

ライフサイエンス分野データの可視化と共有化 at AJACS安芸

情報・システム研究機構 データサイエンス共同利用基盤施設 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 <mailto:bono@dbcls.rois.ac.jp>

2016年7月6日

これは統合データベース講習会AJACS安芸「ライフサイエンス分野データの可視化と共有化」の資料です。

概要

本講習は、データ可視化とそれらを共有化する手段に関して、さまざまなツールを紹介します。とくに、大量の塩基配列データを可視化する手段、ならびに遺伝子発現量などの数値データからそれらの特徴を抽出する方法について説明、実習します。

講習の流れ

今回の講習では、以下の内容について説明します。【実習】と書かれた項目を皆さんと一緒にやって参ります。【発展】はこの場ではやりませんが、早く出来てしまった人は是非取り組んでみてください。

- 研究現場で頻繁に使われるDBやツールを知る: 統合TV
- 配列データの可視化と共有化
 - 多重配列アラインメント
 - ゲノムブラウザ
 - 配列レポジトリ
- 数値データの共有化: 公共データレポジトリ
- 数値データの可視化
 - パイチャート(pie chart)
 - 主成分分析(PCA: Principal Component Analysis)
 - 階層クラスタリング(Hierarchical Clustering)

- ヒートマップ(heatmap)
- ビジネスインテリジェンス(BI)ツールの活用

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://togotv.dbcls.jp/ja/>

- 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://togotv.dbcls.jp/> <http://togotv.dbcls.jp/>

TOGO TV 生命科学系DB・ツール使い倒し系チャンネル

「統合TV」は、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です。

🔍 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

- ゲノム (124)
- 遺伝子 (217)
- タンパク質 (75)
- 配列解析 (116)
- 発現解析 (176)
- NGS (110)
- 文献検索 (72)
- 情報収集 (47)

🔍 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1080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、入力の数ごとに即座に候補の番組が絞り込まれます

GeneTrail2を使って、エンリッチメント解析を行う ▶

マイクロアレイ実験や次世代シーケンシング(NGS)などで得られた大量な遺伝子を扱う際に、どういった機能を持つ遺伝子が変化する傾向にあるかを知るエンリッチメント解析は非常に有用な解析手法です。GeneTrail2は、ウェブ上で多様なエンリッチメント解析ができるツールです。トランスクリプトームだけでなく、タンパク質やmiRNA, SNPのエンリッチメント解析をすることもできます。本動画では、GeneTrail2を使ってNCBI GEOに登録されているマイクロアレイデータのエンリッチメント解析を行う方法について紹介します。

無料統計ソフトEZR (Easy R)を使ってマウス操作だけで多彩な統計解析をする ~導入・基本編~ ▶

Rは、フリーで使える統計解析向けのプログラミング言語です。非常に多彩な解析を行うことができますが、コマンド操作が中心であるため、初心者には敷居が高い面もあります。

<https://gyazo.com/edbabee661757e2a50f6f8ee77c3e778>

- YouTube版もあります <http://youtube.com/togotv/> <http://youtube.com/togotv/>
- 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やすい。
 - 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、「発現解析」タグのついた動画にあります。
 - 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を作ってくれる方、募集中!!
<https://twitter.com/bonohu/status/747954940157399040>

配列データの可視化と共有化

多重配列アラインメント

- Jalview <http://www.jalview.org/> <http://www.jalview.org/>

【発展1】 Jalviewでアミノ酸配列群を可視化する

1. DesktopとApplet版がありますが、Desktop版をJalviewのウェブサイトからインストールしましょう
2. 例として、さまざまな生物種で *PARK7(DJ-1)*のタンパク質配列(**PARK7.fa** <https://raw.githubusercontent.com/AJACS-training/AJACS60/master/bono2/PARK7.fa>)を File -> Input Alignment -> From File から読み込んでみましょう。
3. 読み込んで現れたWindowの中のメニューの Web Service -> Alignment -> ClustalO -> withDefaults を選んでみましょう。しばらく待つと多重配列アラインメントが計算されて返ってきます。
4. さらに Calculate -> Calculate Tree -> Neighbor Joining Using... を選ぶと系統樹が描かれます
5. 配列名がIDばかりで無味乾燥? **PARK7.out2.fa** <https://raw.githubusercontent.com/AJACS-training/AJACS60/master/bono2/PARK7.out2.fa>)を代替りの読み込んでみましょう

