

ゲノム配列とそれに付与されたデータの使い方 at AJACS薩摩

情報・システム研究機構 ライフサイエンス統合データベースセンター

坊農 秀雅 <http://bonohu.jp/> bono@dbcls.rois.ac.jp

2016年1月26日 鹿児島大学桜ヶ丘キャンパス

これは統合データベース講習会AJACS薩摩 「ゲノム配列とそれに付与されたデータの使い方」の資料です。

概要

本講習は、だれでも自由に使うことができる公共データベース(DB)のうちゲノム配列について、それに付与されたデータの使い方を、ウェブツールを活用して、研究のさまざまな場面で使う方策について学びます。

講習の流れ

今回の講習では、コンピュータを使って以下の内容について説明します。

- 研究現場で頻繁に使われるデータベースやツールを知る
- ゲノム配列を探す
 - NCBI
- ゲノム配列から探す
 - Web上で
 - BLAST
 - GGGenome
 - Local BLASTで(紹介のみ)
- アノテーションを足して見る
 - UCSCのtrack
 - ENCODE TF Binding
 - TFBS Conserved

- Resetのやり方
 - Ensemblへジャンプ
 - Synteny
 - Resequencing
-

講習に際しての注意とお願ひ

- みんなで同時にアクセスするとサイトにつながりにくくなることが予想されます。
 - 資料を見ながら自力で進められそうな方はどんどん先に、そうでない方は講師と一緒にすすめていきましょう。
 - サイトの反応が悪い時はタイミングをずらして実行してみてください。
 - 反応が無いからと言って何度もクリックするとますます繋がらなくなってしまいます。おおらかな気持ちで臨みましょう。
 - わからないことがあったら挙手にてスタッフにお知らせください。
 - 遠慮は無用です(そのための講習会です!)。おいてけぼりは楽しくありません。
 - 質問や個別の相談がある方は講習時間が終わったあとでも講師を捕まえて話してください。
-

研究現場で頻繁に使われるデータベースやツールを知る

統合TV <http://tgotv.dbcls.jp/>

- 生命科学分野の有用なデータベースやツールの使い方を動画で紹介するウェブサイト
 - <http://tgotv.dbcls.jp/>

目的別に検索

- AJACS講習会資料
 - ゲノム・核酸配列解析
 - タンパク質配列・構造解析
 - 発現制御解析・可視化
 - 文献・辞書・プログラミング
 - 著名データベース
 - 学会講演・講習会
- 関連するタグから検索
- AJACS (276) BioHackathon (74)
 - Bioconductor (18) CAGE (6)
 - CRISPR (3) ChIP-seq (10)
 - DBCLS (174) DDBJ (37)
 - DDBJing (15) EBI (10)
 - EMBOSS (21) English (108)
 - Ensembl (19) FANTOM (4)
 - GEO (17) Galaxy (31) Google (11)
 - KEGG (14) Linux (12) NBDC (30)
 - NCBI (51) NGS (100)
 - NGS速習・ハンズオン (52)
 - PDB (20) PDBj (8) Perl (8)
 - Python (2) R (24) RDF (24)
 - RNA-seq (15) Reseq (4) UCSC (16)

Q 全番組のリストから、調べたいDBやウェブツールに関するキーワードで検索! (全 1029 件)

番組のタイトルや画像をクリックすると番組の再生ページへ移動します。

表示件数を選ぶ ▾ 検索窓にキーワードを入れると、入力の度ごとに即座に候補の番組が絞り込まれます

生物多様性情報分野からのインプット@Annotathon2015

本日の統合TVは、2015年11月12-13日(木・金)に行われた2015年度 国立遺伝学研究所 研究会 Annotathon2015 (生命科学データベースの利用価値向上のためのアノテーションマラソン) から 国立科学博物館 神保 宇嗣 研究員による「生物多様性情報分野からのインプット」をお送りします。約22分です。

研究会の一連の動画はYouTubeの再生リストからもご覧いただけます。



微生物ゲノムアノテーションの現状と問題点@Annotathon2015

本日の統合TVは、2015年11月12-13日(木・金)に行われた2015年度 国立遺伝学研究所 研究会 Annotathon2015 (生命科学データベースの利用価値向上のためのアノテーションマラソン) から 東京工業大学大学院 生命理工学研究科 生命情報専攻 森 宙史 助教による「微生物ゲノムアノテーションの現状と問題点」をお送りします。約35分です。

研究会の一連の動画はYouTubeの再生リストからもご覧いただけます。



ChIP-seqデータベースのためのメタ情報アノテーション@Annotathon2015

本日の統合TVは、2015年11月12-13日(木・金)に行われた2015年度 国立遺伝学研究所 研究会 Annotathon2015 (生命科学データベースの利用価値向上のためのアノテーションマラソン) から 九州大学大学院 医学院研究科 発生再生医学分野 沢 真弥 助教による「ChIP-seqデータベースのためのメタ情報アノテーション」をお送りします。約15分です。

既報のChIP-Seqデータを簡単な操作で半自動的に可視化するSraTailorや、既報のChIP-Seqデータの可視化と解析を行うChIP-Atlasを開発する上で必要となるメタデータのアノテーションの現状と対策についてのお話です。



<https://gyazo.com/9b890666bc8176d672d1e4f560162d05>

- YouTube版もあります <http://www.youtube.com/user/togotv/videos>

The screenshot shows the TOGO TV YouTube channel page. At the top, there's a banner featuring a cartoon character and the channel name. Below the banner, the channel name 'togotv' is displayed, along with a 'チャンネル登録' (Subscribe) button and a subscriber count of 779. The main content area shows a video thumbnail for '統合TV プロモーションムービー' (Promotional Movie) with 1,913 views. To the right of the video, there's a summary text in Japanese. Below the video, there's a section titled '人気のアップロード' (Popular Uploads) showing another video thumbnail for 'パワーポイントの图形描画機能でイラストをつくる方法' (How to create illustrations using the drawing function of Microsoft PowerPoint).

<http://gyazo.com/4c4ffa07ba8a0ea1846a2e76be02284e>

- ウェブサイトへのアクセスから結果の見方まで、操作の一挙手一投足がわかります。
 - 講義・講習などの参考資料や後輩指導の教材として利用できます。
 - 本講習中、本家サイトが繋がらない時は、統合TVのYouTube版を見ればお

およその内容がわかるようになっています。

- 今回の講習に関連する内容の多くは、統合TVのゲノム・核酸配列解析カテゴリー <http://togotv.dbcls.jp/ja/genome.html>にあります。
 - 過去の講習会の内容はそのほとんどが統合TVに収録 <http://togotv.dbcls.jp/ja/lecture.html>されており、いつでもどこでも繰り返し復習できるようになっています。
 - お探しの動画が見つからない or 統合TV未掲載の場合は、統合TV番組リクエストフォーム <http://togotv.dbcls.jp/ja/contact.html>へどうぞ!!
 - 統合TVを作つてみたい方、募集中です。ご連絡下さい。
-

ゲノム配列を探す

さまざまな生物種のゲノム配列が解読されています。それを自らの研究に利用しない手はありません。その利用の第一歩としてそれを探してくる方法を学びます。

NCBI

- NCBI(National Center for Biotechnology Information)
 - <http://www.ncbi.nlm.nih.gov/>
- アメリカ合衆国の予算(NIH)による分子生物学情報のリソース
- いろんな調べ物、基本ここから

【実習1】 NCBIからゲノム配列を検索、取得する

1. 中東呼吸器症候群の原因ウイルスのゲノム配列を取得しましょう。日本語しかわからない場合は、NBDCのデータベース横断検索 <http://biosciencedbc.jp/dbsearch/>から英語名を調べましょう。その結果、出てくる

Middle East respiratory syndrome でNCBIから検索します

NCBI Resources How To

NCBI National Center for Biotechnology Information

All Databases Middle East respiratory syndrome Search

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Popular Res

- PubMed
- Bookshelf
- PubMed Cent
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

NCBI Annou

April 8th webin
Minute: Introduct

<http://gyazo.com/4b14584d8cae14de8717a5b1b31d21af>

2. 検索結果を見ましょう。それぞれのDBの横にある数字は、検索ヒット数です。ゲノム配列が得たいので、**Genome** の項目(ヒット数1件)をクリックして詳細を取得します

NCBI Resources How To

Sign in to NCBI Help

Search NCBI databases

Middle East respiratory syndrome Search

Results found in 21 databases for "Middle East respiratory syndrome"

