グローバルCOEプログラム系統講義「ベーシックサイエンスコース」

バイオインフォマティクス：
生物系研究者における情報リテラシーとしての配列解析

Bioinformatics: Sequence analyses as information literacy of life scientists

嶋田　誠
（藤田保健衛生大学 遺伝子発現機構学）

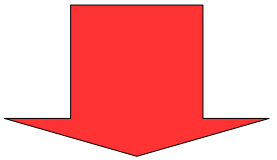
1日目：2010年 9月28日（火） 17:00−18:30　基礎医学研究棟1階会議室
2日目：2010年10月 4日（月） 18:00−19:30　基礎医学研究棟1階会議室

アウトライン

- [第1回]
  - 概論
    - ねらい
    - バイオインフォマティクス：どこまで把握すべきか？
  - データベース
  - ゲノム関連大規模プロジェクト
- [第2回]
  - バイオインフォマティクス独学のコツ
  - ツール群の分類と見つけ方
  - 実験研究者にとってのプログラミング
    - アルゴリズム
    - データ加工変換
    - パース
  - 配列研究関連ツールの紹介

ねらい

- 実験系の研究者が、
- 今後、それぞれの研究活動において、
- 適宜、必要なデータやツールを探し出して、使えるようになること。



基本的に独学:

- 調べ方
- 他人への尋ね方
  - -->基本的用語の解説
- 最初の段階で重要な事は何か
  - データ構造の解釈＆データ入出力の仕方
  - -->将来 in silicoで解析しやすい実験データのまとめ方

バイオインフォマティクス：
どこまで把握しておくべきか？

実験室での作業だけでも時間管理が大変なのに。

- バイオインフォマティクス用語
  - やたらと略語が多い？
- データベースやツール群
  - どこになにがあるのか。
- 計算手法・アルゴリズム
  - ソフト任せのブラックボックス状態でいいのか？

「どこまで」を考えるうえでのヒント1

- 開発者とユーザーとがいる
  - 開発者:
    - プログラムを組んでソフトやDBを作りだす。
    - アルゴリズムを考える。
    - バイオインフォマティクスそのものの未来を考える。
  - ユーザー:
    - 自分のデータを解析する。
    - 解析のアイデアを練る。-->開発者へ
- ユーザーは開発者にならなくてもいい。
  - 開発者（仲介者）とコミュニケーションをとる能力は養えるのでは。
    - ヘルプ、マニュアル文書、メール、打合せ等
    - ユーザーとして必要な用語や概念だけでも理解しておきたい。

「どこまで」を考えるうえでのヒント2

- 研究手法・解析手法における栄枯盛衰
  - DBは激しい。
  - 手法やアルゴリズムにおける原理は様々な分野で繰り返し利用されている。
    - 例）モンテカルロ、パーシモニー、ニューラルネットワーク
-
- DBは必要な時に必要なものを探し出す技術を。
- 手法やアルゴリズムは原理を（一度は）理解し、アレルギーを起こさないようにしておく。

「どこまで」を考えるうえでのヒント3

- 研究は常に新発見を求め、新しきことを試す行為なり。

ならば、

  - 既存のツールでは限界がある。
    - 別の処理と組み合わせたり、修正したり、が必要。
    - 開発といっても、全く無から新規を作り出す必要はない。
      - 「車輪の再発明はするな。」

- 「数こそ力」の課題もある。

ならば、

  - 繰り返し作業の得意なコンピュータを利用しよう。

「どこまで」を考えるうえでのヒント4

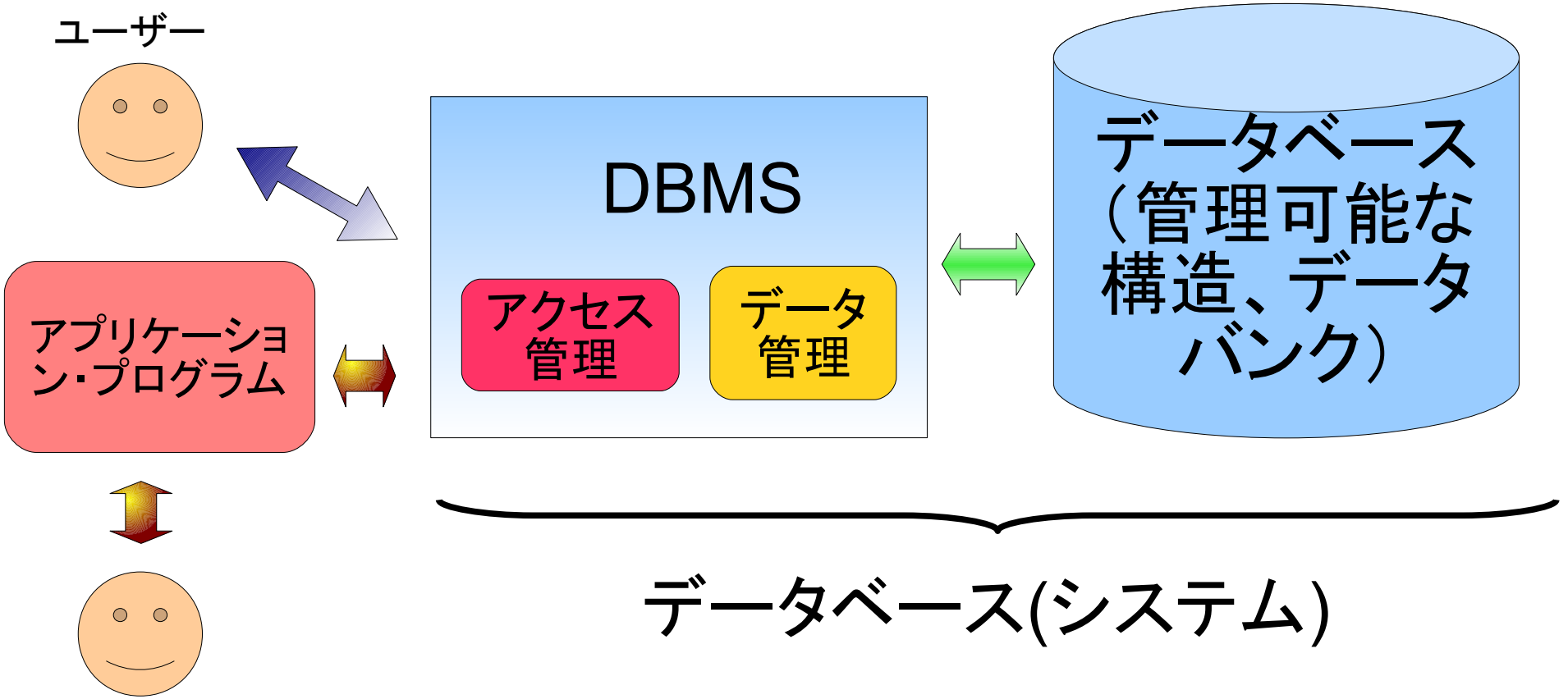
- Publicationは論文だけではない。
  - 論文投稿前に、データをデータベースに登録する必要がある。
    - 配列データ、発現データ、立体構造など
  - データベース管理者側も親切にマニュアルを用意している。
    - eg., 次世代シーケンサーによるデータの登録システムのマニュアル http://trace.ddbj.nig.ac.jp/files/manual/DRA_submission_manual.pdf
    - 技術革新の速さ、研究の多様性から、想定外のことも多い。
    - データの形式を一括変換するのはユーザー（登録者）側。

データベース

- データベース(DB)とは
- 基本的なDB・ブラウザの紹介
  - 文献DBとMeSH
  - 核酸DBとannotation(注釈づけ)
  - タンパク質DBと配列の機能推定
  - ゲノムDBとゲノム・ブラウザ
- 必要なDBを見つけるために

データベースとは？(問い)

- 「～をデータベース化して」vs.「～をリスト化して」
- データベース管理システム (DBMS)



さらにDBMS

- DBMSが無いと、
  - プログラムから直接、ファイルに対してデータの保存や読み込みを行う
  - 個々のファイルのデータ構造を予め知っていないとアクセスできない
  - 複数のユーザーやアプリケーションに対応できない
  - 安全性・安定性に限界
- DBMS：構造の設計から管理・運用を行う
- 現場では意外と聞かないことば
  - リレーショナル・データベース（RDBMS）が一般的

リレーショナル・データベース（RDB: Relational Data Base）の概要（1）

- データを表（テーブル）形式で表し、複数の表にリレーションシップ（関係付け）を結ぶことができる。
- 通常、元データの表はタブやコンマなどで区切ったテキストファイル形式。

HIT00004499413	-2	NHEJ1	rs10498064	A/T	T	fwd/T	Y	na	inconsistent
HIT0000155588	-1	KIAA0802	rs632423	C/G	C	fwd/B	Y	0.497+/-0.037	inconsistent
HIT0000414818	-1	C13orf18	rs2478044	C/G	G	fwd/T	Y	na	consistent
HIT00002846621	-1	IL16	rs4778639	G/T	T	fwd/B	Y	0.188+/-0.242	inconsistent

Column Name of Table S2

#1 HIT ID: H-InvDB Transcript ID
#2 Intron No: _th intron in the transcript
#3 SitePos: see Figure 1a
#4 GeneSybl: HUGO gene symbol
#5 rs ID: rs SNP ID in dbSNP
#6 rsAlleles: alleles in dbSNP
#7 Ancestral Allele: ancestral allele estimated by NCBI and shown in dbSNP
#8 Ori/Str(rs): orientation and strand of the SNP, see ftp://ftp.ncbi.nih.gov/snp/database/Illumina_top_bot_strand.note.txt
#9 number of locations SNP mapped: 'Y'='OneMap' or 'N'='MultiMap', see List of Abbreviation
#10 Het: heterozygosity with standard error shown in dbSNP
#11 AncAlvsGT-AG: whether estimated ancestral allele is consistent with GT-AG rule

record（行）

field, column（列）

Shimada MK *et al. BMC Evol Biol*, 10:122, 2010. Additinal files.

リレーショナル・データベース（RDB: Relational Data Base）の概要（2）

- キー(key)の重要性：
  - 主キー(Primary key)：レコードを一意に指定するフィールド（通常非冗長ID）
  - 外部キー(Foreign key):2つの表を結び付けるキー
    - 応用例（DB間を結び付ける）, http://biodb.jp/
- SQLと呼ばれるデータベース（問い合わせ）言語が標準化されている。
- 代表的実装例
  - 商用：Oracle Database, Microsoft SQL Server
  - オープンソース：MySQL, PostgreSQL, BerkeleyDB

データベースとは(答え)

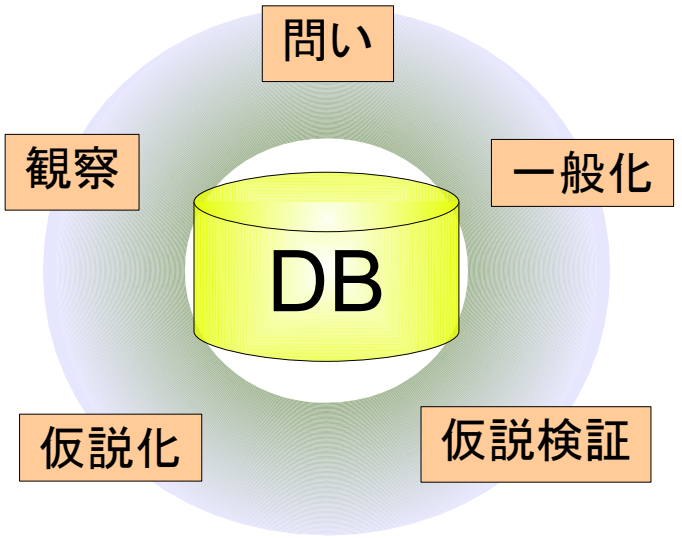
- 「様々な目的を考慮して整理整頓されたデータの集まり」である。
- 設計思想をもつ（DBMSにて体現）。
  - 生物情報DBでは普通、
    - 検索機能、
    - web利用が前提、
    - リンク付け、
    - viewerやtool群とセットになっている。

- 始まりは、第二次大戦後の米軍が点在する情報を集約し、一か所にアクセスするだけで、様々な情報が得られるようにした「情報基地」

DB(生物情報)の生命科学における意義

- データの種類の変化：計算機利用前提-->データ量の急速な増大
  - シークエンサーのデータ産出量 vs. CPU速度
  - 増加率 [2002(ABI3730)以降]： X2/8mo. vs. X1.5/8mo (Moors' Law)

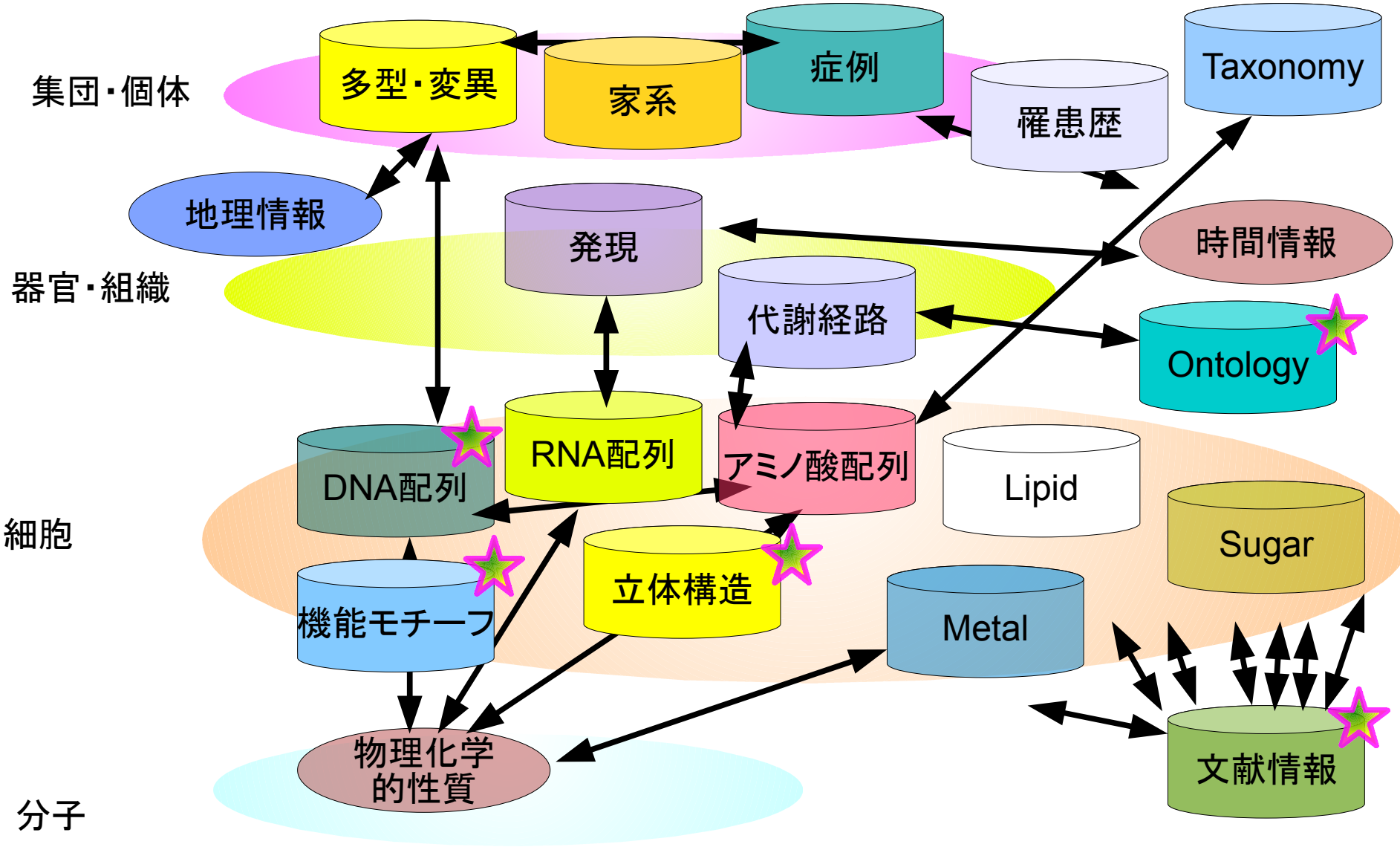
- 研究活動の様々なステップにおいて活用される余地あり。
  - 膨大な量のデータ[DBに蓄積、仮説構築、実験による検証、DBに蓄積]
  - 帰納的(データ->法則)vs.演繹的(法則->検証)(仮説driven vs. Data driven)
    - 情報リテラシーが研究活動の様々な局面で要求される



個々の研究者とDBとの関係

- 研究活動における公開原則と研究者の姿勢
  - 「配列データはDBに登録しなければ論文受理できない」
    - 研究成果は共通財産という考え-->意図的に働きかけた。
      - -->研究活動が活性化した。
    - DBによるデータの質の管理:
      - -->論文著者には厳しいときも。
  - Curatorの言うことは聞きましょう。
  - きまりは守るためにある。。。

生物情報データベース：対象と次元



DB検索に知っとくと得するキーワード

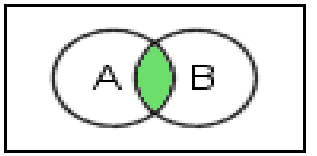
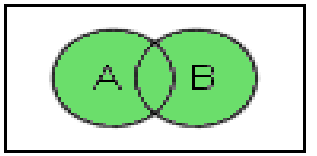
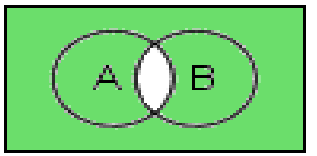
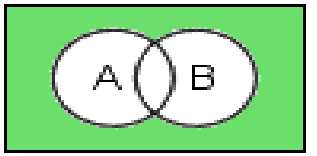
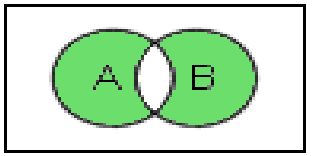
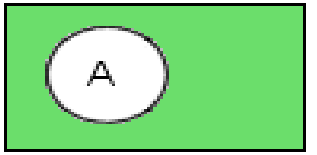
- 演算子(operator)、引数（オペランド operand）、

例）$a + 2$　の例では演算子と引数はそれぞれどれか

答え）演算子= “+”、引数= “a”, “2”

- ブール演算(Boolean Operation)
  - 論理演算(Logical Operation)ともいわれる。
  - 1（真）か0（偽）かの2通りの入力値に対して1つの値を出力する演算のこと。
  - プログラミングだけでなく、データベースを複合検索(eg., AND検索、NOT検索)する際にも用いられる。
    - データベースごとに約束事が違うので注意。

ブール演算(Boolean Operation)

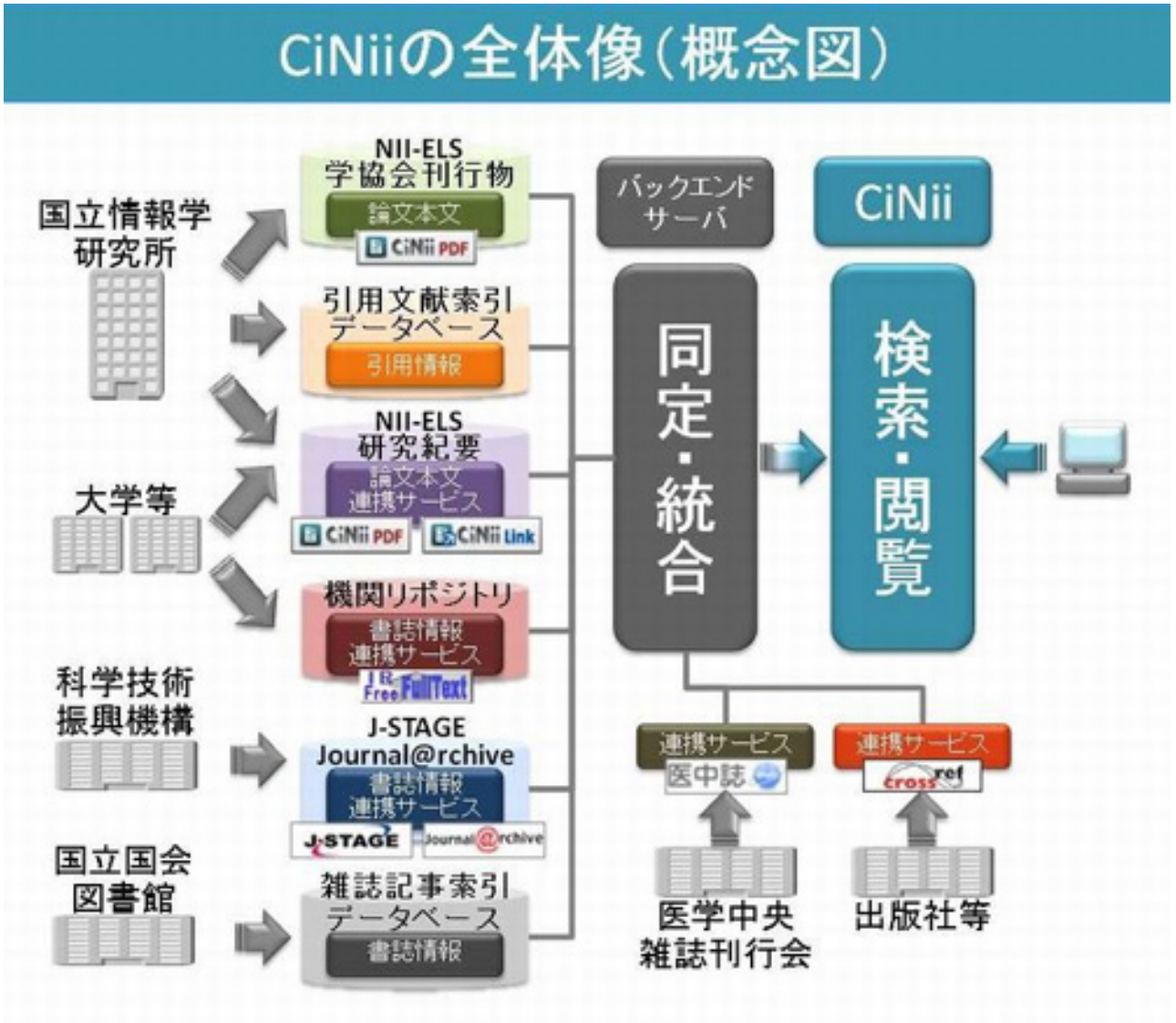
	式	真理値表（Truth Table）			ベン図（Venn Diagram）
		入力A	入力B	出力	
論理積（AND）	A and B A・B A×B A∩B A∧B	0	0	0	
		0	1	0	
		1	0	0	
		1	1	1	
論理和（OR）	A or B A＋B A∪B A∨B	0	0	0	
		0	1	1	
		1	0	1	
		1	1	1	
否定論理積（NAND）	A nand B $\overline{A・B}$ ¬（A∧B）	0	0	1	
		0	1	1	
		1	0	1	
		1	1	0	
否定論理和（NOR）	A nor B $\overline{A＋B}$ ¬（A∨B）	0	0	1	
		0	1	0	
		1	0	0	
		1	1	0	
排他的論理和（EOR、XOR）	A eor B A xor B $\overline{A}・B＋A・\overline{B}$ （A∨B）∧¬（A∧B）	0	0	0	
		0	1	1	
		1	0	1	
		1	1	0	
否定（NOT）	A not B $\overline{A}$ ¬A	0		1	
		1		0	

http://www.asahi-net.or.jp/~ax2s-kmtn/ref/logicope.htmlより改変

文献DBの例

- 生命科学関連DBほとんど生データは英語
  - 2バイト文字、機種依存文字をきらうから。
  - 医学文献の場合:国や文化によって症状、薬が違う。
  - 医中誌（医学中央雑誌刊行会 http://www.jamas.or.jp/）
  - J-STAGE (http://www.jstage.jst.go.jp/browse/-char/ja)
  - CiNii（NII論文情報ナビゲータ[サイニィ], http://ci.nii.ac.jp/)-->概念図参照
- PubMed(NCBI, http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed)
  - 英文綴り（イギリス式、オクスフォード式、アメリカ式）
  - MeSHについて-->詳細に
- OMIM (NCBI,http://www.ncbi.nlm.nih.gov/omim)
  - human curated

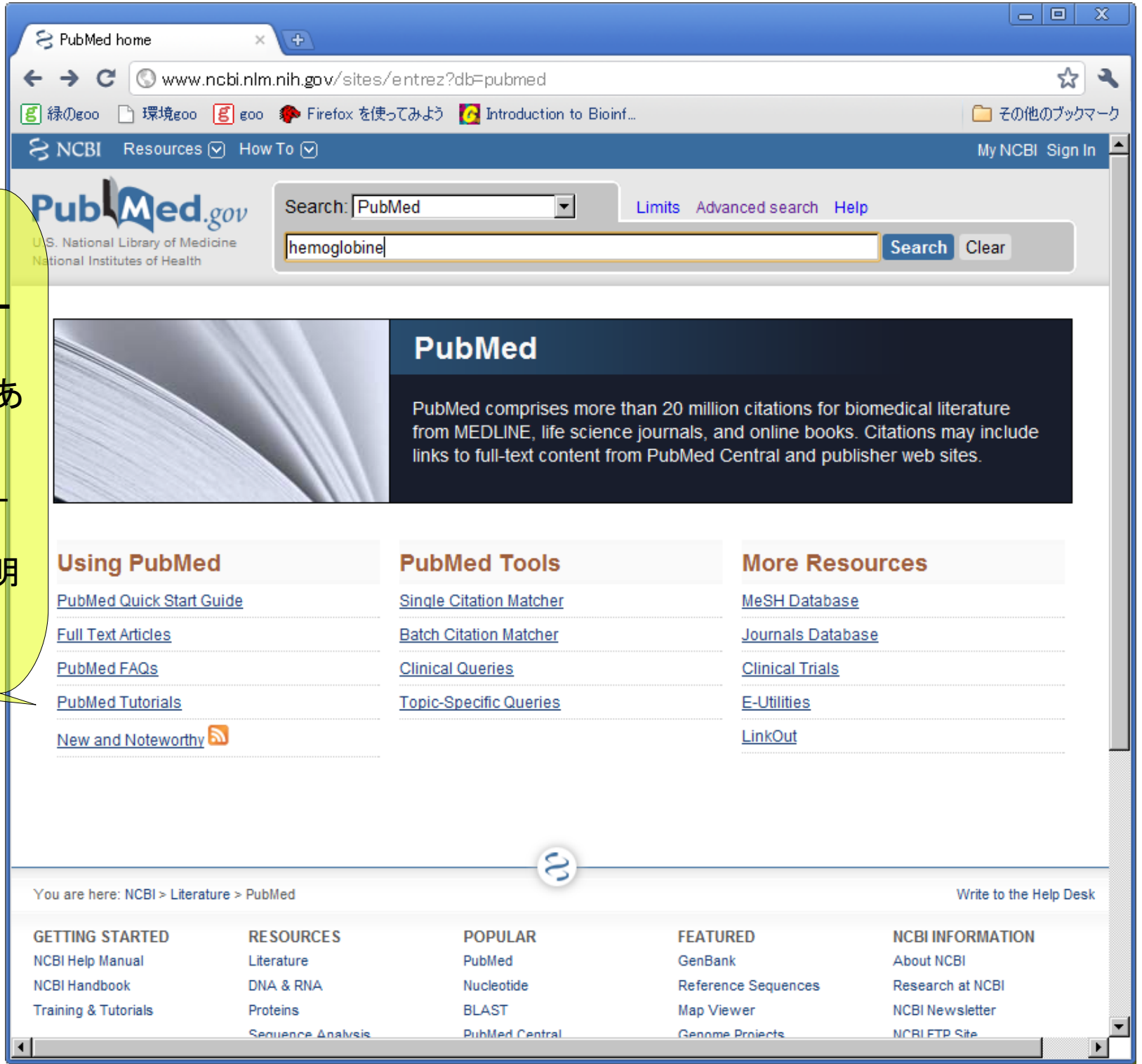
CiNiiの概念図


CiNiiの全体像（概念図）
国立情報学研究所
大学等
科学技術振興機構
国立国会図書館
NII-ELS
学協会刊行物
論文本文
CiNii PDF
引用文献索引データベース
引用情報
NII-ELS
研究紀要
論文本文連携サービス
CiNii PDF
CiNii Link
機関リポジトリ
書誌情報連携サービス
J-STAGE
Journal@rchive
書誌情報連携サービス
雑誌記事索引データベース
書誌情報
バックエンドサーバ
CiNii
同定・統合
検索・閲覧
連携サービス
医中誌
連携サービス
医学中央雑誌刊行会
出版社等

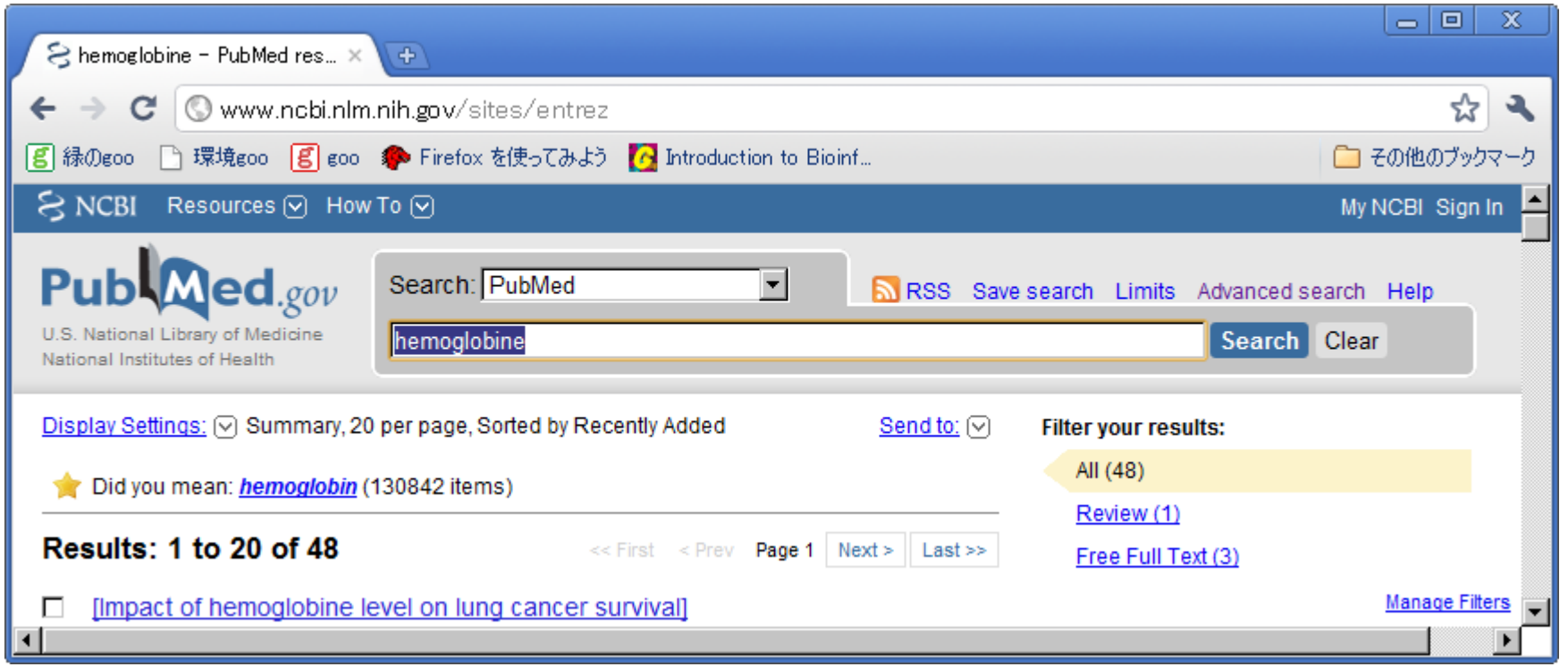
PubMed

教則ビデオ

- 1~5分程度に細かく区切ってある。
- きれいな英語
- お勧め機能：E-mail Alert
- MeSH本日説明



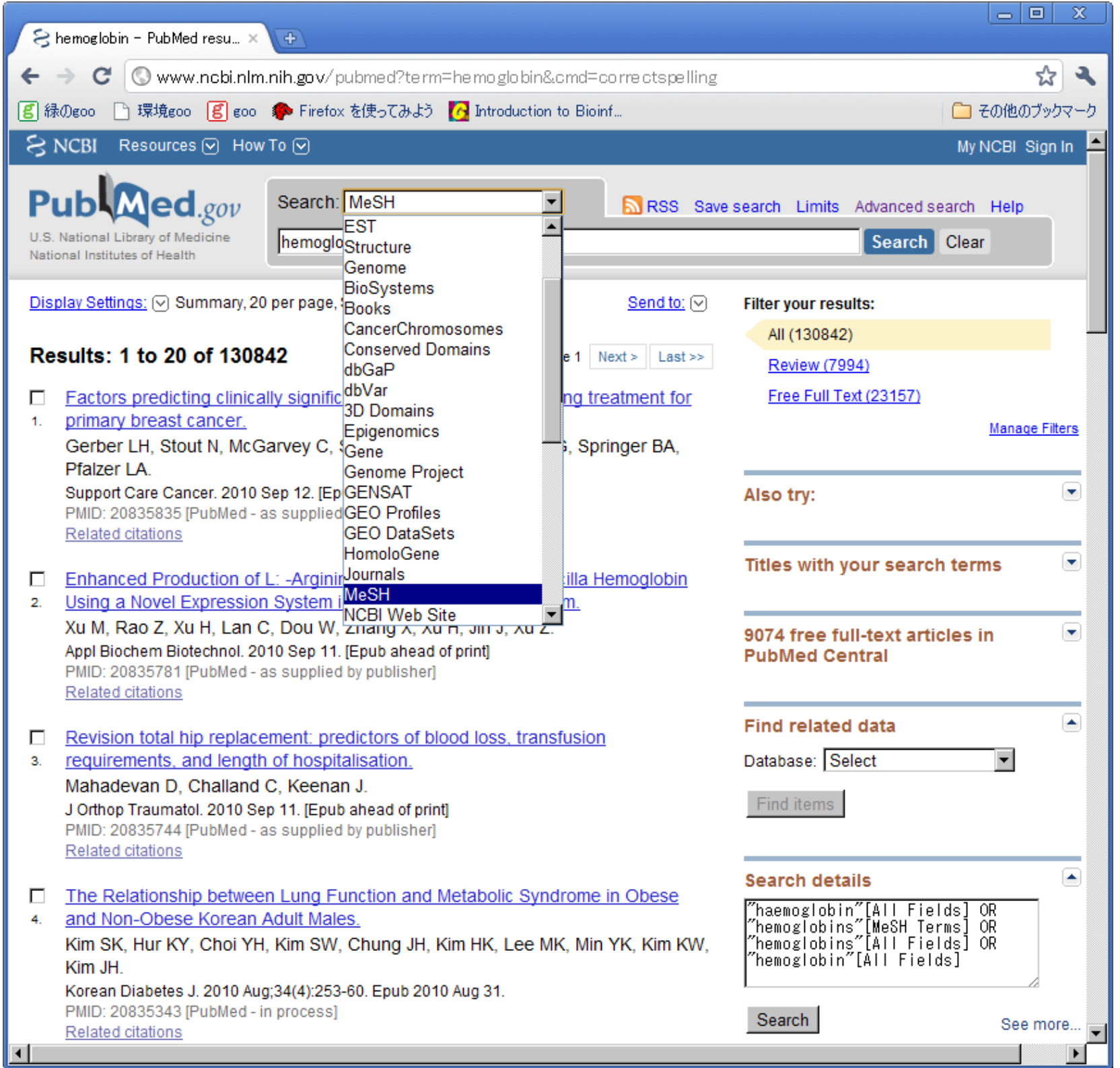
スペルミスの場合


hemoglobine - PubMed res...
www.ncbi.nlm.nih.gov/sites/entrez
緑のgoo
環境goo
goo
Firefox を使ってみよう
Introduction to Bioinf...
その他のブックマーク
NCBI
Resources
How To
My NCBI
Sign In
PubMed.gov
U.S. National Library of Medicine
National Institutes of Health
Search: PubMed
RSS
Save search
Limits
Advanced search
Help
hemoglobine
Search
Clear
Display Settings: Summary, 20 per page, Sorted by Recently Added
Send to:
Filter your results:
All (48)
Review (1)
Free Full Text (3)
Did you mean: hemoglobin (130842 items)
Results: 1 to 20 of 48
<< First
< Prev
Page 1
Next >
Last >>
[Impact of hemoglobine level on lung cancer survival]
Manage Filters