【復習用 統合TV】 Jalviewを使って配列解析・系統樹作成をする2013
<http://doi.org/10.7875/togotv.2013.049>

ゲノムブラウザ

- UCSC Genome Browser <http://genome.ucsc.edu/> <http://genome.ucsc.edu/>

【実習1】 UCSC Genome Browserで公共データを可視化する

1. <http://genome.ucsc.edu/> <http://genome.ucsc.edu/> にアクセスし、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
2. mirrorサーバー選択のページがでますので、 **genome-asia.ucsc.edu** をクリック



You might want to navigate to your nearest mirror - genome-asia.ucsc.edu

- User settings (sessions and custom tracks) will differ between sites. [Read more.](#)
- Take me to genome-asia.ucsc.edu
- Let me stay here genome.ucsc.edu

<https://gyazo.com/378a8f0db4792fbe2b42b5ed46d00484>

3. Humanアイコンをクリックして、Assemblyに Feb.2009 (GRCh37/hg19) を選び、Position/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、GOボタンをクリック

Browse/Select Species

POPULAR SPECIES

Human Mouse Rat Fruitfly Worm Yeast

Enter species or common name

REPRESENTED SPECIES

Human
Chimp
Bonobo
Gorilla
Orangutan
Gibbon
Crab-eating macaque
Rhesus
Baboon (anubis)
Baboon (hamadryas)
Marmoset
Squirrel monkey

Find Position

Human Assembly
Feb. 2009 (GRCh37/hg19)

Position/Search Term
HIF1

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

The February 2009 human reference sequence (GRCh37) was produced by the **Genome Reference Consortium**. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the **User's Guide** for more information.

Homo sapiens
(Graphic courtesy of CBSE)

<https://gyazo.com/6ede06a7f0a4ce62ac054038b6624857>

4. HIF1Aがコードされたゲノム上の領域が表示されます

Genomes Genome Browser Tools Mirrors Downloads My Data View Help About Us

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr14:62,162,119-62,214,977 52,859 bp. enter position, gene symbol or search terms go

Scale chr14: 62,170,000 20 kb hg19 62,180,000 62,190,000 62,200,000 62,210,000

UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)

RefSeq Genes Publications: Sequences in Scientific Articles

Sequences SNPs Human mRNAs from GenBank Human ESTs That Have Been Spliced

Spliced ESTs 100 Layered H3K27ac H3K27ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

DNase Clusters DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V5)

Txn Factor ChIP 4.80 Transcription Factor ChIP-seq (151 factors) from ENCODE with Factorbook Motifs

100 Vert. Cons 0 -4.5 100 vertebrates Basepair Conservation by PhyloP

Multiple Alignments of 100 Vertebrates Rhesus Mouse Dog Chicken X_tropicalis Zebrafish Lamprey

Simple Nucleotide Polymorphisms (dbSNP 144) Found in >= 1% of Samples

Repeating Elements by RepeatMasker

RepeatMasker

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

move start < 2.0 > move end < 2.0 >

track search default tracks default order hide all add custom tracks track hubs configure multi-region reverse resize refresh

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

collapse all expand all

Mapping and Sequencing refresh

Genes and Gene Predictions refresh

UCSC Genes	RefSeq Genes	AceView Genes	Augustus	CCDS	Ensembl Genes
pack ▾	dense ▾	hide ▾	hide ▾	hide ▾	hide ▾
EvoFold	Exoniphy	GENCODE...	Geneid Genes	Genscan Genes	H-Inv 7.0
hide ▾	hide ▾	hide ▾	hide ▾	hide ▾	hide ▾
IKMC Genes	lincRNAs...	LRG Transcripts	MGC Genes	N-SCAN	Old UCSC Genes
hide ▾	hide ▾	hide ▾	hide ▾	hide ▾	hide ▾
ORFome			Retroposed		

<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

Genomes Genome Browser Tools Mirrors Downloads My Data View Help About Us

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr14:62,109,260-62,267,836 158,577 bp.

