

BLAT、In-Silico PCRを使って配列を検索する / Gene Sorterを使って発現データを解析する

BLATとは

BLATは、NCBI-BLASTを高速化したツールとして、UCSC GBに実装されています。localで利用することも容易でバイナリーファイルや整形済みBLAT用のDBもダウンロードすることができます。このような配列検索は、NCBI-BLASTをはじめとしてGGGenomeなどでも可能ですが、**配列検索後の結果と他のゲノムワイドな情報を並べて比較する** 場合は、UCSC GBが便利だと思います。

WebのBLASTよりも有利な点など (UCSC GBサイトの記述より)

- スピード（キューなし、数秒での応答）、その代償として相同性深さは劣る。
- 同時に多数のクエリを書いた長いFastaフォーマットで送信できる。
- 5種類の出力ソートオプション
- UCSC ブラウザへの直接リンク
- アライメントブロックの詳細を自然なゲノム順序で表示
- **カスタムトラックの一部**としてアライメントを後で起動するオプション

BLATの動作

DNA探索の際のBLAT は、ゲノム全体のインデックスをメモリ上に保持することで動作します。このインデックスは、繰り返しに大きく関与するものを除き、重なり合う全ての11-merを5段に分割して構成され、約2GbytesのRAMを消費します。**ゲノム自体はメモリに保持されないため、local machineでも快適に動作**します。タンパク質BLATは、11-merではなく4-merであることを除けば、同様の方法で動作します。**タンパク質のインデックスは2Gbytesを少し超える容量が必要**でです。

BLATにおけるDNA searchは、長さ**25塩基以上の類似度95%以上**の配列を高速に検索するように設計されています。20塩基の完全な配列一致を見つけることができますが、**より分岐の多い配列や短い配列のアラインメント**は見逃すことがあります。タンパク質のBLATは、20アミノ酸以上の長さで、80%以上の類似性のある配列を見つけることができます。したがって今回の**例では34塩基**の塩基配列を検索用の文字列として与えています。

UCSCの上面タブから**Tools -> Blat**を選択してください。

[Home](#) [Genomes](#) [Genome Browser](#) [Tools](#) [Mirrors](#) [Downloads](#) [My Data](#) [Projects](#) [Help](#) [About Us](#)

Human BLAT Search

BLAT Search Genome

Genome: ☐ Search all
Human

Assembly:
Dec. 2013 (GRCh38/hg38)

Query type:
BLAT's guess

Sort output:
query,score

Output type:
hyperlink

> test_SOD1
GTCCTCGGAACCAAGGACCTCGGCGTGGCCTAGCG ←fasta形式

☐ All Results (no minimum matches)

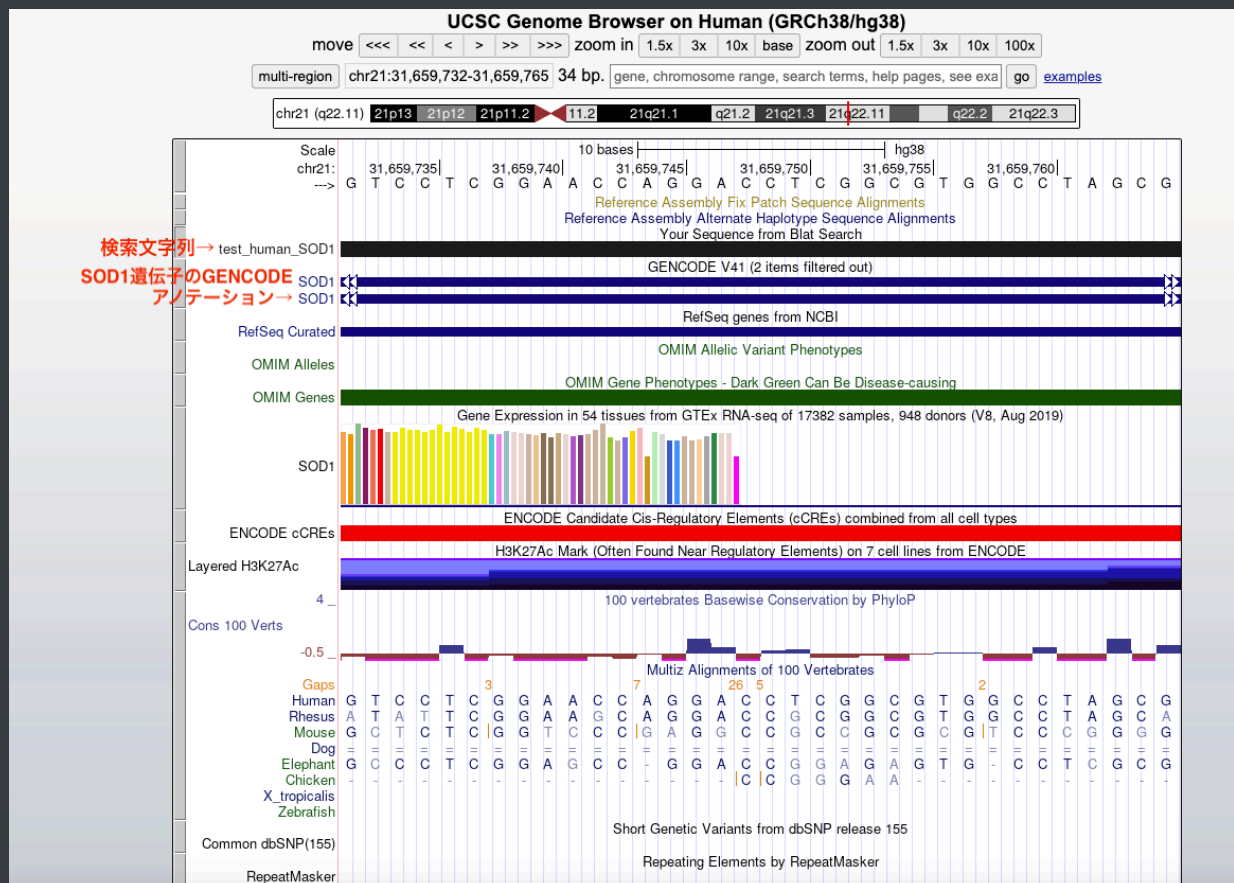
Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