英文綴字の問題とMeSH

MeSH term: Medical Subject Headings, いろいろな表現がある単語を統一して使えるように関連付けした用語集

MeSH

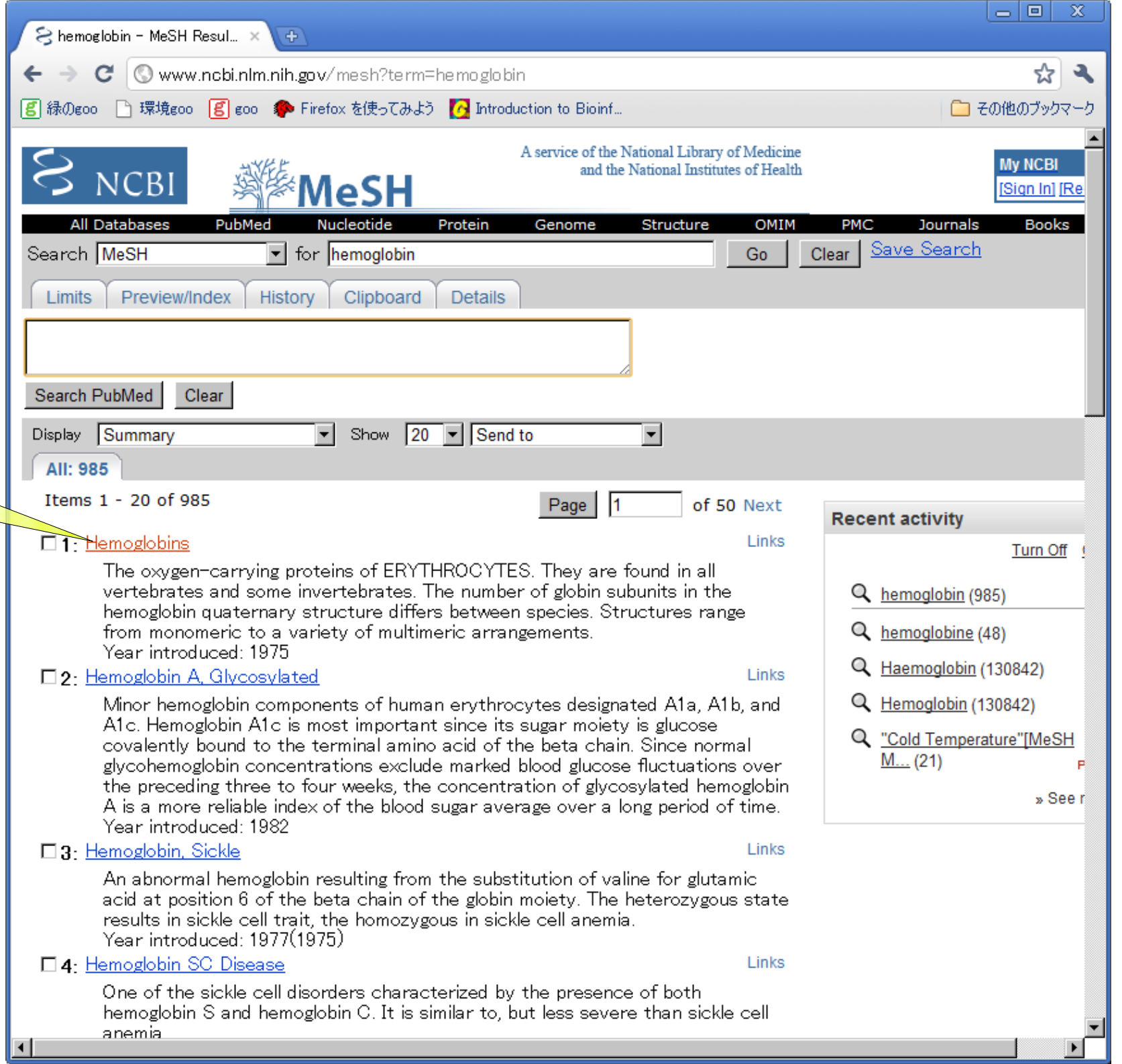

hemoglobin - PubMed resu...
www.ncbi.nlm.nih.gov/pubmed?term=hemoglobin&cmd=correctspelling
NCBI Resources How To
My NCBI Sign In
PubMed.gov
U.S. National Library of Medicine
National Institutes of Health
Search: MeSH
EST
Structure
Genome
BioSystems
Books
CancerChromosomes
Conserved Domains
dbGaP
dbVar
3D Domains
Epigenomics
Gene
Genome Project
GENSAT
GEO Profiles
GEO DataSets
HomoloGene
Journals
MeSH
NCBI Web Site
RSS Save search Limits Advanced search Help
Search Clear
Display Settings: Summary, 20 per page,
Send to:
Results: 1 to 20 of 130842
Next > Last >>
1. Factors predicting clinically significant ... treatment for primary breast cancer.
Gerber LH, Stout N, McGarvey C, ... Springer BA, Pfalzer LA.
Support Care Cancer. 2010 Sep 12. [Epub ahead of print]
PMID: 20835835 [PubMed - as supplied by publisher]
Related citations
2. Enhanced Production of L: -Arginine ... Hemoglobin Using a Novel Expression System ...
Xu M, Rao Z, Xu H, Lan C, Dou W, ...
Appl Biochem Biotechnol. 2010 Sep 11. [Epub ahead of print]
PMID: 20835781 [PubMed - as supplied by publisher]
Related citations
3. Revision total hip replacement: predictors of blood loss, transfusion requirements, and length of hospitalisation.
Mahadevan D, Challand C, Keenan J.
J Orthop Traumatol. 2010 Sep 11. [Epub ahead of print]
PMID: 20835744 [PubMed - as supplied by publisher]
Related citations
4. The Relationship between Lung Function and Metabolic Syndrome in Obese and Non-Obese Korean Adult Males.
Kim SK, Hur KY, Choi YH, Kim SW, Chung JH, Kim HK, Lee MK, Min YK, Kim KW, Kim JH.
Korean Diabetes J. 2010 Aug;34(4):253-60. Epub 2010 Aug 31.
PMID: 20835343 [PubMed - in process]
Related citations
Filter your results:
All (130842)
Review (7994)
Free Full Text (23157)
Manage Filters
Also try:
Titles with your search terms
9074 free full-text articles in PubMed Central
Find related data
Database: Select
Find items
Search details
"haemoglobin"[All Fields] OR "hemoglobins"[MeSH Terms] OR "hemoglobins"[All Fields] OR "hemoglobin"[All Fields]
Search
See more...

MeSH
hemoglobin

関連するMeSH term：それぞれの定義を比較、選択できる。

クリック

鎌形赤血球は別タームとなっていることに注意。



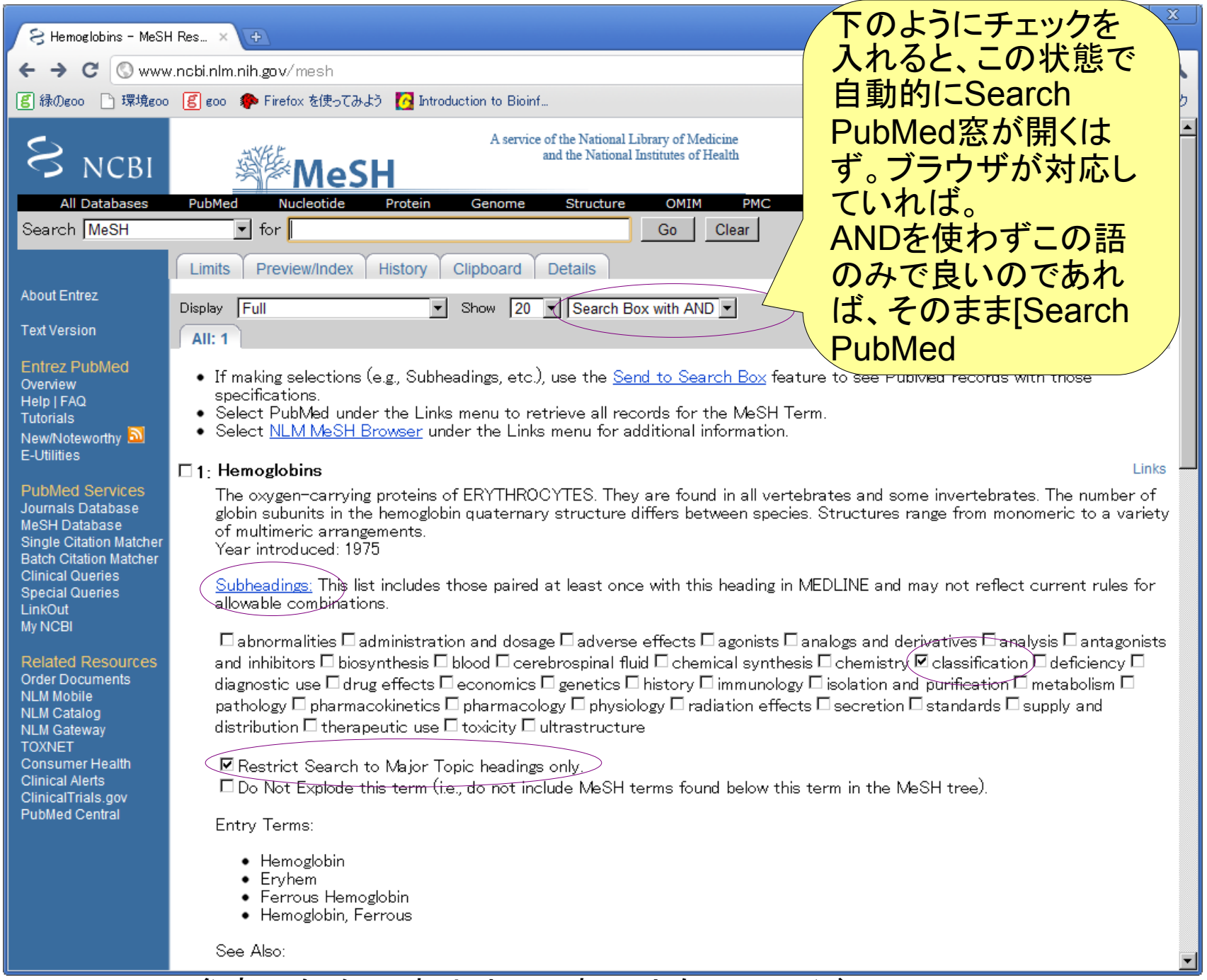
MeSH

hemoglobins

ページ上部

Subheadingsを絞ることで、特定の意味で検索語が使われている文献に絞って検索できる。

単に検索語が含まれるだけの文献を排除できる。



参考のため、これより下の表示は次のスライドで。

MeSH hemoglobins ページ下部

Hemoglobins - MeSH Res...

ncbi.nlm.nih.gov/mesh

goo Firefox を使ってみよう Introduction to Bioinf... その他のブックマーク

Do Not Explode this term (i.e., do not include MeSH terms found below this term in the MeSH tree).

Entry Terms:

- Hemoglobin
- Eryhem
- Ferrous Hemoglobin
- Hemoglobin, Ferrous

See Also:

- Iron Chelating Agents

All MeSH Categories
Chemicals and Drugs Category
Amino Acids, Peptides, and Proteins
Proteins
Blood Proteins
Hemoglobins
Carboxyhemoglobin
Erythrocruorins
Fetal Hemoglobin
Hemoglobin A
Hemoglobin A, Glycosylated
Hemoglobin A2
Hemoglobin Subunits
alpha-Globins +
beta-Globins +
Hemoglobins, Abnormal
Hemoglobin C
Hemoglobin E
Hemoglobin H
Hemoglobin J
Hemoglobin M
Hemoglobin, Sickle
Methemoglobin
Oxyhemoglobins
Sulfhemoglobin

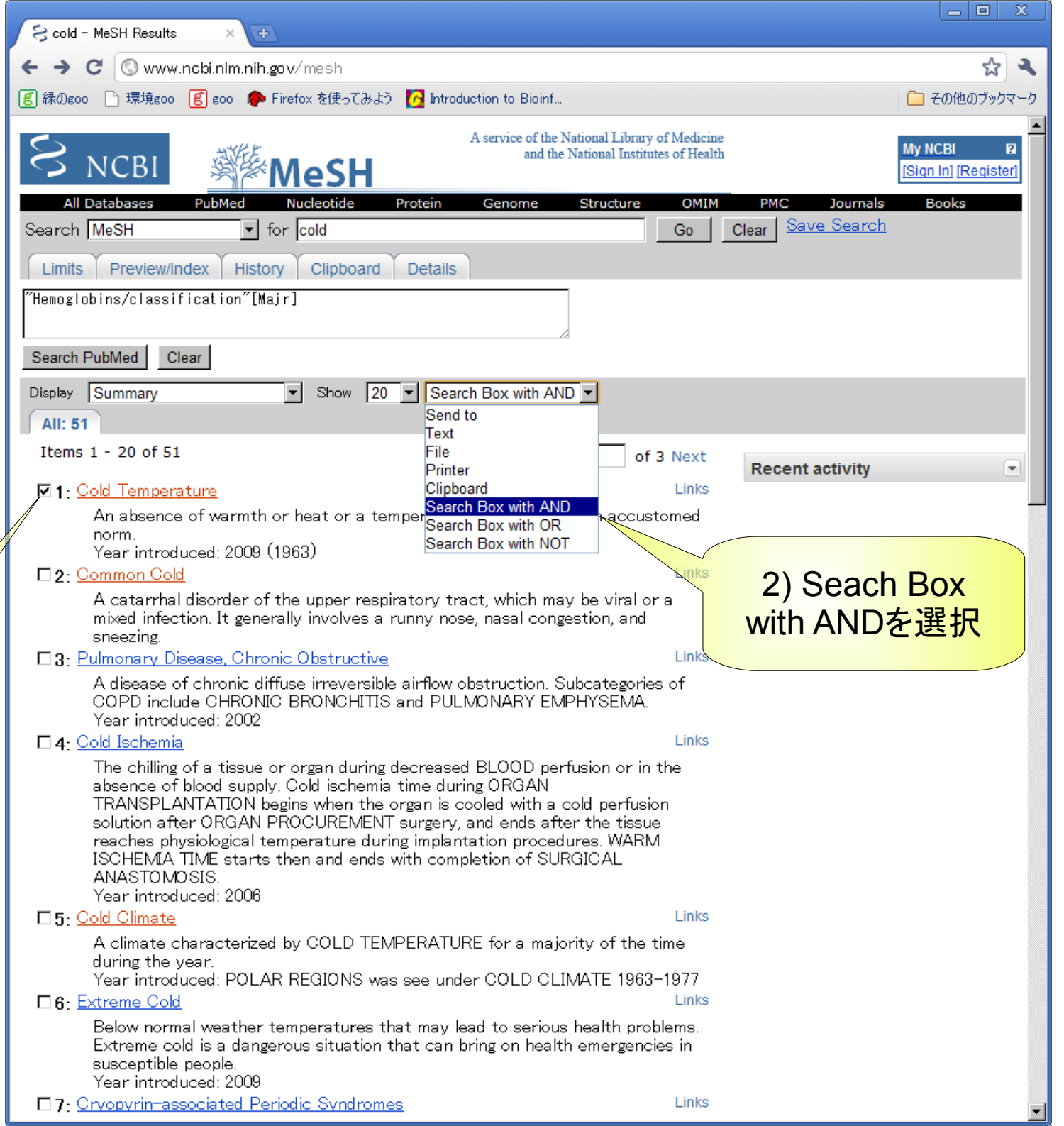
All MeSH Categories
Chemicals and Drugs Category
Amino Acids, Peptides, and Proteins
Proteins
Hemeproteins
Globins
Hemoglobins
Carboxyhemoglobin
Erythrocruorins

MeSH termのカテゴリ
調べたい概念により近いMeSH termが下部の階層にあるならば、なるべく下部階層の語を使う。

“cold”と検索窓に入力して[Go]

MeSH term vs. 自然語

「低温」や「風邪」といった自然語もMeSH検索で概念を絞ることができる。



再び別タームを同様に選択

checkいれてから

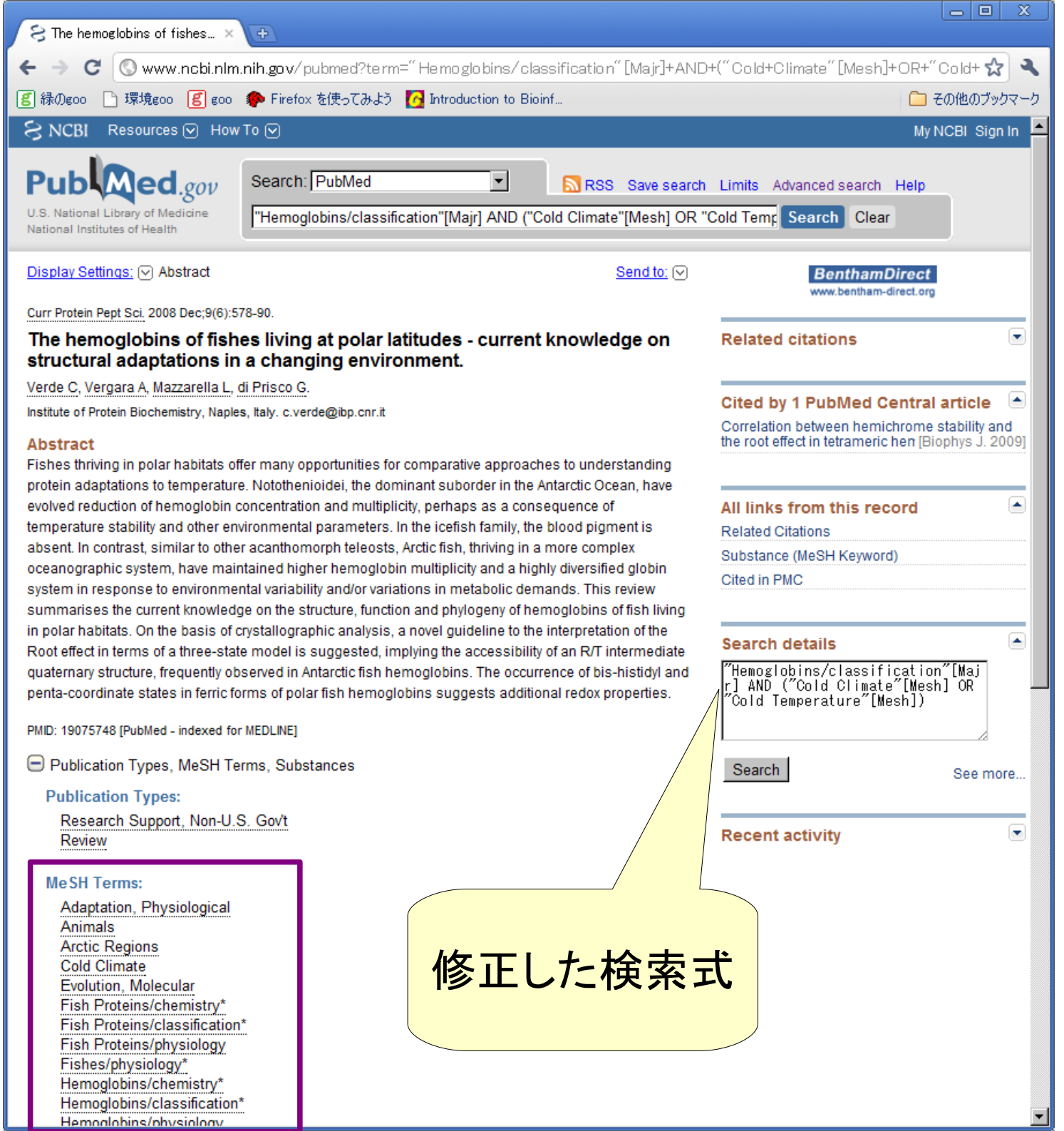
検索窓を確認

括弧の位置が意図したもの(Cold Temp OR Cold Climate)と違う
-->手入力で修正

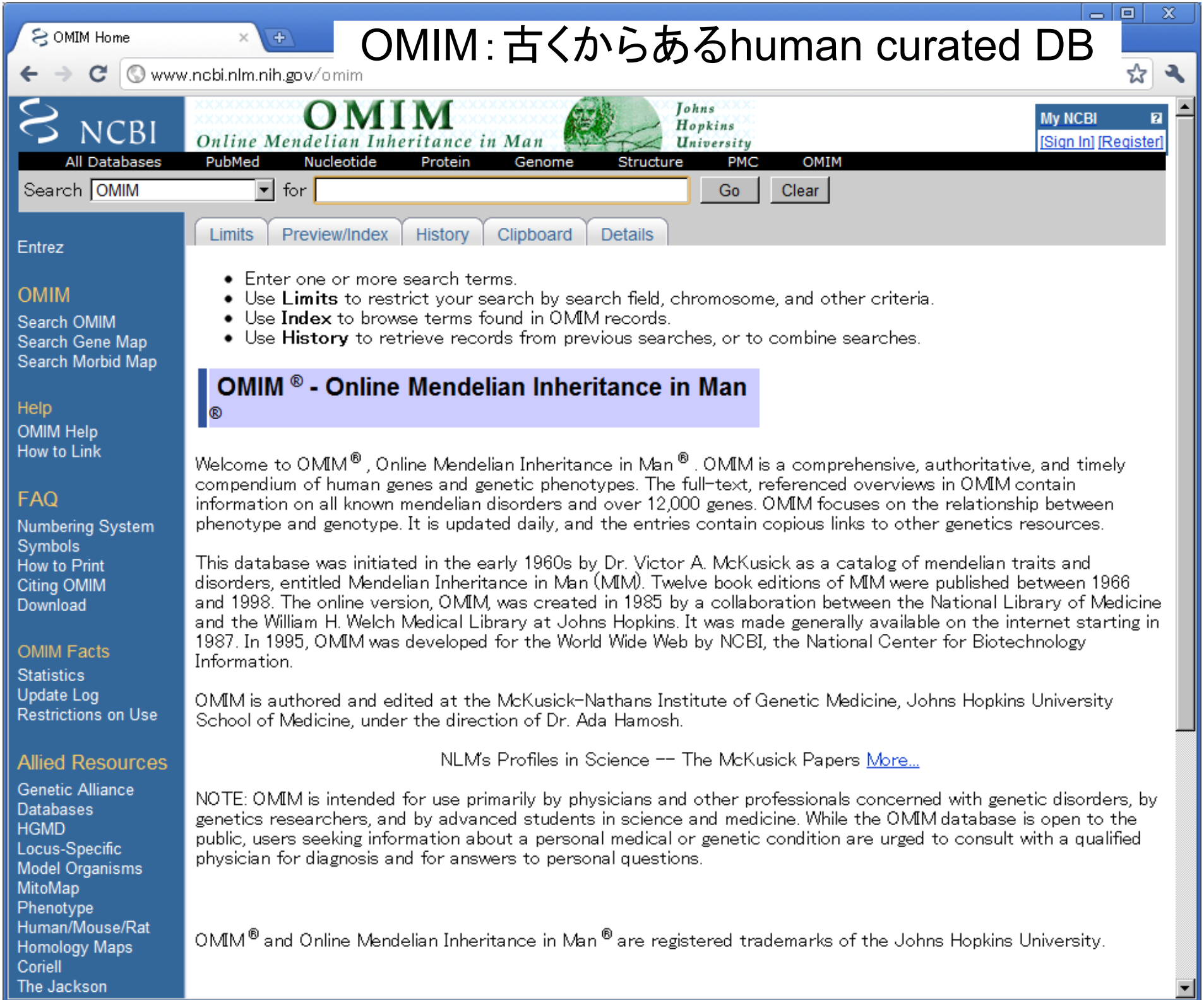
注意:
MeSH termを検索し、検索語を順次追加-->最後にPubMed検索すること。
逆に、PubMed検索した後にMeSH検索しても、両者の検索語の結合は自動的にはできない。

検索結果
1件の場合

どんなMeSH term
で検索されるか



OMIM：古くからあるhuman curated DB



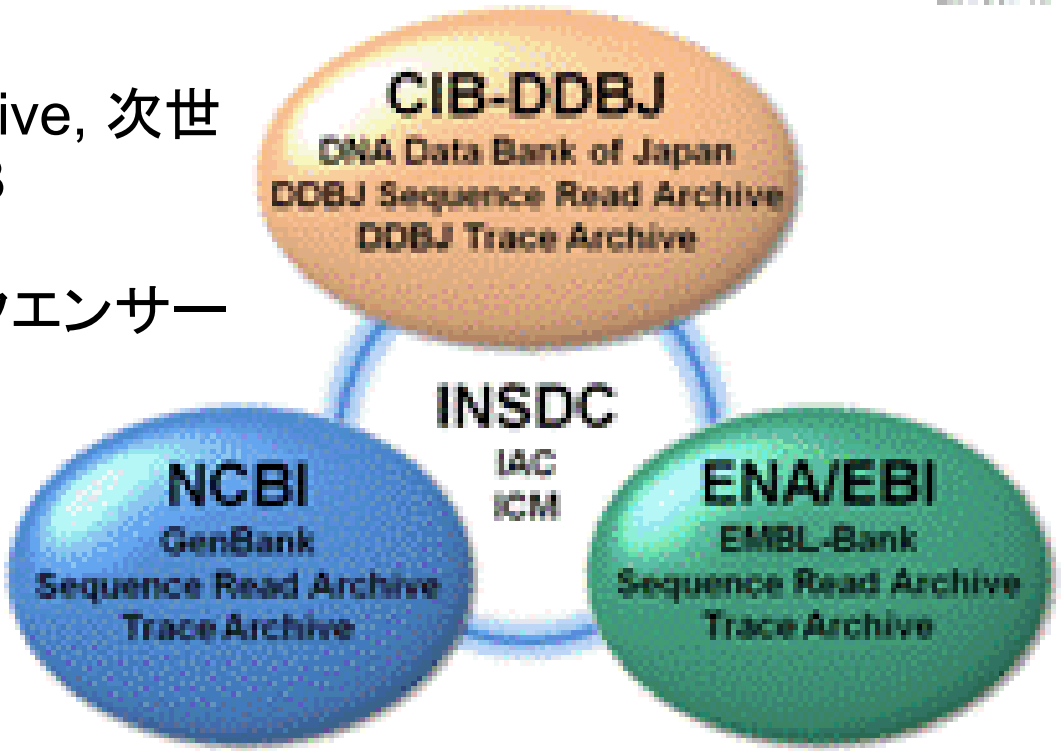
核酸DBと遺伝子アノテーション

- 国際塩基配列データベース(INSDC)
- データ構造
- データのカテゴリ
- INSDC核酸DBでのID命名法
- 遺伝子アノテーション
- 遺伝子推定法とID命名法

•国際塩基配列データベース(INSDC)

- ３センターの違い:
  - データアクセス(共通)とデータ登録方法(別々)
- 図:それぞれの機関3行になっている意味

•1行目:従来からの核酸DB
•2行目:SRA (Sequence Read Archive, 次世代シークエンサーのreadの保存用DB
  • DDBJ SRA(DRA)
•3行目:Trace Archive, 従来型シークエンサーのreadの保存用DB
  • DDBJ Trace Archive (DTA)



データ構造
核酸DB
フラット・ファイル

フィールド

フラットファイルの
形式はEMBLだけ
少し異なる。

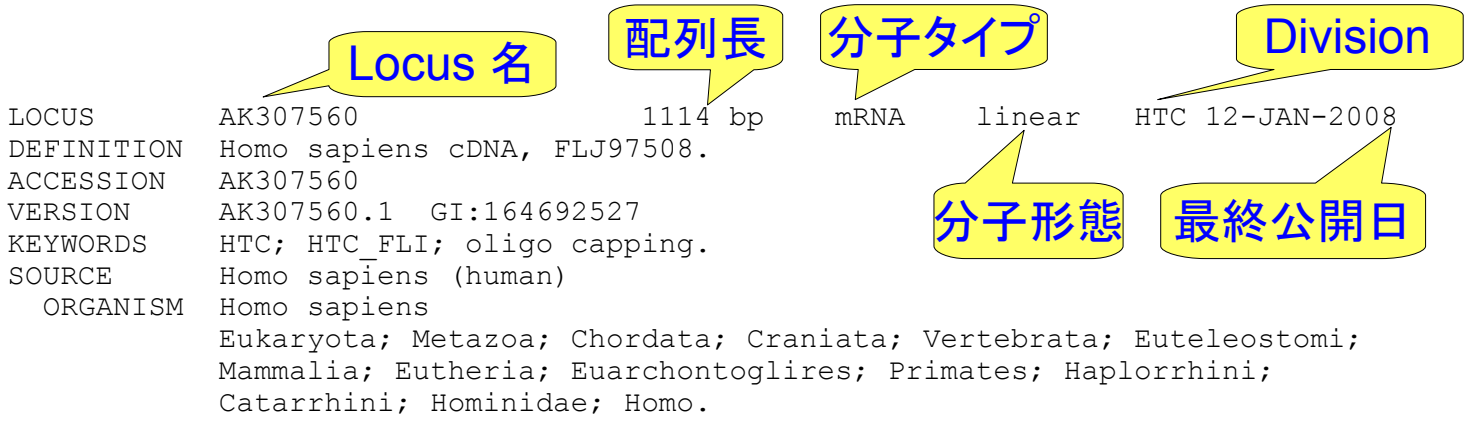
識別子

内容

識別子はDB内で統一
-->キーとして重要
核酸DBの主キーは
アクセッション

```
LOCUS       AK307560               1114 bp    mRNA    linear   HTC 12-JAN-2008
DEFINITION  Homo sapiens cDNA, FLJ97508.
ACCESSION   AK307560
VERSION     AK307560.1  GI:164692527
KEYWORDS    HTC; HTC_FLI; oligo capping.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1
  AUTHORS   Wakamatsu,A., Yamamoto,J., Kimura,K., Ishii,S., Watanabe,K.,
            Sugiyama,A., Murakawa,K., Kaida,T., Tsuchiya,K., Fukuzumi,Y.,
            Kumagai,A., Oishi,Y., Yamamoto,S., Ono,Y., Komori,Y., Yamazaki,M.,
            Kisu,Y., Nishikawa,T., Sugano,S., Nomura,N. and Isogai,T.
  TITLE     NEDO human cDNA sequencing project
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1114)
  AUTHORS   Isogai,T. and Yamamoto,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (11-JAN-2008) Contact:Takao Isogai Reverse Proteomics
            Research Institute; 1-9-11 Kaji-cho, Chiyoda-ku, Tokyo 101-0044,
            Japan E-mail :flj-cdna@nifty.com
COMMENT     Human cDNA sequencing project focused on splicing variants of mRNA
            in NEDO functional analysis of protein and research application
            project supported by Ministry of Economy, Trade and Industry,
            Japan; cDNA selection for complete cds sequencing: Reverse
            Proteomics Research Institute (REPRORI), Hitachi, Ltd., Japan
            (Hitachi) and Japan Biological Informatics Consortium, Japan
            (JBIC); cDNA complete cds sequencing: JBIC; cDNA library
            construction: Helix Research Institute supported by Japan Key
            Technology Center, Japan (HRI); cDNA 5'- & 3'-end sequencing:
            Research Association for Biotechnology, Japan, Biotechnology
            Center, National Institute of Technology and Evaluation, Japan and
            HRI; cDNA mapping to human genome: Central Research Laboratory,
            Hitachi; evaluation and annotation: REPRORI.
FEATURES             Location/Qualifiers
     source          1..1114
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /clone="NETRP2000337"
                     /cell_type="neutrophils"
                     /clone_lib="NETRP2"
                     /note="cloning vector: pME18SFL3;
                     primary culture, neutrophils"
ORIGIN
        1 agtgtcgacg gcagcggcgg cggcgggtgg gaaatggcgg agtatctggc ctccatcttc
```