Scale chr14: 59 kb hg19 62,150,000 62,200,000 62,250,000

UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNA & Comparative Genomics)

RefSeq Genes

Publications: Sequences in Scientific Articles

Human mRNAs from GenBank

Human ESTs That Have Been Spliced

Layered H3K27Ac

H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

DNase Clusters

DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)

Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs

100 Vert. Cons

100 vertebrates Basewise Conservation, by PhyloP

MultiZ Alignments of 100 Vertebrates

Rhesus Mouse Dog Elephant Chicken X_tropicalis Zebrafish Lamprocy

Simple Nucleotide Polymorphisms (SNP 144) Found in >= 1% of Samples

Repeating Elements by RepeatMasker

move start < 2.0 > move end < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

track search default tracks default order hide all add custom tracks track hubs configure multi-region reverse resize refresh

collapse all Use drop-down controls below and press refresh to alter tracks displayed. expand all

Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing

Genes and Gene Predictions

UCSC Genes	RefSeq Genes	AceView Genes	Augustus	CCDS	Ensembl Genes
pack	dense	hide	hide	hide	hide
EvoFold	Exoniphy	GENCODE...	Geneid Genes	Genscan Genes	H-Inv 7.0
hide	hide	hide	hide	hide	hide
KMC Genes	lincRNAs...	LRG Transcripts	MGC Genes	N-SCAN	Old UCSC Genes
Mapped	hide	hide	hide	hide	hide
hide	hide	hide	hide	hide	hide

<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** ボタンを押してみてください

Yale
Pseudo60
hide

Phenotype and Literature refresh

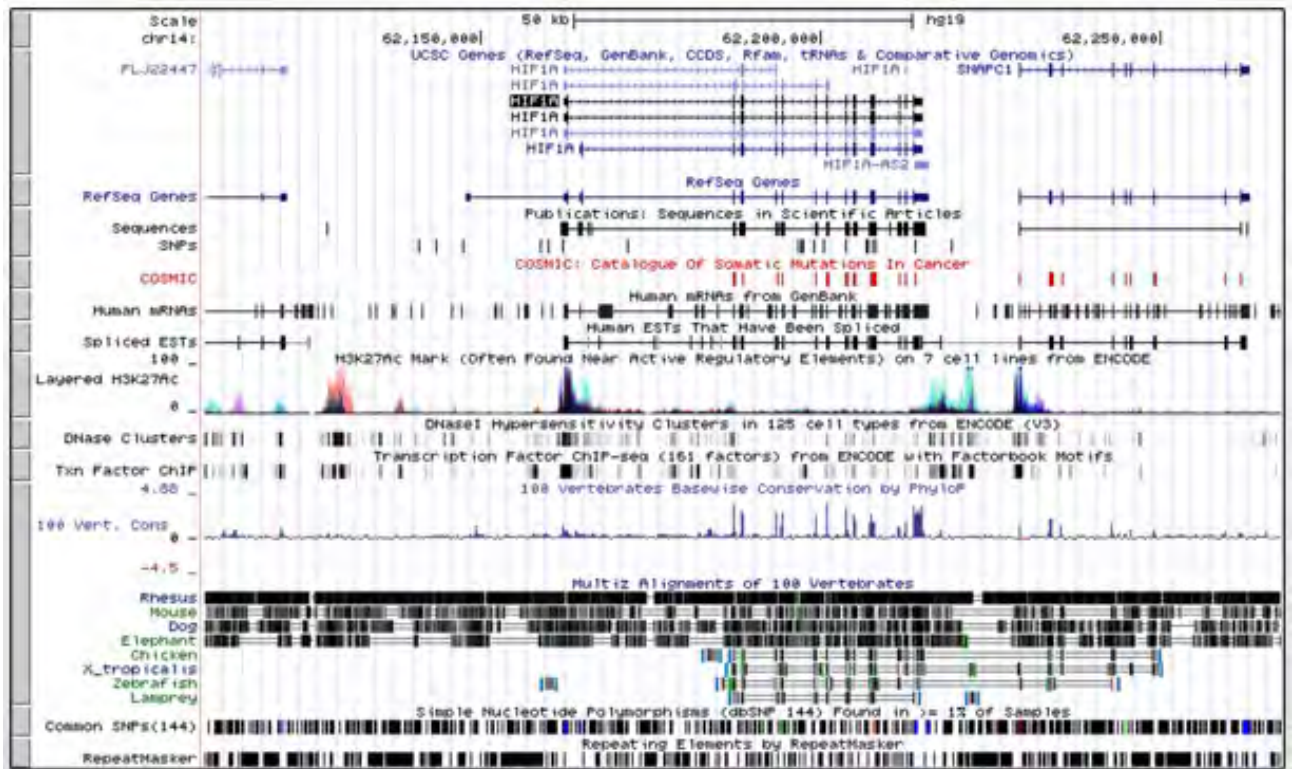
Publications dense	ClinGen CNVs hide	ClinVar Variants hide	Coriell CNVs hide	COSMIC hide	DECIPHER hide
Development Delay hide	GAD View hide	GeneReviews hide	GWAS Catalog hide	HGMD Variants hide	LOVD Variants hide
MGI Mouse QTL hide	OMIM AV SNPs hide	OMIM Genes hide	OMIM Pheno Loci hide	RGD Human QTL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