Literature			Genes		
Books	0	books and reports	EST	0	expressed sequence tag sequences
MeSH	3	ontology used for PubMed indexing	Gene	20	collected information about gene loci
NLM Catalog	185	books, journals and more in the NLM Collections	GEO DataSets	1,477	functional genomics studies
PubMed	9,075	scientific & medical abstracts/citations	GEO Profiles	0	gene expression and molecular abundance profiles
PubMed Central	11,613	full-text journal articles	HomoloGene	0	homologous gene sets for selected organisms
Health			PopSet	24	sequence sets from phylogenetic and population studies
ClinVar	0	human variations of clinical significance	UniGene	0	clusters of expressed transcripts
dbGaP	27	genotype/phenotype interaction studies	Proteins		
GTR	0	genetic testing registry	Conserved Domains	0	conserved protein domains
MedGen	1	medical genetics literature and links	Protein	3,032	protein sequences
OMIM	15	online mendelian inheritance in man	Protein Clusters	0	sequence similarity-based protein clusters
PubMed Health	189	clinical effectiveness, disease and drug reports	Structure	37	experimentally-determined biomolecular structures
Genomes			Chemicals		
Assembly	0	genome assembly information	BioSystems	0	molecular pathways with links to genes, proteins and chemicals
BioProject	12	biological projects providing data to NCBI	PubChem BioAssay	5	bioactivity screening studies
BioSample	180	descriptions of biological source materials	PubChem Compound	0	chemical information with structures, information and links
Clone	24	genomic and cDNA clones	PubChem Substance	9	deposited substance and chemical information
dbVar	0	genome structural variation studies			
Epigenomics	0	epigenomic studies and display tools			
Genome	1	genome sequencing projects by organism			
GSS	0	genome survey sequences			
Nucleotide	746	DNA and RNA sequences			
Probe	0	sequence-based probes and primers			
SNP	0	short genetic variations			
SRA	323	high-throughput DNA and RNA sequence read archive			
Taxonomy	0	taxonomic classification and nomenclature catalog			

<https://gyazo.com/59d401eee38017a64719e9ae631ff113>

3. Middle East respiratory syndrome coronavirusのゲノム構造と各種リンクが表示されま

す。**RefSeq** のところにあるIDをクリックするとRefSeq(リファレンス配列(Reference Sequence)のデータベース)のレコードが表示されます

The screenshot shows the NCBI RefSeq genome page for Middle East respiratory syndrome coronavirus (NC_019843.3). The main content includes:

- Organism Overview:** Middle East respiratory syndrome coronavirus
- Lineage:** Viruses[4614]; ssRNA viruses[1244]; ssRNA positive-strand viruses, no DNA stage[982]; Nidovirales[62]; Coronaviridae[44]; Coronavirinae[40]; Betacoronavirus[13]; unclassified Betacoronavirus[6]; Middle East respiratory syndrome coronavirus[1]
- Publications:** 1. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. van Boheemen S, et al. MBio 2012 Nov 20
- Representative (genome information for reference and representative genomes):** Reference genome: [see all organisms] Middle East respiratory syndrome coronavirus ViralProj183710 Submitter: NCBI RefSeq Genome Project
- Genome Region:** A visualization showing the genome structure with segments labeled from 1 to 30,119 bp.

The sidebar on the right contains:

- Tools:** Genome Resource
- Related information:** Assembly, BioProject, Gene, Other genomes for species, Components, Protein, PubMed, Taxonomy
- Search details:** Middle[All Fields] AND East[All Fields] AND respiratory[All Fields] AND syndrome[All Fields]
- Recent activity:** Turn Off, Clear

<http://gyazo.com/fcc3c14b8158238233cd4b6087866db8>

- 得られるRefSeqのレコードを見ましょう。ゲノムにコードされているタンパク質配列などを確認しましょう
- そのレコードの一番下にゲノム配列が記述されています。ゲノム塩基配列を抜き出すには、最上部の **FASTA** をクリックします

The screenshot shows the NCBI Nucleotide search results page for the Middle East respiratory syndrome coronavirus genome. The results include:

- NCBI Reference Sequence: NC_019843.3
- Links: **FASTA**, **Graphics**

<http://gyazo.com/cd454ae926b1299e37680aaa7dc1906f>

- MERSコロナウイルスのゲノム配列が得られました!

Nucleotide

Nucleotide

Advanced

Display Settings: FASTA

Middle East respiratory syndrome coronavirus, complete genome

NCBI Reference Sequence: NC_019843.3

[GenBank](#) [Graphics](#)

```
>gi|667489388|ref|NC_019843.3| Middle East respiratory syndrome coronavirus, complete genome
GATTTAAGTGAATAGCTGGCTATCTCACTTCCCTCGTCTCTGCAGAACTTGATTTAACGAACTT
AAATAAAAAGCCCTGTGTTAGCGTATCGTGCACCTGTCTGGTGGATTGTCGGCATTAAATTGCCCTGCT
CATCTAGGCAGTGGACATATGCTCAACACTGGGTATAATTCTAATTGAATACTATTTTACGTAGAGCG
TCGTGTCTCTTGTACGTCTCGGTACAATACACGGTTCGTCCGGTGCACGGCAATTGGGGCACATCAT
GTCTTCGTGGCTGGTGTGACCGCGCAAGGTGCGCGCGTACGTATCGAGCAGCGCTCAACTCTGAAAAAA
CATCAAGACCAGTGTCTCTAAGTGTGCCACTCTGTGGTTAGGAAACCTGGTTGAAAAACTTACCAT
GGTTCATGGATGGCGAAAATGCCTATGAAGTGGTGAAGGCCATGTTACTTAAAAGGAGGCCATTCTA
TGTGCCCATCCGGCTGGCTGGACACACTAGACACCTCCAGGTCTCGTGTACCTGGTTGAGAGGCTC
ATTGCTTGTGAAAATCCATTGTTAACCAATTGGCTTATAGCTCTAGTGCAAATGGCAGCCTGGTTG
GCACAACTTGCAGGGCAAGCCTATTGGTATGTTCTCCCTTATGACATCGAACTTGTACAGGAAAGCA
AAATATTCTCCTGCAGTATGGCGTGGTTATCACTACACCCATTCCACTATGAGCGAGACAAC
ACCTCTGCCCTGAGTGGATGGACGATTTGAGCGGATCCTAAAGGCAAATATGCCAGAATCTGCTTA
AGAAGTTGATTGGCGGTGATGTCACTCCAGTTGACCAATACATGTTGCGTTGATGGAAAACCCATTAG
TGCCTACGCTATTAAATGGCCAAGGATGGAATAACAAACTGGCTGATGTTGAAGCGGACGTCAGCA
CGTGCCTGATGACGAAGGCTTCATCACATTAAAGAACAACTATAGATTGGTTGGCATGTTGAGCGTA
AAGACGTTCCATATCTAACGAACTCTATTAACTATTAAAGTGTGGTCAAAAGGATGGTGTGAAAAA
CACTCCCTCACTATTAACTCTGGATGCAAATTTAACGCTCACCCACGCAACAAGTGGAGTGGC
GTTTCTGACTTGTCCCTCAAACAAAAACTCCTTACACCTTCTATGGTAAGGAGTCACCTGAGAACCAA
CCTACATTACCACTCCGCATTGAGTGTGGAAGTTGTGGTAATGATTCCCTGGCTTACAGGGAAATGC
TATCCAAGGGTTGCTGTGGATGTGGGCATCATACAGCTAATGATGTCAGTCCAATCATCTGGC
ATGATTAAGCCAATGCTTCTTGTGCTACTTGCCTTGTCAAGGGTGTAGCTGTTCTTCTAATT
GCAAACATTGCTCAGTTGGTAGTTACCTTCTGAACGCTGTAATGTTATTGCTGATTCTAAAGTC
CTTCACACTTATCTTGGTGGCTAGCTACGCCTACTTGGATGTGAGGAAGGTACTATGTTACTTGTG
CCTAGAGCTAAGTGTGTCAGGATTGGAGACTCCATCTTACAGGCTGACTGGCTTGGAAACA
AGGTCACTCAAATTGCTAACATGTTCTGGAACAGACTCAGCATTCCCTAACCTTGTGGAGAGTC
TGTCAACGATGTTGTCCTCGCAATTCTCTGGAACCCACAATGTTGACAAAATACGCCAGCTCTC
AAAGGTGTCACCCCTGACAAGTTGCGTATTAGTACTATGACGTAGCAGTCAGTGCAGCCGGCCAT
TCATGGATAATGCTATTAAATGTTGGTGTACAGGATTACAGTATGCCGCAATTACTGCACCTTATGAGT
TCTCACTGGCTTAGGTGAGTCCTTAAGAAAGTTGCAACCATAACCGTATAAGGTTGCAACTCTGTTAAG
GATACTCTGGCTTATTATGCTCACAGCGTGTGTACAGAGTTTCCCTTATGACATGGATTCTGGTGTG
CATCCTTACTGAACTACTTTGATTGCGTTGATCTTCAGTAGCTTCTACCTATTTTACTGCGCAT
CTTGCAGATAAGACTGGCGACTTTATGTCATAATTACTCCCTGCCAAACTGCTGTTAGTAAGCTT
```

