=1]
ATTCTCTTCGCCGCCATATCTGCACAGATGTGAGCCTTAGGCGCCAGCCACCCTGCTAGG
AGCGCACACCACTCTGGCTAGGCTTTCTCAGGGTTCTGCAAACTCCATCTCTGACTTGGC
TTTGGGAGCCAGGGGTCGCGCCCCTTAGGCCGTGAGGGGCTGGGACTCCCTGACCACGCC
CCCGTGTCTCCGCCTCCCATTGGCGGCTGCAGGGGGCGGAGGCGAGGACTTGTTCTTCAA
CGCTGCAGTCGCCCTCCGATCGGCAAGGCTTCTCTCGGCTCCGTTCGGCTCGGCTCGCCC
ATTTCAGTTCAGTTCGGGTCAGTTAAGCTCCGCTCAGTTCAGCATGCTGCGCGCCGCACT
```

参照... ファイルが選択...ていません。

Use the example | Clear sequence | More example inputs

Parameters

OUTPUT FORMAT

pair

More options

Submit

Title

EMBOSS Needle's job

Submit

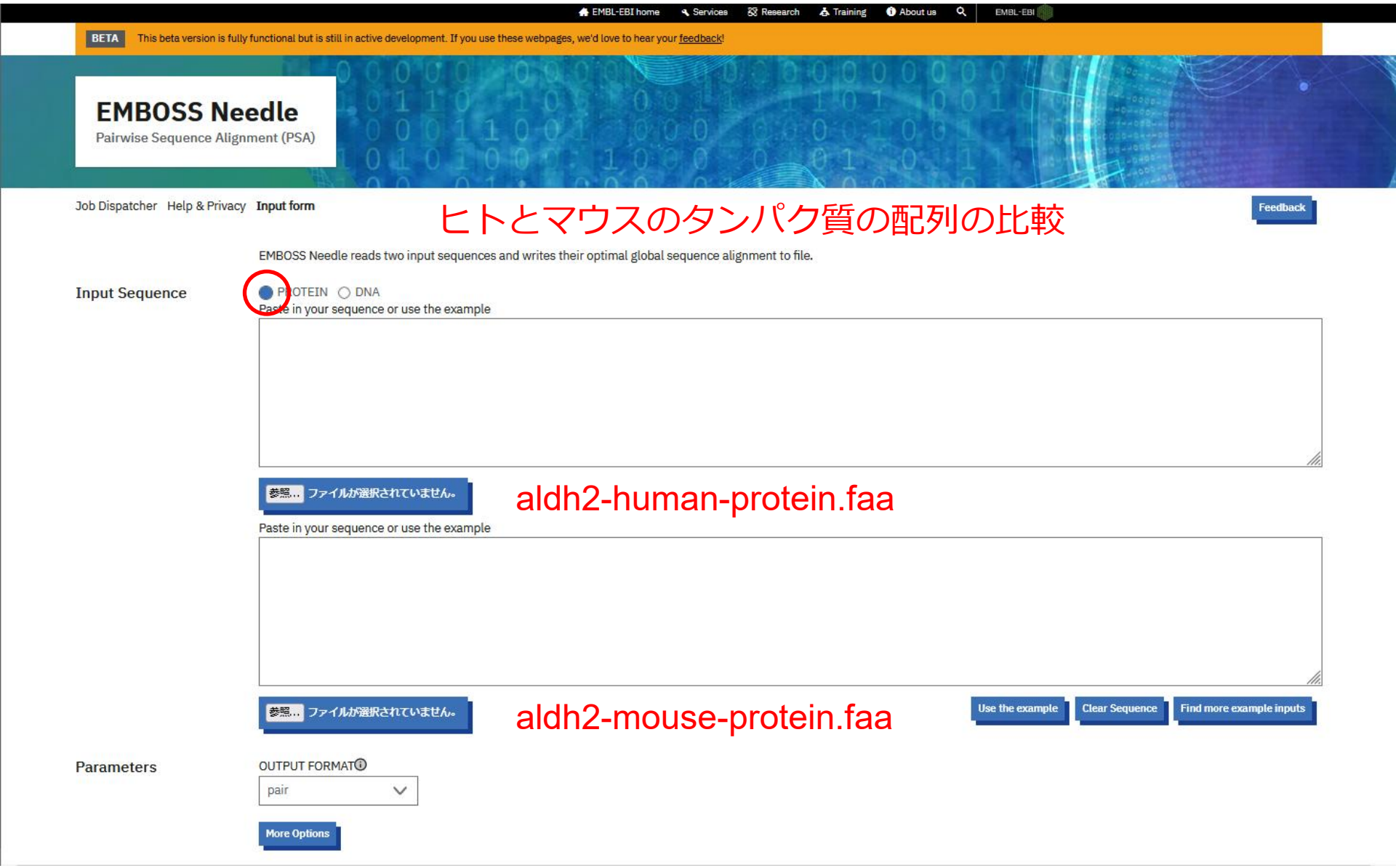
配列の比較

Results for job: emboss_needle-I20231009-150124-0957-45762922-p1m

Tool Output | Result Files | Submission Details