mRNA and EST refresh

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

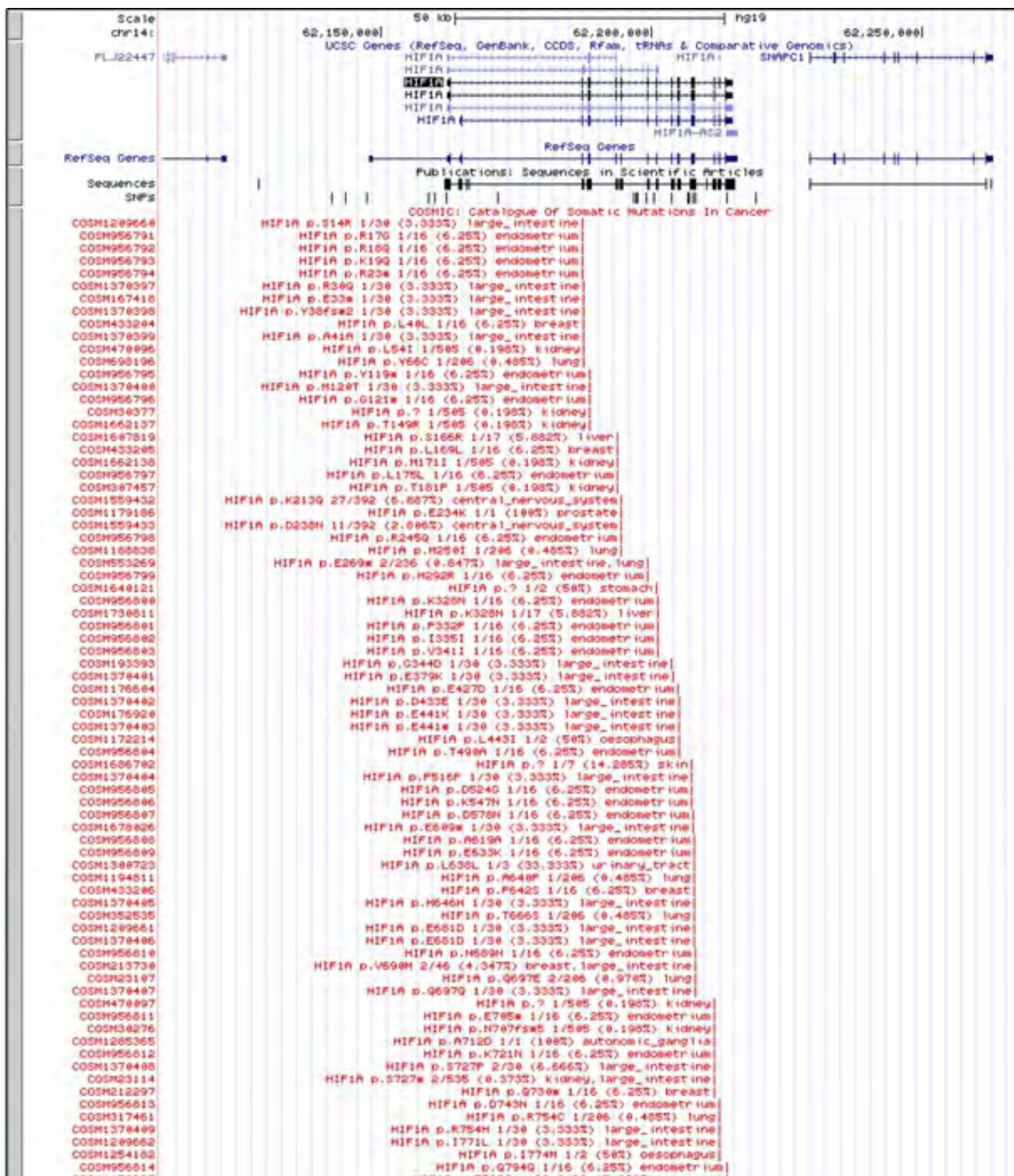
<https://gyazo.com/a4130a4abc68b93803584ce5aa2ff8bb>

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/8ae374468cf7f99b684496068319fd9>

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://gyazo.com/4375db31f539aba1af1af1777546123f>

- 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、その個別のアノテーションの詳細情報が見えます。

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

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>

- 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

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

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足して行って、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

move start move end

track search **default tracks** default order hide all add custom tracks track hubs configure multi-region reverse resize refresh

collapse all

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing

Genes and Gene Predictions

UCSC Genes RefSeq Genes AceView Genes Augustus CCDS Ensembl Genes

pack dense hide hide hide

<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...