<http://gyazo.com/20fe76da0752e3b5abafcb632c3f69db>

蛇足: GOLD(Genomes Online Database) <https://gold.jgi-psf.org/> というゲノム塩基配列解読プロジェクトを集めたデータベースを使うと、すでに終了したプロジェクトだけでなく、現在進行中のゲノムプロジェクトやメタゲノムプロジェクトについても調べることができます

【参考】 GOLD -Genomes Online Database- の使い方(統合TV)

<http://tgotv.dbcls.jp/ja/20150515.html>

##ゲノム配列から探す ゲノム配列が得られてもその中から欲しい領域を目で見つけてくること(眼grep(「めグレップ」)と呼んでいます)は容易ではありません。そこで、コンピュータの力を借りることになります。

- BLAST(Basic Local Alignment Search Tool)
 - <http://blast.ncbi.nlm.nih.gov/>
 - 前世紀から使われている配列類似性検索のデ・ファクト・スタンダード
 - かつては遠縁の配列相同性を検出するためのツール
 - 今はほぼ完全一致を探すために用いられることが多い



1. 配列解析(主に類似性)の歴史

1970	Needleman-Wunsch法
1977	バクテリオファージ(ϕ X174) ゲノム解読(初)
1981	Smith-Waterman法
1988	FASTA論文, NCBI設立
1990	BLAST論文
1995	<i>H. influenzae</i> ゲノム解読 (free-living organism初)
1997	BLAST2(Gapped BLAST, PSI-BLAST)論文
2002	BLAT(BLAST Like Alignment Tool)論文
2003	<i>Homo sapiens</i> ゲノム解読 (最初のヒトゲノム)
2009	BWA, bowtie論文
2012	GGRNA論文

5

© 2014 DBCLS Licensed under CC BY 2.1 JAPAN

<http://gyazo.com/0805e3f1ea046192de5a52cac47dce58> (スライド出典: 「配列解析基礎」 by 坊農秀雅 <http://www.slideshare.net/sayamatcher/140905bono>)

【余談】なぜ相同性じゃなく類似性か

- 遺伝学では、相同性という言葉はタンパク質のアミノ酸配列や遺伝子の塩基配列が共通の祖先をもつときに用いる
- バイオインフォマティクスでは、タンパク質やDNAでの相同性は、配列類似性に基づいて判断される
 - <http://togetter.com/li/307635>

BLAST® Basic Local Alignment Search Tool My NCBI [Sign In] [Register]

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

New DELTA-BLAST, a more sensitive protein-protein search [Go](#)

BLAST Assembled Genomes

Find Genomic BLAST pages:
Enter organism name or id—completions will be suggested [GO](#)

<input type="checkbox"/> Human	<input type="checkbox"/> Rabbit	<input type="checkbox"/> Zebrafish
<input type="checkbox"/> Mouse	<input type="checkbox"/> Chimp	<input type="checkbox"/> Clawed frog
<input type="checkbox"/> Rat	<input type="checkbox"/> Guinea pig	<input type="checkbox"/> Arabidopsis
<input type="checkbox"/> Cow	<input type="checkbox"/> Fruit fly	<input type="checkbox"/> Rice
<input type="checkbox"/> Pig	<input type="checkbox"/> Honey bee	<input type="checkbox"/> Yeast
<input type="checkbox"/> Dog	<input type="checkbox"/> Chicken	<input type="checkbox"/> Microbes

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontiguous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast, delta-blast</i>
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

Your Recent Results [New!](#)

All Recent results...

News

BLAST XML
The NCBI is now making a new version of the BLAST XML available for testing.
Wed, 29 Apr 2015 18:00:00 EST [More BLAST news...](#)

Tip of the Day [More tips...](#)

<http://gyazo.com/883c9871790d3a201513809d9eb38adc>

- 現在(2016年1月)のBLASTウェブインターフェース
 - Assembled Genomesに対する生物種ごとのBLAST検索が上位に
 - Basic BLAST(以下に説明)が下の方に

DBCLS

配列類似性検索

- query: 質問配列
 - 核酸配列
 - アミノ酸配列
- DB: 検索対象のDB
 - 核酸配列
 - アミノ酸配列
- 閾値などのパラメータ
 - 期待値(E)

DB → query ↓	核酸 配列	アミノ酸 配列
核酸 配列	blastn tblastx	blastx
アミノ酸 配列	tblastn	blastp

*blastnだけが核酸配列レベルでの比較。
残り全てはアミノ酸配列レベルの比較

- GGGenome

- 「じえじえじえのむ」と読みます(参考: GGRNA <http://ggrna.dbcls.jp/>(「ぐぐるな」))
- <http://gggenome.dbcls.jp/>
- 上記のAssembled Genomesに対する高速検索のツール

【実習2】 GGGenomeを使ってゲノム配列から探す

1. <http://gggenome.dbcls.jp/> にアクセスします

超絶高速ゲノム配列検索 [Help](#) | [English](#)

GGGenome

検索窓 検索 Human genome, GRCh37/hg19 (Feb, 2009) フォント選択

許容するミスマッチ/ギャップの数: 0 (検索する塩基配列の長さの25%まで)

双方向を検索 +方向のみ検索 -方向のみ検索

検索例:

- [[TTCATTGACAACATT](#)] 塩基配列を検索
- 詳細な使い方

検索結果へのリンク:

- [http\[s\]://GGGenome.dbcls.jp/db/k/\[strand\]/sequence\[.format\]\[.download\]](http://GGGenome.dbcls.jp/db/k/[strand]/sequence[.format][.download])
 - db → hg19, mm10, rn5, calJac3, susScr3, galGal4, xenTro3, Xenla7, danRer7, ci2, dm3, ce10, TAIR10, rice, sorBic, bmor1, sacCer3, pombe, refseq, hs_refseq, mm_refseq, prok, ddbj。省略時は hg19
 - k → 許容するミスマッチ/ギャップの数。あまり大きいとしぼうする。省略時は 0
 - strand → '+' ('plus') または '-' ('minus') で特定の方向のみ検索。省略時は両方向を検索
 - sequence → 塩基配列。大文字・小文字は区別しない
 - format → html, txt, csv, bed, gff, json。省略時は html
 - download → URLの最後に付加すると検索結果をファイルとしてダウンロードできる
- 例1:<http://GGGenome.dbcls.jp/TTCATTGACAACATT>
 - ヒトゲノム [hg19](#) (省略可)
 - ミスマッチ/ギャップを許容せず (省略可)
 - [TTCATTGACAACATT](#) を検索し
 - html 形式 (省略可) で結果を返す
- 例2:<http://GGGenome.dbcls.jp/mm10/2/+//TTCATTGACAACATTGCGT.txt>
 - マウスゲノム [mm10](#) で
 - 2 ミスマッチ/ギャップまで許容して
 - + 方向に限定して
 - [TTCATTGACAACATTGCGT](#) を検索し
 - txt 形式 (タブ区切りテキスト) で結果を返す

<http://gyazo.com/f0ffa1901878ade96d9f4274b235eb86>

2. **CTGACGGTCA** (10塩基)を入力して「検索」ボタンを押します。この塩基配列がヒトゲノム中で何回出てくるか、簡単にわかります

CTGACGGTCA

検索

Human genome, GRCh37/hg19 (Feb, 2009)

許容するミスマッチ/ギャップの数: 0 (検索する塩基配列の長さの25%まで)

双方向を検索 +方向のみ検索 -方向のみ検索

2015-06-14 09:43:51, GGGenome : Human genome, GRCh37/hg19 (Feb, 2009)

Summary:

- CTGACGGTCA (370)
- TGACCGTCAG (363)
- **TOTAL (733)**

Results:

+鎖および-鎖それぞれ50件まで表示。検索語に色がつきます(ミスマッチ・挿入欠失)。

chr1:762000-762009 ▼762000
GGAATTAGGCTCTGCTGCCCTCTGCTA**CTGACGGTCA**AGGCCTCCTATTGTATTCTGTCCATA

chr1:1245036-1245045 ▼1245036
GGAGGGGGCGGGCAGGCCGGCCCCACCCCT**CTGACGGTCA**CCCTGGTCCCTGAAGCTGCCTGGATATGGT

chr1:1398304-1398313 ▼1398304
CCAGCAAGAAAGGTGGGGCATGTCA**CTGACGGTCA**CAGGTAGGAAGCCAGTGCAGCCTCTAGT

chr1:1730896-1730905 ▼1730896
TGATTACCAATGGAAACAGATGCACAGACT**CTGACGGTCA**TGGGAAGGGACTGCTAACATAGAGAACAC

chr1:2401367-2401376 ▼2401367
GGGTGAAGACCCCAGAGGGGCCCTGTGG**CTGACGGTCA**CTGGAGACAACAGTGAGCCACTTGGTGGG

chr1:3292004-3292013 ▼3292004
GGTCTGAGTCACAGTGTGCTCTTTATT**CTGACGGTCA**GGGCTGCAGCCATGGCTACAAGCCAAAC

chr1:8003172-8003181 ▼8003172
TGAGCTGCCGTCAAATGTGTTCCAAATA**CTGACGGTCA**TCTGAGAGTTACCTGTGACATCTGT

<http://gyazo.com/ca47196af2a9e98a291a7a68ff69177d>

3. 検索ボタンの右にある生物種を **S.cerevisiae** にして検索しなおしましょう。ヒット数はどう変化するでしょうか?