フラット・ファイルの説明・Locus field



- Locus名：
  - 各DB内でユニークな名前、DBごとに独自の命名法をとる。DDBJでは1996年7月以降アクセッション番号と同一に。
- 分子タイプ：
  - DNA, RNA, mRNA, rRNA, tRNA, cRNA のいずれか。
  - DDBJで登録する場合、source feature の /mol_type qualifier に記載する。
- 分子形態：
  - 線状（linear）か，環状（circular）か。ただし、circularはエントリが完全長の場合。
- division：
  - 配列登録者が選ぶものではない。
  - 生物種に基づくdivision
  - データ形式に基づくdivision
  - データ形式として良く使われる略語がdivisionではない場合もあるので注意
    - WGS(whole genome shotgun),
    - TPA(Third Party Annotation),
    - MGA(Mass sequence for Genome Annotation)
    - SRA (Sequence Read Archive, 次世代シークエンサーのreadの保存用DB
    - Trace Archive, 従来型シークエンサーのreadの保存用DB
  - これらの略語は相同性検索の際、ターゲットを絞るために知っておくとよい。

HOME > 塩基配列の登録 > DDBJ のデータ公開形式 (flat file) の説明 > LOCUS
http://www.ddbj.nig.ac.jp/sub/locus-j.html

Division

*	HUM	ヒト
	PRI	霊長類（ヒトを除く）
	ROD	齧歯類
	MAM	哺乳類（上記を除く）
	VRT	脊椎動物（MAMを除く）
	INV	無脊椎動物
	PLN	植物・真菌類　など
	BCT	バクテリア
	VRL	ウイルス
	PHG	バクテリオファージ

* GenBankでは無い。

BLASTなど相同性検索の際の参考に。

	PAT	特許出願に含まれる塩基配列データ
	ENV	環境上のサンプルに由来した配列
	SYN	synthetic constructs
		人為的に構成された配列
	EST	expressed sequence tags
*	TSA	transcriptome shotgun assemblies, すでに登録された転写物配列をassembleして得られた長い配列でータ
	GSS	genome survey sequences short single pass のゲノム配列
	HTC	high throughput cDNA sequences, 配列が finish した後，生物種による division に移される場合あり。
	HTG	high throughput genomic sequences, ゲノムプロジェクトに由来，以下の 3 phase に分類されKEYWORD 行に記載。finish した後は生物種による division に移る。
		phase0: piece contig が構築される以前の配列
		phase1: 構築された piece contig の向きや順序が未確定の配列
		phase2: piece contig の向きや順序が確定した unfinished の配列
	STS	sequence tagged sites
*	CON	Contig / Constructed, ゲノムプロジェクトのように個々に登録された一連の配列データをデータバンク側が結合し，1つのアクセッション番号を付与した長大なデータ。

http://www.ddbj.nig.ac.jp/sub/locus-j.html

INSDC核酸DBでのID

- INSDC核酸DB
  - 通常division
    - アルファベット１文字＋５桁数字または、アルファベット2文字＋６桁数字
    - アルファベット部分はDBおよびdivisionによって割り振りが決まっている。
    - 参照：http://www.ddbj.nig.ac.jp/sub/prefix.htm
  - WGS data
    - アルファベット4文字＋８桁数字
  - MGA data
    - アルファベット5文字＋７桁数字

遺伝子アノテーション

•配列決定後、配列に対して有用な記述を追加する作業

•個々の塩基配列データでは通常、配列決定者が最低限のアノテーションをする。

- 例）配列全体に対して
  - 名称、遺伝子シンボル、定義など
- 例）配列中の部分領域に対して
  - [ゲノム配列に対し]遺伝子、反復配列など
  - [遺伝子配列に対し]遺伝子構造、多型など
  - [アミノ酸配列に対し]ドメイン、活性部位、修飾部位など

登録の際は、自分が何に対してどういうアノテーションが求められているか把握することが大切。
年々マニュアルが更新されているので注意。

•ゲノム配列アノテーション・３段階

- 反復配列探索
  - 前処理として反復配列をマスクする (eg., RepeatMasker)
- 遺伝子探索(cf., 次のスライド)
  - 計算機処理によって遺伝子(候補)領域を抽出する
- 機能探索
  - 機能に関係することが知られている配列（タンパク質機能ドメイン）を抽出する

転写物のアノテーション

•cDNAおよび成熟mRNAは、ゲノム配列とマッピングさせることによって、そのまま遺伝子アノテーションになる。
•cDNAのおもな種類

- Expressed Sequence Tags (ESTs)
  - cDNAクローンをランダムに選び末端400-600bp程度をone-passで配列決定
  - 比較的不正確(約2%のエラー)な傾向
  - 網羅的に発現位置を決める目的で使用
  - 核酸DBにESTというdivisionがある
- Full-length cDNA（完全長cDNA）
  - 配列の正確性向上。
  - ASパターンも予測可能。
  - 長い遺伝子ではデータが少なくなる傾向

H-InvDB	完全長cDNA配列にアノテーションをつけることを出発点とし、アミノ酸配列、AS、発現などの関連情報と連動したDB
RefSeq	非冗長的に、遺伝子、転写物、タンパク質、それぞれの代表配列をモデル生物で決め、整備したDB
Ensembl	ゲノムデータの自動アノテーションとブラウザ表示を出発点とし、アミノ酸配列、AS、発現などの関連情報と連動したDB

primary配列DBとannotated DB

GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common	Single records for each molecule of major organisms
Records can contradict each other	
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to **primary literature**	Akin to **review articles**
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

遺伝子探索のおもな戦略

- 転写産物比較
  - 最も信頼度が高く、ゲノム配列と同時に転写物配列も網羅的に読むプロジェクトも多い
- 比較ゲノム
  - 他の生物ゲノムと比較し、保存性の高い領域を遺伝子候補とする
- *ab initio*法
  - 既知の知見によって遺伝子の特徴を備えた配列を遺伝子候補とする
  - 信頼性は最も低い
  - *“ab initio”*とは「始めから」の意。ゲノムのアッセンブル法でもこの語がつかわれるので混乱注意。

RefSeq、H-InvDBにおけるIDの付け方

- RefSeqのID
  - アルファベット2文字＋アンダーバー＋数字
  - アルファベット1文字目：
    - A: Alternate assembly or annotation
    - N: kNown (A以外のゲノム配列とキュレータによりレヴューされたRNA&protein)
    - X: predictiction(prediXon)
    - Z: NZ_accession(WGS)上についたproteinのアノテーション
  - アルファベット2文字目：
    - M: mRNA
    - R: noncoding RNA
    - P: protein
    - Z: genomic, whole genome shotgun (WGS) sequence data
    - W: genomic, Intermediate assemblies of BAC or WGS sequence data
    - T: genomic, Intermediate assemblies of BAC and/or WGS sequence data
    - S: genomic, unplaced scaffolds, etc
    - G: genomic, incomplete
    - C: genomic, complete

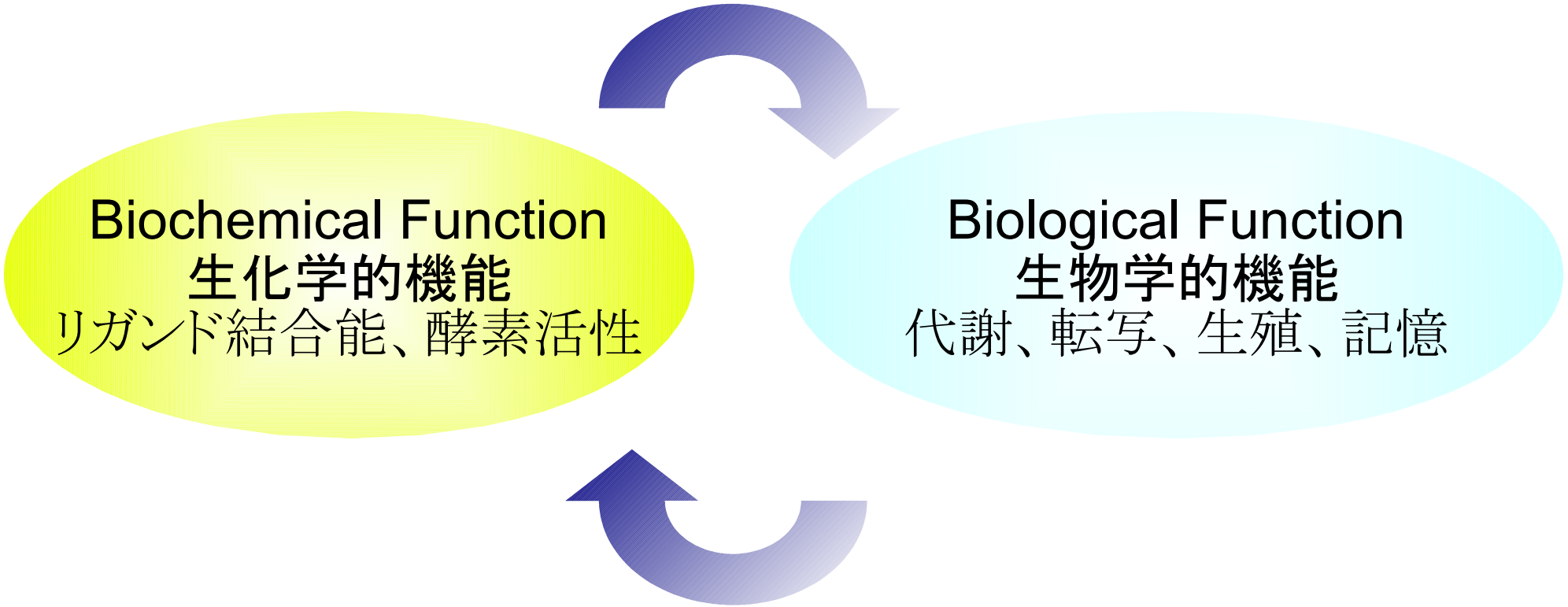
- H-InvDBのID
  - HIT (H-Invitational transcript):HIT + 9桁の数字 + version番号　例）HIT000000001.1
  - HIX (H-Invitational cluster): HIX + 7桁の数字 + version番号　例）HIX0000001.1
  - HIP (H-Invitational protein): HIP + 9桁の数字 + version番号　例）HIP000000001.1
  - HIF (H-Invitational gene family/group):　HIF + 7桁の数字　例）HIF0000001

http://hinv.jp/hinv/help/help_index_jp.html　http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accessions

タンパク質データベースと機能推定

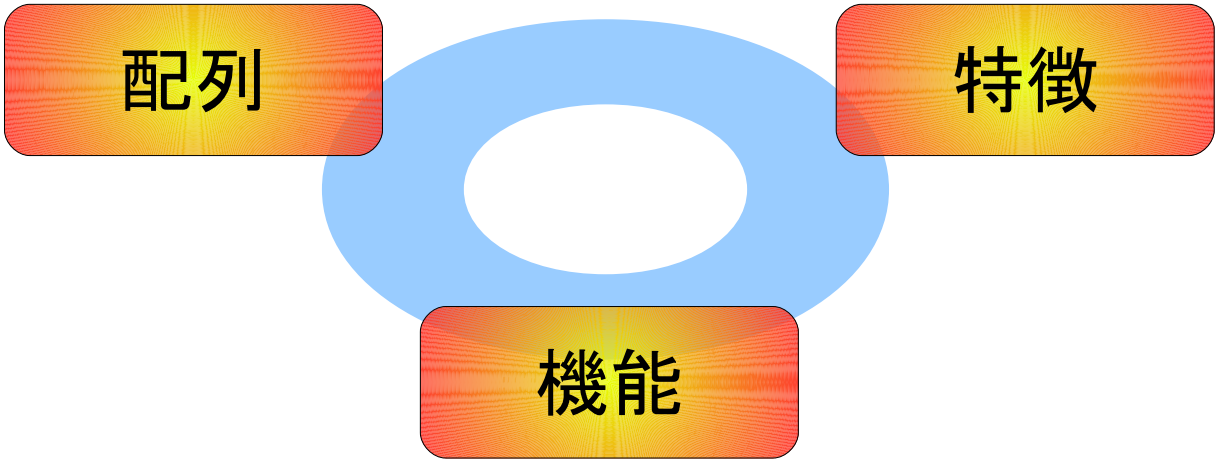
- 「機能」と配列について
- 機能ドメインデータベース
  - UniProt
  - InterPro
- 立体構造データベース
  - PDB

「機能」という語の多義性



作業のscopeは何かが重要

機能と配列と特徴



- 配列に隠された機能の単位（ドメイン）。
- 機能ドメインは進化的に保存されやすい。
- 保存領域(Sequence signature)の長さ・規模は様々
  - 翻訳後修飾部位、モチーフ、ドメイン
- 配列上の特徴が機能ドメインゆえであることが多い。
- 決まった立体構造を通じて機能する

UniProt

UniProt - Google 検索 | UniProt

www.uniprot.org

HotMail の無料サービス | Windows Media | Windows | リンクの変更 | その他のブックマーク

UniProt

Downloads · Contact · Documentation/Help

Search | Blast | Align | Retrieve | ID Mapping

Search in Protein Knowledgebase (UniProtKB)

Query Search Clear Fields »

WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: ★ Swiss-Prot, which is manually annotated and reviewed. ☆ TrEMBL, which is automatically annotated and is **not** reviewed. Includes Complete Proteome Sets.
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords and more.

Getting started

- Text search
- Sequence similarity searches (BLAST)
- Sequence alignments
- Batch retrieval

NEWS

UniProt release 2010_09 – Aug 10, 2010

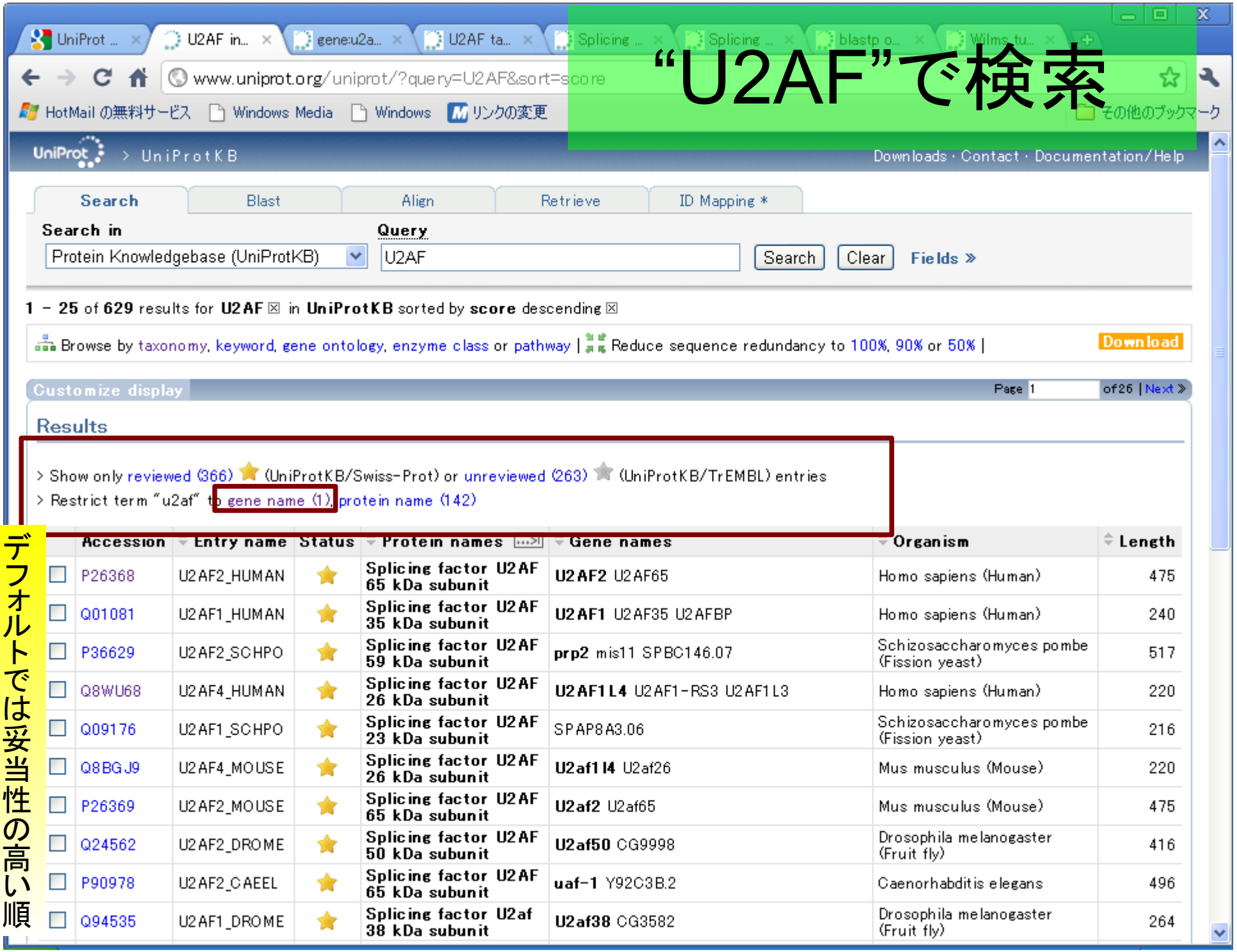
'De-merge' of multi-gene entries derived from a single species in UniProtKB/Swiss-Prot

> Statistics for UniProtKB: Swiss-Prot · TrEMBL
> Forthcoming changes
> News archives

SITE TOUR

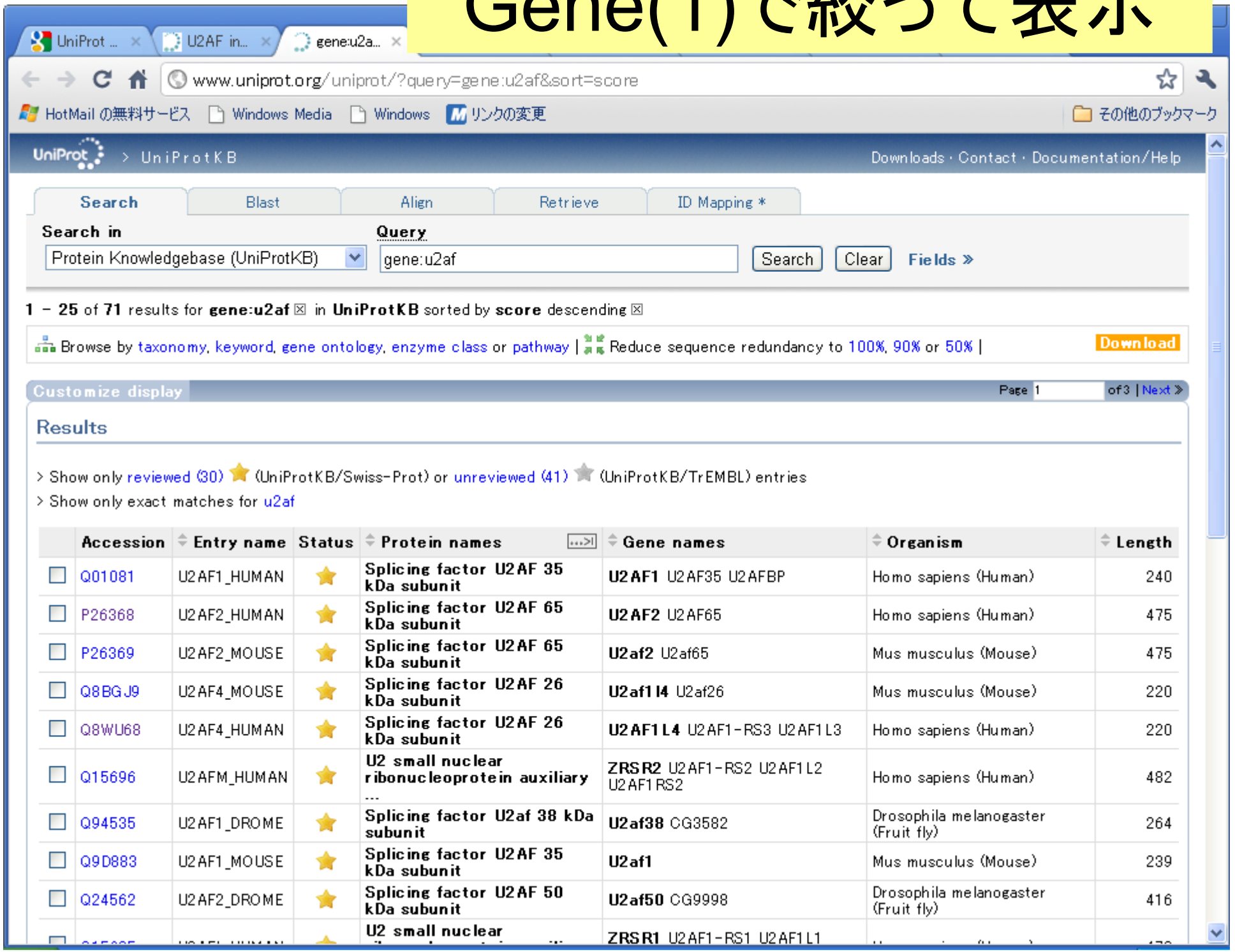
Learn how to make best use of the tools and data on this site.

PROTEIN SPOTLIGHT



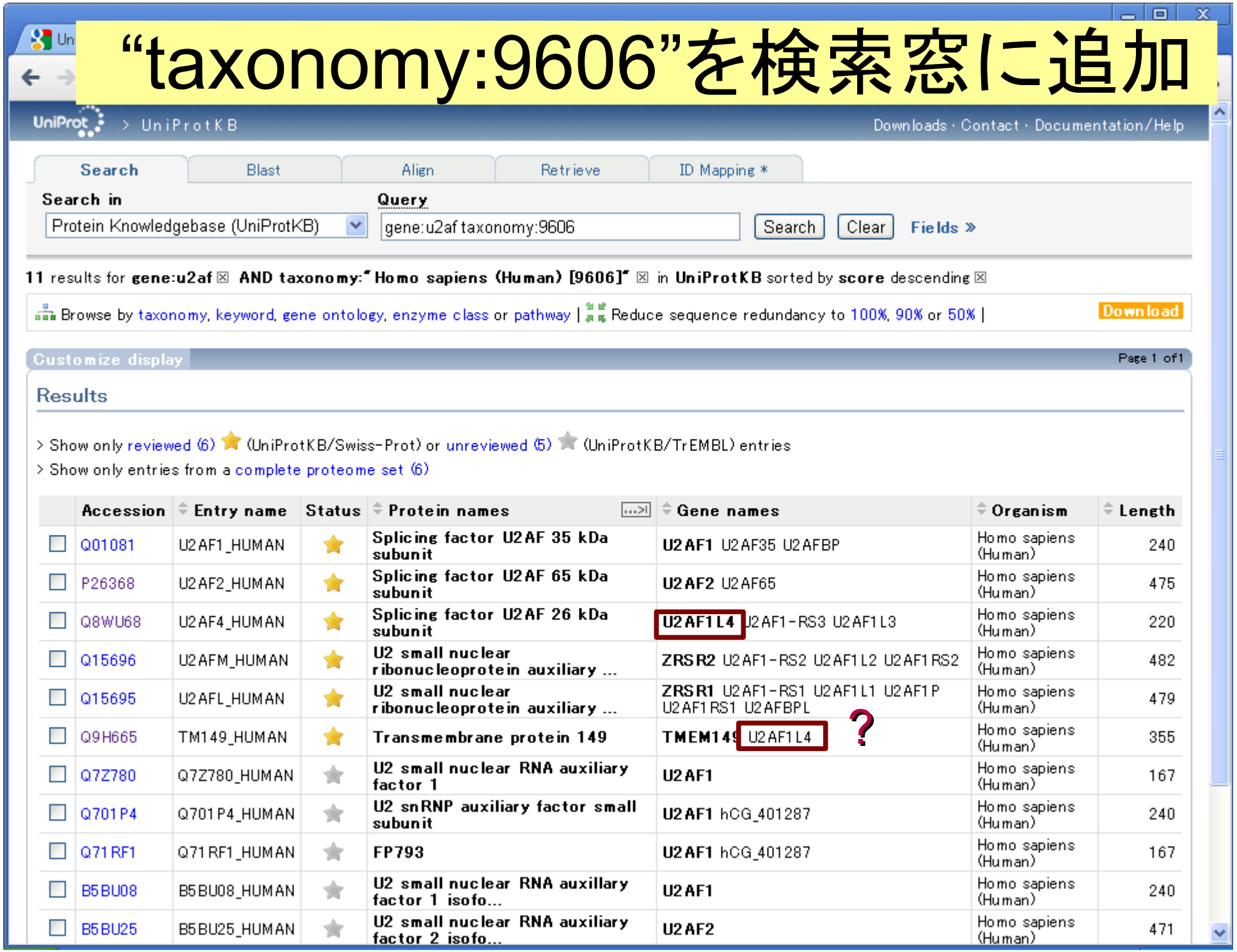
	Accession	Entry name	Status	Protein names	Gene names	Organism	Length
☐	P26368	U2AF2_HUMAN	★	Splicing factor U2AF 65 kDa subunit	**U2AF2** U2AF65	Homo sapiens (Human)	475
☐	Q01081	U2AF1_HUMAN	★	Splicing factor U2AF 35 kDa subunit	**U2AF1** U2AF35 U2AFBP	Homo sapiens (Human)	240
☐	P36629	U2AF2_SCHPO	★	Splicing factor U2AF 59 kDa subunit	**prp2** mis11 SPBC146.07	Schizosaccharomyces pombe (Fission yeast)	517
☐	Q8WU68	U2AF4_HUMAN	★	Splicing factor U2AF 26 kDa subunit	**U2AF1L4** U2AF1-RS3 U2AF1L3	Homo sapiens (Human)	220
☐	Q09176	U2AF1_SCHPO	★	Splicing factor U2AF 23 kDa subunit	SPAP8A3.06	Schizosaccharomyces pombe (Fission yeast)	216
☐	Q8BGJ9	U2AF4_MOUSE	★	Splicing factor U2AF 26 kDa subunit	**U2af1l4** U2af26	Mus musculus (Mouse)	220
☐	P26369	U2AF2_MOUSE	★	Splicing factor U2AF 65 kDa subunit	**U2af2** U2af65	Mus musculus (Mouse)	475
☐	Q24562	U2AF2_DROME	★	Splicing factor U2AF 50 kDa subunit	**U2af50** CG9998	Drosophila melanogaster (Fruit fly)	416
☐	P90978	U2AF2_CAEEL	★	Splicing factor U2AF 65 kDa subunit	**uaf-1** Y92C3B.2	Caenorhabditis elegans	496
☐	Q94535	U2AF1_DROME	★	Splicing factor U2af 38 kDa subunit	**U2af38** CG3582	Drosophila melanogaster (Fruit fly)	264

Gene(1)で絞って表示



	Accession	Entry name	Status	Protein names	Gene names	Organism	Length
☐	Q01081	U2AF1_HUMAN	★	**Splicing factor U2AF 35 kDa subunit**	**U2AF1** U2AF35 U2AFBP	Homo sapiens (Human)	240
☐	P26368	U2AF2_HUMAN	★	**Splicing factor U2AF 65 kDa subunit**	**U2AF2** U2AF65	Homo sapiens (Human)	475
☐	P26369	U2AF2_MOUSE	★	**Splicing factor U2AF 65 kDa subunit**	**U2af2** U2af65	Mus musculus (Mouse)	475
☐	Q8BGJ9	U2AF4_MOUSE	★	**Splicing factor U2AF 26 kDa subunit**	**U2af1l4** U2af26	Mus musculus (Mouse)	220
☐	Q8WU68	U2AF4_HUMAN	★	**Splicing factor U2AF 26 kDa subunit**	**U2AF1L4** U2AF1-RS3 U2AF1L3	Homo sapiens (Human)	220
☐	Q15696	U2AFM_HUMAN	★	**U2 small nuclear ribonucleoprotein auxiliary** ...	**ZRSR2** U2AF1-RS2 U2AF1L2 U2AF1RS2	Homo sapiens (Human)	482
☐	Q94535	U2AF1_DROME	★	**Splicing factor U2af 38 kDa subunit**	**U2af38** CG3582	Drosophila melanogaster (Fruit fly)	264
☐	Q9D883	U2AF1_MOUSE	★	**Splicing factor U2AF 35 kDa subunit**	**U2af1**	Mus musculus (Mouse)	239
☐	Q24562	U2AF2_DROME	★	**Splicing factor U2AF 50 kDa subunit**	**U2af50** CG9998	Drosophila melanogaster (Fruit fly)	416
☐			★	**U2 small nuclear**	**ZRSR1** U2AF1-RS1 U2AF1L1		

“taxonomy:9606”を検索窓に追加



UniProt > UniProtKB

Downloads · Contact · Documentation/Help

Search | Blast | Align | Retrieve | ID Mapping *

Search in: Protein Knowledgebase (UniProtKB)

Query: gene:u2af taxonomy:9606

Search | Clear | Fields »

11 results for gene:u2af AND taxonomy:"Homo sapiens (Human) [9606]" in UniProtKB sorted by score descending

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% | Download

Customize display — Page 1 of 1

Results

> Show only reviewed (6) (UniProtKB/Swiss-Prot) or unreviewed (5) (UniProtKB/TrEMBL) entries

> Show only entries from a complete proteome set (6)

Accession	Entry name	Status	Protein names	Gene names	Organism	Length
Q01081	U2AF1_HUMAN	★	Splicing factor U2AF 35 kDa subunit	U2AF1 U2AF35 U2AFBP	Homo sapiens (Human)	240
P26368	U2AF2_HUMAN	★	Splicing factor U2AF 65 kDa subunit	U2AF2 U2AF65	Homo sapiens (Human)	475
Q8WU68	U2AF4_HUMAN	★	Splicing factor U2AF 26 kDa subunit	U2AF1L4 U2AF1-RS3 U2AF1L3	Homo sapiens (Human)	220
Q15696	U2AFM_HUMAN	★	U2 small nuclear ribonucleoprotein auxiliary ...	ZRSR2 U2AF1-RS2 U2AF1L2 U2AF1RS2	Homo sapiens (Human)	482
Q15695	U2AFL_HUMAN	★	U2 small nuclear ribonucleoprotein auxiliary ...	ZRSR1 U2AF1-RS1 U2AF1L1 U2AF1P U2AF1RS1 U2AFBPL	Homo sapiens (Human)	479
Q9H665	TM149_HUMAN	★	Transmembrane protein 149	TMEM149 U2AF1L4 ?	Homo sapiens (Human)	355
Q7Z780	Q7Z780_HUMAN	☆	U2 small nuclear RNA auxiliary factor 1	U2AF1	Homo sapiens (Human)	167
Q701P4	Q701P4_HUMAN	☆	U2 snRNP auxiliary factor small subunit	U2AF1 hCG_401287	Homo sapiens (Human)	240
Q71RF1	Q71RF1_HUMAN	☆	FP793	U2AF1 hCG_401287	Homo sapiens (Human)	167
B5BU08	B5BU08_HUMAN	☆	U2 small nuclear RNA auxillary factor 1 isofo...	U2AF1	Homo sapiens (Human)	240
B5BU25	B5BU25_HUMAN	☆	U2 small nuclear RNA auxiliary factor 2 isofo...	U2AF2	Homo sapiens (Human)	471

個別データ(ブラウザ表示)例

www.uniprot.org/uniprot/Q8WU68

Names and origin

Protein names	*Recommended name:* **Splicing factor U2AF 26 kDa subunit** *Alternative name(s):* U2 auxiliary factor 26 U2 small nuclear RNA auxiliary factor 1-like protein 4 U2(RNU2) small nuclear RNA auxiliary factor 1-like protein 3 Short name=U2 small nuclear RNA auxiliary factor 1-like protein 3 Short name=U2AF1-like protein 3
Gene names	Name: **U2AF1L4** Synonyms:U2AF1-RS3, U2AF1L3
Organism	**Homo sapiens (Human)** [Complete proteome]
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota > Metazoa > Chordata > Craniata > Vertebrata > Euteleostomi > Mammalia > Eutheria > Euarchontoglires > Primates > Haplorrhini > Catarrhini > Hominidae > Homo

Protein attributes

Sequence length	220 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is further processed into a mature form.
Protein existence	Evidence at protein level.

General annotation (Comments)

Function	RNA-binding protein that function as a pre-mRNA splicing factor. Plays a critical role in both constitutive and enhancer-dependent splicing by mediating protein-protein interactions and protein-RNA interactions required for accurate 3'-splice site selection. Acts by enhancing the binding of U2AF2 to weak pyrimidine tracts. Also participates in the regulation of alternative pre-mRNA splicing. Activates exon 5 skipping of PTPRC during T cell activation; an event reversed by GFI1. Binds to RNA at the AG dinucleotide at the 3'-splice site (By similarity).
Subunit structure	Interacts with GFI1 and U2AF2 (By similarity).
Subcellular location	Nucleus (By similarity). Nucleus speckle (By similarity).
Tissue specificity	Isoform 2 is widely expressed. Isoform 3 is highly expressed in heart, brain and lung, lower expressed in thymus and much lower expressed in peripheral blood leukocytes. (Ref.1)
Domain	The second zinc finger is necessary for interaction with GFI1 and for alternative pre-mRNA splicing events (By similarity).