File Upload: Rather than pasting a sequence, you can choose to upload a text file containing the sequence.

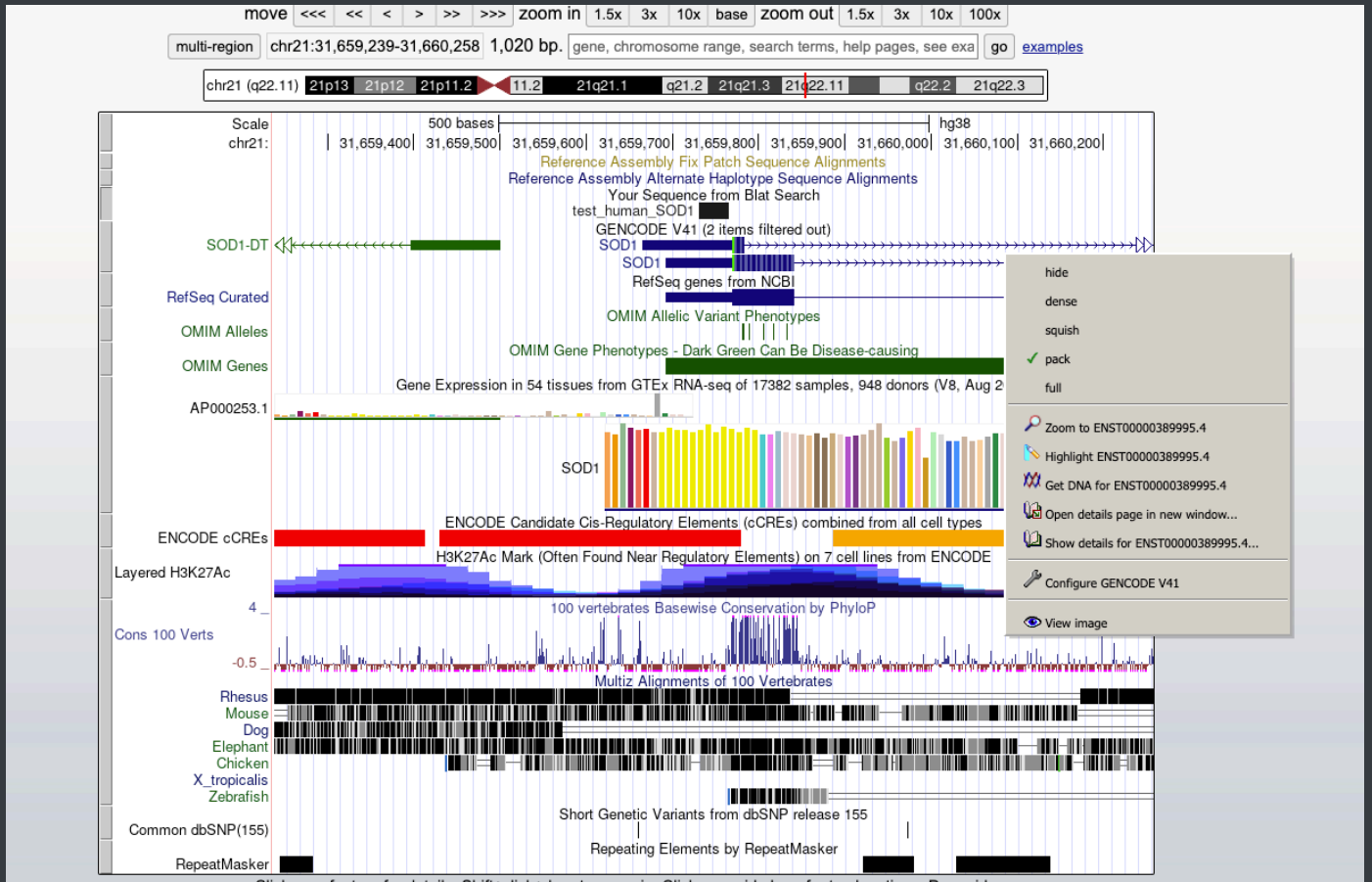
Upload sequence: 選択されていません

Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 10000 or fewer letters will be processed. Up to 25 sequences can be submitted at the same time. The total limit for multiple sequence submissions is 50,000 bases or 25,000 letters.
A valid example is GTCCTCGGAACCAAGGACCTCGGCGTGGCCTAGCG (human SOD1). ←この例の文字列を入力しましょう

Blatに検索する文字列を入力する画面になります。今回はこのページで紹介されている**SOD1**の塩基配列を例に検索してみようと思います。この際に、fasta形式で入力するとゲノムブラウザで検索配列の名前が表示されるので、お勧めします。その後、submitもしくはIm feeling luckyを押します。Im feeling luckyは、googleの検索と同様に、検索結果のtopヒットのリンク先をダイレクトに表示してくれます。今回の例の配列はSOD1にtopヒットすることが自明なので、まず**Im feeling lucky**を押してみましょう。



結果のtopヒットのリンク先であるゲノム地図がダイレクトに表示されました。検索文字列は黒いバーで表示されています。この地図の窓枠全体が検索文字列（34塩基）になっています。また、その次の行のGENCODE V41のアノテーションが濃い青で表示されていますが、ゲノムをズームアップしすぎて遺伝子構造まで見えないので、ズームアウトしましょう。zoom outを利用してx30倍くらいまでズームアウトしてください。



x30倍ズームアウトすると、今回検索した文字列がSOD1遺伝子の5'UTR付近にアライメントしていることがわかります。また、シス領域を示すヒストン修飾とのオーバーラップ、配列の種間保存性などとの比較が直ぐにできます。地図の画面上で、興味のあるトラックで右クリックを押すとその周囲の情報の表示変更やトラックの設定を変更することができます。

BLATの結果の解釈

少し戻り、BLATのtopヒット以外の結果を眺めます。