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Search for Tracks in the Human Feb. 2009 (GRCh37/hg19) Assembly

Search

return to browser (0 of 100 selected) Listing 1 - 100 of 306 tracks 1 2 3 4 > >>

Visibility	Track Name	Description
<input type="checkbox"/> hide	COSMIC	COSMIC: Catalogue Of Somatic Mutations In Cancer
<input type="checkbox"/> hide	Ag Can 4x180k	Agilent SurePrint G3 Human CGH+SNP Cancer Microarray 4x180K AMADID 030587
<input type="checkbox"/> hide	CGAP SAGE	CGAP Long SAGE
<input type="checkbox"/> hide	MCF-7 ZNF217	MCF-7 TFBS Uniform Peaks of ZNF217 from ENCODE/USC/Analysis
<input type="checkbox"/> hide	SK-N-SH_RA DNase	SK-N-SH_RA DNase HS Uniform Peaks from ENCODE/Analysis
<input type="checkbox"/> hide	MCF7 Z217 UCD	MCF-7 ZNF217 UC Davis ChIP-seq Signal from ENCODE/SYDH
<input type="checkbox"/> hide	MCF7 Z217 UCD	MCF-7 ZNF217 UC Davis ChIP-seq Peaks from ENCODE/SYDH
<input type="checkbox"/> hide	SK-N-SH_RA EP300	SK-N-SH_RA TFBS Uniform Peaks of p300 from ENCODE/HudsonAlpha/Analysis
<input type="checkbox"/> hide	SK-N-SH_RA Sig	SK-N-SH_RA DNase DGF Per-base Signal from ENCODE/UW
<input type="checkbox"/> hide	SK-N-SH_RA 2	SK-N-SH_RA Exon array Signal Rep 2 from ENCODE/UW
<input type="checkbox"/> hide	SK-N-SH_RA 1	SK-N-SH_RA Exon array Signal Rep 1 from ENCODE/UW
<input type="checkbox"/> hide	SK-N-SH_RA	SK-N-SH_RA Methylation 450K Bead Array from ENCODE/HAIB
<input type="checkbox"/> hide	SK-N-SH_RA Pk	SK-N-SH_RA DNase DGF Peaks from ENCODE/UW
<input type="checkbox"/> hide	SK-N-SH_RA Hot	SK-N-SH_RA DNase DGF Hotspots from ENCODE/UW
<input type="checkbox"/> hide	SK-N-SH_RA 2	SK-N-SH BC RA Methyl-RRBS Rep 2 from ENCODE/HudsonAlpha
<input type="checkbox"/> hide	SK-N-SH_RA 1	SK-N-SH BC RA Methyl-RRBS Rep 1 from ENCODE/HudsonAlpha
<input type="checkbox"/> hide	SK-N-SH_RA Raw	SK-N-SH_RA DNase DGF Raw Signal from ENCODE/UW
<input type="checkbox"/> hide	SK-N-SH_RA 2	SK-N-SH_RA Copy number variants Replicate 2 from ENCODE/HAIB
<input type="checkbox"/> hide	SK-N-SH_RA 1	SK-N-SH_RA Copy number variants Replicate 1 from ENCODE/HAIB
<input type="checkbox"/> hide	SKNSHRA Pk 2	SK-N-SH RA DNase HS Peaks Rep 2 from ENCODE/UW
<input type="checkbox"/> hide	SKNSHRA Pk 1	SK-N-SH RA DNase HS Peaks Rep 1 from ENCODE/UW
<input type="checkbox"/> hide	SKNSHRA Ht 2	SK-N-SH RA DNase HS HotSpots Rep 2 from ENCODE/UW
<input type="checkbox"/> hide	SKNSHRA Ht 1	SK-N-SH RA DNase HS HotSpots Rep 1 from ENCODE/UW
<input type="checkbox"/> hide	SKRA cell pA+	SK-N-SH RA whole cell polyA+ CAGE TSS HMM from ENCODE/RIKEN
<input type="checkbox"/> hide	SKSH cell pA+ + 2	SK-N-SH_RA whole cell polyA+ CAGE Plus start sites Rep 2 from ENCODE/RIKEN
<input type="checkbox"/> hide	SKSH cell pA+ + 1	SK-N-SH_RA whole cell polyA+ CAGE Plus start sites Rep 1 from ENCODE/RIKEN
<input type="checkbox"/> hide	SKSH cell pA+ - 2	SK-N-SH_RA whole cell polyA+ CAGE Minus start sites Rep 2 from ENCODE/RIKEN
<input type="checkbox"/> hide	SKSH cell pA+ - 1	SK-N-SH_RA whole cell polyA+ CAGE Minus start sites Rep 1 from ENCODE/RIKEN
<input type="checkbox"/> hide	SKSH cell pA+ A 2	SK-N-SH_RA whole cell polyA+ CAGE Alignments Rep 2 from ENCODE/RIKEN
<input type="checkbox"/> hide	SKSH cell pA+ A 1	SK-N-SH_RA whole cell polyA+ CAGE Alignments Rep 1 from ENCODE/RIKEN
<input type="checkbox"/> hide	SKNSHRA Sq 2	SK-N-SH RA DNase HS Raw Signal Rep 2 from ENCODE/UW
<input type="checkbox"/> hide	SKNSHRA Sq 1	SK-N-SH RA DNase HS Raw Signal Rep 1 from ENCODE/UW