```
########################################
# Program: needle
# Rundate: Mon  9 Oct 2023 15:01:33
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20231009-150124-0957-45762922-p1m.asequence
#    -bsequence emboss_needle-I20231009-150124-0957-45762922-p1m.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: NM_000690.4
# 2: NM_009656.4
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 10177
# Identity:    2739/10177 (26.9%)
# Similarity:  2739/10177 (26.9%)
# Gaps:        6910/10177 (67.9%)
# Score: 8004.5
#
#
#=======================================

NM_000690.4       1 ----------------------------------------------------      0

NM_009656.4       1 ATTCTCTTCGCCGCCATATCTGCACAGATGTGAGCCTTAGGCGCCAGCCA     50

NM_000690.4       1 ----------------------------------------------------      0

NM_009656.4      51 CCCTGCTAGGAGCGCACACCACTCTGGCTAGGCTTTCTCAGGGTTCTGCA    100

NM_000690.4       1 ----------------------------------------------------      0

NM_009656.4     101 AACTCCATCTCTGACTTGGCTTTGGGAGCCAGGGGTCGCGCCCCTTAGGC    150

NM_000690.4       1 ----------------------------------------------------      0

NM_009656.4     151 CGTGAGGGGCTGGGACTCCCTGACCACGCCCCCGTGTCTCCGCCTCCCAT    200

NM_000690.4       1 ----------------------------------------------------      0

NM_009656.4     201 TGGCGGCTGCAGGGGGCGGAGGCGAGGACTTGTTCTTCAACGCTGCAGTC    250

NM_000690.4       1 GAGACCCT----------AG------CTCTGCTC---TCGG-TCCGCTCG     30
                    |  ||||         ||      |||.||||   |||| ||.|||||
NM_009656.4     251 G---CCCTCCGATCGGCAAGGCTTCTCTCGGCTCCGTTCGGCTCGGCTCG    297

NM_000690.4       31 C---TGTCCGCT----------------AGC-CCGCT-------GCGATG     53
                     |  |.||.|.|               ||| |||||       || |||
```

配列の比較


EMBL-EBI home
Services
Research
Training
About us
EMBL-EBI
BETA
This beta version is fully functional but is still in active development. If you use these webpages, we'd love to hear your feedback!
EMBOSS Needle
Pairwise Sequence Alignment (PSA)
Job Dispatcher
Help & Privacy
Input form
ヒトとマウスのタンパク質の配列の比較
Feedback
EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.
Input Sequence
PROTEIN
DNA
Paste in your sequence or use the example
参照...
ファイルが選択されていません。
aldh2-human-protein.faa
Paste in your sequence or use the example
参照...
ファイルが選択されていません。
aldh2-mouse-protein.faa
Use the example
Clear Sequence
Find more example inputs
Parameters
OUTPUT FORMAT
pair
More Options

配列の比較

Results for job: emboss_needle-I20231009-150412-0014-63955803-p1m

Tool Output | Result Files | Submission Details

```
########################################
# Program: needle
# Rundate: Mon  9 Oct 2023 15:04:17
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20231009-150412-0014-63955803-p1m.asequence
#    -bsequence emboss_needle-I20231009-150412-0014-63955803-p1m.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: NP_000681.2
# 2: NP_033786.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 520
# Identity:     493/520 (94.8%)
# Similarity:   504/520 (96.9%)
# Gaps:           4/520 ( 0.8%)
# Score: 2564.0
#
#
#=======================================

NP_000681.2         1 MLRAA---ARFGPRLGRRLLSAAATQAVPAPNQQPEVFCNQIFINNEWHD     47
                      |||||   .|.||||. |||||||.|||||.|||||||||||||||||
NP_033786.1         1 MLRAALTTVRRGPRLS-RLLSAAATSAVPAPNHQPEVFCNQIFINNEWHD     49

NP_000681.2        48 AVSRKTFPTVNPSTGEVICQVAEGDKEDVDKAVKAARAAFQLGSPWRRMD     97
                      ||||||||||||||||||||||:||||||||||||||||||||||||||
NP_033786.1        50 AVSRKTFPTVNPSTGEVICQVAEGNKEDVDKAVKAARAAFQLGSPWRRMD     99

NP_000681.2        98 ASHRGRLLNRLADLIERDRTYLAALETLDNGKPYVISYLVDLDMVLKCLR    147
                      ||.|||||.|||||||||||||||||||||||||||||||||||||||||
NP_033786.1       100 ASDRGRLLYRLADLIERDRTYLAALETLDNGKPYVISYLVDLDMVLKCLR    149

NP_000681.2       148 YYAGWADKYHGKTIPIDGDFFSYTRHEPVGVCGQIIPWNFPLLMQAWKLG    197
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
NP_033786.1       150 YYAGWADKYHGKTIPIDGDFFSYTRHEPVGVCGQIIPWNFPLLMQAWKLG    199