CTGACGGTCA

検索

S. cerevisiae (S288C) genome, sacCer3 (Apr, 2011)

許容するミスマッチ/ギャップの数: 0 (検索する塩基配列の長さの25%まで)
 双方向を検索 +方向のみ検索 -方向のみ検索

2015-06-14 09:45:31, GGGenome : S. cerevisiae (S288C) genome, sacCer3 (Apr, 2011)

Summary:

- CTGACGGTCA (11)
- TGACCGTCAG (5)
- **TOTAL (16)**

Results:

+鎖および-鎖それぞれ50件まで表示。検索語に色がつきます(ミスマッチ・挿入欠失)。

chrX:121260-121269 ▼121260 AAACCACTGCCGCAGCTGTTCCAAATTACTGACGGTCAAGTTCAAGCCACTACAAAAACCACTCAAGC

chrXI:144696-144705 ▼144696 ATATCACCCTGCTGCTGTTCTCAAATAACTGACGGTCAAGTTCAAGCTGCTAAGTCTACTGCCGCTGC

chrXI:144750-144759 ▼144750 AGTCTACTGCCGTGCTGTTCCAAATAACTGACGGTCAAGTTCAAGCTGCTAAGTCTACTGCCGCTGC

chrXI:144804-144813 ▼144804 AGTCTACTGCCGTGCGTTCTCAAATAACTGACGGTCAAGTTCAAGCTGCTAAGTCTACTGCCGCTGC

chrXI:259341-259350 ▼259341 CTGAATCCGCTGCCGCATTCTCAAATCACTGACGGTCAAAATCCAAGCTACTACCAACTGCTACCAACCGA

chrXI:652655-652664 ▼652655 ATATCTTCCGGTTTCATCTTACTAGAACTGACGGTCAAGGTCTTGAGAGTACTCGCTTTTTGA

chrXII:28033-28042 ▼28033 TTAAAGTACTCTAGCATAGCTACAGTGGCTGACGGTCAATTCCAGCACTGTTCAAAATTTTAAAA

<http://gyazo.com/0d519de6d1c0fa8982ff55d8e3b9eae3>

4. 染色体と位置のところにリンクがあり、これをクリックするとUCSC Genome Browserの該当箇所になります。その使い方は次節で説明します

- 【復習用】高速配列検索 GGGenome 《ゲゲゲノム》の使い方(統合TV)
<http://togotv.dbcls.jp/ja/20131025.html>
- English version available here -> GGGenome: a fast and simple DNA sequence search engine(TogoTV) <http://togotv.dbcls.jp/ja/20150514.html>

(参考)LocalBLASTで

検索する質問配列の数が多くなると、いちいちウェブブラウザを開いて貼り付けてクリックして...が大変になります。それを克服する方法として、Local(自分のパソコン)にBLASTのプログラムをインストールしてBLASTを実行する方法(LocalBLAST)があります。詳しくはこちらの資料 <http://motdb.dbcls.jp/?AJACS32%2Fbono#e17b6eed>を御覧ください。実際にセットアップする統合TVもそこから紹介しています。

アノテーションを足して見る

ゲノム解読がなされて終わりではありません。どこに遺伝子がコードされているか、転写因子の結合領域はどこかなど、ゲノム上の座標に対して注釈付けがなされていきます。それが
ゲノムアノテーションです。

ゲノム配列にはバージョンがある

当然最新のものが一番ゲノム配列としては正確なものになっているはずなのですが、**ゲノムアノテーションが最新のものに対しても遅滞なくきちんとなされているわけではなく、利用目的によってうまく選択しないといけないのが現状です。**

- Human
 - GRCh38/hg38
 - GRCh37/hg19
- Mouse
 - GRCm38/mm10
 - NCBI37/mm9

ゲノムブラウザ

それを見る方法としてゲノムブラウザが使われます。ここではウェブブラウザ上で使えるゲノムブラウザとして有名なUCSC Genome Browser(「ゆーしーえすしー げのむ ぶらうざー」)とEnsembl Genome Browser(「あんさんぶる」)の使い方を学びます。

【実習3】 UCSC & Ensembl Genome Browserに隠されたアノテーションを発掘する

- Track HubでFANTOM5, cancer
1. <http://genome.ucsc.edu/> にアクセスします
 2. 上部のメニューバーの **Genomes** にマウスをのせると上述のヒトとマウスのアッセンブリが選択できますが、ここでは気にせずそのままクリックします。
 3. groupに **Mammal**、genomeに **Human**、assemblyに **Feb. 2009 (GRCh37/hg19)** を選び、search termに **HIF1** と入力すると入力補完されるので、一番上の **HIF1A** を選び、submitボタンを押しましょう

The UCSC Genome Browser was created by the [Genome Bioinformatics Group of UC Santa Cruz](#).
Software Copyright (c) The Regents of the University of California. All rights reserved.

group	genome	assembly	position	search term
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr14:62,109,260-62,267,836	HIF1A

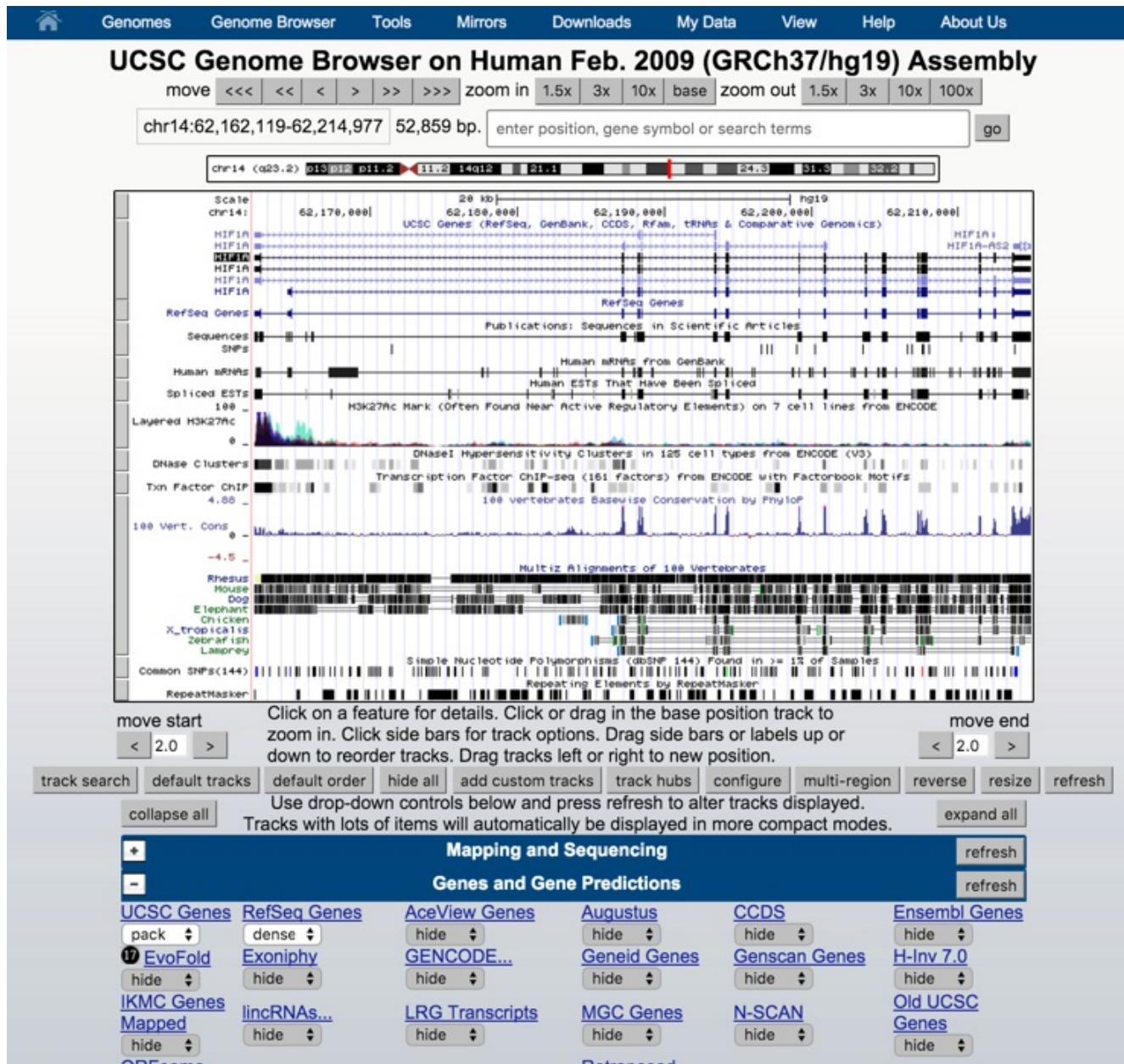
[Click here to reset](#) the browser user interface settings to default