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

【発展2】 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
<input type="button" value="Connect"/>	Cancer genome polyA site & usage	An in-depth map of polyadenylation sites in cancer (matched-pair tissues and cell lines)	hg19
<input type="button" value="Connect"/>	ENCODE Analysis Hub	ENCODE Integrative Analysis Data Hub	hg19
<input type="button" value="Connect"/>	miRcode microRNA sites	Predicted microRNA target sites in GENCODE transcripts	hg19
<input type="button" value="Connect"/>	Translation Initiation Sites (TIS)	Translation Initiation Sites (TIS) track	hg19
<input type="button" value="Connect"/>	Broad Improved Canine Annotation v1	Broad Institute CanFam3 Improved Annotation Data v1	canFam3
<input type="button" value="Connect"/>	CPTAC Hub v1	CPTAC Hub v1	hg19

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

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

- 【復習用 統合TV】 [UCSC Genome Browserの使い方～配列取得編～2013] (<http://doi.org/10.7875/togotv.2013.087>)
- [【復習用】 UCSC Genome Browserの使い方～表示+ENCODE編～ 2012(統合TV)] (<http://togotv.dbcls.jp/ja/20120528.html>)

- Ensembl Genome Browser <http://www.ensembl.org/> <http://www.ensembl.org/>
- IGV(Integrative Genomics Viewer) <https://www.broadinstitute.org/igv/>
<https://www.broadinstitute.org/igv/>
 - 【復習用 統合TV】 Integrative Genomics Viewer IGVを使い倒す ～基本編～ <http://doi.org/10.7875/togotv.2014.027>
 - 【復習用 統合TV】 Integrative Genomics Viewer IGVを使い倒す ～マッピングデータを可視化する～ <http://doi.org/10.7875/togotv.2014.038>

配列レポジトリ

- DBCLS SRA <http://sra.dbcls.jp/> <http://sra.dbcls.jp/>
- AOE <http://aoe.dbcls.jp/> <http://aoe.dbcls.jp/>

数値データの共有化: 公共データレポジトリ

- Figshare <http://figshare.com/>
- DRYAD <http://datadryad.org/>

数値データの可視化

1. パイチャート (pie chart)

【実習2】 Rを起動し以下のファイルに記述されたRのコマンド(`01piechart.r`)を実行しなさい。#(シャープ)から始まる行は、コメント行で、プログラムの実行には関係ありません(=入力しなくてよい)。データファイルとして必要な `srabystudy.txt` をgithubのサイトからダウンロードし、現在作業しているディレクトリにおいておく必要があります。実行後、そのディレクトリに `pie1.png` というファイルが新たに生成されていることを確認しなさい。

```
#出力する画像のファイル名を指定します
png("pie1.png")
#タブ区切りのデータを読み込みます
dat <- read.delim2("srabystudy.txt", header=F)
#columnにお名前を
names(dat) <- c("Study", "freq")
#列を結合
dat <- cbind(dat, serial=seq(dim(dat)[1]))
#値が0じゃないデータだけを使います
dat2 <- dat[ dat$freq != 0, ]
#パイチャートを描きます
pie(dat2$freq, labels=paste(dat2$Study, dat2$freq),
    col=rainbow(dim(dat[1])[dat2$serial]), main="SRA by Study")
#画像を完成させるおまじない。忘れずに!
dev.off()
```