Names · Attributes · General annotation · Ontologies · Alt products · Sequence annotation · Sequences · References · Cross-refs · Entry info · Documents

www.uniprot.org/uniprot/Q9H665

	Recommended name: **Transmembrane protein 149** *Alternative name(s):* U2 small nuclear RNA auxiliary factor 1-like 4
	Name: **TMEM149** Synonyms:U2AF1L4
	Homo sapiens (Human) [Complete proteome]
	9606 [NCBI]
	Eukaryota > Metazoa > Chordata > Craniata > Vertebrata > Euteleostomi > Mammalia > Eutheria > Euarchontoglires > Primates > Haplorrhini > Catarrhini > Hominidae > Homo
	355 AA.
	Complete.
	The displayed sequence is further processed into a mature form.
	Evidence at transcript level.

n (Comments)

	Membrane; Single-pass type I membrane protein (Potential).
	Membrane
rsity	Alternative splicing Polymorphism
	Signal Transmembrane Transmembrane helix
	Complete proteome
	integral to membrane Inferred from electronic annotation. Source: UniProtKB-KW

General annotation · Ontologies · Alt products · Sequence annotation · nces · Cross-refs · Entry info · Documents

field(項目)

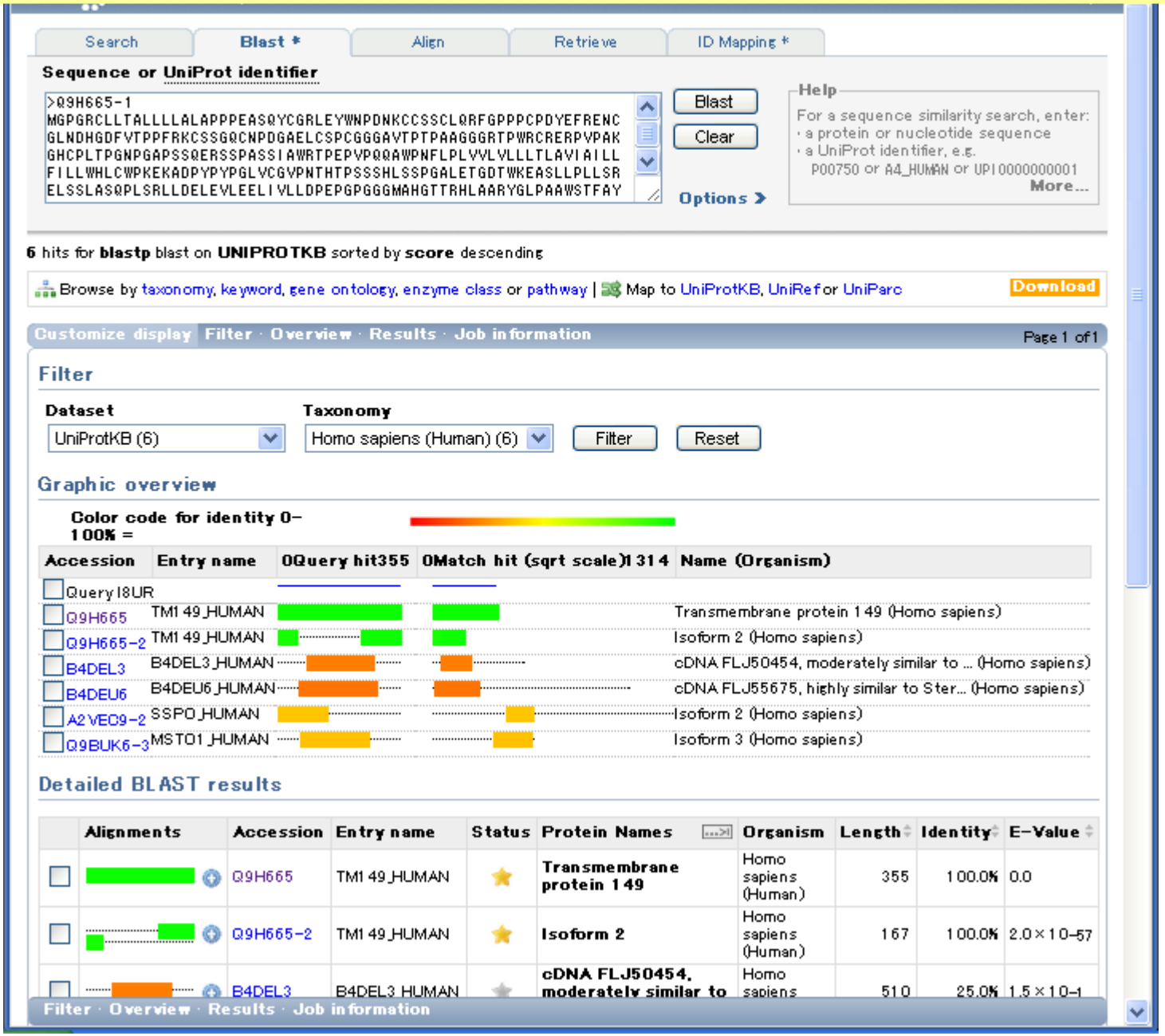
テキスト表示例（部分）

```
ID   U2AF4_HUMAN             Reviewed;         220 AA.
AC   Q8WU68; A6NKI8; Q56UU3;
DT   13-NOV-2007, integrated into UniProtKB/Swiss-Prot.
DT   13-NOV-2007, sequence version 2.
DT   10-AUG-2010, entry version 81.
DE   RecName: Full=Splicing factor U2AF 26 kDa subunit;
DE   AltName: Full=U2 auxiliary factor 26;
DE   AltName: Full=U2 small nuclear RNA auxiliary factor 1-like protein 4;
DE   AltName: Full=U2(RNU2) small nuclear RNA auxiliary factor 1-like protein 3;
DE            Short=U2 small nuclear RNA auxiliary factor 1-like protein 3;
DE            Short=U2AF1-like protein 3;
GN   Name=U2AF1L4; Synonyms=U2AF1-RS3, U2AF1L3;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 2), AND TISSUE SPECIFICITY.
RX   PubMed=17312947; DOI=10.1080/10425170600807744;
RA   Chen F., Ji C., Dou T., Zheng N., Qiu R., Peng J., Fang W., Feng C.,
RA   Xie Y., Mao Y.;
RT   "Cloning and characterization of a novel splice variant of human
RT   U2AF1L3 gene.";
RL   DNA Seq. 17:282-286(2006).
```

アクセッション番号は、統合などのため別名が複数ある

Q9H665-1でBLASTし、結果をヒトに絞って表示

自分自身TM149とそのisoformはあるが、U2AF4は無い。
両者は配列では無関係。



InterPro

interpro - Google 検索 | InterPro protein sequence ...

www.ebi.ac.uk/interpro/

- InterPro:Home
- Advanced Search
  - InterProScan
  - BioMart
- Help / Documentation
  - About InterPro
  - Release Notes
  - BioMart Manual
  - Tutorial
  - Publications
  - Contributors
- Web Services
- Downloads
- Protein Focus
  - Collagen

EBI > Databases > InterPro

InterPro protein sequence analysis & classification

InterPro is an integrated database of predictive protein "signatures" used for the classification and automatic annotation of proteins and genomes. InterPro classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites. InterPro adds in-depth annotation, including GO terms, to the protein signatures.

Current release: **28.0 11th August 2010** (see Release Notes for further details)

Search InterPro: >

Do a sequence search of InterPro, via InterProScan

Extract large datasets by querying our BioMart

You can access our data programmatically, via Web Services

If you have any questions or feedback please contact us.

InterPro 28.0

検索窓にどのようなIDや語、あるいはそのリストを入れたらよいか、指示が表示される。

- タンパク質(おもに機能ドメイン)DBの**統合DBである。**
  - ファミリー分類
  - **機能ドメイン**
  - リピート
  - 翻訳後修飾など機能サイト
  - 機能推定に必要な情報
- 20837エントリーのデータ。

(InterPro release 28.0)

統合されたデータベース

“dopamine”
を検索窓に

検索結果の一部。
この中から目的のエントリを選択。

Contents and coverage of the current release

InterPro protein matches are now calculated for all UniProtKB and UniParc proteins. The following statistics are for all UniProtKB proteins.
InterPro release 28.0 contains 20837 entries, representing:

Active site	89
Binding site	64
Conserved site	586
Domain	5608
Family	13023
PTM	16
Region	1188
Repeat	263

Last entry: IPR022869

32320 publications in PUBMED are referenced from InterPro.

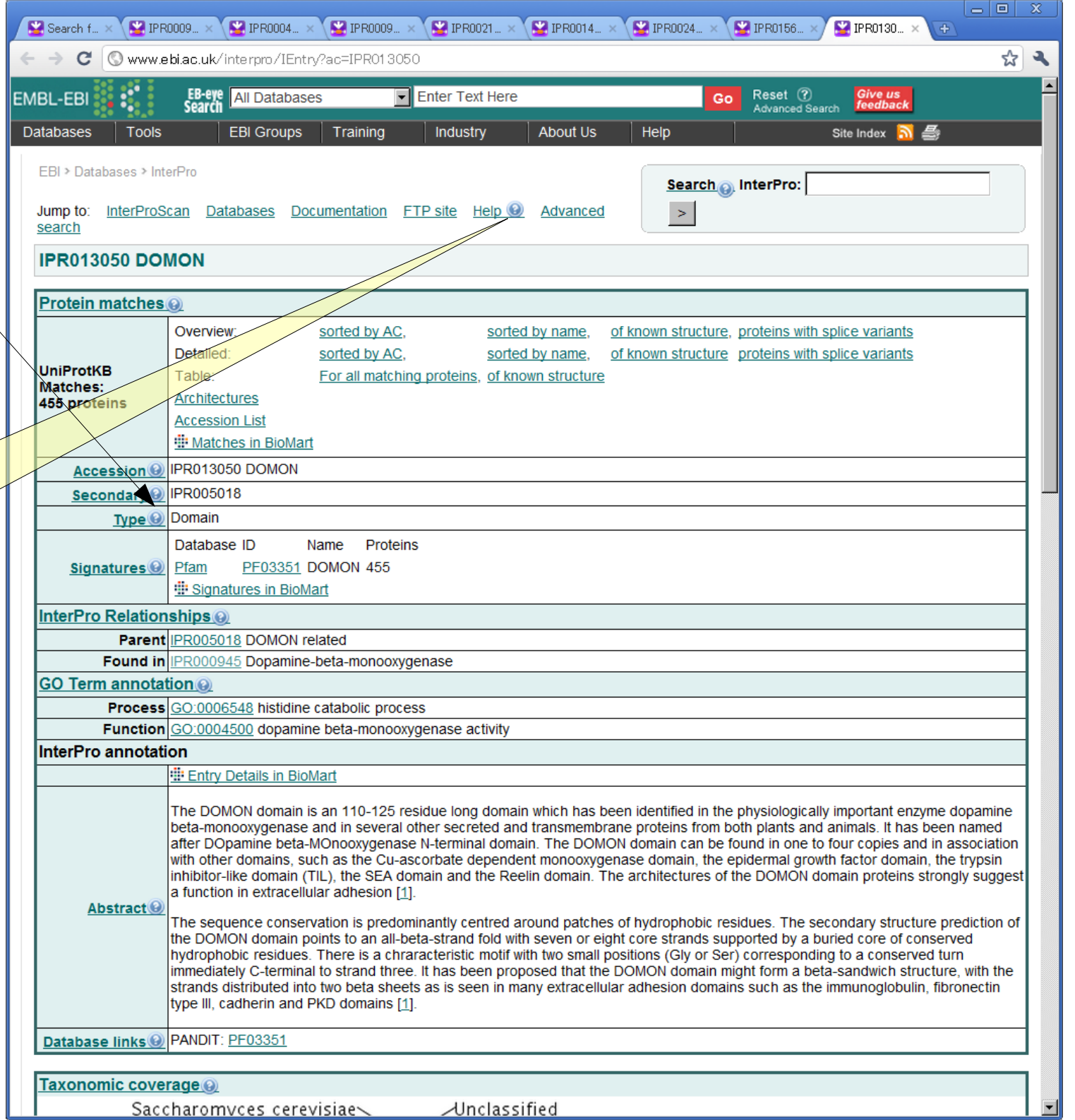
Member database information

Signature Database	Version	Signatures	Integrated Signatures
PANTHER	6.1	30128	2260
Pfam	24.0	11912	11431
PIRSF	2.72	3221	2751
PRINTS	40.0	2000	1962
ProDom	2006.1	1894	1008
PROSITE patterns	20.52	1308	1292
PROSITE profiles	20.52	860	837
SMART	6.0	809	804
TIGRFAMs	9.0	3808	3795
GENE3D	3.0.0	2147	1025
SUPERFAMILY	1.69	1538	1094
HAMAP	180510	1656	733
PfamB	24.0	142303	0

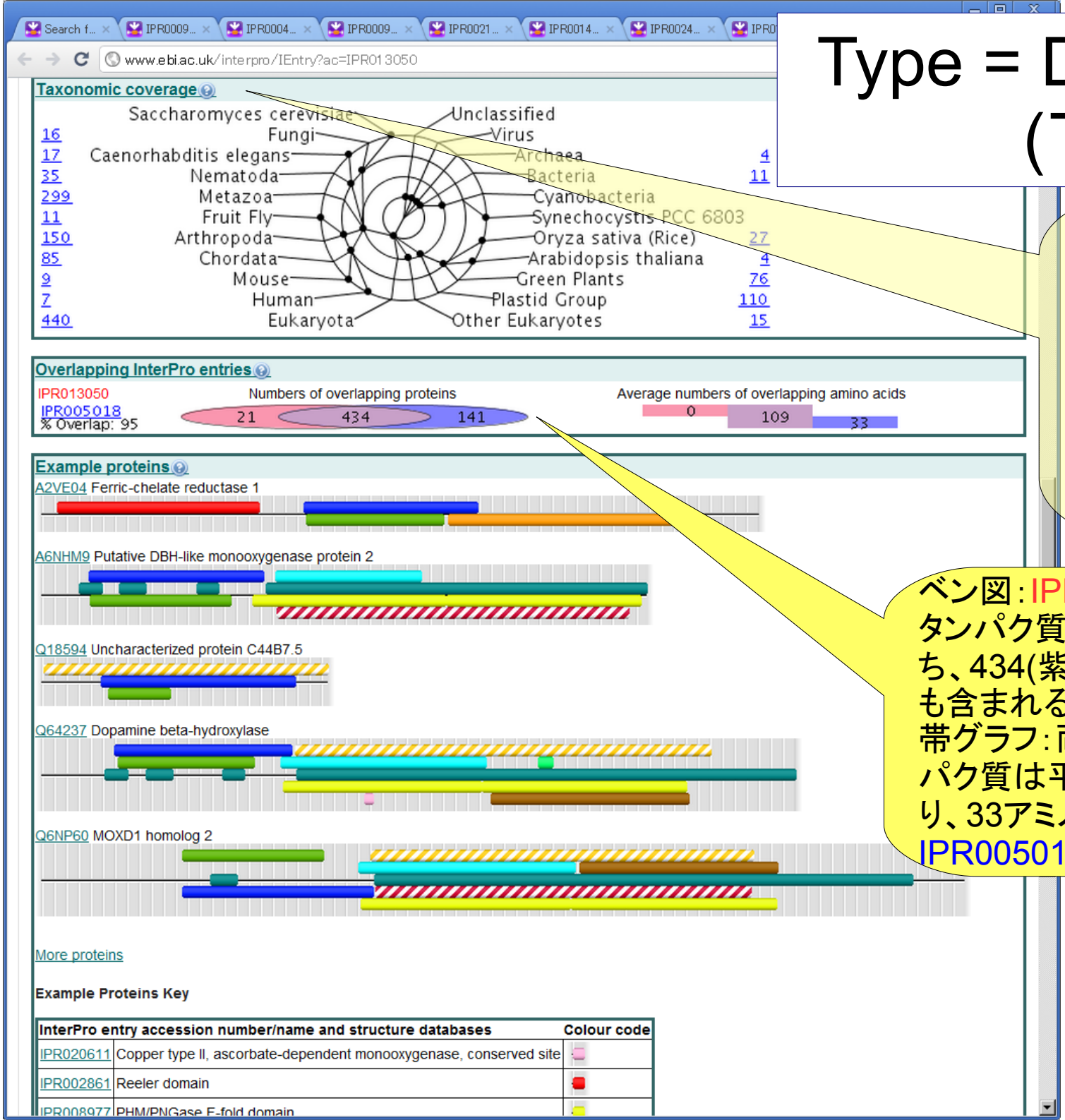
検索結果のページ下部には、DB全体の統計情報が表示されている。

Type = Domainの例(上部)

Help(?)をclickしてから、各field名の脇(?)をclick。
-->それぞれのfiledの説明が表示される。



Type = Domainの例(下部)



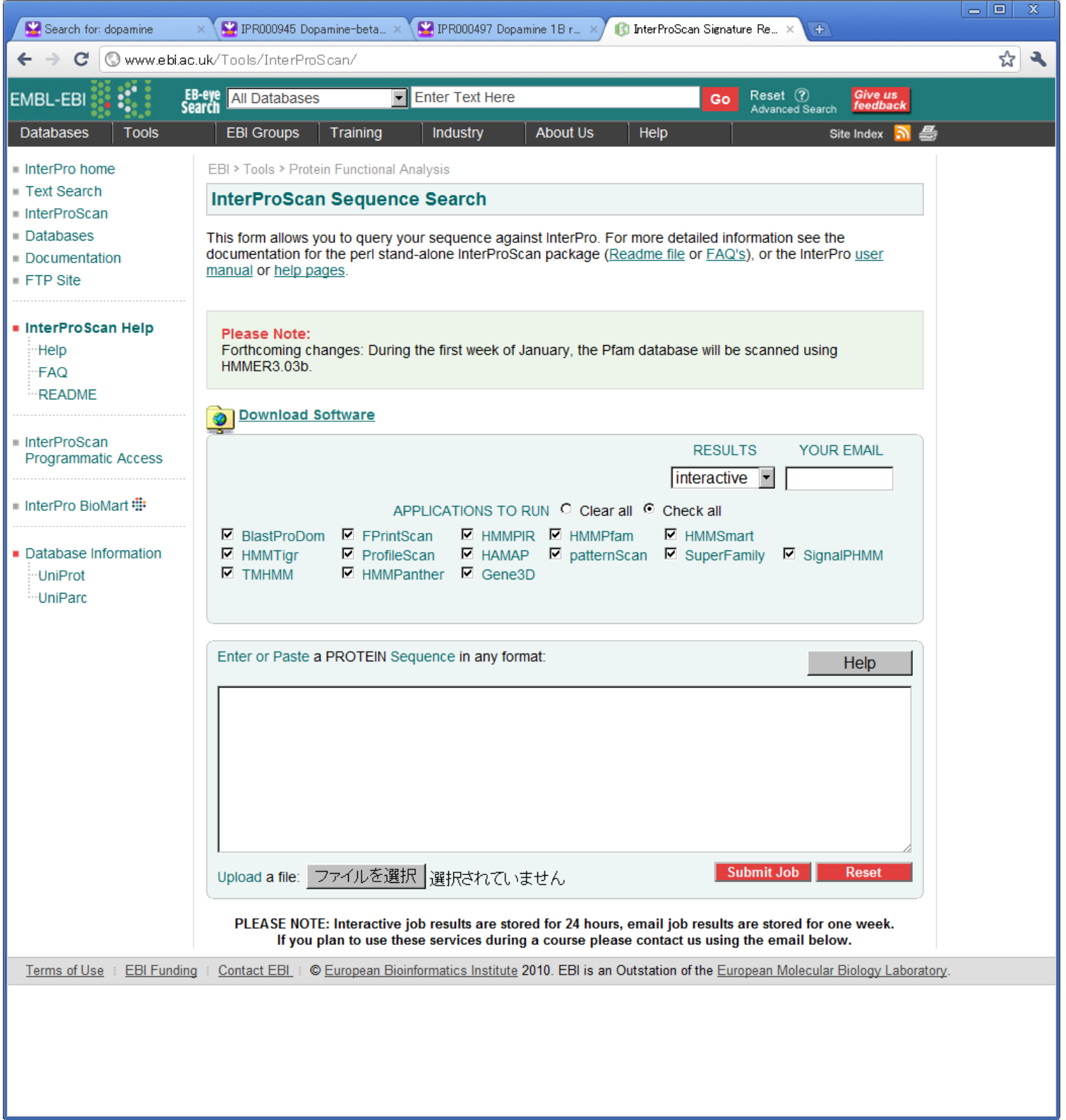
系統ごとに該当するエントリ(DOMON domain, IPR013050)が含まれているタンパク質の数が表示されている。
最も内側の同心円が最も深い分類群を表す。
Eukaryota(440)+Archaea(4)+Bacteria(11)=455

ベン図:IPR01350に含まれる455タンパク質(ピンク部分)のうち、434(紫部分)はIPR005018にも含まれる。
帯グラフ:両者に共通する434タンパク質は平均109アミノ酸が重なり、33アミノ酸は重ならずにIPR005018のみに見られる。

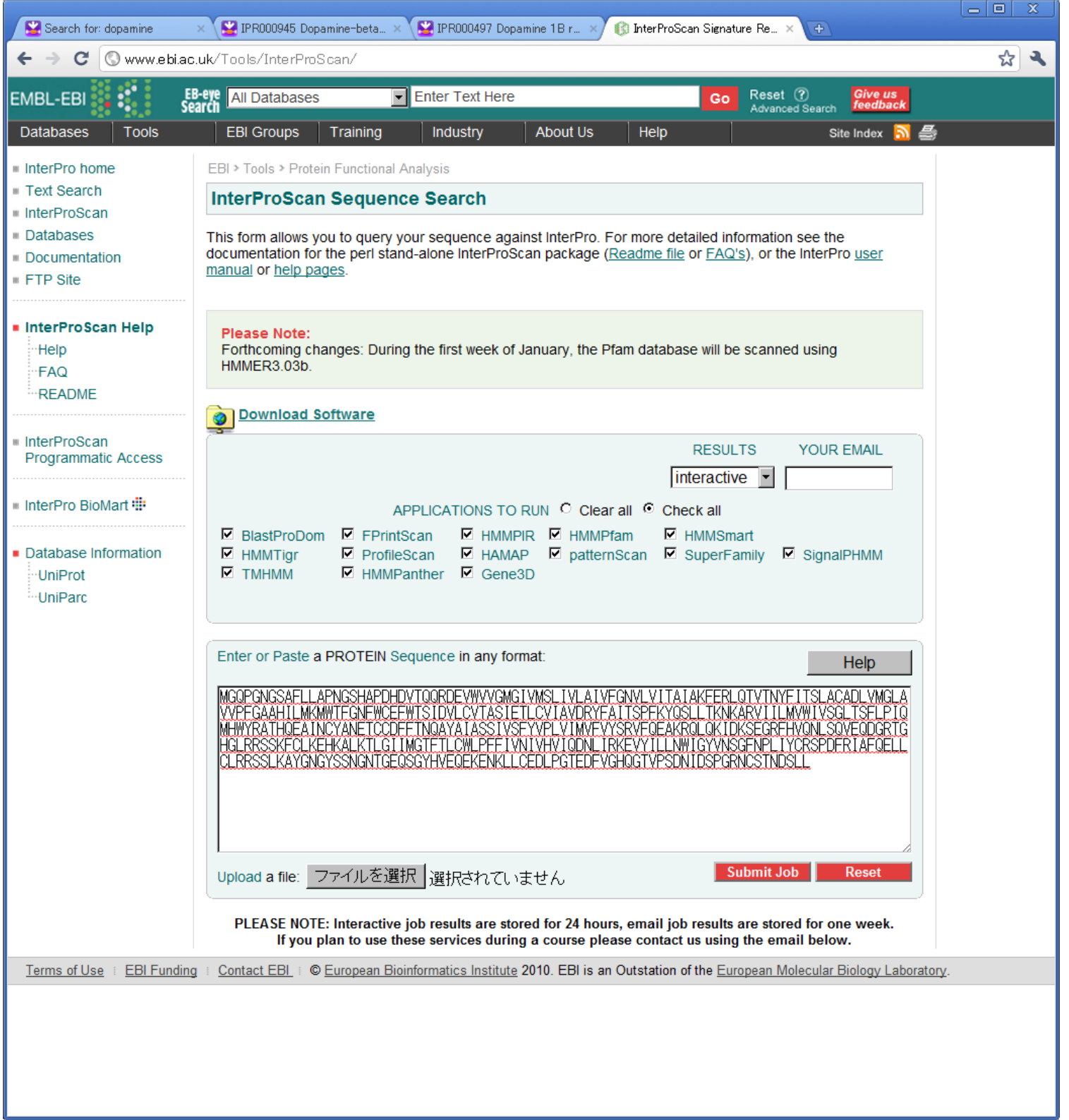
InterProを用いた配列解析ツール InterProScan

- InterProに登録された各種機能ドメインを検索する。
- 機能未知配列中の各種機能関連sequence signatureを抽出して、特徴を発見する。
- ツールをダウンロードしてローカルで走らせることも可能。
- http://www.ebi.ac.uk/Tools/InterProScan/

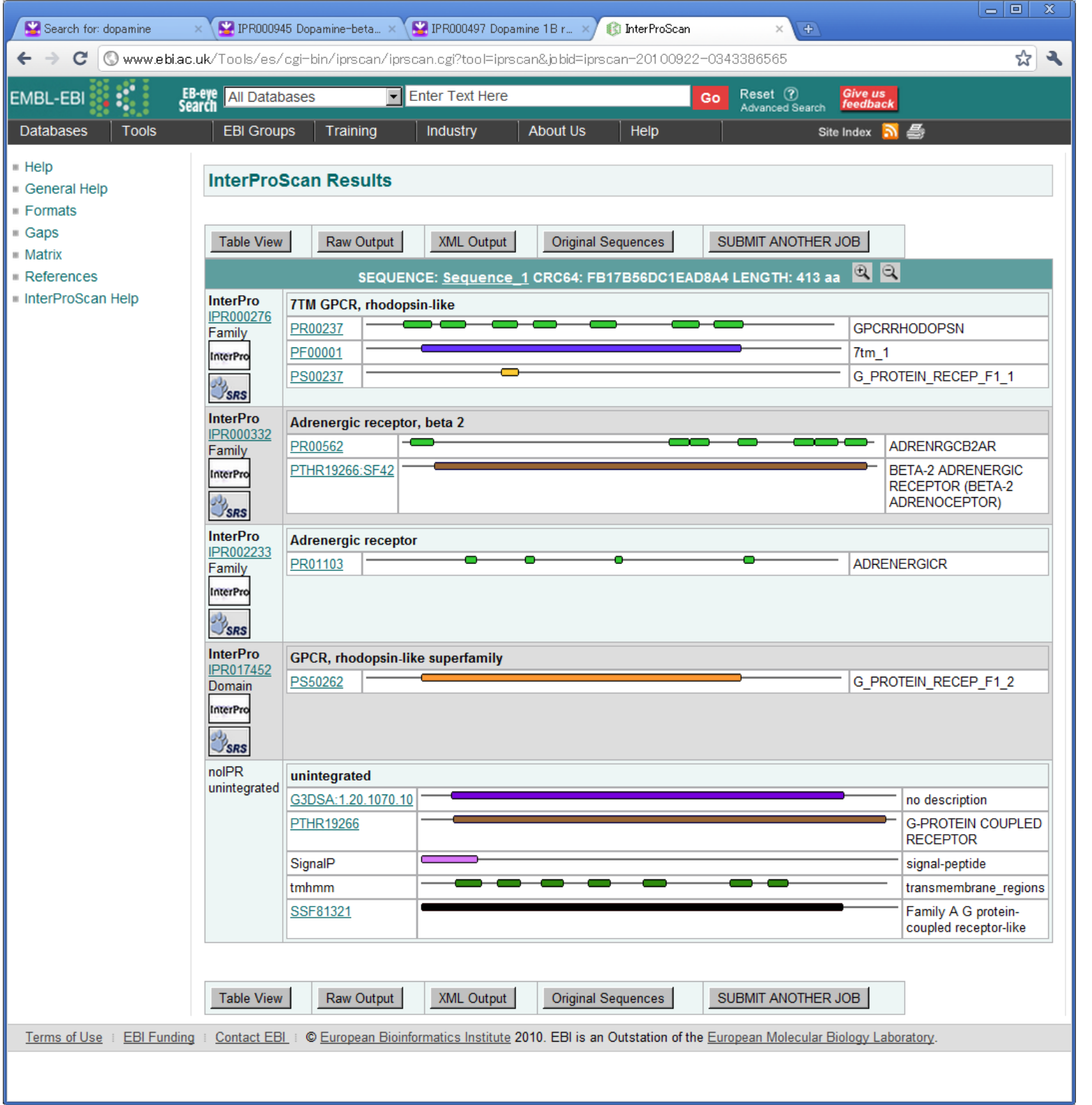
InterProScan トップ画面



配列をペーストして、実行



InterPro 実行結果



タンパク質立体構造のデータベース

http://www.wwpdb.org/

PDBとは

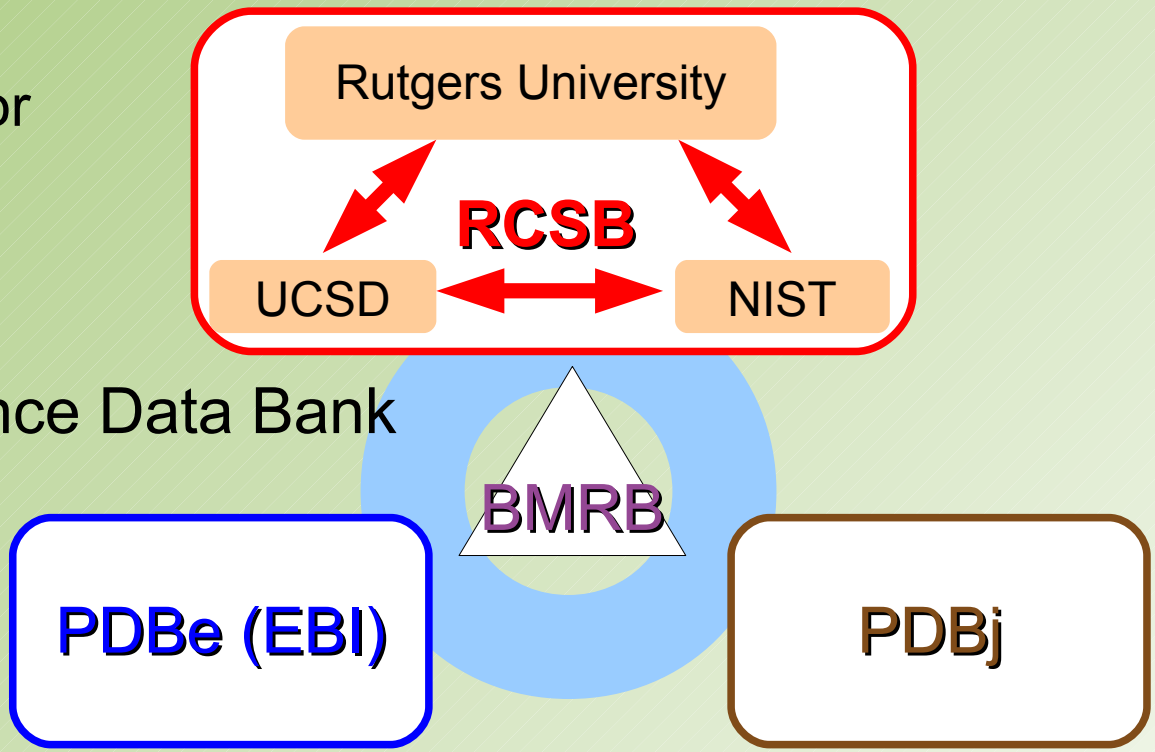
- タンパク質および核酸の三次元構造のデータベース
- X線結晶解析法やNMR(核磁気共鳴）法実験によって得られた三次元データを登録者本人とアノテータとでそれぞれ審査・検証を経て公開される。
- wwPDBにおける共同関係は核酸DBにおける国際塩基配列データベース(INSDC)に似ている。
  - データ・アーカイブは唯一で共通である。
  - 各機関はそれぞれにデータの登録受付とおよびブラウザ、ツール、webサービスを開発し公開する。

wwPDBの組織

Research Collaboratory for Structural Bioinformatics (RCSB), USA

Biological Magnetic Resonance Data Bank (BMRB), USA

Protein Databank in Europe (PDBe), Europe



日本蛋白質構造データバンク (PDBj), Japan

PDBデータの検索・取得

- http://www.pdbj.org/の「検索」を使う。
- 詳しい使い方は以下（今年8月名古屋大情報文化学部棟での講習会資料）参照のこと。
- http://www.pdbj.org/workshop/201008/Kinjo.pdf
- ただし、PDBのIDが既知である場合の検索の仕方で説明してある。未知の場合でもキーワードで検索可能。