HIF1A (Homo sapiens hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor) (HIF1A), transcript variant 1, mRNA).
 HIF1A-AS2 (Homo sapiens HIF1A antisense RNA 2 (HIF1A-AS2), antisense RNA).
 HIF1AN (Homo sapiens hypoxia inducible factor 1, alpha subunit inhibitor (HIF1AN), mRNA).

Human Genome Browser – hg19 assembly (sequences)

<https://gyazo.com/6b98e53d802a833793e674ff37eb21bf>

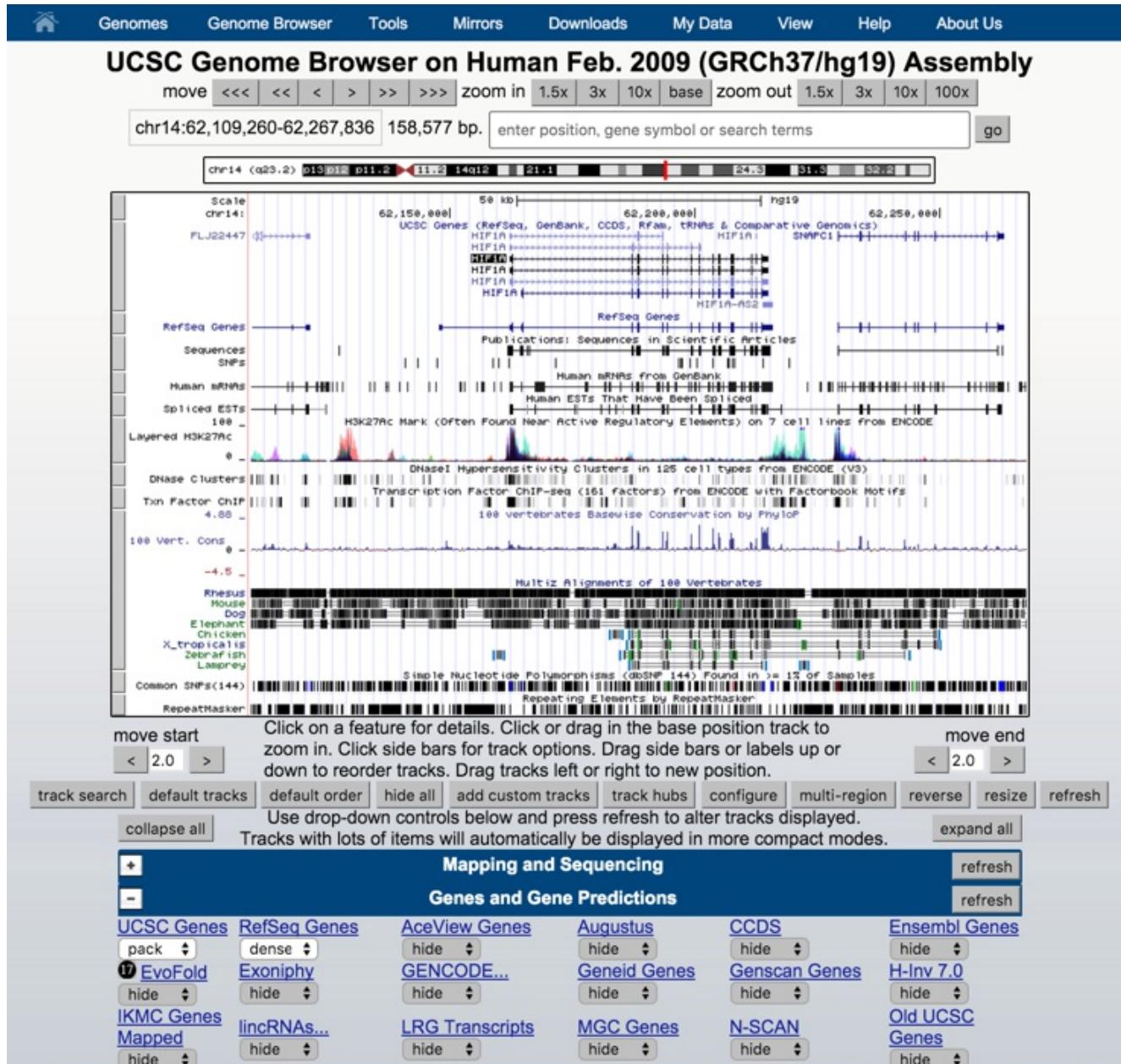
4. HIF1Aがコードされたゲノム上の領域が表示されます



<https://gyazo.com/bdbbb1a3cb3835e4db052b0c5843528f> これがゲノムブラウザの基本画面です。このページにあるさまざまなボタンなどいろいろと操作して必要な情報を追加したり、見たくない情報を削除したりします。

5. 上部のnavigationボタンでmoveやzoom in/out等できますが、遺伝子名検索でたどり着いた場合、mRNAの領域に拡大されて表示されるので、zoom out **3x** しておきましょ

う



<https://gyazo.com/1d6b096e08ec68e38ee8fae6e53f0e72>

6. 画面下の方にあるのがアノテーションです。Phenotype and Literature カテゴリー中の **COSMIC** が'hide'になっているのを **dense** に変えて、**refresh** ボタンを押してみましょう

[Yale](#)
[Pseudo60](#)

hide ▲

Phenotype and Literature

refresh

Publications	ClinGen CNVs	ClinVar Variants	Coriell CNVs	COSMIC	DECIPHER
dense ▲	hide ▲	hide ▲	hide ▲	hide ▲	hide ▲
Development Delay	GAD View	GeneReviews	GWAS Catalog	HGMD Variants	LOVD Variants
hide ▲	hide ▲	hide ▲	hide ▲	hide ▲	hide ▲
18 MGI Mouse QTL	OMIM AV SNPs	OMIM Genes	OMIM Pheno Loci	18 RGD Human QTL	18 RGD Rat QTL
hide ▲	hide ▲	hide ▲	hide ▲	hide ▲	hide ▲
UniProt Variants	Web Sequences				
hide ▲	hide ▲				