NP_000681.2       198 PALATGNVVVMKVAEQTPLTALYVANLIKEAGFPPGVVNIVPGFGPTAGA    247
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
NP_033786.1       200 PALATGNVVVMKVAEQTPLTALYVANLIKEAGFPPGVVNIVPGFGPTAGA    249

NP_000681.2       248 AIASHEDVDKVAFTGSTEIGRVIQVAAGSSNLKRVTLELGGKSPNIIMSD    297
                      |||||||.||||||||||:|.:|||||||||||||||||||||||||||
NP_033786.1       250 AIASHEGVDKVAFTGSTEVGHLIQVAAGSSNLKRVTLELGGKSPNIIMSD    299

NP_000681.2       298 ADMDWAVEQAHFALFFNQGQCCCAGSRTFVQEDIYDEFVERSVARAKSRV    347
                      ||||||||||||||||||||||||||||||::|||||||||||||||||
NP_033786.1       300 ADMDWAVEQAHFALFFNQGQCCCAGSRTFVQENVYDEFVERSVARAKSRV    349
```

課題 1

ヒトとクジラ、ヒトとマグロのミオグロビンの配列について、アラインメントを比較してみよう。

ヒト　　myoglobin-human.fasta

クジラ　　myoglobin-whale.fasta

マグロ　　myoglobin-tuna.fasta

余裕があれば、ヒトのアザラシのミオグロビンについても比較してみよう

アザラシ　myoglobin-seal.fasta

課題 1の解説

ヒトとクジラ

```
########################################
# Program: needle
# Rundate: Fri 10 Oct 2025 09:16:10
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20251010-091604-0708-43449506-p1m.asequence
#    -bsequence emboss_needle-I20251010-091604-0708-43449506-p1m.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: MYG_HUMAN
# 2: A0A8C0D084_BALMU
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 154
# Identity:     132/154 (85.7%)
# Similarity:   143/154 (92.9%)
# Gaps:           0/154 ( 0.0%)
# Score: 710.0
#
#
#=======================================
```

課題 1の解説

ヒトとクジラのミオグロビン

```
MYG_HUMAN           1 MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHL     50
                      |.|||.|||||||:|.|||||:.||||::||||||||||||||||||||
A0A8C0D084_BA       1 MVLSDAEWQLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHL     50

MYG_HUMAN          51 KSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKI    100
                      |:|.||||||||||||.||||||||||||||||:||||||||||||||
A0A8C0D084_BA      51 KTEAEMKASEDLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKI    100

MYG_HUMAN         101 PVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKEL    150
                      |:|||||||:.||.|||.|:||||||||||.||||||||||||:|:.|||||
A0A8C0D084_BA     101 PIKYLEFISDAIIHVLHSRHPGDFGADAQAAMNKALELFRKDIAAKYKEL    150

MYG_HUMAN         151 GFQG    154
                      ||||
A0A8C0D084_BA     151 GFQG    154
```

タンパク質の表面

分子表面の**正味電荷が＋に寄る**
ミオグロビンを超高濃度に貯蔵しても凝集しにくいよう、表面の同電荷反発を強める？

課題 1の解説

ヒトとマグロのミオグロビン

```
MYG_HUMAN           1 MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHL     50
                        ..:...||..||.||||....|..||.||||.||||.:.|.||..:
MYG_THUTH           1 ----MADFDAVLKCWGPVEADYTTIGGLVLTRLFKEHPETQKLFPKFAGI     46

MYG_HUMAN          51 KSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKI    100
                      ...| :..:...:..||||||..||.:||.||.|.|.:||||.||||||||
MYG_THUTH          47 AQAD-IAGNAAVSAHGATVLKKLGELLKAKGSHAAILKPLANSHATKHKI     95

MYG_HUMAN         101 PVKYLEFISECIIQVLQSKHPGDFGADAQG--AMNKALELFRKDMASNYK    148
                      |:...:.|||.:::|:..|    .|.||.|  |:...:.:...|:.:|||
MYG_THUTH          96 PINNFKLISEVLVKVMHEK----AGLDAGGQTALRNVMGIIIADLEANYK    141

MYG_HUMAN         149 ELGFQG    154
                      ||||.|
MYG_THUTH         142 ELGFSG    147
```

課題 1の解説

ヒトとアザラシのミオグロビン

```
MYG_HUMAN           1 MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHL     50
                      |||||||.|||||||||.|:.||||||||||||.||||||||||||||
MYG_PHOVI           1 MGLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKFKHL     50

MYG_HUMAN          51 KSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKI    100
                      ||||:|:.||||:|||.||||||||||||||||||:||||||||||||||
MYG_PHOVI          51 KSEDDMRRSEDLRKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKI    100

MYG_HUMAN         101 PVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKEL    150
                      |:||||||||.||.||.||||.:||||||.||.|||||||.||:|:.||||
MYG_PHOVI         101 PIKYLEFISEAIIHVLHSKHPAEFGADAQAAMKKALELFRNDIAAKYKEL    150

MYG_HUMAN         151 GFQG    154
                      ||.|
MYG_PHOVI         151 GFHG    154
```