PDBデータのキーワード検索1

情報リッチな画面であっても、大概のDBは検索窓がトップページにあるので、落ち着いて探してみよう。

PDBデータのキーワード検索2

分子名称が同一であっても、著者や実験手法（つまりエントリー）によって、3Dデータが異なるので注意。
サムネール画像の向きが違うだけかもしれない。

entry:エントリー、エントリ。ひとつの分子種につき、ひとつの実験で得られた立体構造情報。

個別データ表示例


service.pdbj.org/mine/Detail?PDBID=2U2F&PAGEID=Summary
PDBj
Japanese
統計情報 ヘルプ FAQ お問い合わせ
トップページ
データ登録 >>
ADIT: PDB Deposition
ADIT-NMR
検索 >>
Search PDB (Mine/xPSSS)
Latest Released Search
Sequence-Navigator
Structure-Navigator
SeSAW
Ligand Binding Sites (GIRAF)
EM Navigator
Search NMR Data (BMRB)
Status Search
サービス&ソフトウェア >>
jV: Graphic Viewer
Protein Globe
ASH
MAFFTash
Structure Prediction >>
CRNPRED
Spanner
SFAS
二次データベース >>
eF-site/eF-seek/eF-surf
eProtS
ProMode
Molecule of the Month
ダウンロード >>
PDB Archive/Snapshot Archive
リンク集
日本蛋白質構造データバンク(PDBj: Protein Data Bank Japan)は、JST-BIRDの支援を受け、米国RCSB、BMRB、および欧州PDBeと協力して、生体高分子の立体構造データベースを国際的に統一化されたアーカイブとして運営するとともに、様々な解析ツールを提供しております。
Mine
概要[2u2f]
日本語ページについて
PDBj Mineについて
更新情報
概 要
構造情報
実験情報
機能情報
相同蛋白質
ダウンロード/画面表示
外部データベース
検索
3次元構造ビューア
jV3 / Jmol
jV3と Jmol には Java(TM)Plug-in 1.5以上が必要です。
x軸周りに90度回転
250X250 500X500
y軸周りに90度回転
250X250 500X500
エントリーID (PDB ID)
2u2f 配列情報 (FASTA形式)
PDBファイルのダウンロード
関連構造のPDB ID
1u2f
分子名称
SPLICING FACTOR U2AF 65 KD SUBUNIT
タイトル
SOLUTION STRUCTURE OF THE SECOND RNA-BINDING DOMAIN OF HU2AF65
機能のキーワード
SPLICING, U2 SNRNP, RBD, RNA-BINDING PROTEIN, RIKEN Structural Genomics/Proteomics Initiative, RSGI, Structural Genomics
由来する生物種
Homo sapiens (human)
細胞内の位置
[UNP - U2AF2_HUMAN] Nucleus
由来する組織
[UNP - U2AF2_HUMAN] Lymph, and Skin
[UNP - U2AF2_HUMAN] Colon carcinoma
[UNP - U2AF2_HUMAN] Hepatocyte
[UNP - U2AF2_HUMAN] Epithelium
ポリマー鎖の合計数
1
分子量の合計
9096.5 (詳細は 構造情報のページ)
著者
Ito, T. , Muto, Y. , Green, M.R. , Yokoyama, S. , RIKEN Structural Genomics/Proteomics Initiative (RSGI) (登録日 : 1999-05-26, 公開日 : 1999-08-20)
引用文献
Ito, T. , Muto, Y. , Green, M.R. , Yokoyama, S.
Solution structures of the first and second RNA-binding domains of human U2 small nuclear ribonucleoprotein particle auxiliary factor (U2AF(65)).
MP900433188.JPG
MP900438881.JPG
MP900438753.JPG
すべてのダウンロードを表示

PDBのentry dataには3種類の書式がある

書式	説明	完全性	人による判別	プログラムでの処理
PDB	Flat file形式。情報に欠落があり、厳密な解析には向かない。	－	＋＋	－
mmCIF	国際結晶学連合(IUCr)が定めた(Crystallographic Information File) CIF形式を拡張した。http://mmcif.pdb.org/ Data parsing tools(C++)がRCSBから用意されている(http://sw-tools.rcsb.org/)。	＋	＋	＋
PDBML	mmCIFに基づくXML形式。http://pdbml.pdb.org/	＋	－	＋＋

どの形式でもダウンロードできる。

PDB書式の例（ヘッダー部分抜粋）

field名

同一field内で2行以上になるときには行番号がついている。

```
HEADER    RNA BINDING PROTEIN                     08-AUG-06   2HZC
TITLE     CRYSTAL STRUCTURE OF THE N-TERMINAL RRM OF THE U2AF LARGE
TITLE    2 SUBUNIT
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: SPLICING FACTOR U2AF 65 KDA SUBUNIT;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: RRM 1;
COMPND   5 SYNONYM: U2 AUXILIARY FACTOR 65 KDA SUBUNIT, U2 SNRNP
COMPND   6 AUXILIARY FACTOR LARGE SUBUNIT, HU2AF65;
COMPND   7 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGANISM_TAXID: 9606;
SOURCE   5 GENE: U2AF2, U2AF65;
```

同一field内で2行以上になるときには行番号がついている。

```
JRNL        AUTH   K.R.THICKMAN,E.A.SICKMIER,C.L.KIELKOPF
JRNL        TITL   ALTERNATIVE CONFORMATIONS AT THE RNA-BINDING
JRNL        TITL 2 SURFACE OF THE N-TERMINAL U2AF(65) RNA RECOGNITION
JRNL        TITL 3 MOTIF.

REMARK   1 REFERENCE 1
REMARK   1  AUTH   E.A.SICKMIER,K.E.FRATO,S.PARANAWITHANA,H.SHEN,
REMARK   1  AUTH 2 M.R.GREEN,C.L.KIELKOPF
REMARK   1  REF    MOL.CELL                      V.  23    49 2006
REMARK   2 RESOLUTION.    1.47 ANGSTROMS.
```

'REMARK 数字' で一field扱い

PDB書式の例(分子データ部分抜粋)

アミノ酸番号

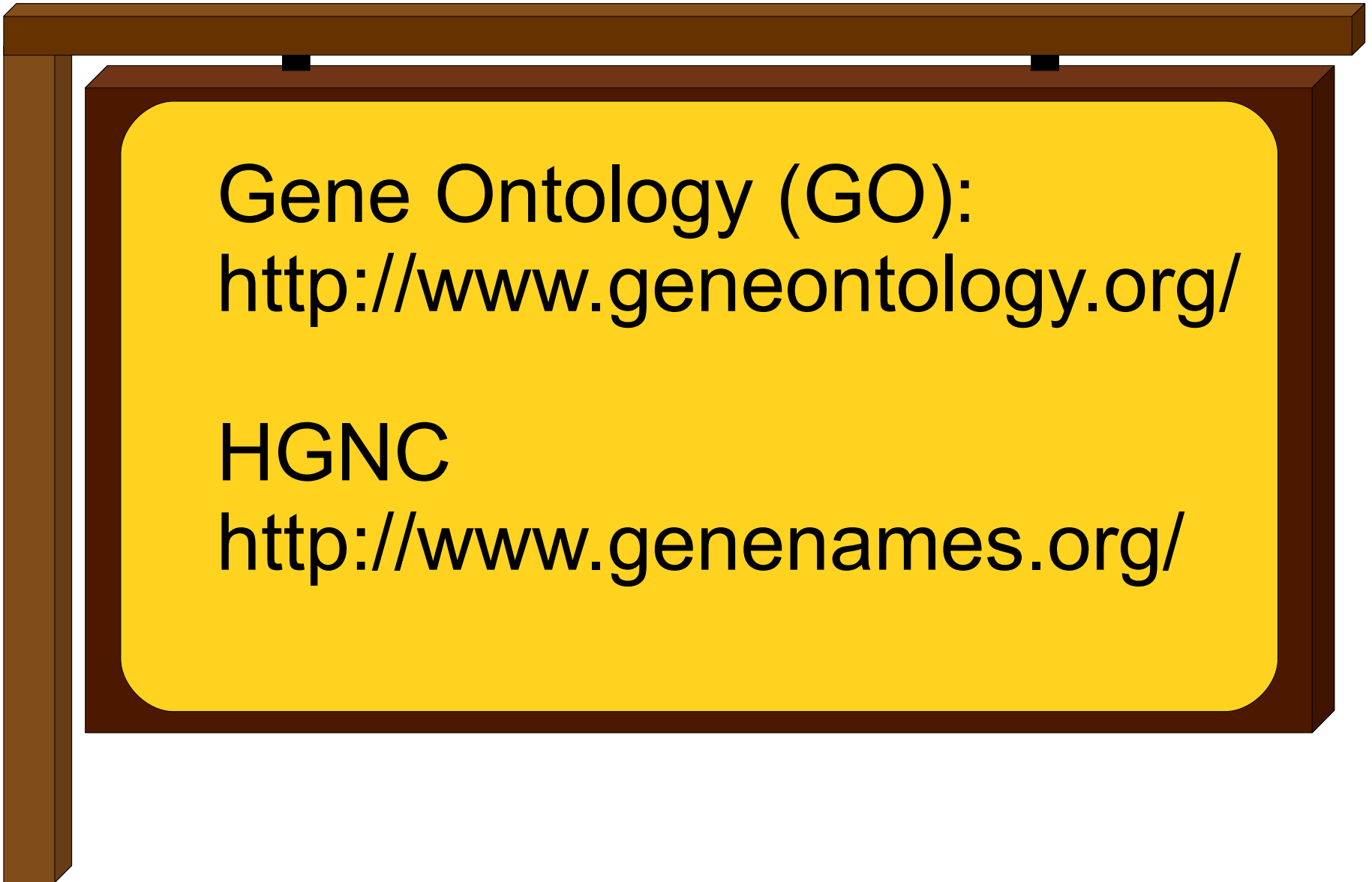
```
ATOM      1  N   GLY A 143       1.397  10.715  63.542  1.00 79.91           N
ANISOU    1  N   GLY A 143    17305   4118   8949   3151  -2891  -3871        N
ATOM      2  CA  GLY A 143       0.525  10.264  62.474  1.00 76.48           C
ANISOU    2  CA  GLY A 143    14905   5306   8858   3443  -2947  -2851        C
```

```
HETATM  925  O   HOH A 492      21.491  -8.646  69.133  1.00 45.06           O
ANISOU  925  O   HOH A 492     5806   4899   6418    309   -542  -1086        O
HETATM  926  O   HOH A 493      11.582  -0.505  36.484  1.00 38.57           O
ANISOU  926  O   HOH A 493     5534   4232   4885   -492  -1658   -847        O
HETATM  927  O   HOH A 494      28.316  -8.557  58.005  1.00 54.25           O
ANISOU  927  O   HOH A 494     7247   5534   7837   2015   1449  -1110        O
HETATM  928  O   HOH A 495      27.091  -6.522  50.635  1.00 41.14           O
ANISOU  928  O   HOH A 495     5567   4645   5421    851   1342    163        O
CONECT  163  715
CONECT  174  716
CONECT  207  716

MASTER      324    0    3    4    6    0    7    6  927    1   28    7
END
```

The ATOM records present the atomic coordinates for standard amino acids and nucleotides.
The ANISOU records present the anisotropic temperature factors
詳細はhttp://www.wwpdb.org/docs.html

名称、用語、概念の共通化に関するデータベース



Gene Ontologyとは

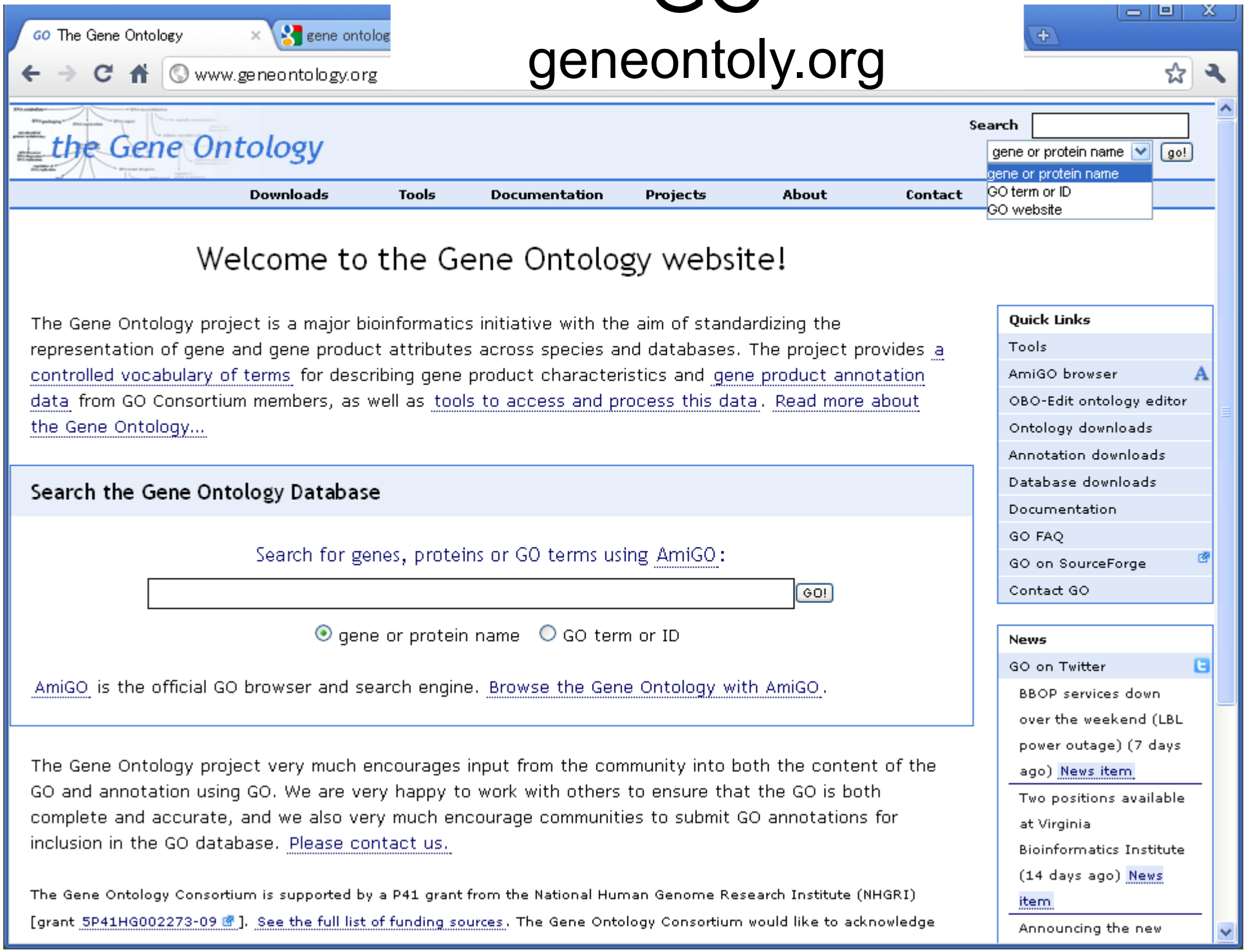
- “Ontology”＝”on”存在＋”logy”
  - 元々、哲学のひとつ存在論：entity（実体）の存在のあり方や他の実体との関係性（階層性、分類・体系性）について問う学問。
  - 情報学において、「概念化の明示的な仕様」と定義。
    - 同表記異義語の問題を解決、
    - 文章：単語の集まり-->意味のある実体として、コンピュータ処理が可能となった。（発展；セマンティック・ウェッブ）
- Gene Ontology Project:生物種やDBさらに分野の枠を超え、遺伝子（産物）関連用語を標準化。
  - 異祖同機能の問題解決、
  - -->DB間でのリンクや統合が容易に。
  - おもなゲノム関連研究機関が参加。
- GOは網羅的な遺伝子解析結果集計に多用。

GO termの構造

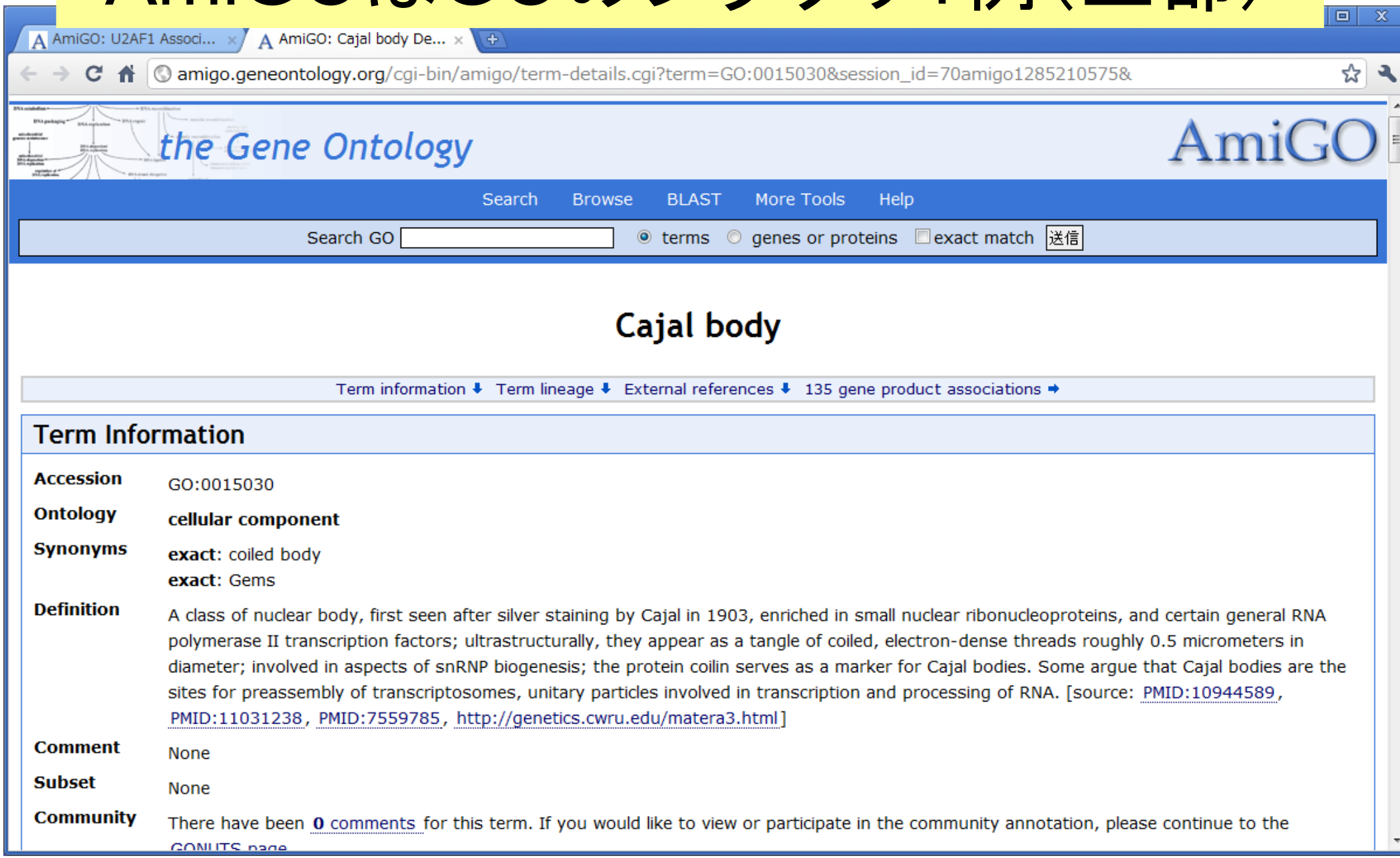
- GO term全体に階層性を持つ。
- GOは3つのdomainをカバーする。
  - cellular component, 遺伝子産物細胞内外分布
  - molecular function, 遺伝子産物の機能
  - biological process, 生体内における役割

例 :
Accession GO:0015030
Ontology cellular component

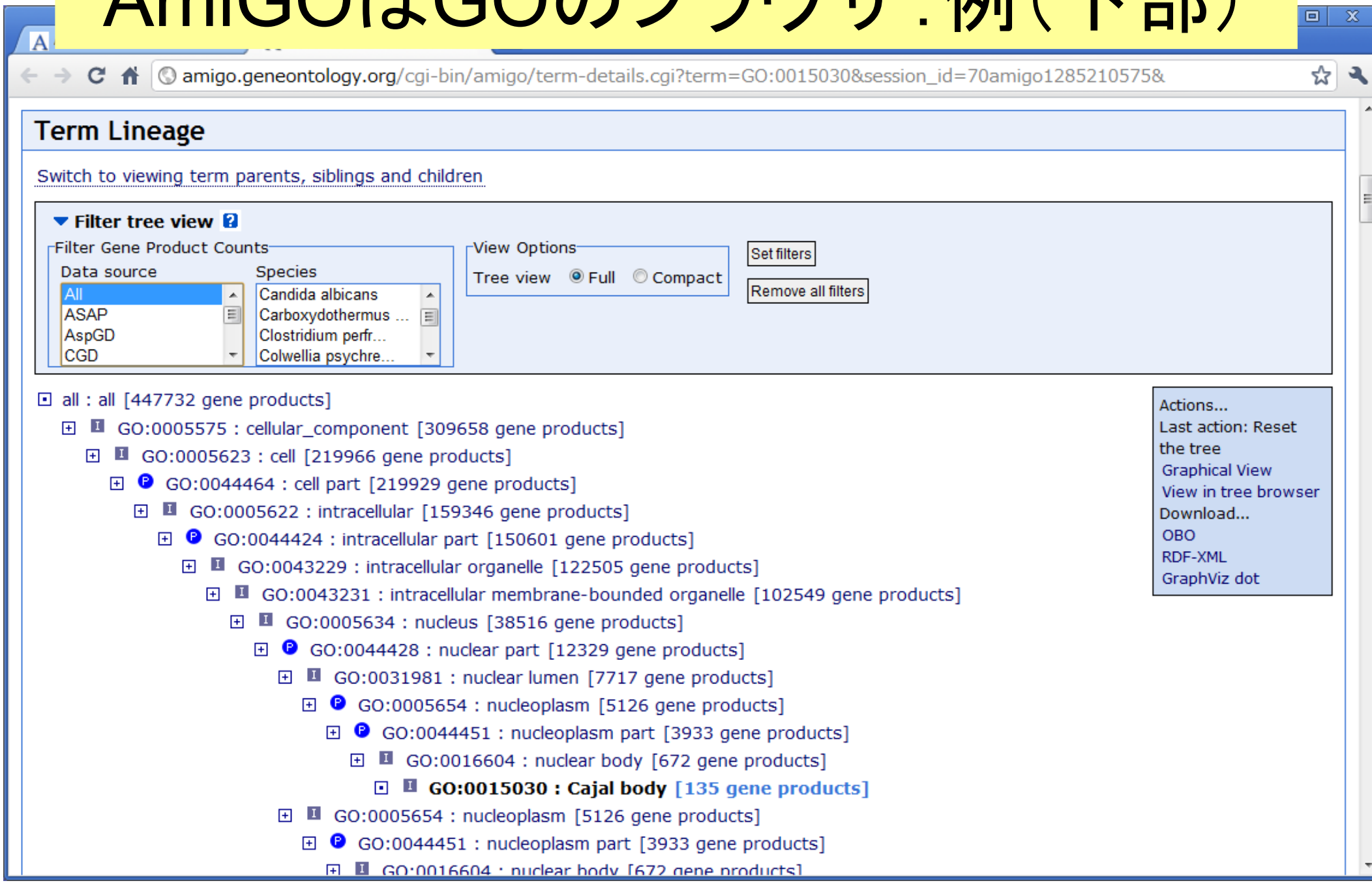
GO
geneontoly.org

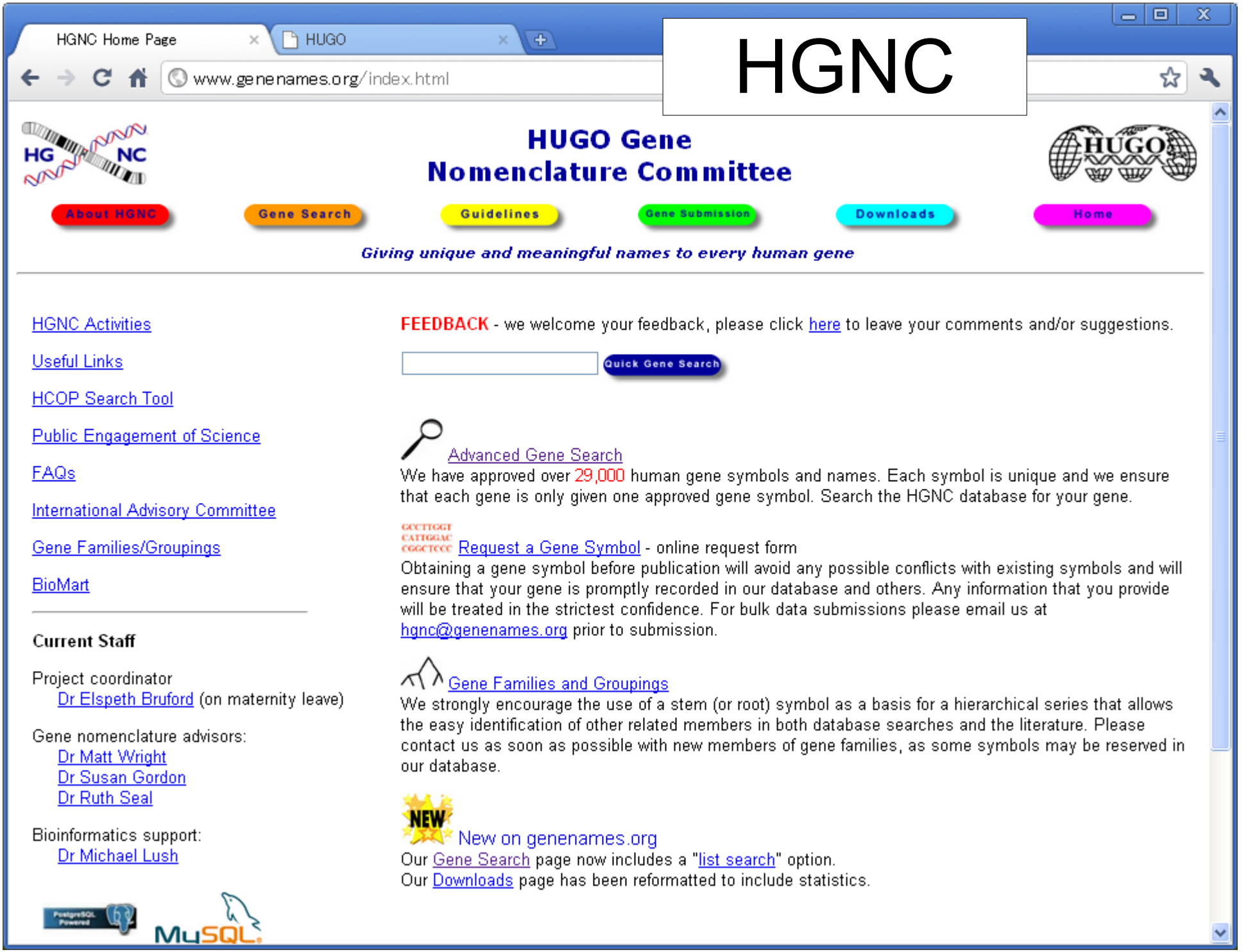


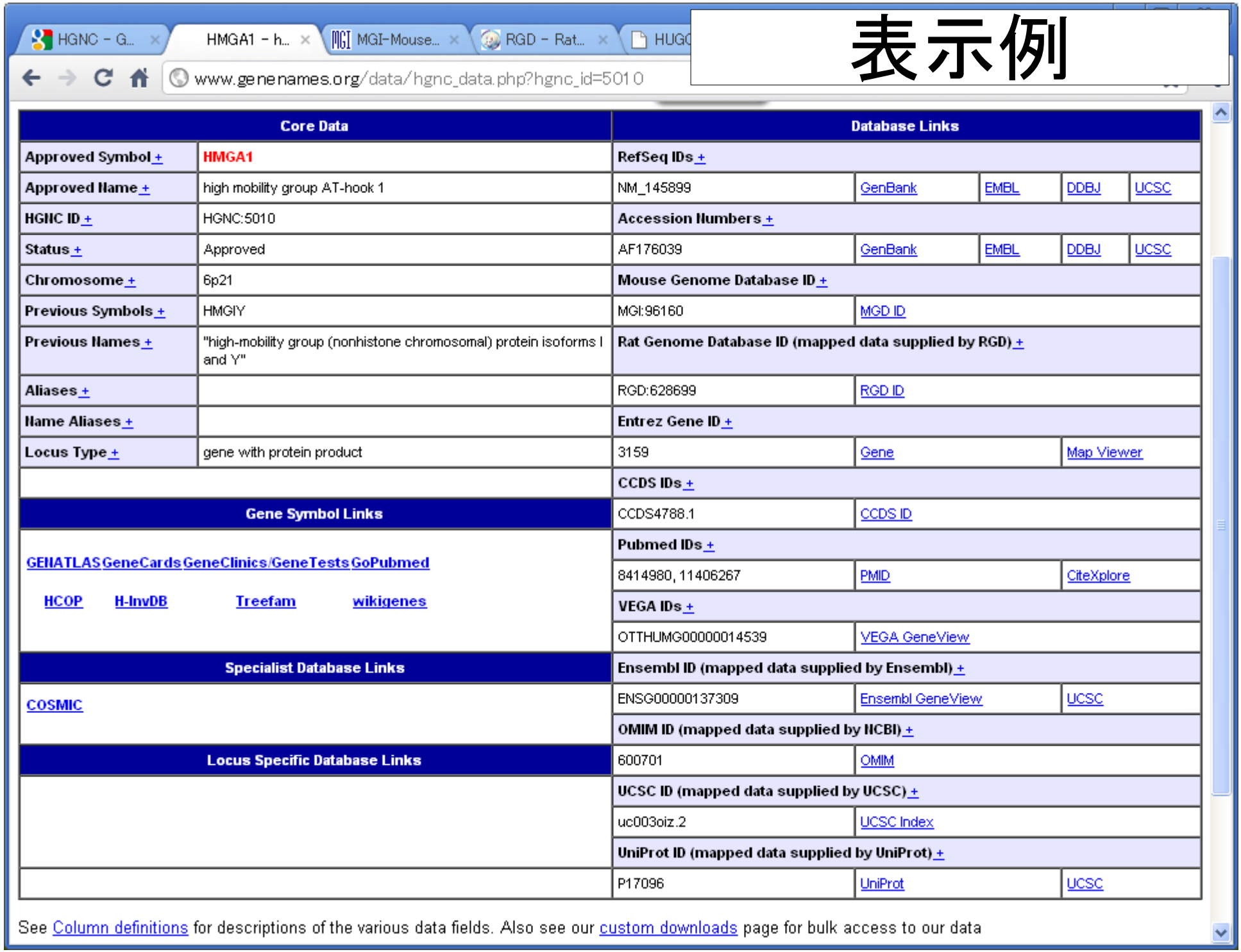
AmiGOはGOのブラウザ：例（上部）



AmiGOはGOのブラウザ：例（下部）



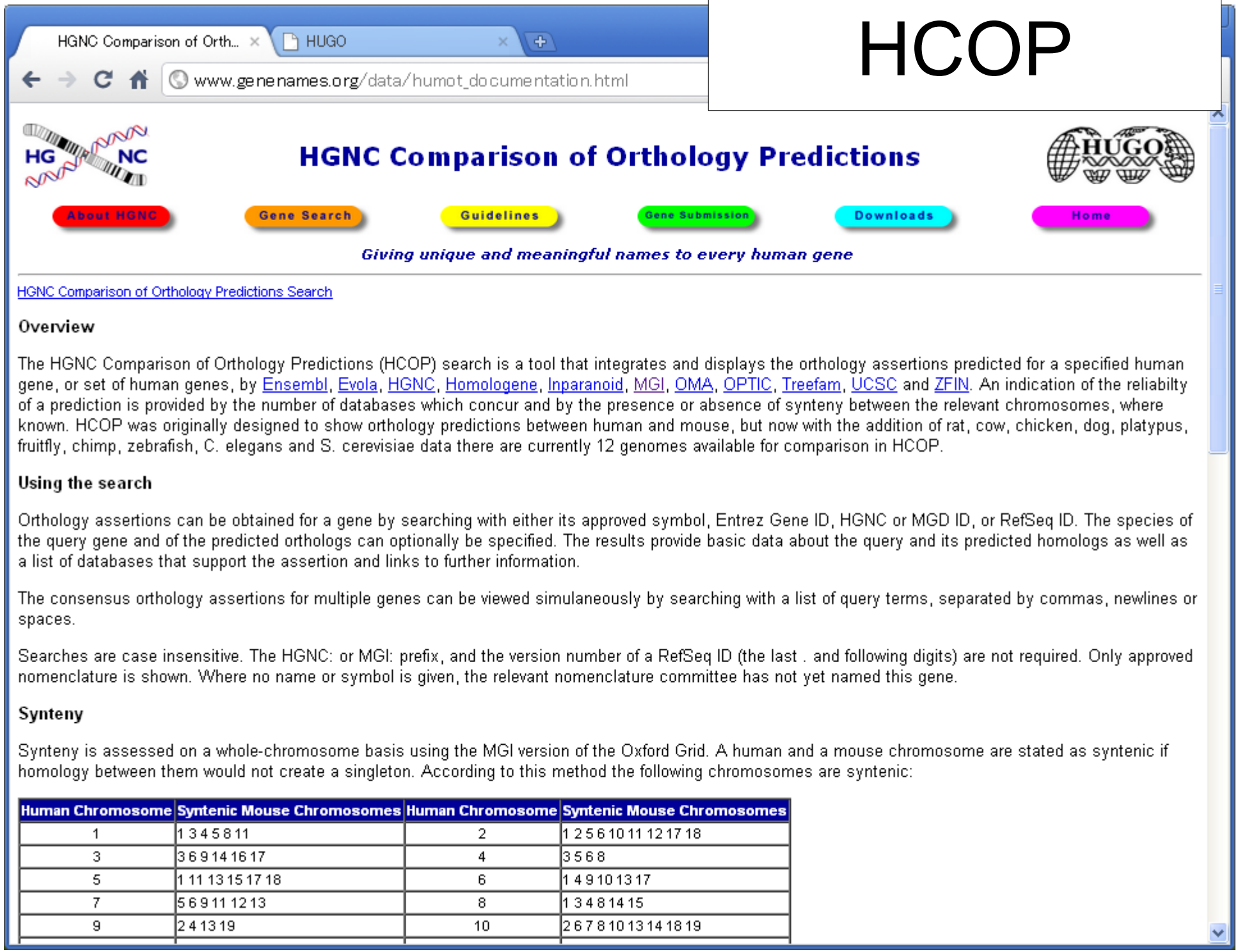

HGNC Home Page
HUGO
www.genenames.org/index.html
HGNC
HUGO Gene Nomenclature Committee
About HGNC
Gene Search
Guidelines
Gene Submission
Downloads
Home
Giving unique and meaningful names to every human gene
HGNC Activities
Useful Links
HCOP Search Tool
Public Engagement of Science
FAQs
International Advisory Committee
Gene Families/Groupings
BioMart
Current Staff
Project coordinator
Dr Elspeth Bruford (on maternity leave)
Gene nomenclature advisors:
Dr Matt Wright
Dr Susan Gordon
Dr Ruth Seal
Bioinformatics support:
Dr Michael Lush
FEEDBACK - we welcome your feedback, please click here to leave your comments and/or suggestions.
Quick Gene Search
Advanced Gene Search
We have approved over 29,000 human gene symbols and names. Each symbol is unique and we ensure that each gene is only given one approved gene symbol. Search the HGNC database for your gene.
Request a Gene Symbol - online request form
Obtaining a gene symbol before publication will avoid any possible conflicts with existing symbols and will ensure that your gene is promptly recorded in our database and others. Any information that you provide will be treated in the strictest confidence. For bulk data submissions please email us at hgnc@genenames.org prior to submission.
Gene Families and Groupings
We strongly encourage the use of a stem (or root) symbol as a basis for a hierarchical series that allows the easy identification of other related members in both database searches and the literature. Please contact us as soon as possible with new members of gene families, as some symbols may be reserved in our database.
NEW
New on genenames.org
Our Gene Search page now includes a "list search" option.
Our Downloads page has been reformatted to include statistics.