**Aligned Blocks with gaps ≤ 8 bases are merged for this display when only one sequence has a gap, or when gaps in both sequences are of the same size.*

与えた配列の情報と、そのゲノム周囲の配列情報（200塩基+与えた配列の長さ）、そしてアライメントの結果が表示されます。

In-Silico PCR

BLATの条件を振り返って... qPCR primerのような短い配列の場合はどうするのか？

qPCRプライマーのようなエキソンジャンクションにかかるような **非常に短い配列**は、BLATではゲノムに存在しないため見つけることができません。このような場合は、In-Silico PCRを使い、ターゲットとなる遺伝子セットを選択してみてください。**In-Silico PCRの方が感度が高い**ので、プライマーのペアにはこちらを利用する方がいいと思います。

また、現在の実験科学の環境下では、プライマー設計は受託会社の設計プログラムを用いて行うことが一般的だと思います。融解温度やPCR条件について詳細な情報が得られ、PCRの設計に時間をかけないで済むため、フリーサイトで作成することがほとんどありません。しかしながら、受託会社のプログラムに慣れた人でも時折おとずれるPCRがかからない現象や、条件がマッチせず**手作業で作らなくてはいけない状況**、**作成したPCR primerをゲノム上に図示した図を作りたい**状況は未だに遭遇します。そういったときに、このツールが使えるようになっておくと便利と思います。

Tools -> In-Silico PCRを選択してください。

[Genomes](#)[Genome Browser](#)[Tools](#)[Mirrors](#)[Downloads](#)[My Data](#)[Projects](#)[Help](#)[About Us](#)

UCSC In-Silico PCR

Genome:

Human

Assembly:

Feb. 2009 (GRCh37/hg19)

Target:

genome assembly

Forward Primer:

Reverse Primer:

submit

Max Product Size:

4000

Min Perfect Match:

15

Min Good Match:

15

Flip Reverse Primer:

☐

About In-Silico PCR

In-Silico PCR searches a sequence database with a pair of PCR primers, using an indexing strategy for fast performance. See an example [video](#) on our YouTube channel.