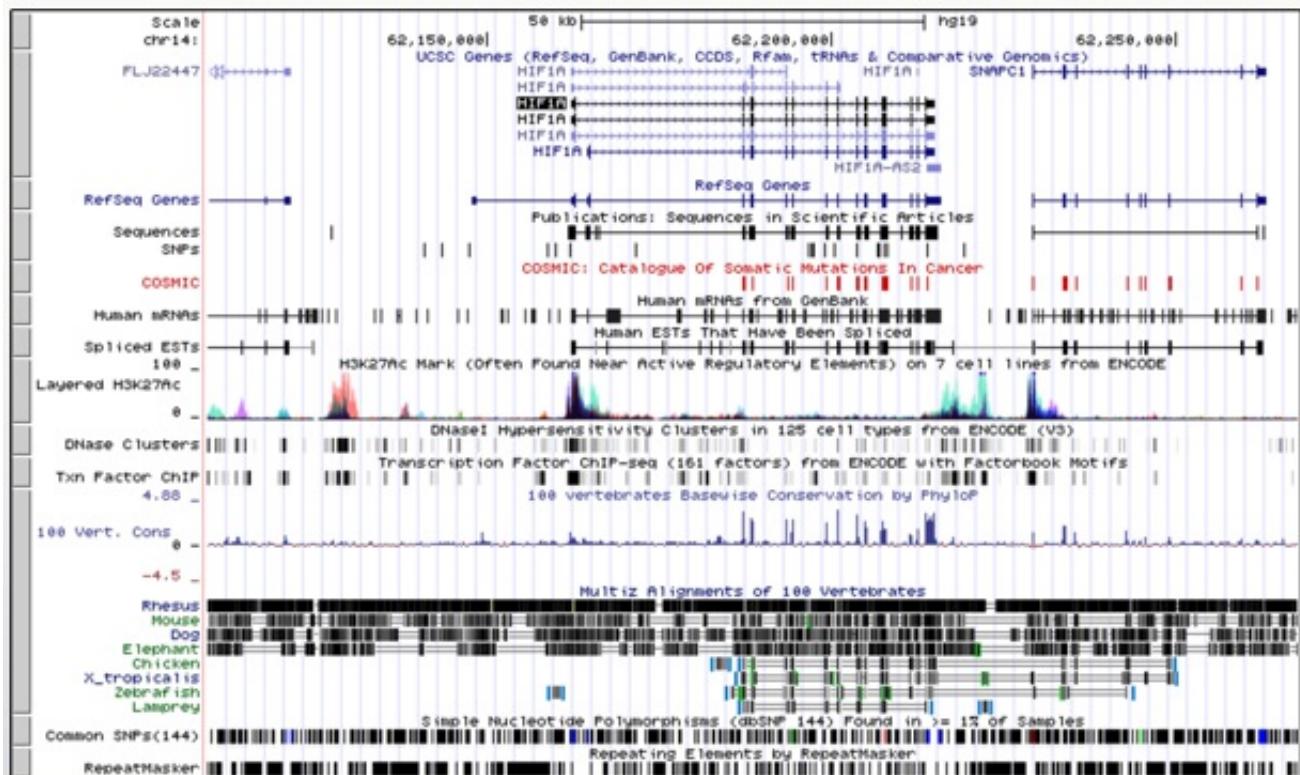
mRNA and EST

refresh

Human mRNAs	Spliced ESTs	18 CGAP SAGE	Gene Bounds	18 H-Inv	Human ESTs
dense ▲	dense ▲	hide ▲	hide ▲	hide ▲	hide ▲

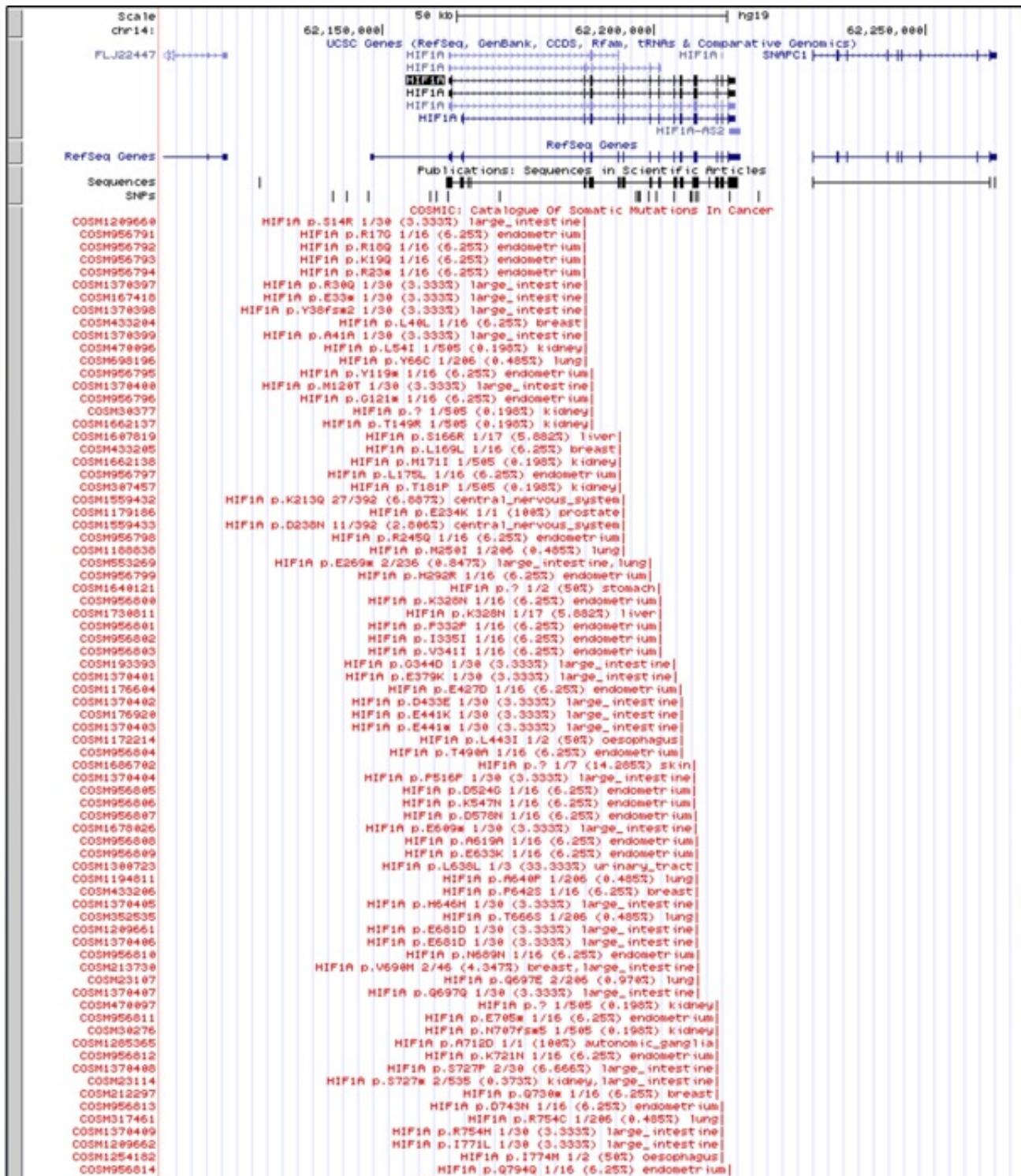
<https://gyazo.com/a4130a4abc68b93803584ce5aa2ff8bb>

7. そうすると、上部のゲノム領域にこのゲノムアノテーション(COSMIC)が付加されて表示されます(画像中央、赤色)



<https://gyazo.com/8ae374468cf7f99b684496068319ffd9>

8. dense,squich, pack, fullとなるに連れて、情報量が多くなります。**full**にしてみると...



<https://qyazo.com/4375db31f539aba1af1af1777546123f>

9. この図は見たい情報のところをクリックすると詳細情報が得られます。たとえば、COSMICの気になる部分のIDをクリックすると

**COSMIC: Catalogue Of Somatic Mutations In Cancer (COSM1209660)**

COSMIC ID: [1209660](#) (details at COSMIC site)

Gene Name: HIF1A

Accession Number: ENST00000337138

Genomic Position: chr14:62187104-62187104

Mutation Description: Substitution - Missense

Mutation Syntax CDS: c.40A>C

Mutation Syntax AA: p.S14R

Mutation NT: a>c

Mutation AA: S>R

Tumor Site: large_intestine

Mutated Samples: 1

Examined Samples: 30

Mutation Frequency: 3.33

Total Mutated Samples: 1

Total Examined Samples: 30

Total Mutation Frequency: 3.333%

Position: [chr14:62187104-62187104](#)

Band: 14q23.2

Genomic Size: 1

[View table schema](#)

[Go to COSMIC track controls](#)

Data version: 68

Data last updated: 2014-02-25

<https://gyazo.com/2b52bd8f59a6b4256f1d069c3b69eff5> その個別のアノテーションの詳細情報が見れます。

10. このゲノムアノテーションそのものについて詳しく知りたい場合、さきほどshowに切り替えた選択画面の上にあったリンクをクリックしてみましょう。詳しい説明が得られます

COSMIC Track Settings

COSMIC: Catalogue Of Somatic Mutations In Cancer ([All Phenotype and Literature tracks](#))

Display mode:

[View table schema](#)

Data version: 68

Data last updated: 2014-02-25

Description

[COSMIC](#), the "Catalogue Of Somatic Mutations In Cancer," is an online database of somatic mutations found in human cancer. Focused exclusively on non-inherited acquired mutations, COSMIC combines information from a range of sources, curating the described relationships between cancer phenotypes and gene (and genomic) mutations. This data is then made available in a number of ways including here in the UCSC genome browser, on the COSMIC website with custom analytical tools, via a federated [Biomart](#), or offline via datasheets downloaded from the [FTP site](#). Publications using COSMIC as a data source may cite any of our references below.