Core Data		Database Links				
Approved Symbol +	HMGA1	RefSeq IDs +				
Approved Name +	high mobility group AT-hook 1	NM_145899	GenBank	EMBL	DDBJ	UCSC
HGNC ID +	HGNC:5010	Accession Numbers +				
Status +	Approved	AF176039	GenBank	EMBL	DDBJ	UCSC
Chromosome +	6p21	Mouse Genome Database ID +				
Previous Symbols +	HMGIY	MGI:96160	MGD ID			
Previous Names +	"high-mobility group (nonhistone chromosomal) protein isoforms I and Y"	Rat Genome Database ID (mapped data supplied by RGD) +				
Aliases +		RGD:628699	RGD ID			
Name Aliases +		Entrez Gene ID +				
Locus Type +	gene with protein product	3159	Gene		Map Viewer	
		CCDS IDs +				
Gene Symbol Links		CCDS4788.1	CCDS ID			
GENATLAS GeneCards GeneClinics/GeneTests GoPubmed HCOP H-InvDB Treefam wikigenes		Pubmed IDs +				
		8414980, 11406267	PMID		CiteXplore	
		VEGA IDs +				
		OTTHUMG00000014539	VEGA GeneView			
Specialist Database Links		Ensembl ID (mapped data supplied by Ensembl) +				
COSMIC		ENSG00000137309	Ensembl GeneView		UCSC	
		OMIM ID (mapped data supplied by NCBI) +				
Locus Specific Database Links		600701	OMIM			
		UCSC ID (mapped data supplied by UCSC) +				
		uc003oiz.2	UCSC Index			
		UniProt ID (mapped data supplied by UniProt) +				
		P17096	UniProt		UCSC	

See Column definitions for descriptions of the various data fields. Also see our custom downloads page for bulk access to our data

HCOP



HGNC Comparison of Orthology Predictions

About HGNC | Gene Search | Guidelines | Gene Submission | Downloads | Home

Giving unique and meaningful names to every human gene

HGNC Comparison of Orthology Predictions Search

Overview

The HGNC Comparison of Orthology Predictions (HCOP) search is a tool that integrates and displays the orthology assertions predicted for a specified human gene, or set of human genes, by Ensembl, Evola, HGNC, Homologene, Inparanoid, MGI, OMA, OPTIC, Treefam, UCSC and ZFIN. An indication of the reliabilty of a prediction is provided by the number of databases which concur and by the presence or absence of synteny between the relevant chromosomes, where known. HCOP was originally designed to show orthology predictions between human and mouse, but now with the addition of rat, cow, chicken, dog, platypus, fruitfly, chimp, zebrafish, C. elegans and S. cerevisiae data there are currently 12 genomes available for comparison in HCOP.

Using the search

Orthology assertions can be obtained for a gene by searching with either its approved symbol, Entrez Gene ID, HGNC or MGD ID, or RefSeq ID. The species of the query gene and of the predicted orthologs can optionally be specified. The results provide basic data about the query and its predicted homologs as well as a list of databases that support the assertion and links to further information.

The consensus orthology assertions for multiple genes can be viewed simulaneously by searching with a list of query terms, separated by commas, newlines or spaces.

Searches are case insensitive. The HGNC: or MGI: prefix, and the version number of a RefSeq ID (the last . and following digits) are not required. Only approved nomenclature is shown. Where no name or symbol is given, the relevant nomenclature committee has not yet named this gene.

Synteny

Synteny is assessed on a whole-chromosome basis using the MGI version of the Oxford Grid. A human and a mouse chromosome are stated as syntenic if homology between them would not create a singleton. According to this method the following chromosomes are syntenic:

Human Chromosome	Syntenic Mouse Chromosomes	Human Chromosome	Syntenic Mouse Chromosomes
1	1 3 4 5 8 11	2	1 2 5 6 10 11 12 17 18
3	3 6 9 14 16 17	4	3 5 6 8
5	1 11 13 15 17 18	6	1 4 9 10 13 17
7	5 6 9 11 12 13	8	1 3 4 8 14 15
9	2 4 13 19	10	2 6 7 8 10 13 14 18 19

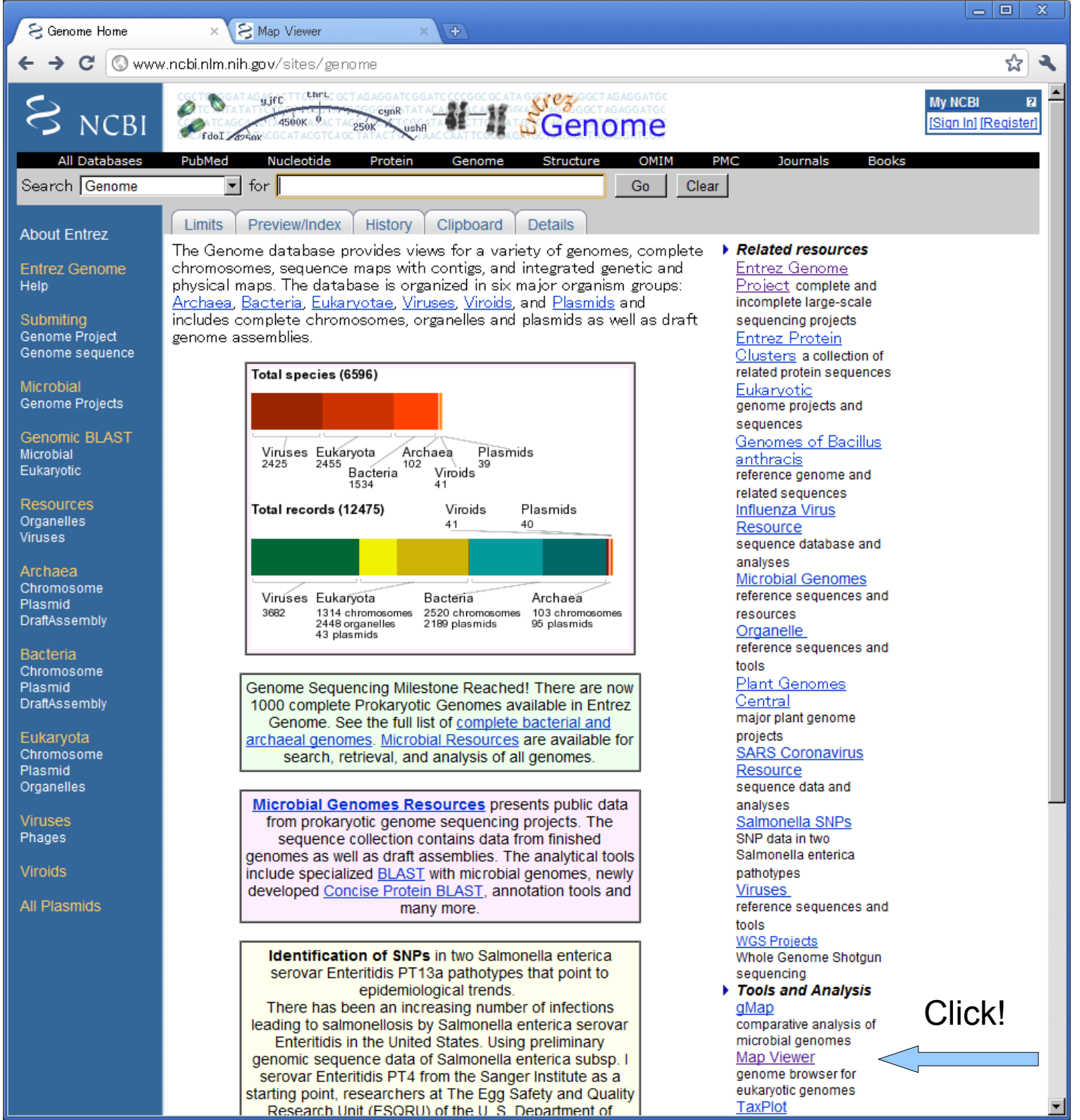
ゲノムDB・ゲノムbrowser

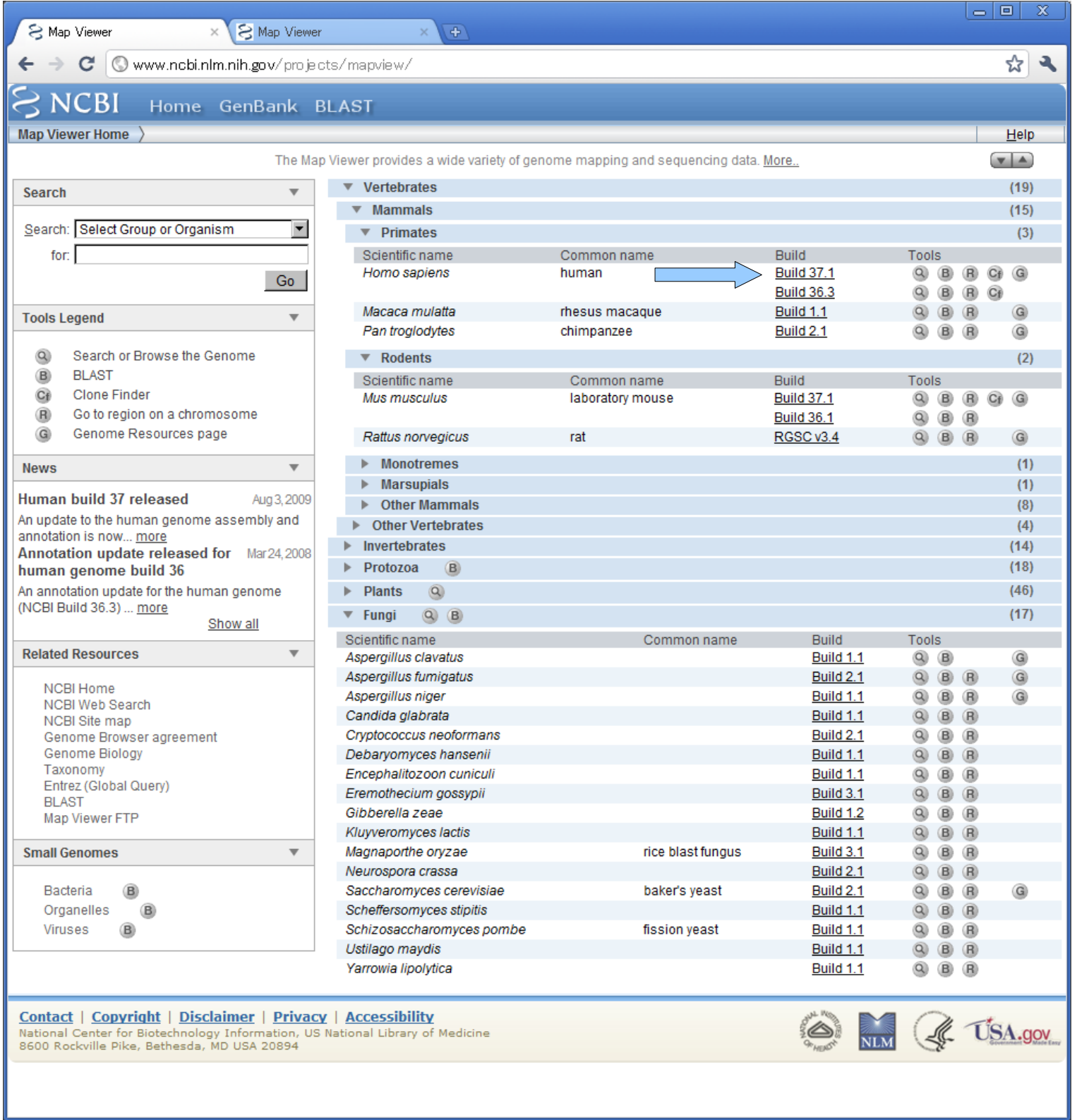
- NCBI
  - NCBI Map viewer
  - Model Maker (mm)
- Ensembl
  - Ensembl Genome browser
  - BioMart
- UCSC
  - UCSC genome browser
- J.Craig Venter Institute (JCVI) HuRef
- バージョンについて

NCBI home


National Center for Biotec...
Map Viewer
www.ncbi.nlm.nih.gov
NCBI Resources How To
My NCBI Sign In
NCBI
National Center for Biotechnology Information
Search All Databases
Search Clear
Resources
NCBI Home
All Resources (A-Z)
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
More about the NCBI | Mission | Organization | Research | RSS
PubMed Central
Free Full Text. Over 1,500,000 articles from over 450 journals. Linked to PubMed and fully searchable.
1 2 3 4
How To...
Determine conserved synteny between the genomes of two organisms
Find a homolog for a gene in another organism
Obtain the full text of an article
Design PCR primers and check them for specificity
See all ...
Popular Resources
BLAST
Bookshelf
Gene
Genome
Nucleotide
OMIM
Protein
PubChem
PubMed
PubMed Central
SNP
NCBI News
MyNCBI supports OpenID and InCommons IDs
22 Sep 2010
A list of supported organizations can be found on the MyNCBI
Personalized settings in My NCBI
30 Aug 2010
Personalized preferences such as record format and results display
More...
You are here: NCBI
Write to the Help Desk
GETTING STARTED
NCBI Help Manual
NCBI Handbook
Training & Tutorials
RESOURCES
Literature
DNA & RNA
Proteins
Sequence Analysis
Genes & Expression
Genomes & Maps
Domains & Structures
Genetics & Medicine
POPULAR
PubMed
Nucleotide
BLAST
PubMed Central
Gene
Bookshelf
Protein
OMIM
Genome
FEATURED
GenBank
Reference Sequences
Map Viewer
Genome Projects
Human Genome
Mouse Genome
Influenza Virus
Primer-BLAST
Sequence Read Archive
NCBI INFORMATION
About NCBI
Research at NCBI
NCBI Newsletter
NCBI FTP Site
www.ncbi.nlm.nih.gov を待機しています...

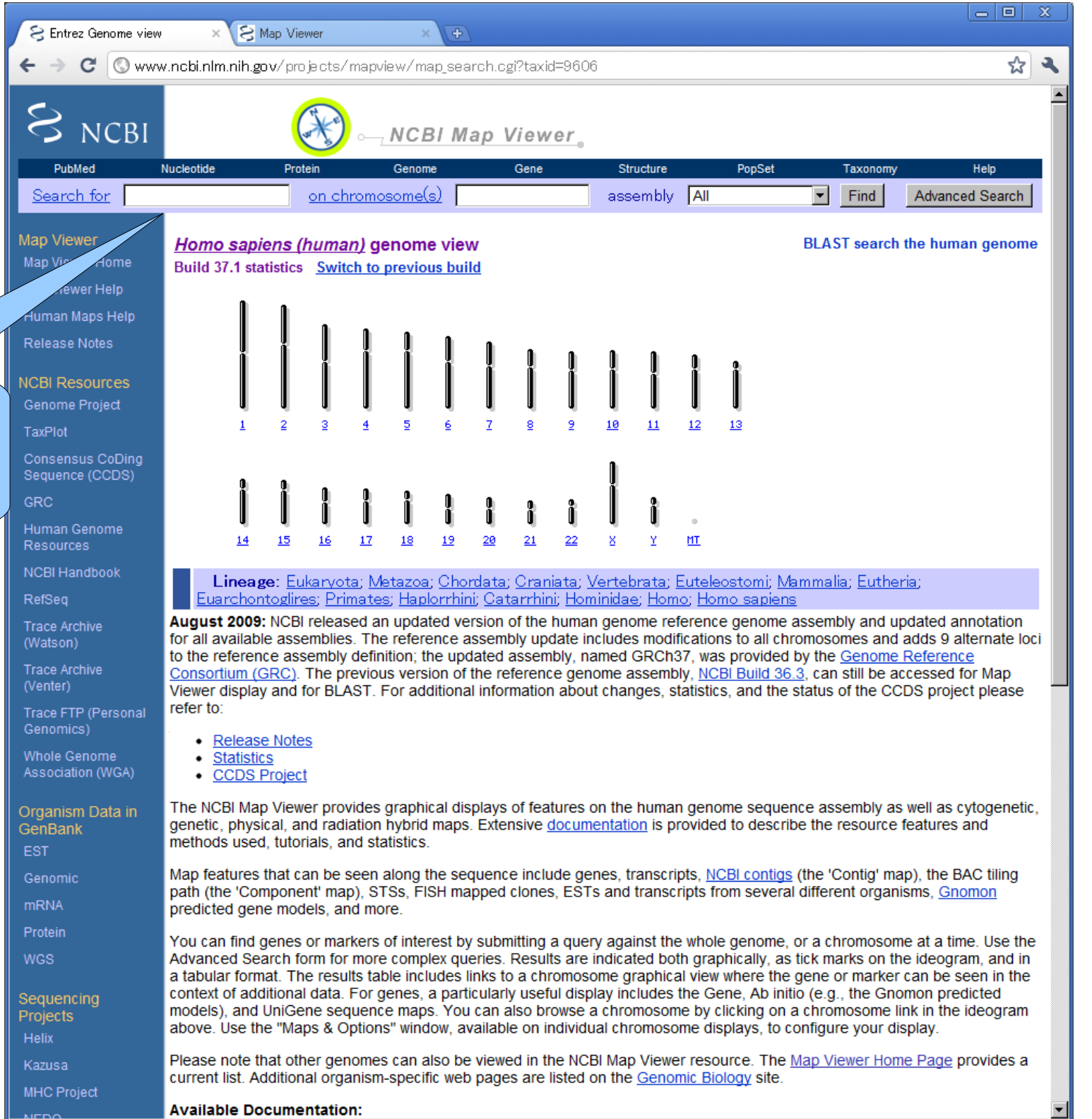
NCBI > genome

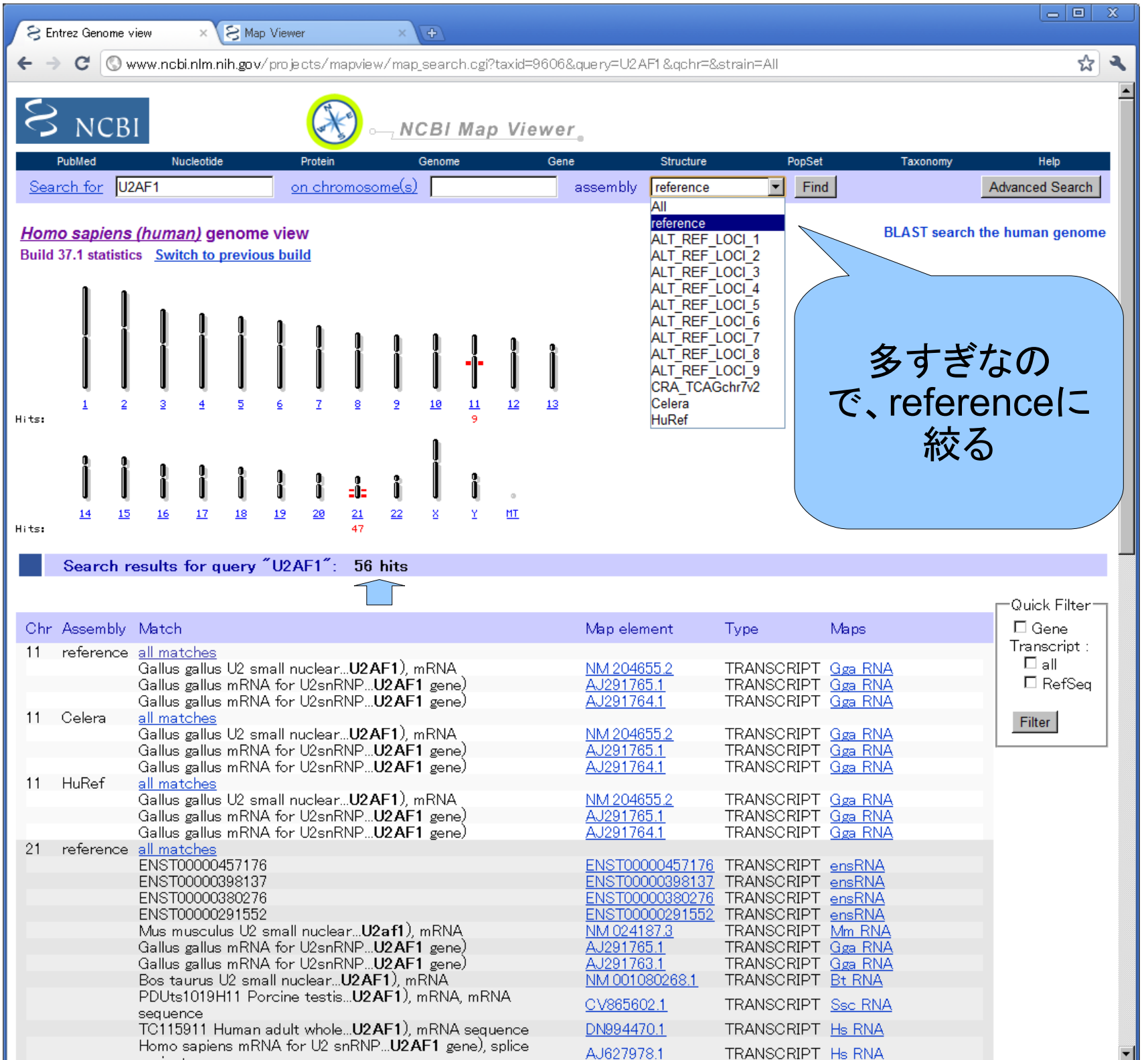

Genome Home
Map Viewer
www.ncbi.nlm.nih.gov/sites/genome
NCBI
Entrez Genome
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search Genome for Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Entrez Genome Help
Submiting Genome Project Genome sequence
Microbial Genome Projects
Genomic BLAST Microbial Eukaryotic
Resources Organelles Viruses
Archaea Chromosome Plasmid DraftAssembly
Bacteria Chromosome Plasmid DraftAssembly
Eukaryota Chromosome Plasmid Organelles
Viruses Phages
Viroids
All Plasmids
The Genome database provides views for a variety of genomes, complete chromosomes, sequence maps with contigs, and integrated genetic and physical maps. The database is organized in six major organism groups: Archaea, Bacteria, Eukaryotae, Viruses, Viroids, and Plasmids and includes complete chromosomes, organelles and plasmids as well as draft genome assemblies.
Total species (6596)
Viruses 2425
Eukaryota 2455
Bacteria 1534
Archaea 102
Viroids 41
Plasmids 39
Total records (12475)
Viruses 3682
Eukaryota 1314 chromosomes 2448 organelles 43 plasmids
Bacteria 2520 chromosomes 2189 plasmids
Archaea 103 chromosomes 95 plasmids
Viroids 41
Plasmids 40
Genome Sequencing Milestone Reached! There are now 1000 complete Prokaryotic Genomes available in Entrez Genome. See the full list of complete bacterial and archaeal genomes. Microbial Resources are available for search, retrieval, and analysis of all genomes.
Microbial Genomes Resources presents public data from prokaryotic genome sequencing projects. The sequence collection contains data from finished genomes as well as draft assemblies. The analytical tools include specialized BLAST with microbial genomes, newly developed Concise Protein BLAST, annotation tools and many more.
Identification of SNPs in two Salmonella enterica serovar Enteritidis PT13a pathotypes that point to epidemiological trends.
There has been an increasing number of infections leading to salmonellosis by Salmonella enterica serovar Enteritidis in the United States. Using preliminary genomic sequence data of Salmonella enterica subsp. I serovar Enteritidis PT4 from the Sanger Institute as a starting point, researchers at The Egg Safety and Quality Research Unit (ESQRU) of the U. S. Department of
Related resources
Entrez Genome Project complete and incomplete large-scale sequencing projects
Entrez Protein Clusters a collection of related protein sequences
Eukaryotic genome projects and sequences
Genomes of Bacillus anthracis reference genome and related sequences
Influenza Virus Resource sequence database and analyses
Microbial Genomes reference sequences and resources
Organelle reference sequences and tools
Plant Genomes Central major plant genome projects
SARS Coronavirus Resource sequence data and analyses
Salmonella SNPs SNP data in two Salmonella enterica pathotypes
Viruses reference sequences and tools
WGS Projects Whole Genome Shotgun sequencing
Tools and Analysis
gMap comparative analysis of microbial genomes
Map Viewer genome browser for eukaryotic genomes
TaxPlot
Click!


Map Viewer
Map Viewer
www.ncbi.nlm.nih.gov/projects/mapview/
NCBI Home GenBank BLAST
Map Viewer Home
Help
The Map Viewer provides a wide variety of genome mapping and sequencing data. More..
Search
Search: Select Group or Organism
for:
Go
Tools Legend
Search or Browse the Genome
BLAST
Clone Finder
Go to region on a chromosome
Genome Resources page
News
Human build 37 released Aug 3, 2009
An update to the human genome assembly and annotation is now... more
Annotation update released for human genome build 36 Mar 24, 2008
An annotation update for the human genome (NCBI Build 36.3) ... more
Show all
Related Resources
NCBI Home
NCBI Web Search
NCBI Site map
Genome Browser agreement
Genome Biology
Taxonomy
Entrez (Global Query)
BLAST
Map Viewer FTP
Small Genomes
Bacteria
Organelles
Viruses
Vertebrates (19)
Mammals (15)
Primates (3)
Scientific name Common name Build Tools
Homo sapiens human Build 37.1
Build 36.3
Macaca mulatta rhesus macaque Build 1.1
Pan troglodytes chimpanzee Build 2.1
Rodents (2)
Scientific name Common name Build Tools
Mus musculus laboratory mouse Build 37.1
Build 36.1
Rattus norvegicus rat RGSC v3.4
Monotremes (1)
Marsupials (1)
Other Mammals (8)
Other Vertebrates (4)
Invertebrates (14)
Protozoa (18)
Plants (46)
Fungi (17)
Scientific name Common name Build Tools
Aspergillus clavatus Build 1.1
Aspergillus fumigatus Build 2.1
Aspergillus niger Build 1.1
Candida glabrata Build 1.1
Cryptococcus neoformans Build 2.1
Debaryomyces hansenii Build 1.1
Encephalitozoon cuniculi Build 1.1
Eremothecium gossypii Build 3.1
Gibberella zeae Build 1.2
Kluyveromyces lactis Build 1.1
Magnaporthe oryzae rice blast fungus Build 3.1
Neurospora crassa Build 2.1
Saccharomyces cerevisiae baker's yeast Build 2.1
Scheffersomyces stipitis Build 1.1
Schizosaccharomyces pombe fission yeast Build 1.1
Ustilago maydis Build 1.1
Yarrowia lipolytica Build 1.1
Contact | Copyright | Disclaimer | Privacy | Accessibility
National Center for Biotechnology Information, US National Library of Medicine
8600 Rockville Pike, Bethesda, MD USA 20894
NLM
USA.gov

Map viewer Top画面

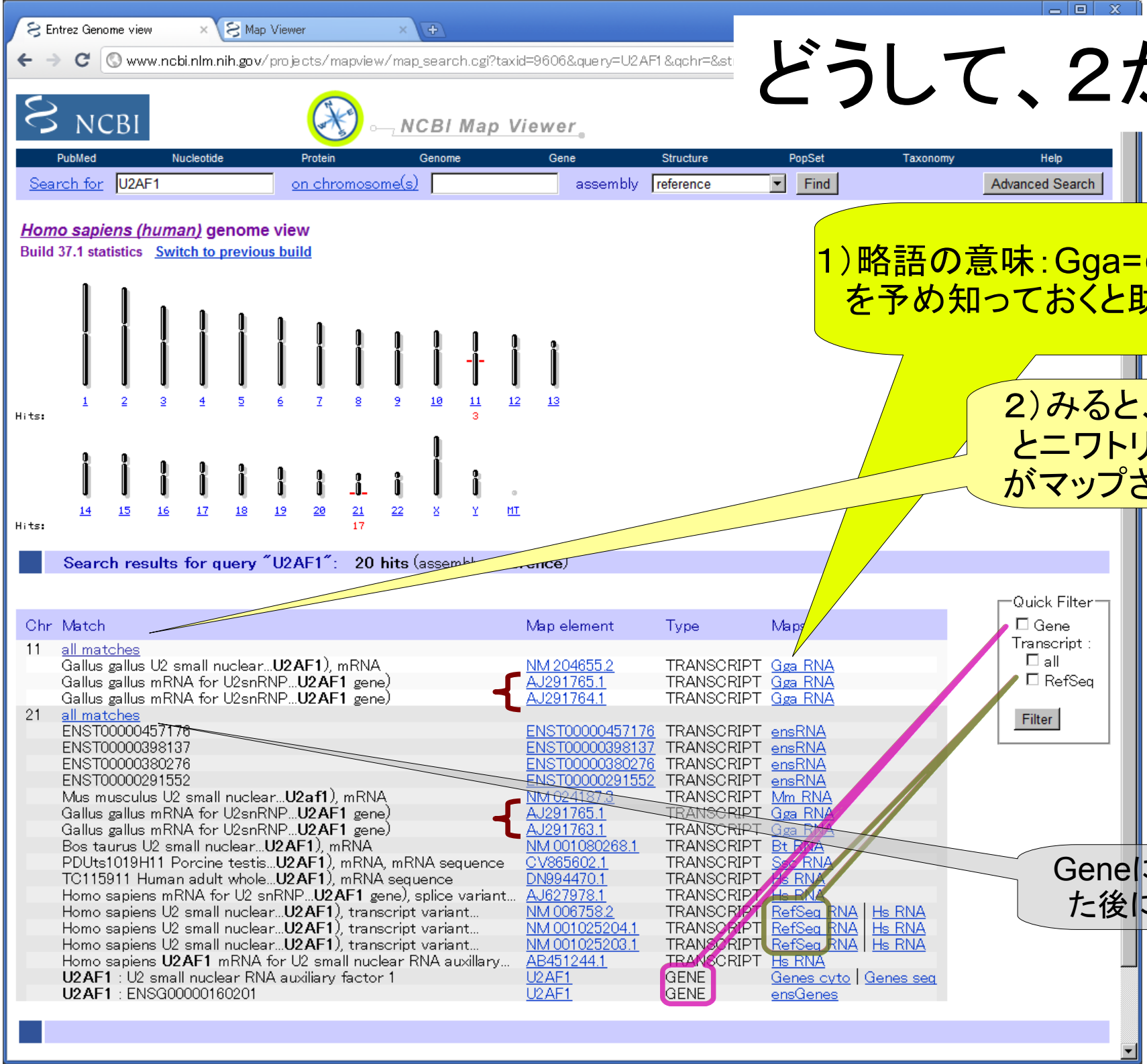
U2AF1
と入力





Chr	Assembly	Match	Map element	Type	Maps
11	reference	all matches			
		Gallus gallus U2 small nuclear...**U2AF1**), mRNA	NM_204655.2	TRANSCRIPT	Gga_RNA
		Gallus gallus mRNA for U2snRNP...**U2AF1** gene)	AJ291765.1	TRANSCRIPT	Gga_RNA
		Gallus gallus mRNA for U2snRNP...**U2AF1** gene)	AJ291764.1	TRANSCRIPT	Gga_RNA
11	Celera	all matches			
		Gallus gallus U2 small nuclear...**U2AF1**), mRNA	NM_204655.2	TRANSCRIPT	Gga_RNA
		Gallus gallus mRNA for U2snRNP...**U2AF1** gene)	AJ291765.1	TRANSCRIPT	Gga_RNA
		Gallus gallus mRNA for U2snRNP...**U2AF1** gene)	AJ291764.1	TRANSCRIPT	Gga_RNA
11	HuRef	all matches			
		Gallus gallus U2 small nuclear...**U2AF1**), mRNA	NM_204655.2	TRANSCRIPT	Gga_RNA
		Gallus gallus mRNA for U2snRNP...**U2AF1** gene)	AJ291765.1	TRANSCRIPT	Gga_RNA
		Gallus gallus mRNA for U2snRNP...**U2AF1** gene)	AJ291764.1	TRANSCRIPT	Gga_RNA
21	reference	all matches			
		ENST00000457176	ENST00000457176	TRANSCRIPT	ensRNA
		ENST00000398137	ENST00000398137	TRANSCRIPT	ensRNA
		ENST00000380276	ENST00000380276	TRANSCRIPT	ensRNA
		ENST00000291552	ENST00000291552	TRANSCRIPT	ensRNA
		Mus musculus U2 small nuclear...**U2af1**), mRNA	NM_024187.3	TRANSCRIPT	Mm_RNA
		Gallus gallus mRNA for U2snRNP...**U2AF1** gene)	AJ291765.1	TRANSCRIPT	Gga_RNA
		Gallus gallus mRNA for U2snRNP...**U2AF1** gene)	AJ291763.1	TRANSCRIPT	Gga_RNA
		Bos taurus U2 small nuclear...**U2AF1**), mRNA	NM_001080268.1	TRANSCRIPT	Bt_RNA
		PDUts1019H11 Porcine testis...**U2AF1**), mRNA, mRNA sequence	CV865602.1	TRANSCRIPT	Ssc_RNA
		TC115911 Human adult whole...**U2AF1**), mRNA sequence	DN994470.1	TRANSCRIPT	Hs_RNA
		Homo sapiens mRNA for U2 snRNP...**U2AF1** gene), splice	AJ627978.1	TRANSCRIPT	Hs_RNA