先ほど説明のあったSRA(Sequence Read Archive)のメタデータを取りまとめてStudyのタイプで分けた結果がこのデータ(`srabystudy.txt`)で、それを元にpie chartを描いてみるのが本課題でした。

【発展3】 上記の `srabystudy.txt` を別のデータに変えて実行してみましょう。

- `srabystudy_orig.txt` : SRAのStudyの分類をoriginalデータ

- `srabyorganism.txt` : SRAに登録されている生物種で分類したデータ

これらの結果は、お配りしたファイル群の `results` というフォルダの中においてあります(それぞれ、 `pie1.png` , `pie2.png` , `pie3.png`)。実は、このスクリプトはBono H et al. Genome Res 2003 <http://genome.cshlp.org/content/13/6b/1318.full>の図1を作成するために使ったものです。実際には、このRコードを生成するPerlのプログラムを作成してからそれをRで実行していました。現在よく使われているGSEA(Gene Set Enrichment Analysis)のハシリで、遺伝子にアノテーションされたGene Ontologyの高次(根っこに近い)タームで集計してその分布を見ていました

2. 階層的クラスタリング(hierachical clustering)

【実習3】 Rを起動し以下のファイルに記述されたRのコマンド(`02hclust.r`)を実行下さい。データファイルとして必要な `matrix.txt` をgithubのサイトからダウンロードし、現在作業しているディレクトリにおいておく必要があります。実行後、そのディレクトリに `hclust.png` というファイルが新たに生成されていることを確認下さい。

```
#出力する画像のファイル名を指定します
png("hclust.png")
#データを読み込みます(スペース区切り)
d <- read.table('matrix.txt')
#UPGMA法で階層的クラスタリングを実行
c <- hclust(as.dist(d), method = 'average')
#結果をプロット
plot(c, hang=-1)
#画像を完成させるおまじない。忘れずに!
dev.off()
```

このプログラムはかつての蛋白質・核酸・酵素に寄稿したレビューで紹介したもので、当時階層的クラスタリングをフリーウェアで実行する手段として重宝された(はず)。Bono H and Nakao MC, PNE 2003 <http://www.ncbi.nlm.nih.gov/pubmed/12638180>

3. 主成分分析(PCA)

R/Bioconductorのaffyパッケージで発現データを定量化して数値データにして、それを使ってPCAを実行します。

Affymetrixデータ正規化

RMAの計算が重く、パソコンによってはメモリ不足となり実行不可能となることが予想されますので、できれば。

【参考】 ぼうのブログ: justRMAでnormalize <http://bonohu.jp/blog/2013/06/17/justrma/>

【発展4】 Bioconductorを利用します。Rを起動して以下のコマンドでaffy libraryをインストール
しなさい。

```
source("http://bioconductor.org/biocLite.R")
biocLite("affy")
```

そして、 affy library中のjustRMA関数を利用して、RMA(Robust Multichip Average)正規化してみ
ましょう。自らのAffymetrix Genechip データ(CELファイル)がある人はそれを、ない人は
GSE17264 Comparative transcriptome analysis of dedifferentiation in porcine mature adipocytes
and follicular granulosa cellsを例に実行しましょう。GEO(**GSE17264**)
<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17264>もしくはArrayExpress(
E-GEOD-17264) <https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-17264/>から生データ
(CELファイル)をダウンロードできます。 実行する際、 Session → Set Working Directory を、
CELファイルをダウンロードしてきたディレクトリに指定することがポイントです。それができ
たら、以下のコード(**03justRMA.r**)を実行します。

```
#Bioconductorを使うときの呪文
source("http://bioconductor.org/biocLite.R")
#affyライブラリをインストール
biocLite("affy")
#affyライブラリを召喚
library(affy)
#justRMAを実行、RMA.txtというファイルに出力
write.exprs(justRMA(), file="RMA.txt")
```

その結果生成される **RMA.txt** ファイルがRMAによって正規化された遺伝子発現データになりま
す。

【発展5】 上記の実習のサンプルデータは **GPL3533 [Porcine] Affymetrix Porcine Genome
Array** と呼ばれる(カタログ)マイクロアレイ(プラットフォーム)を使ったデータでした。同じマイ
クロアレイ(プラットフォーム)のデータであればjustRMAを実行することが出来ます。同じプラ
ットフォームのデータの中からiPS細胞のものを探しだして上記のデータに混ぜてjustRMAを実行し
なさい。