Methods

The data in COSMIC is curated from a number of high-quality sources and combined into a single resource. The sources include:

- Peer-reviewed journal articles
- [CGP laboratories at the Sanger Institute, UK](#)
- [TCGA data portal](#)
- [The ICGC data portal](#)
- [IARC p53 database](#)

Information on known cancer genes, selected from the [Cancer Gene Census](#) is curated manually to maximise its descriptive content. Data from large scale systematic screens are curated semi-automatically, using Vagrent software to reannotate mutant genomic positions (version 0.1 is described at [VAGRENT: Variation Annotation Generator](#)). The full curated dataset is exported from the COSMIC database in CSV format for uploading to UCSC for each bimonthly release; this file is also available on the COSMIC FTP site.

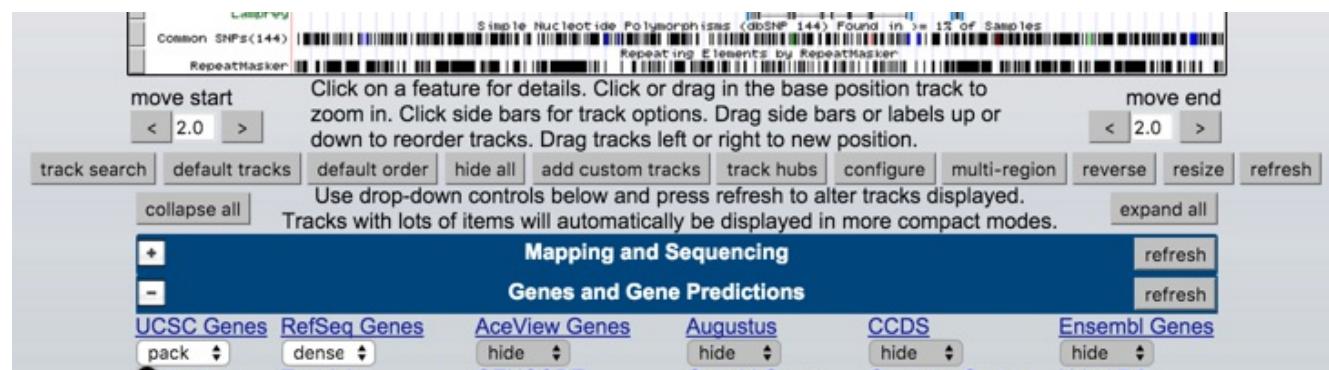
Display

- Dense - Indicate the positions where COSMIC mutations have been annotated in a single horizontal track
- Squish - Indicate each mutation, in vertical pileups where appropriate, whilst minimizing screen space used.
- Pack - Indicate each mutation with COSMIC identifier (COSMnnnnn), mutation annotation, overall mutation frequency and tissues affected, with a link to further details.
- Full - Show each mutation in detail, one per line, with COSM identifier, mutation annotation, overall mutation frequency and tissues

<https://gyazo.com/7c21fba54853479d3d648d878f2fd11c> このようにして必要な情報を見ていくと、自分のほしい情報を得ます。

11. いろいろいじってしまうと元に戻したい時があります。その場合は、

default tracks ボタンを押すとResetされ、元のゲノムアノテーションに簡単に戻せます



<https://gyazo.com/e37518349806c036f070a079b6dfa9cb>

12. このページにあるゲノムアノテーションの検索は、その左の **track search** ボタン

から可能です。cancerをキーワードに検索してみると…

The screenshot shows the UCSC Genome Browser interface. At the top, there's a blue header bar with links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us. Below the header is a light blue search bar titled "Search for Tracks in the Human Feb. 2009 (GRCh37/hg19) Assembly". In the search bar, the word "cancer" is typed. Below the search bar are three buttons: "search", "clear", and "cancel". The main content area has a yellow background. At the top left of this area, it says "return to browser (0 of 100 selected)". On the right, it says "Listing 1 - 100 of 306 tracks 1 2 3 4 »". There's also a "Sort" dropdown menu with options: "by Relevance" (selected), "Alphabetically", and "by Hierarchy". The main list contains 100 entries, each with a "hide" button and a track name. The track names include "COSMIC", "Ag Can 4x180k", "CGAP SAGE", "MCF-7 ZNF217", "SK-N-SH_RA DNase", "MCF7 Z217 UCD", "MCF7 Z217 UCD", "SK-N-SH_RA EP300", "SK-N-SH_RA Sig", "SK-N-SH_RA 2", "SK-N-SH_RA 1", "SK-N-SH_RA", "SK-N-SH_RA Pk", "SK-N-SH_RA Hot", "SK-N-SH_RA 2", "SK-N-SH_RA 1", "SK-N-SH_RA Raw", "SK-N-SH_RA 2", "SK-N-SH_RA 1", "SKNSHRA Pk 2", "SKNSHRA Pk 1", "SKNSHRA Ht 2", "SKNSHRA Ht 1", "SKRA cell pA+", "SKSH cell pA+ + 2", "SKSH cell pA+ + 1", "SKSH cell pA+ - 2", "SKSH cell pAP - 1", "SKSH cell pA+ A 2", "SKSH cell pA+ A 1", "SKNSHRA Sq 2", and "SKNSHRA Sq 1". Most track names have a small downward arrow to their right.

<https://gyazo.com/24ef8682285d7879a8025c078f75bede>

13. さらに、UCSC Genome Browserのサーバー上ではなく、外部で管理されているデータをゲノムブラウザ上に表示することもできます。上部のメニューの My Data の中にあるTrack Hubsをクリックした画面で **cancer** で検索すると以下の様な外部データが利用可能とわかります。

Track Data Hubs

Track data hubs are collections of external tracks that can be imported into the UCSC Genome Browser. Hub tracks show up under the hub's own blue label bar on the main browser page, as well as on the configure page. For more information, see the [User's Guide](#). To import a public hub click its "Connect" button below.

NOTE: Because Track Hubs are created and maintained by external sources, UCSC is not responsible for their content.

Public Hubs
My Hubs

Enter search terms to find in public track hub description pages:

Search Public Hubs

Displayed list **restricted by search terms: cancer** Show All Hubs

Clicking Connect redirects to the gateway page of the selected hub's default assembly.

Display	Hub Name	Description	Assemblies
Connect	Cancer genome polyA site & usage	An in-depth map of polyadenylation sites in cancer (matched-pair tissues and cell lines)	hg19
Connect	ENCODE Analysis Hub	ENCODE Integrative Analysis Data Hub	hg19
Connect	miRcode microRNA sites	Predicted microRNA target sites in GENCODE transcripts	hg19
Connect	Translation Initiation Sites (TIS)	Translation Initiation Sites (TIS) track	hg19
Connect	Broad Improved Canine Annotation v1	Broad Institute CanFam3 Improved Annotation Data v1	canFam3
Connect	CPTAC Hub v1	CPTAC Hub v1	hg19

Contact genome@soe.ucsc.edu to add a public hub.

<https://gyazo.com/0b434ffd362d5f9a846c1de2632a2c4c>

14. また、上部のメニューのうち、Viewをクリックして出てくる **Ensembl** をクリックすると、今見ている領域のEnsembl Genome Browserの該当領域へジャンプします

 ASIA
[Login/Register](#)

BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Human (GRCh38.p5) ▾ Location: 14:62,109,260-62,267,836

Location-based displays

- Whole genome
- Chromosome summary
- Region overview
- **Region in detail**
- Comparative Genomics
- Alignments (image)
- Alignments (text)
- Region Comparison
- Synteny
- Genetic Variation
- Resequencing
- Linkage Data
- Markers
- Other genome browsers
- UCSC
- NCBI
- Vega
- Ensembl GRCh37

Configure this page

Add your data

Export data

Share this page

Bookmark this page

Chromosome 14: 62,109,260-62,267,836

Assembly exceptions Chr. 14 p13 p11.2 q11.2 q12 q21.1 q24.3 q31.3 q32.2

Region in detail

Chromosome bands Contigs Genes (Comprehensive set from GENCODE 24)

Gene Legend merged Ensembl/Havana processed transcript pseudogene

Location: 14:62109260-62267836 Go Gene: Go

Drag>Select: ↕

<https://gyazo.com/07a9546a5bbf54abc0488348c71cafa9> が、HIF1A遺伝子が見当た

りません。なぜでしょうか?