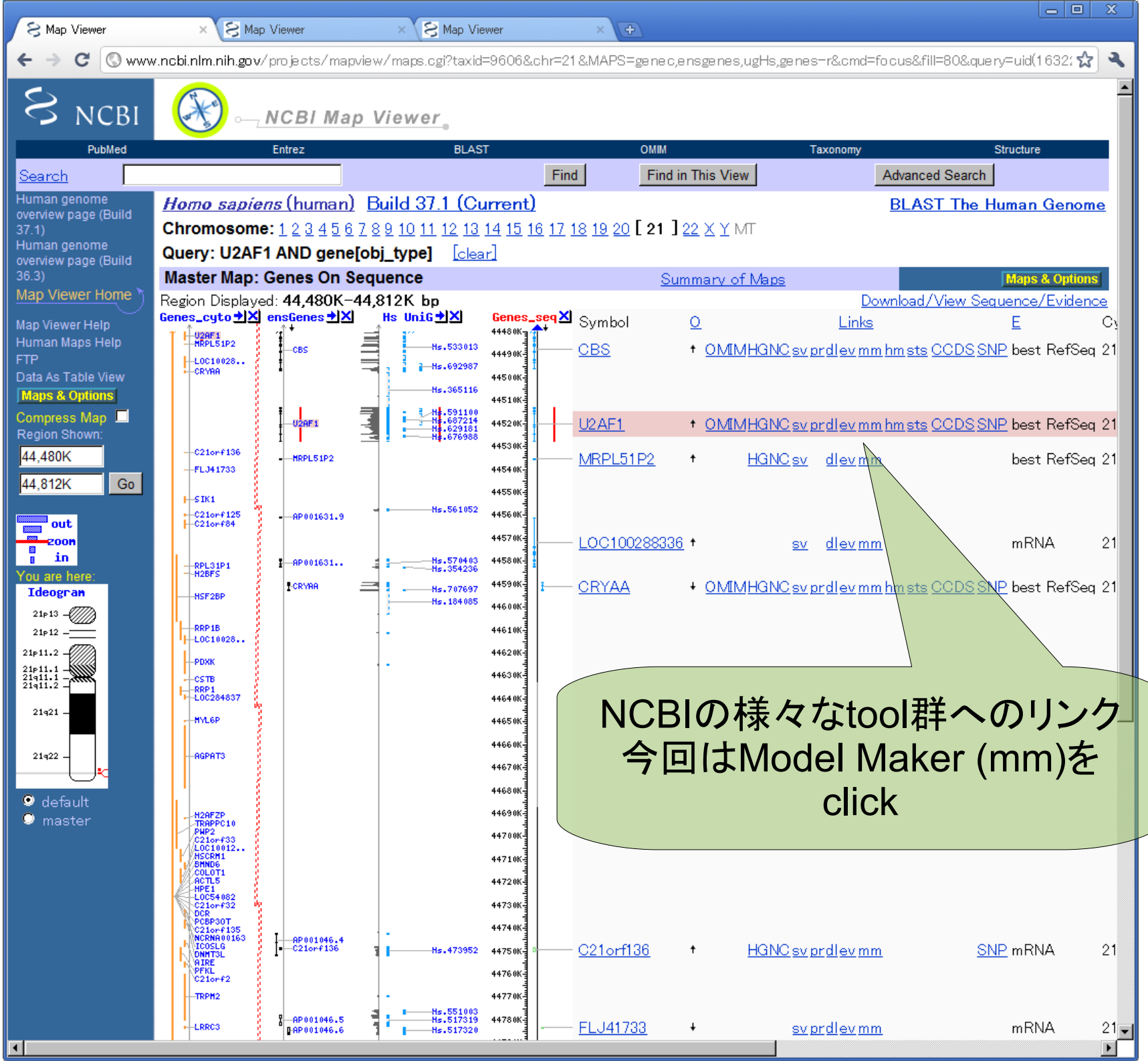
どうして、2か所？

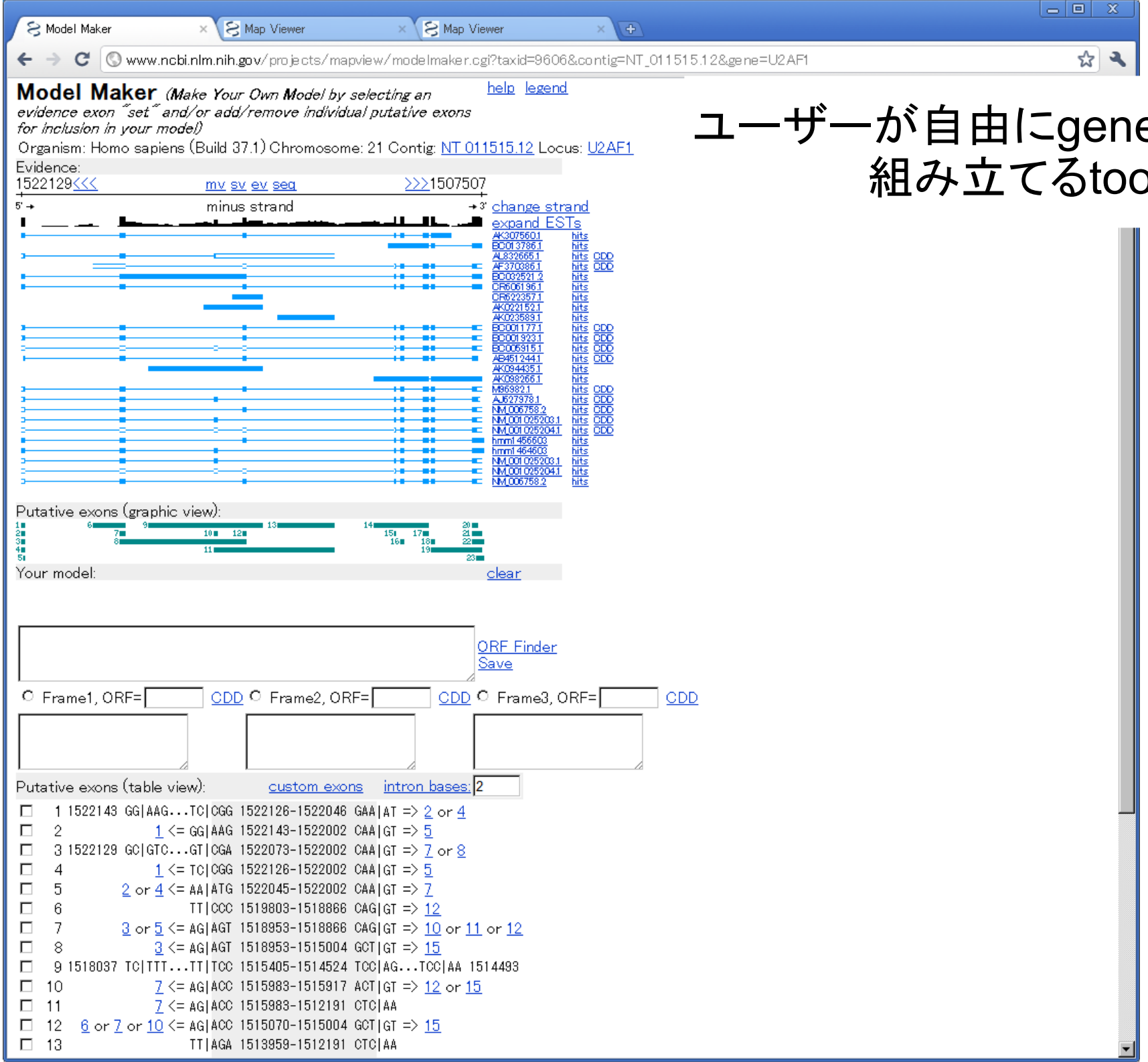


1）略語の意味：Gga=*Gallus gallus*を予め知っておくと助けになる。

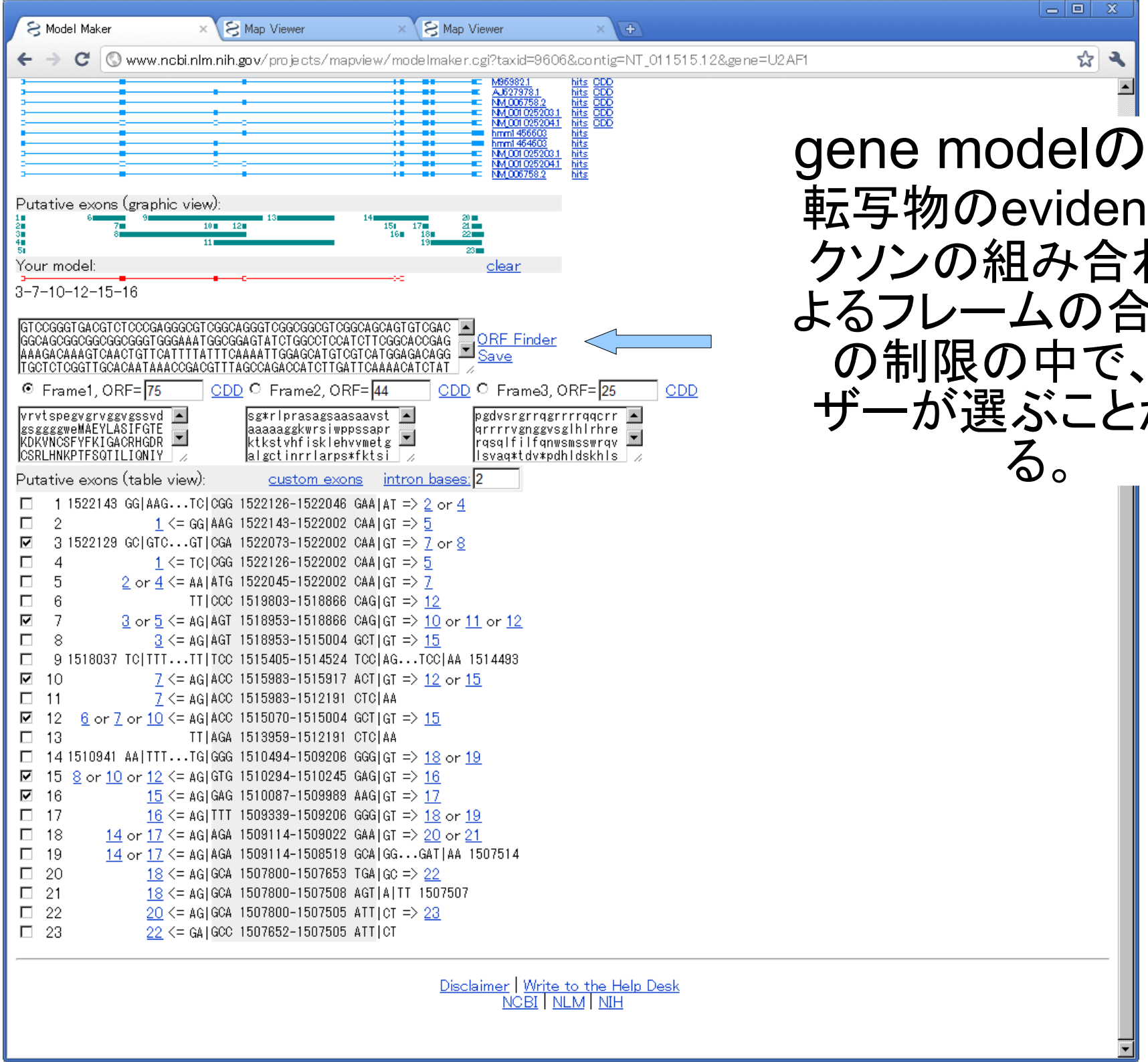
2）みると、ヒト11q13とニワトリのU2AF1がマップされている。

Geneにcheckを入れた後に、ここをclick


Map Viewer
www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?taxid=9606&chr=21&MAPS=genec,ensgenes,ugHs,genes-r&cmd=focus&fill=80&query=uid(1632
NCBI
NCBI Map Viewer
PubMed
Entrez
BLAST
OMIM
Taxonomy
Structure
Search
Find
Find in This View
Advanced Search
Human genome overview page (Build 37.1)
Human genome overview page (Build 36.3)
Map Viewer Home
Map Viewer Help
Human Maps Help
FTP
Data As Table View
Maps & Options
Compress Map
Region Shown:
44,480K
44,812K
Go
out
zoom
in
You are here:
Ideogram
default
master
Homo sapiens (human) Build 37.1 (Current)
BLAST The Human Genome
Chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 [21] 22 X Y MT
Query: U2AF1 AND gene[obj_type] [clear]
Master Map: Genes On Sequence
Summary of Maps
Maps & Options
Region Displayed: 44,480K–44,812K bp
Download/View Sequence/Evidence
Genes_cyto
ensGenes
Hs UniG
Genes_seq
Symbol
O
Links
E
CBS
U2AF1
MRPL51P2
LOC100288336
CRYAA
C21orf136
FLJ41733
NCBIの様々なtool群へのリンク
今回はModel Maker (mm)を
click



ユーザーが自由にgene modelを組み立てるtool



gene modelの構築：転写物のevidenceとエクソンの組み合わせによるフレームの合致などの制限の中で、ユーザーが選ぶことができる。

Model Maker | ORF Finder | Map Viewer | Map Viewer

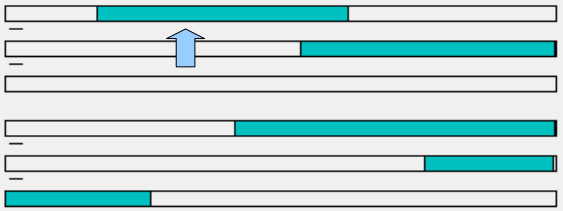
www.ncbi.nlm.nih.gov/gorf/orfig.cgi

ORF Finder (Open Reading Frame Finder)

PubMed Entrez BLAST OMIM Taxonomy Structure

Anonymous

View | 1 GenBank | Redraw | 100 | SixFrames



Frame	from	to	Length
−1	209	498	291
+2	269	498	231
+1	85	312	228
−3	1	134	134
−2	382	498	117

最上段をクリックすることによって、アミノ酸配列が表示された。

Ensembl (www.ensembl.org)

Search: All species for Go

e.g. human gene BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes (Log in to customize this list)

Human
GRCh37

Mouse
NCBIM37

Zebrafish
Zv8

All genomes

-- Select a species --

View full list of all Ensembl species

Other species are available in Ensembl *Pre!* and EnsemblGenomes

Ensembl is a joint project between EMBL - EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

EMBL-EBI

Ensembl receives major funding from the Wellcome Trust. Our acknowledgements page includes a list of additional current and previous funding bodies.

New to Ensembl?

Did you know you can:

- **Learn how to use Ensembl**
  with our video tutorials and walk-throughs
- **Add custom tracks**
  using our new Control Panel
- **Upload and analyse your data**
  and save it to your Ensembl account
- **Search for a DNA or protein sequence**
  using BLAST or BLAT
- **Fetch only the data you want**
  from our public database, using the Perl API
- **Download our databases via FTP**
  in FASTA, MySQL and other formats
- **Mine Ensembl with BioMart**
  and export sequences or tables in text, html, or Excel format

Still got questions? Try our FAQs or glossary

What's New in Release 59 (5 August 2010)

- **New human regulation data and displays** (Human)
- **COSMIC somatic mutations** (Human)
- **Changes to masthead** (all species)
- **Details of data updates, API changes, etc**

More news...

Latest blog posts

- 2010-09-24: Ensembl Events in October 2010
- 2010-09-21: Changing to LASTZ
- 2010-09-09: A.mel vs A.mel

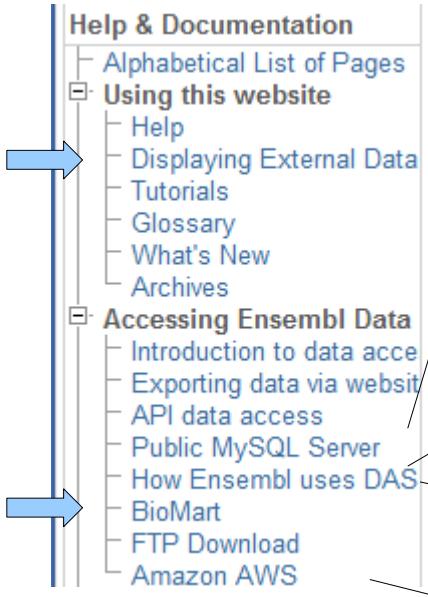
Go to Ensembl blog →

Navigation tips from our blog

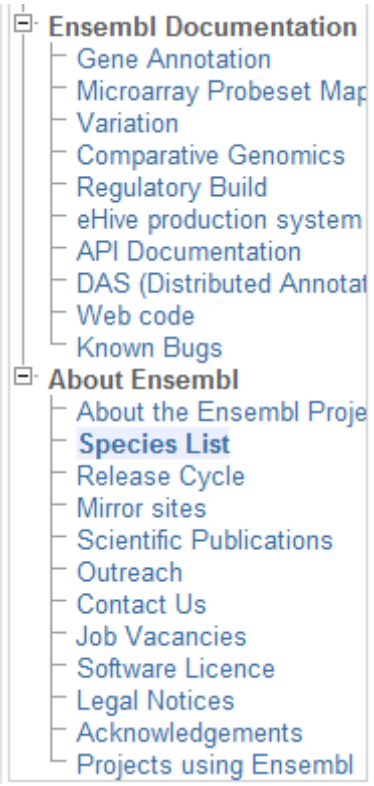
Species List in Ensembl

このページでは51種。Ensemble genome（bacteria, protists, fungi, plants and invertebrate metazoa）では234種

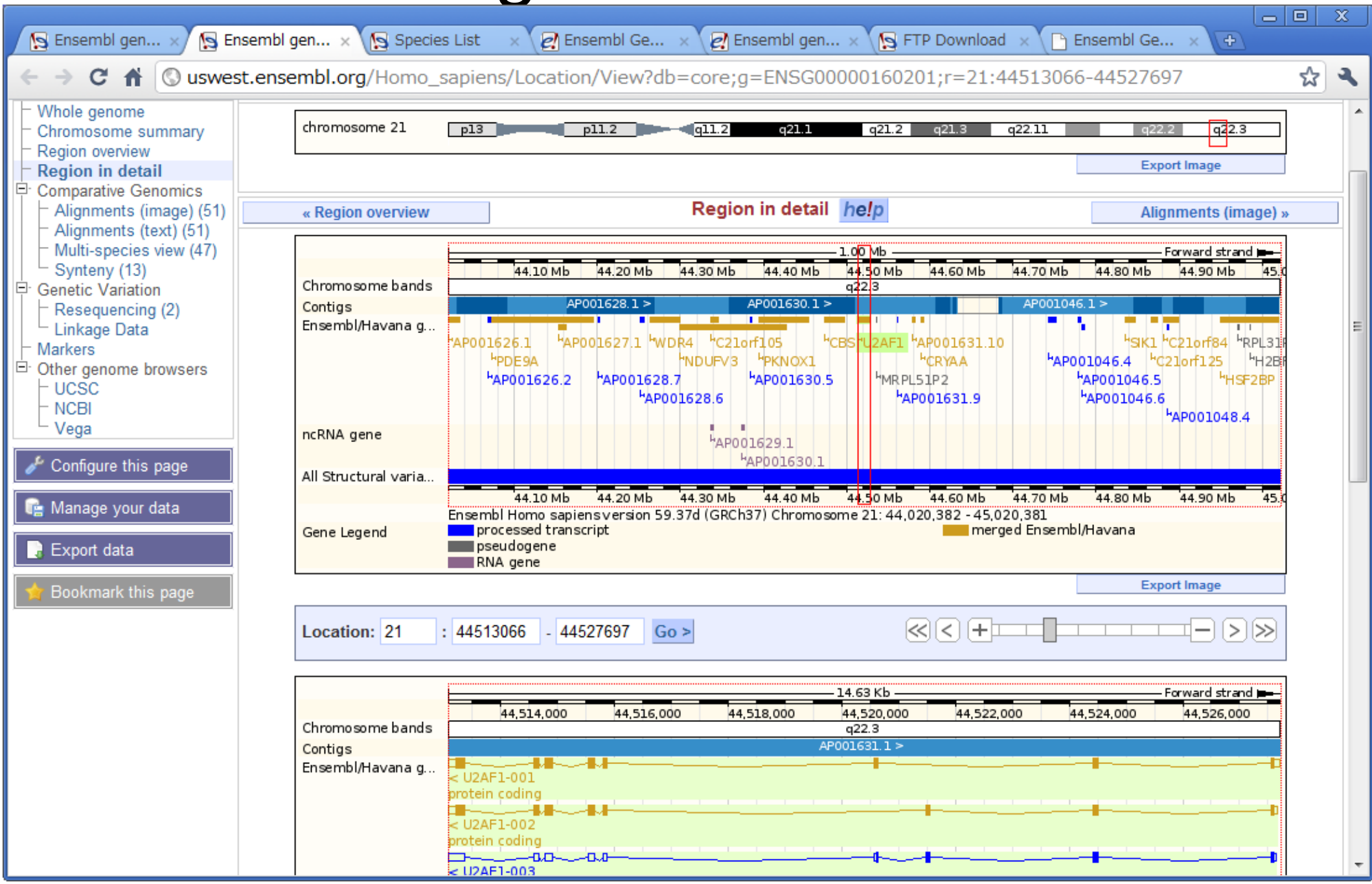
Ensemblのdocument一覧



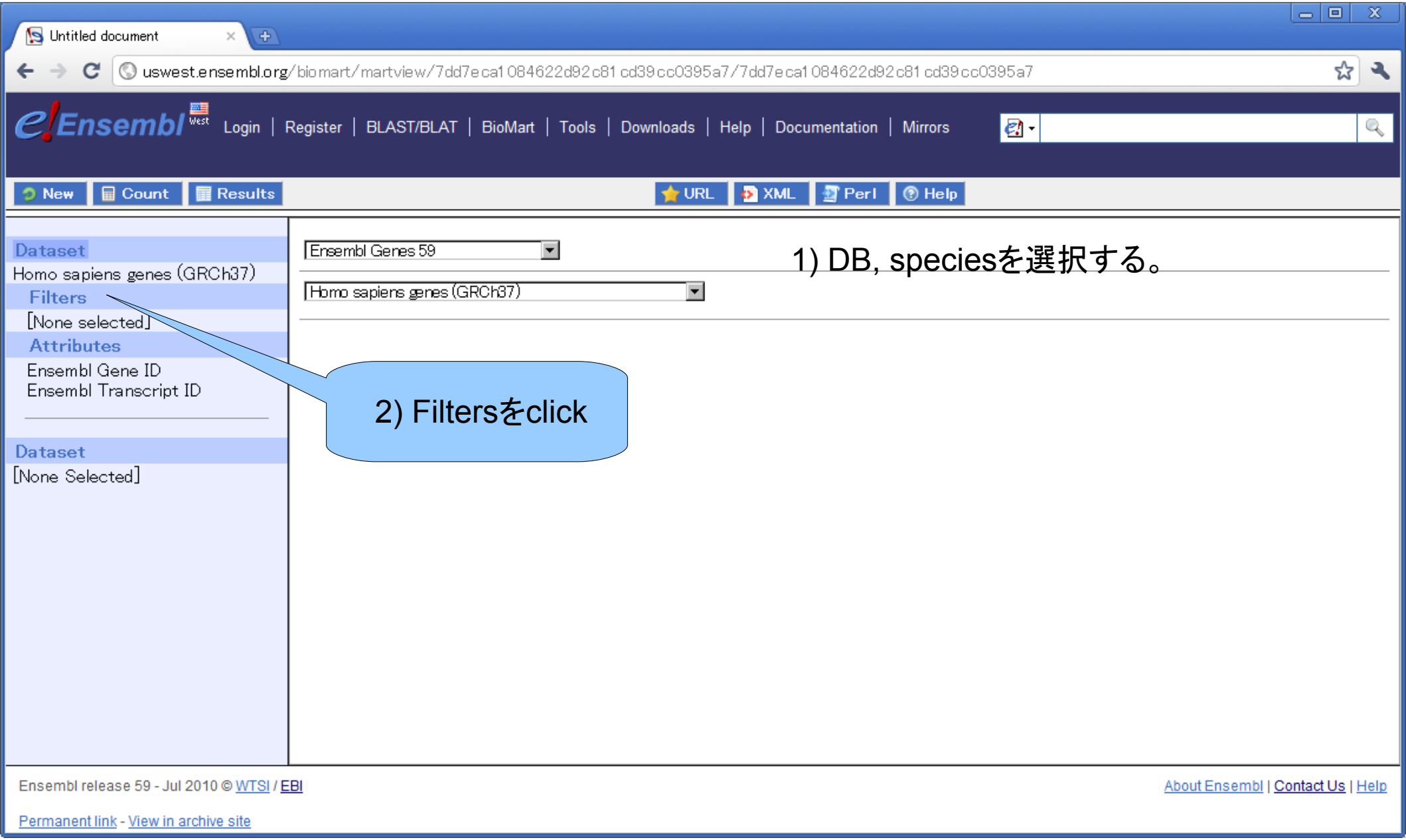
- API: Application Programming Interface webアプリケーションの機能の一部を外部のプログラムから利用できるようにするための手順や書式を定めた仕様。
- EnsemblのMySQLサーバーにアクセス。
- DAS: Distributed Annotation System
- AmzaonのCloud (EC2) serviceを介したデータの共有。