解答例: GSE15472 Induced Pluripotent Stem Cells from the Pig Somatic Cells
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15472> の3つのサンプル
(GSM388154,GSM388155,GSM388156)がそれです。 CELファイルをダウンロードして
同じdirectoryに置き、再度justRMAを実行してみましょう

このマイクロアレイデータは以下の論文で我々が発表したもので、全く同じ手段で正規
化を行いました。また、発展課題にある公共遺伝子発現データを足したデータ解釈はや
はりこの論文で実際に発表したもので、我々のマイクロアレイデータの生物学的解釈
(DFATがMA(Mature AdipocyteよりもiPSに近い)に役だっています。 Ono H, Oki Y, Bono

主成分分析(PCA)

PCA(Primary Component Analysis)。このパートはぼうのブログ「RでPCA」<http://bonohu.jp/blog/2013/07/13/pcaonr/>を参考に。

【実習4】 Rを起動し以下のファイルに記述されたRのコマンド(`03pca.r`)を実行しなさい。データファイルとして必要な `RMA.txt` は【発展4】で作成したものを利用すればよいのですが、現在作業しているディレクトリにおいておく必要があります。

この発展課題をやらなくても実行できるように用意しており、 `RMA.txt.zip` というファイルをダブルクリックすることで解凍(圧縮を解く)すると `RMA.txt` というファイルが生成されるようになっています。

PCAを実行するには、以下のRスクリプトを実行します。

```
#タブ区切りのデータを読み込み
data <- read.table("RMA.txt", header=TRUE, row.names=1, sep="\t",
quote="")
#主成分分析を実行
data.pca <- prcomp(t(data))
#名前を得る
names(data.pca)
#標準偏差をプロット
plot(data.pca$sdev, type="h", main="PCA s.d.")
#行列を転置
data.pca.sample <- t(data) %*% data.pca$rotation[,1:2]
#PCAの結果をプロット
plot(data.pca.sample, main="PCA")
#ラベルに色付け
text(data.pca.sample, colnames(data), col = c(rep("red", 3),
rep("blue", 3), rep("green", 3), rep("black", 3)))
```

図に.CEL.gzの文字がいっぱい入っていて見づらい?そう思った方はこちらを参照
<http://bonohu.jp/blog/2014/09/10/afterjustрма/>

4. ヒートマップ(cuffdiffの結果可視化)

RNA-seqデータ解析プログラムの1つであるcufflinksパッケージに `cuffdiff` という発現データの差分を計算するプログラムがあります。それを実行したあとのデータを可視化する手段として使うBioconductorのパッケージに `cummeRbund` があります(cuffdiffの使用例 <http://dx.doi.org/10.1371%2Fjournal.pone.0104283>)。

- 参考: bonohuの発表資料「NGS解析(RNAseq)」
<http://dx.doi.org/10.6084/m9.figshare.1216717>

【発展6】 Rを起動し以下のファイルに記述されたRのコマンドを実行しなさい。データファイル

として必要なcuffdiffディレクトリ以下のファイルは手持ちのものか、なければサンプルをUSBメモリでお渡しします。現在作業しているディレクトリにおいておく必要があります。

```
#Bioconductorを使うときの呪文
source("http://bioconductor.org/biocLite.R")
#cummeRbundをインストール
biocLite("cummeRbund")
#cummeRbundを召喚
library("cummeRbund")
#cuffdiffの結果のディレクトリを指定
cuff.dir <- "cuffdiff"
#cuffdiffの結果の読み込む
cuff <- readCufflinks(dir=cuff.dir)
```

準備はここまでで、

```
cuff
```

で、読み込んだデータのサマリが見れます。以下、解析事例です。

```
#発現密度分布のプロット
dens <- csDensity(genes(cuff))
dens

# サンプル方向のデンドログラム
dend <- csDendro(genes(cuff))

#CSV形式でFPKM値を出力
gene.matrix <- fpkmMatrix(genes(cuff))
write.csv(gene.matrix, file="fpkm.csv")
```

詳しくは、マニュアルを読みましょう。

```
?cummeRbund
```

ビジネスインテリジェンス(BI)ツールの活用

- Spotfire <http://spotfire.tibco.jp/> <http://spotfire.tibco.jp/>
- Yellowfin <http://yellowfin.jp/> <http://yellowfin.jp/>
- Tableau <http://www.tableau.com/> <http://www.tableau.com/>