15. これはEnsemblに飛んだ先がGRCh38のゲノムアッセンブリで、**UCSC側**とバージョンが違っていて座標がズれているからです。しょうがないので、真ん中のGene:のところで **HIF1A** を検索しましょう。

Location-based displays

- Whole genome
- Chromosome summary
- Region overview
- Region in detail
- Comparative Genomics
 - Alignments (image)
 - Alignments (text)
 - Region Comparison
 - Synteny
- Genetic Variation
 - Resequencing
 - Linkage Data
 - Markers
- Other genome browsers
 - UCSC
 - NCBI
 - Vega
 - Ensembl GRCh37.p5

Configure this page

Add your data

Export data

Share this page

Bookmark this page

Human (GRCh38.p5) Location: 14:61,695,513-61,748,259

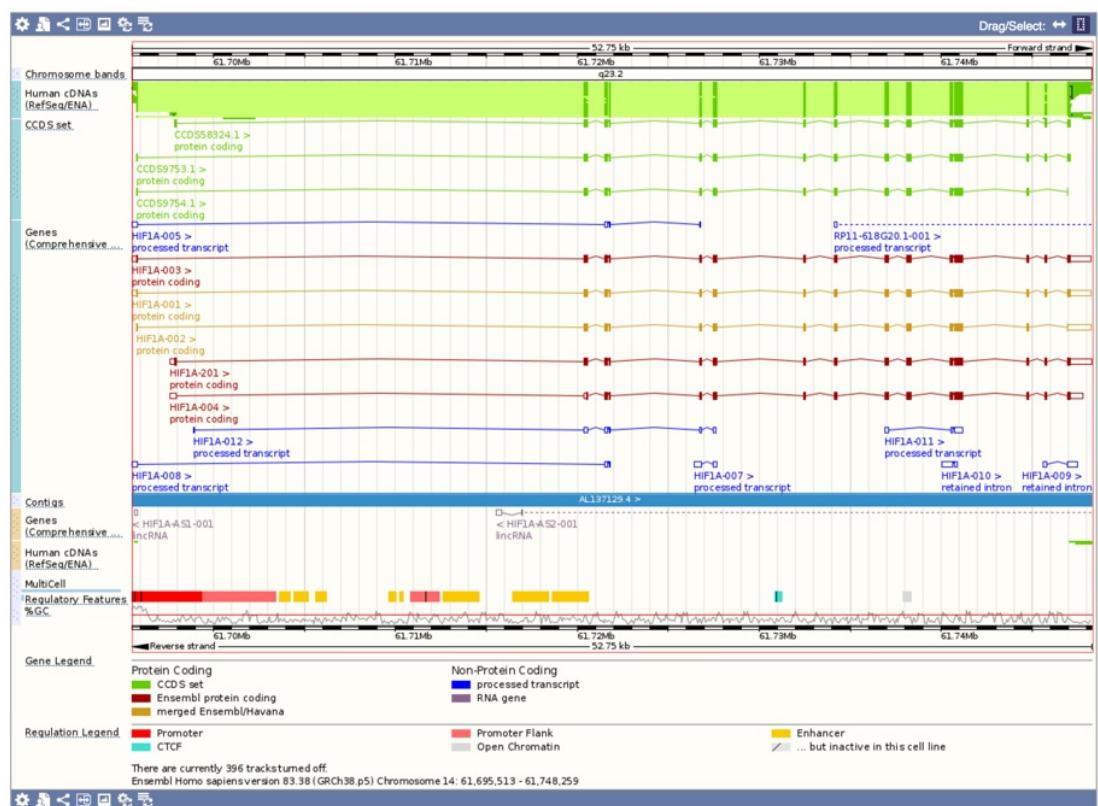
Chromosome 14: 61,695,513-61,748,259

Region in detail

Chromosome 14: 61,695,513-61,748,259

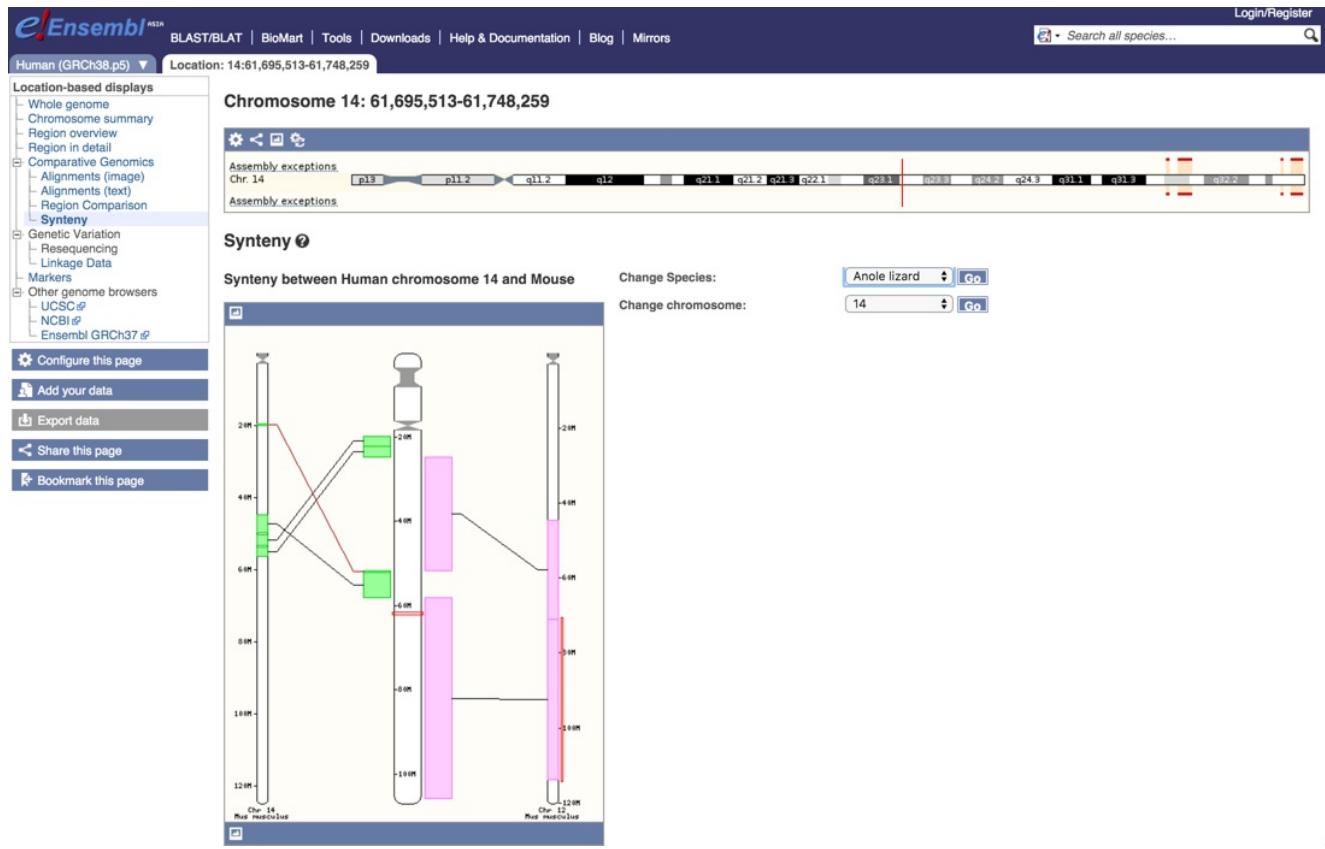
Location: 14:61695513-61748259 Gene: HIF1A

<https://gyazo.com/5663fa4c073b5e119ac01a547c0ce1bb>



<https://gyazo.com/f09fea3c99669a41bc3f334ba1022ceb> 【重要】ゲノム配列はバージョンが同じならどこのサイトでも同一ですが、アノテーションは提供サイトで異なります

16. 左カラムのメニュー中のComparative Genomicsの **Synteny** をクリックすると、ヒトとマウスの間のシンテニーマップが表示されます



<https://gyazo.com/7266536d5e87a533918d7e95449e4fee>

- 【復習用】UCSC Genome Browserの使い方～表示+ENCODE編～2012(統合TV)
<http://tогotv.dbcls.jp/ja/20120528.html>

ここまで出来て時間のある方は、自分の興味のある遺伝子やtrackを試してみましょう。余裕のある方は、ウイルスの持ち出した宿主の遺伝子配列がコードされている領域をアミノ酸配列レベルでゲノム中から探し当てる2012(統合TV)

<http://tогotv.dbcls.jp/ja/20121030.html>で説明されているラウス肉腫ウイルスゲノム配列を取得、その痕跡をニワトリゲノムから探してみるのをやってみましょう。

- 【発展編】UCSC genome browserの使い方～配列取得編～2013
<http://tогotv.dbcls.jp/ja/20131113.html>
- 【発展編】UCSC Genome Browserの使い方～wig形式のファイルをトラックとして追加する～(統合TV) <http://tогotv.dbcls.jp/ja/20120116.html>