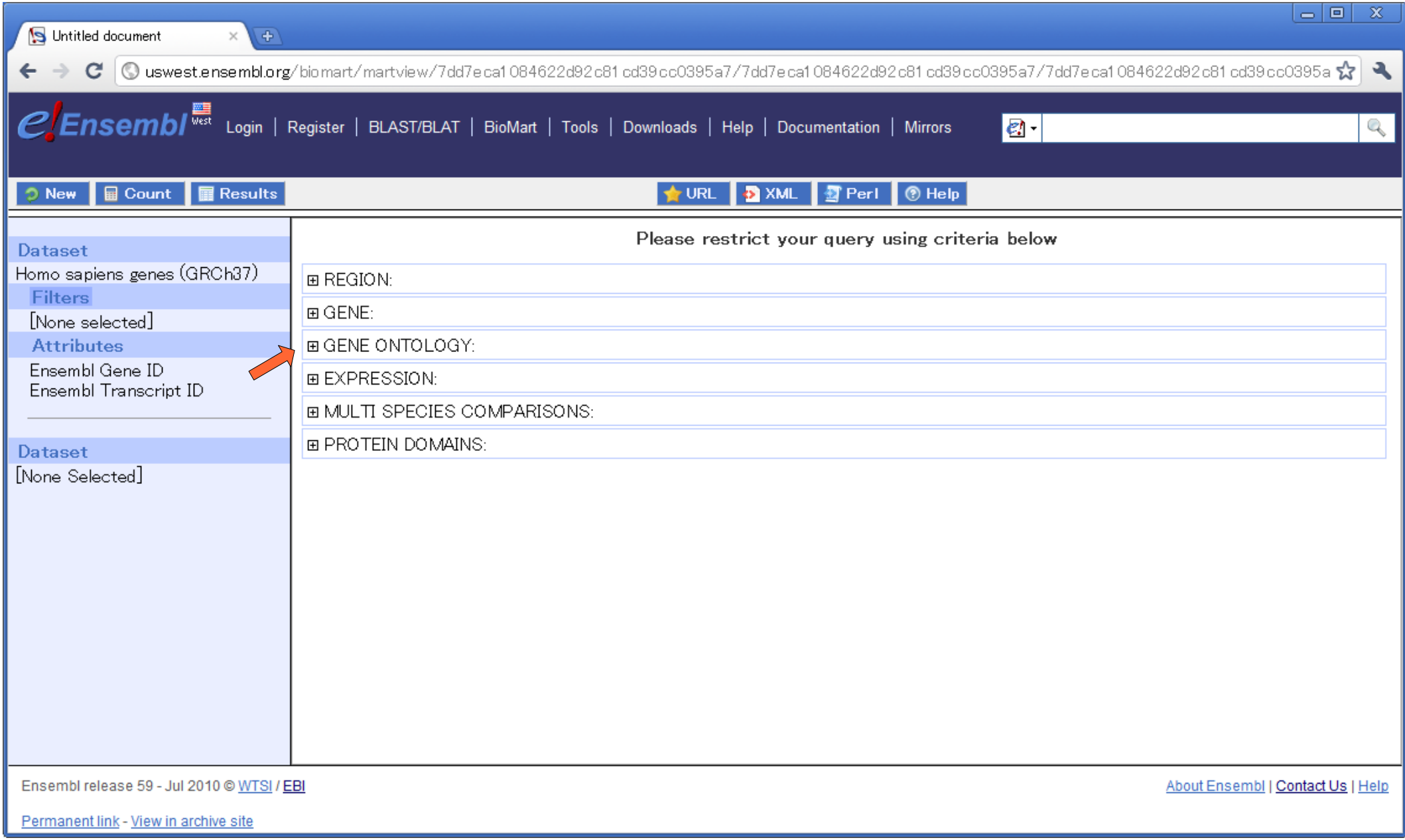
Ensembl: genome browserの例



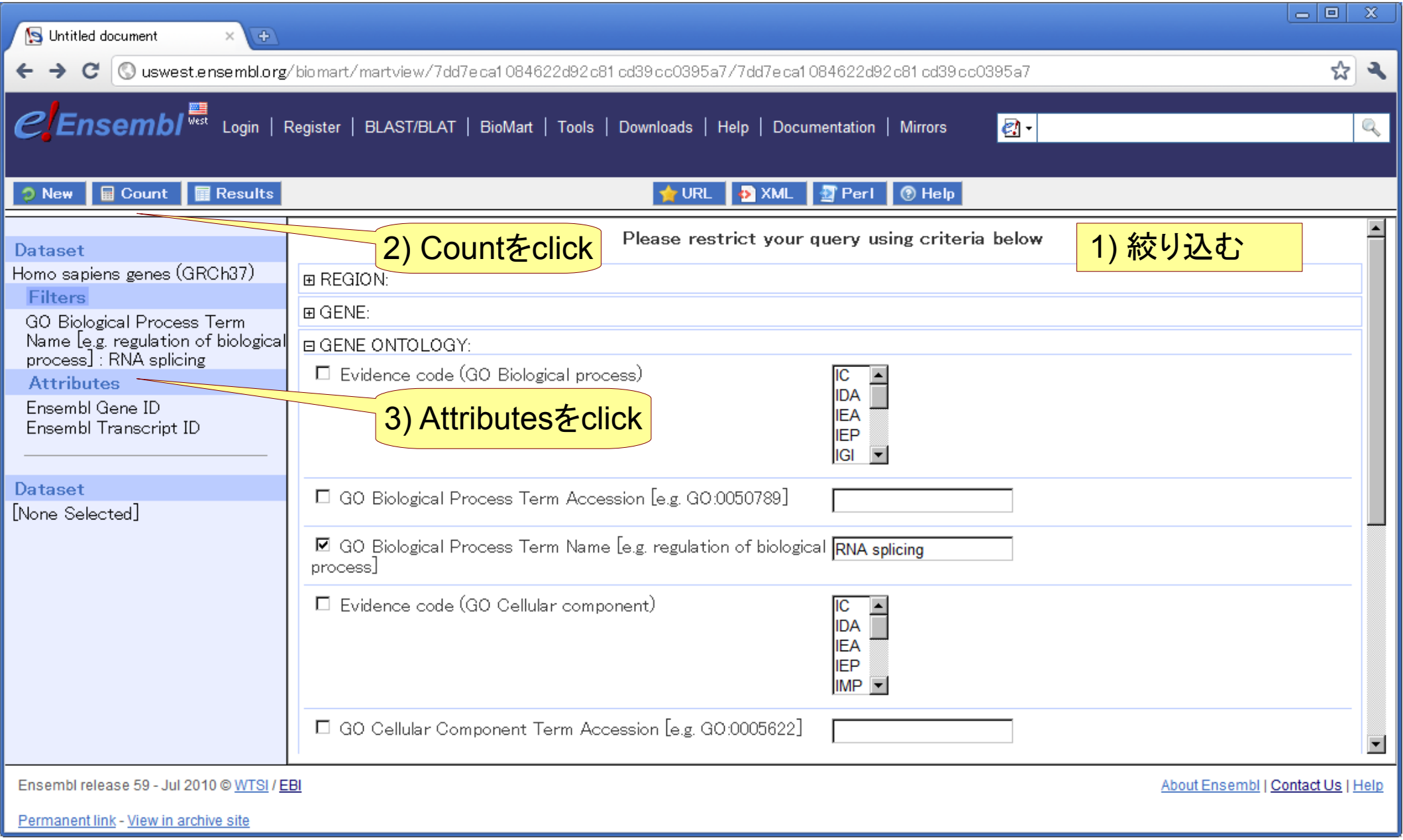
BioMart top page



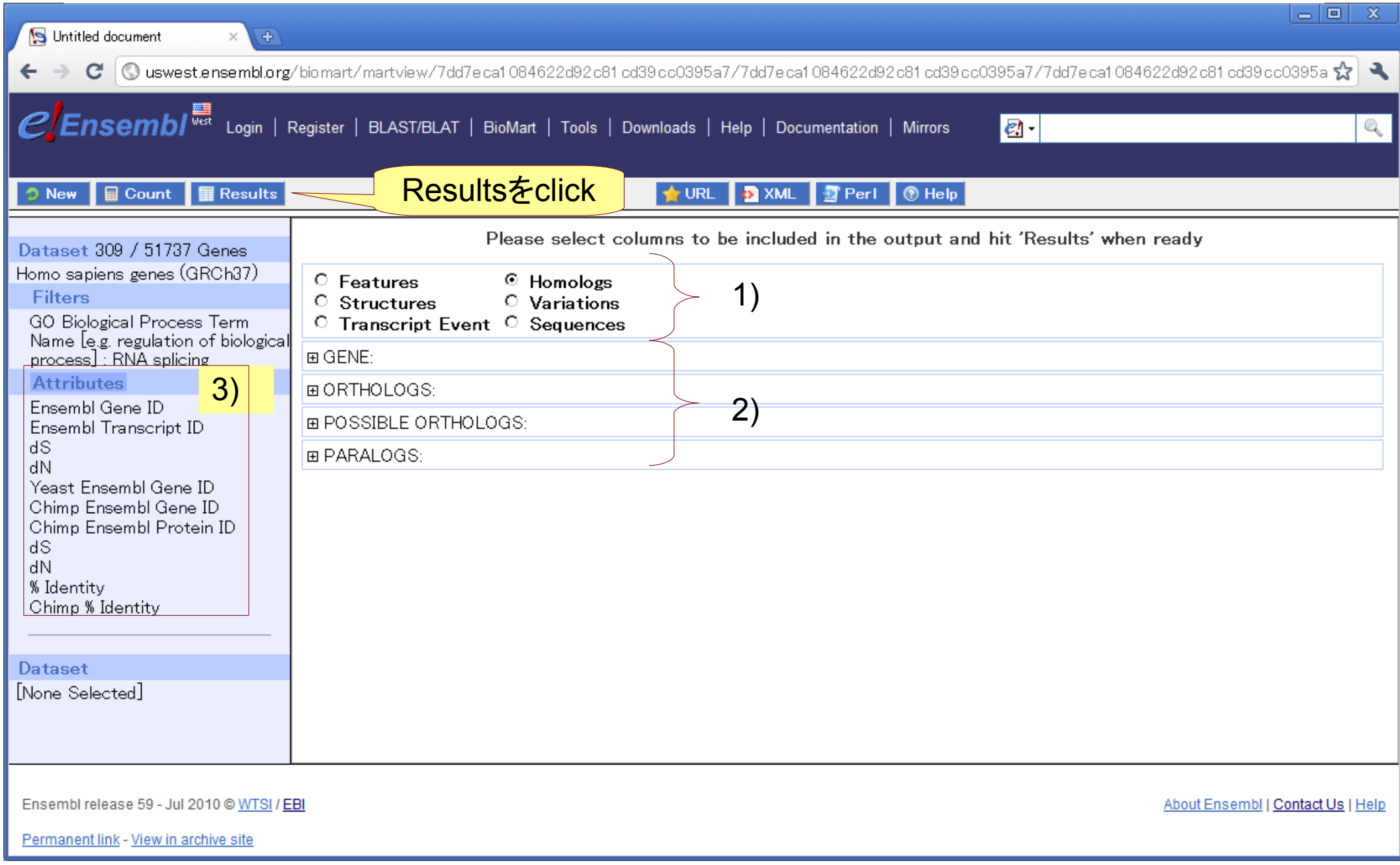
BioMart:Filters;欲しい情報に絞り込む1



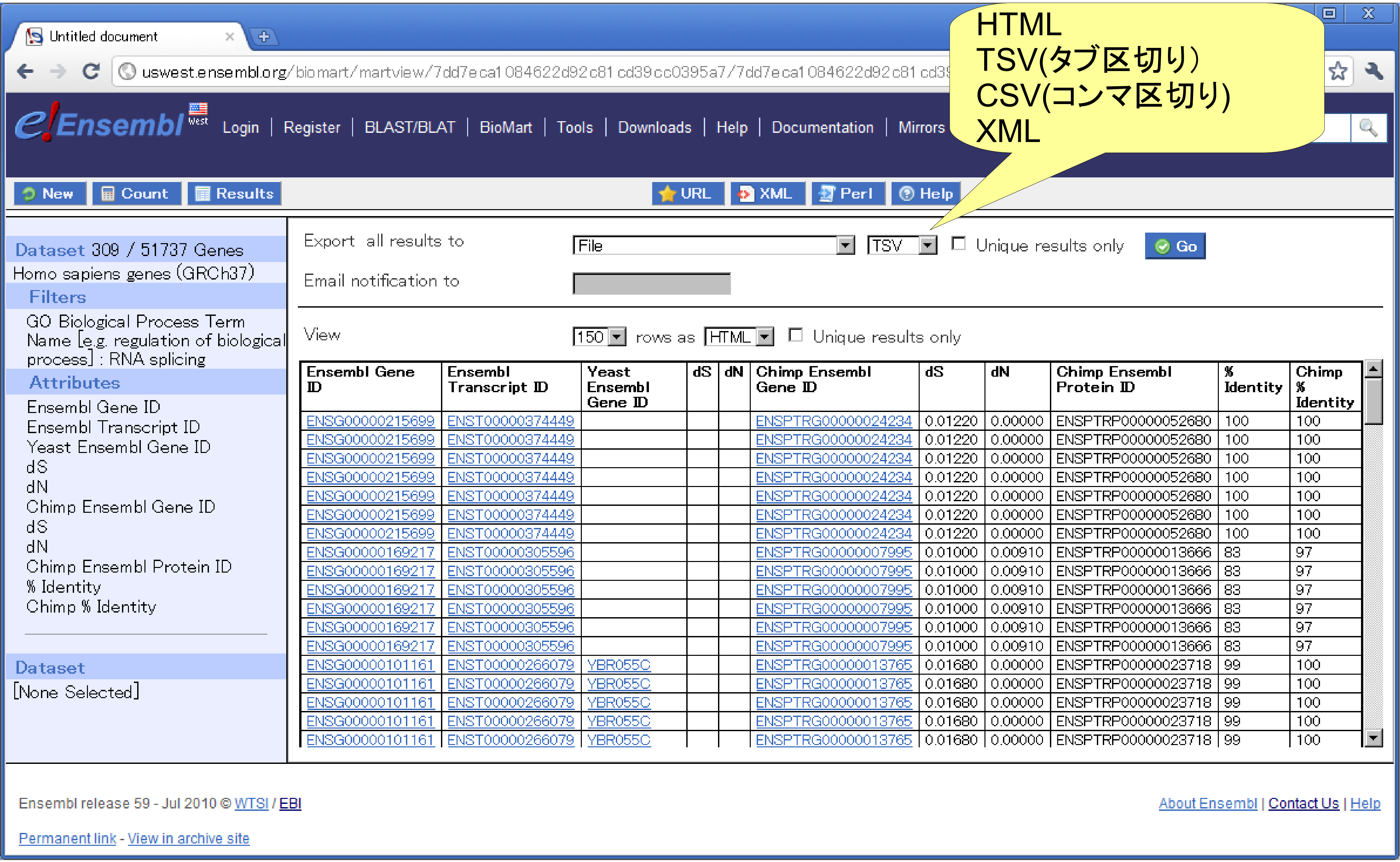
BioMart:Filters;欲しい情報に絞り込む2



BioMart:Attributes;出力したい項目に絞り込む

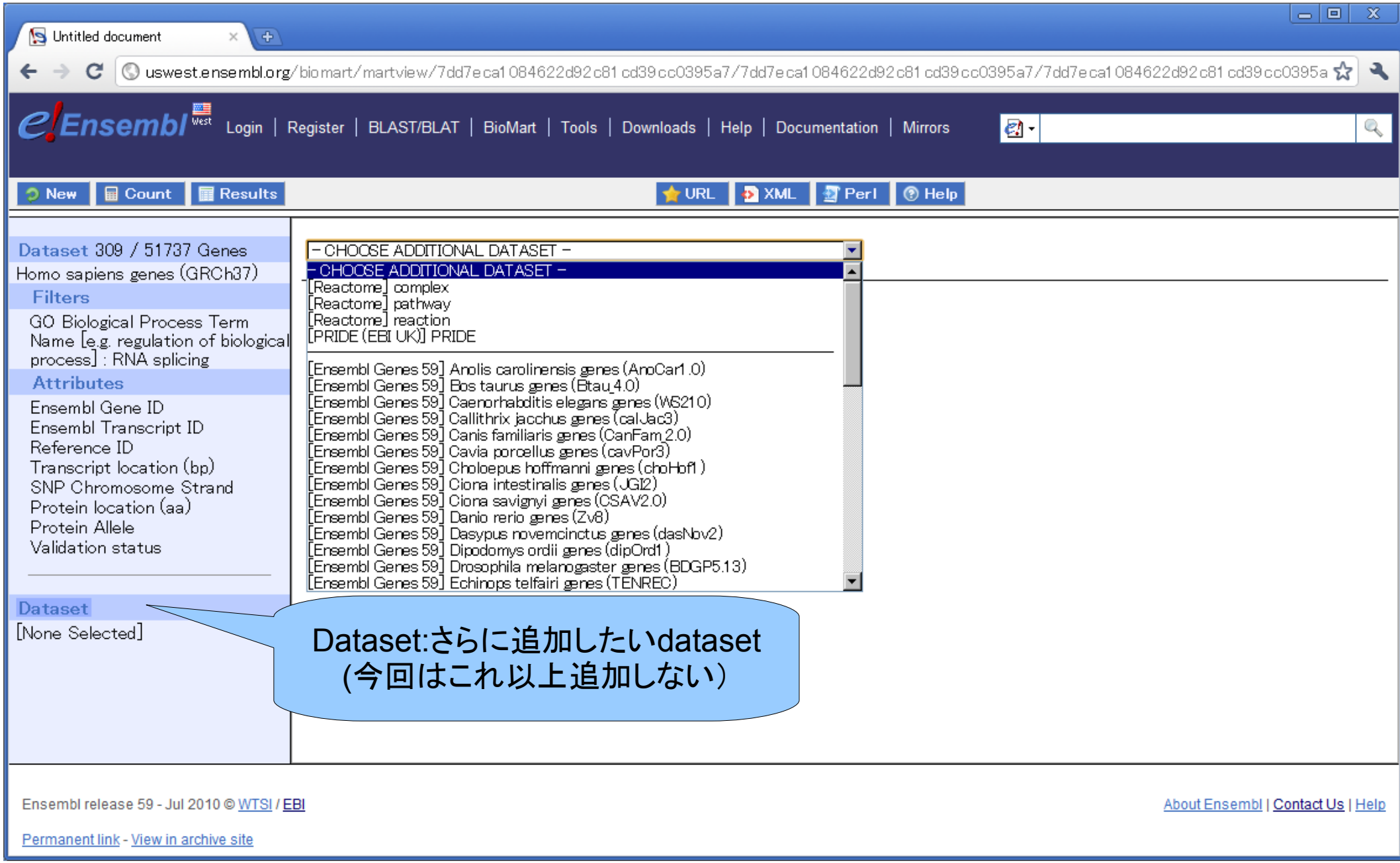


BioMart: Result表示例

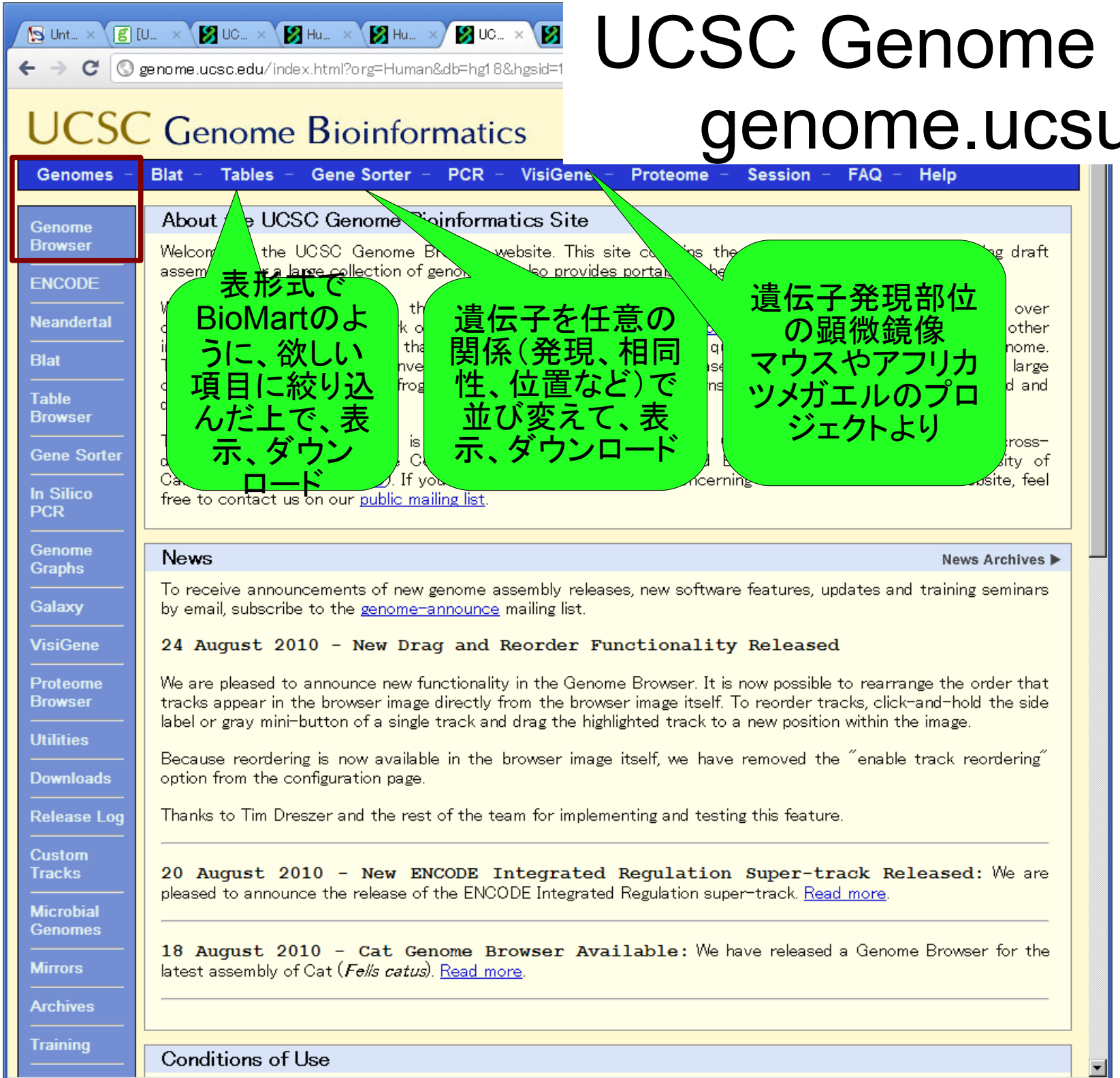


Ensembl Gene ID	Ensembl Transcript ID	Yeast Ensembl Gene ID	dS	dN	Chimp Ensembl Gene ID	dS	dN	Chimp Ensembl Protein ID	% Identity	Chimp % Identity
ENSG00000215699	ENST00000374449				ENSPTRG00000024234	0.01220	0.00000	ENSPTRP00000052680	100	100
ENSG00000215699	ENST00000374449				ENSPTRG00000024234	0.01220	0.00000	ENSPTRP00000052680	100	100
ENSG00000215699	ENST00000374449				ENSPTRG00000024234	0.01220	0.00000	ENSPTRP00000052680	100	100
ENSG00000215699	ENST00000374449				ENSPTRG00000024234	0.01220	0.00000	ENSPTRP00000052680	100	100
ENSG00000215699	ENST00000374449				ENSPTRG00000024234	0.01220	0.00000	ENSPTRP00000052680	100	100
ENSG00000215699	ENST00000374449				ENSPTRG00000024234	0.01220	0.00000	ENSPTRP00000052680	100	100
ENSG00000215699	ENST00000374449				ENSPTRG00000024234	0.01220	0.00000	ENSPTRP00000052680	100	100
ENSG00000169217	ENST00000305596				ENSPTRG00000007995	0.01000	0.00910	ENSPTRP00000013666	83	97
ENSG00000169217	ENST00000305596				ENSPTRG00000007995	0.01000	0.00910	ENSPTRP00000013666	83	97
ENSG00000169217	ENST00000305596				ENSPTRG00000007995	0.01000	0.00910	ENSPTRP00000013666	83	97
ENSG00000169217	ENST00000305596				ENSPTRG00000007995	0.01000	0.00910	ENSPTRP00000013666	83	97
ENSG00000169217	ENST00000305596				ENSPTRG00000007995	0.01000	0.00910	ENSPTRP00000013666	83	97
ENSG00000169217	ENST00000305596				ENSPTRG00000007995	0.01000	0.00910	ENSPTRP00000013666	83	97
ENSG00000101161	ENST00000266079	YBR055C			ENSPTRG00000013765	0.01680	0.00000	ENSPTRP00000023718	99	100
ENSG00000101161	ENST00000266079	YBR055C			ENSPTRG00000013765	0.01680	0.00000	ENSPTRP00000023718	99	100
ENSG00000101161	ENST00000266079	YBR055C			ENSPTRG00000013765	0.01680	0.00000	ENSPTRP00000023718	99	100
ENSG00000101161	ENST00000266079	YBR055C			ENSPTRG00000013765	0.01680	0.00000	ENSPTRP00000023718	99	100
ENSG00000101161	ENST00000266079	YBR055C			ENSPTRG00000013765	0.01680	0.00000	ENSPTRP00000023718	99	100

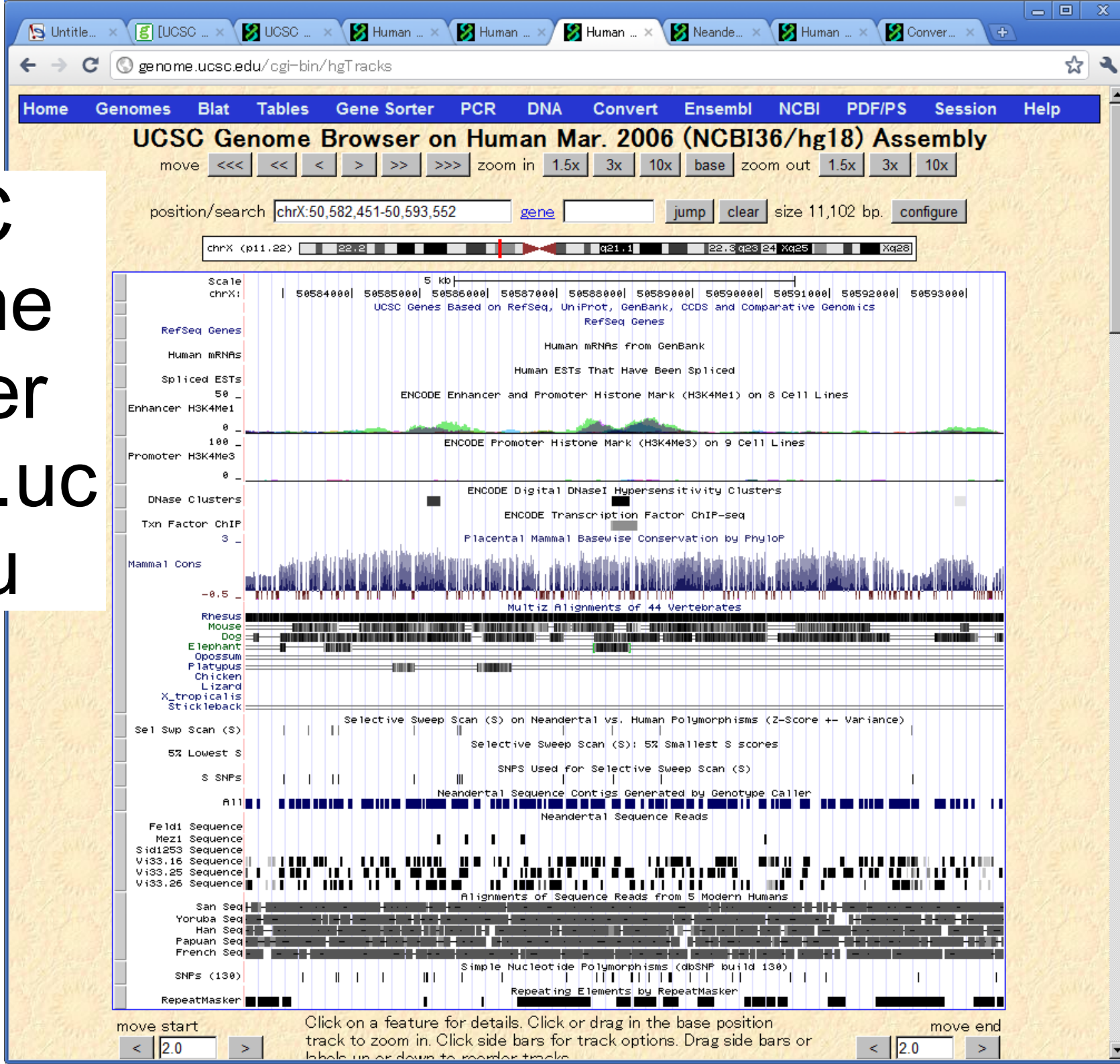
さらなるdataset追加



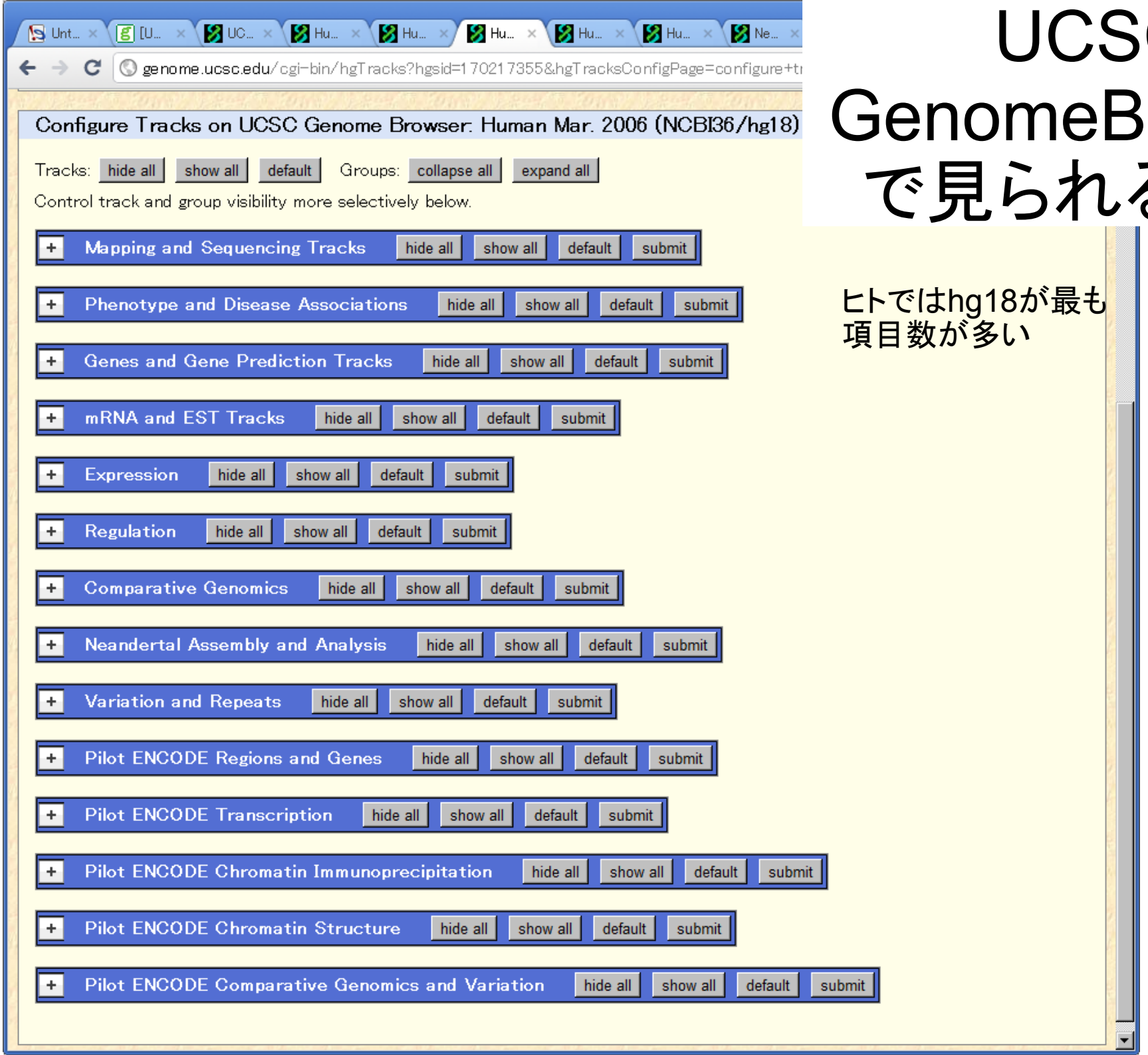
UCSC Genome Browser
genome.ucsu.edu



UCSC Genome Browser genome.ucsu.edu



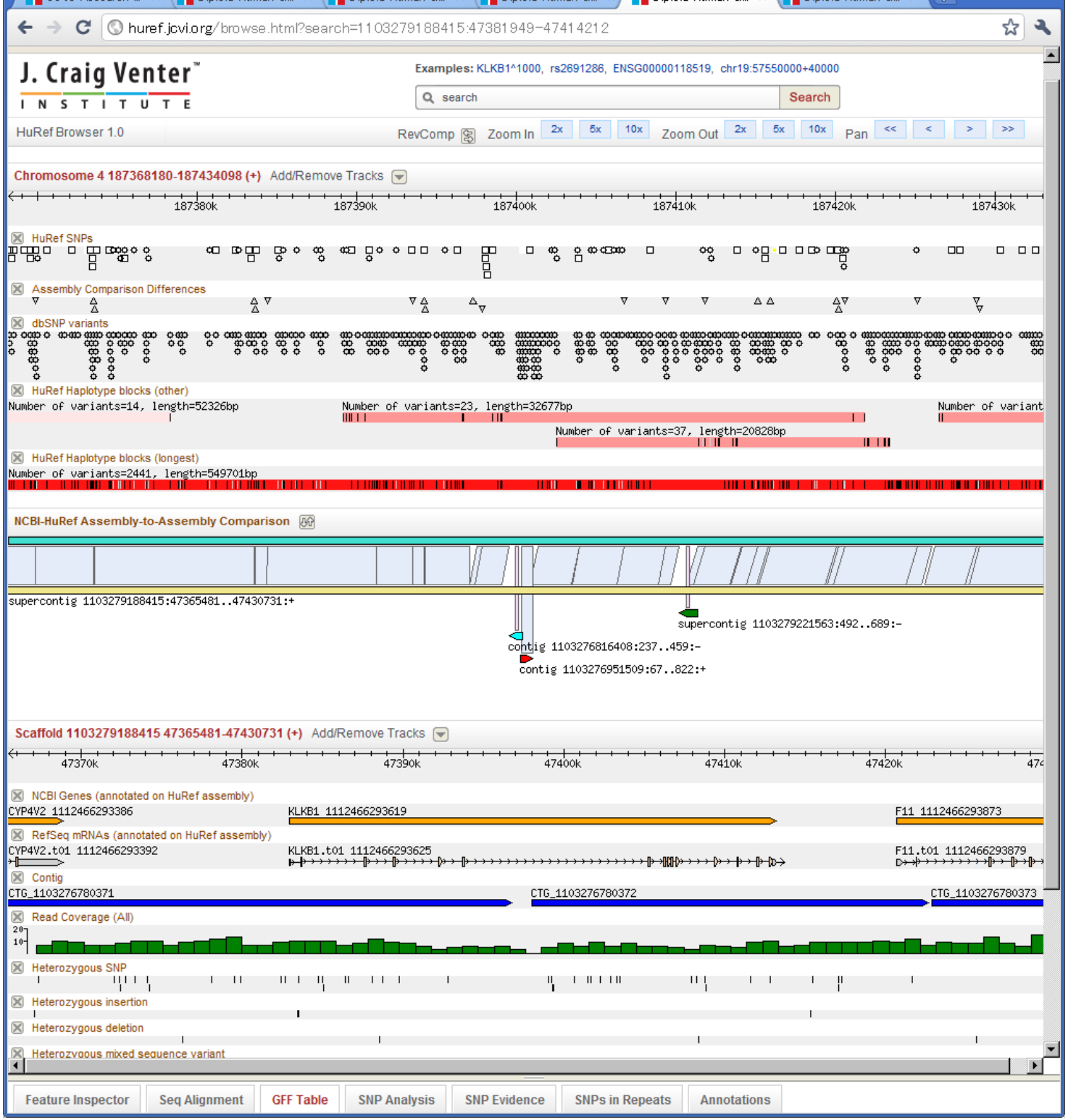
UCSC GenomeBrowserで見られる項目



ヒトではhg18が最も項目数が多い

JCVI HuRef Browser huref.jcvi.org/

INDIVIDUAL HUMAN DIPLOID GENOME



genome versionとは

http://genome.ucsc.edu/FAQ/FAQreleases.html

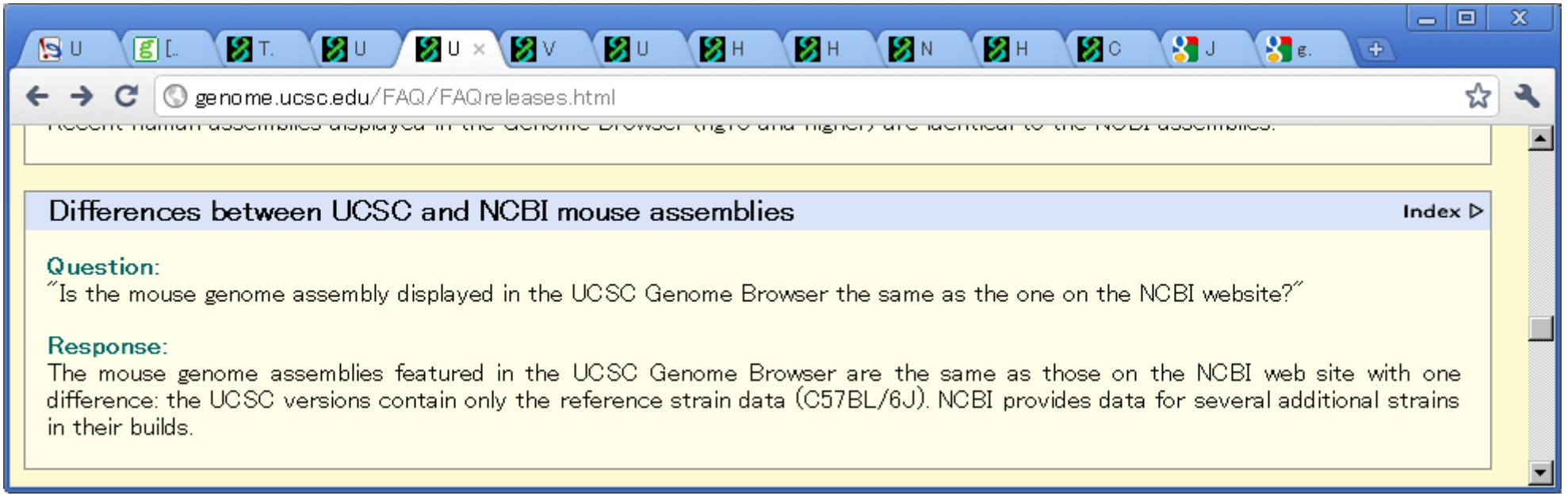
List of UCSC genome releases

Question:
"How do UCSC's release numbers correspond to those of other organizations, such as NCBI?"

Response:

EnsemblもGRChを使用

SPECIES	UCSC VERSION	RELEASE DATE	RELEASE NAME	STATUS
VERTEBRATES				
Human	hg19	Feb. 2009	Genome Reference Consortium GRCh37	Available
	hg18	Mar. 2006	NCBI Build 36.1	Available
	hg17	May 2004	NCBI Build 35	Available
	hg16	Jul. 2003	NCBI Build 34	Available
	hg15	Apr. 2003	NCBI Build 33	Archived
	hg13	Nov. 2002	NCBI Build 31	Archived
	hg12	Jun. 2002	NCBI Build 30	Archived
	hg11	Apr. 2002	NCBI Build 29	Archived
	hg10	Dec. 2001	NCBI Build 28	Archived
	hg8	Aug. 2001	UCSC-assembled	Archived
	hg7	Apr. 2001	UCSC-assembled	Archived
	hg6	Dec. 2000	UCSC-assembled	Archived
	hg5	Oct. 2000	UCSC-assembled	Archived
	hg4	Sep. 2000	UCSC-assembled	Archived
	hg3	Jul. 2000	UCSC-assembled	Archived
	hg2	Jun. 2000	UCSC-assembled	Archived (data set only)
	hg1	May 2000	UCSC-assembled	Archived (data set only)
Cat	felCat4	Dec. 2008	NHGRI catChrV17e	Available
	felCat3	Mar. 2006	Broad Institute Release 3	Available
Chicken	galGal3	May 2006	WUSTL Gallus-gallus-2.1	Available
	galGal2	Feb. 2004	WUSTL Gallus-gallus-1.0	Available
Chimp	panTro2	Mar. 2006	CGSC Build 2 Version 1	Available
	panTro1	Nov. 2003	CGSC Build 1 Version 1	Available
Cow	bosTau4	Oct. 2007	Baylor College of Medicine HGSC Btau_4.0	Available


genome.ucsc.edu/FAQ/FAQreleases.html
Differences between UCSC and NCBI mouse assemblies
Index ▷
Question:
"Is the mouse genome assembly displayed in the UCSC Genome Browser the same as the one on the NCBI website?"
Response:
The mouse genome assemblies featured in the UCSC Genome Browser are the same as those on the NCBI web site with one difference: the UCSC versions contain only the reference strain data (C57BL/6J). NCBI provides data for several additional strains in their builds.

必要なDBを見つけるために

- HGNC Useful Links
  - http://www.genenames.org/useful.html
- WINGpro（データベースポータルサイト）by JST
  - http://wingpro.lifesciencedb.jp/
- 統合DBプロジェクト
  - http://lifesciencedb.jp/
  - 生命科学系 データベース カタログ
    - http://lifesciencedb.jp/lsdb.cgi?pg=1