Configuration Options

Genome and Assembly - The sequence database to search.
Target - If available, choose to query transcribed sequences.
Forward Primer - Must be at least 15 bases in length.
Reverse Primer - On the opposite strand from the forward primer. Minimum length of 15 bases.
Max Product Size - Maximum size of amplified region.
Min Perfect Match - Number of bases that match exactly on 3' end of primers. Minimum match size is 15.
Min Good Match - Number of bases on 3' end of primers where at least 2 out of 3 bases match.
Flip Reverse Primer - Invert the sequence order of the reverse primer and complement it.

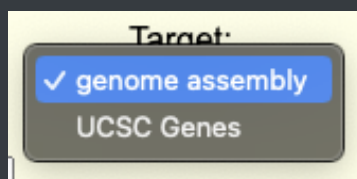
Output

When successful, the search returns a sequence output file in fasta format containing all sequence in the database that lie between and include the primer pair. The fasta header describes the region in the database and the primers. The fasta body is capitalized in areas where the primer sequence matches the database sequence and in lower-case elsewhere. Here is an example from human:

```
>chr22:31000551+31001000 TAACAGATTGATGATGATGAAATGGG CCCATGAGTGGCTCTTAAAGCAGCTGC
TtACAGATTGATGATGATGAAATGGGgggtggcagggtgggggtga
gactgcagagaaaggcaggctggttcataacaagctttgtgctccaa
tatgacagctgaagtttccagggctgatggtgagccagtgaaggtaag
tacacagaacatcttagagaaacctcattcttaagattaaaaataaa
gacttgctgtctgaaggatggattatctctatttgagaaattctgtta
tccagaatggcttacccccacaatgctgaaagtggtaccgtaactcaa
agcaagctctcctcagacagagaaacaccagccgtcacaggaagcaag
aaattggcttacttttaagtgtaatccagaaccagatgtcagagctcc
aagcatttgctctcagctccacGCAGCTGCTTTAGGAGCCACTCATGaG
```

The + between the coordinates in the fasta header indicates this is on the positive strand.

In-Silico PCRは、PCRの経験者なら特に迷うことはないシンプルな入力内容になっていると思います。目的とする生物種を選択し、両端のプライマー配列を入力して、実行するだけです。一方で、先ほどもお伝えした通り、Targetには一考の余地があります。スプライシングアイソフォームを考慮したPCR設計ではエキソンのジャンクションをまたぐ設計が多くなされます。PCRプライマーの合成サービスが連携するようなプライマー作成ツールがそのような設計を採用している場合、ゲノムに探索すると上手くいかないのが、**ターゲットをUCSC GenesもしくはENCODE Genes**にすることで、転写産物を対象とした探索に切り替える方が上手くいきます。



実践的にPCRプライマーを設計する

では、解析する候補遺伝子名だけ明らかになったようなまっさらな状態からスタートして、UCSC GBを使って**Primerを設計するにはどのような順序が良いか**考えてみました。

1. ターゲットの遺伝子をUCSC GBで確認する
2. ターゲットの遺伝子のmRNA配列情報を取得する

1. UCSC GBのツールに慣れるために、**NCBIを介した方法、UCSC GB Table browserの2つの方法**を紹介します。

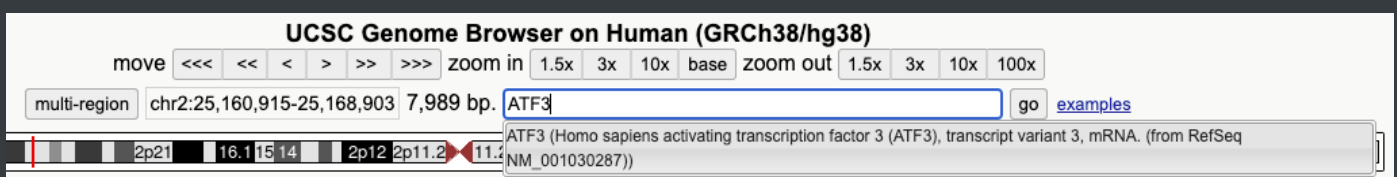
3. mRNAのエキソンジャンクションを調べるためのBLATの実行

1. qPCRなどの場合に備え、特異性をもたせる目的でエキソンジャンクションを挟んだ設計にするためにBLATを実行して、配列をハイライトします。

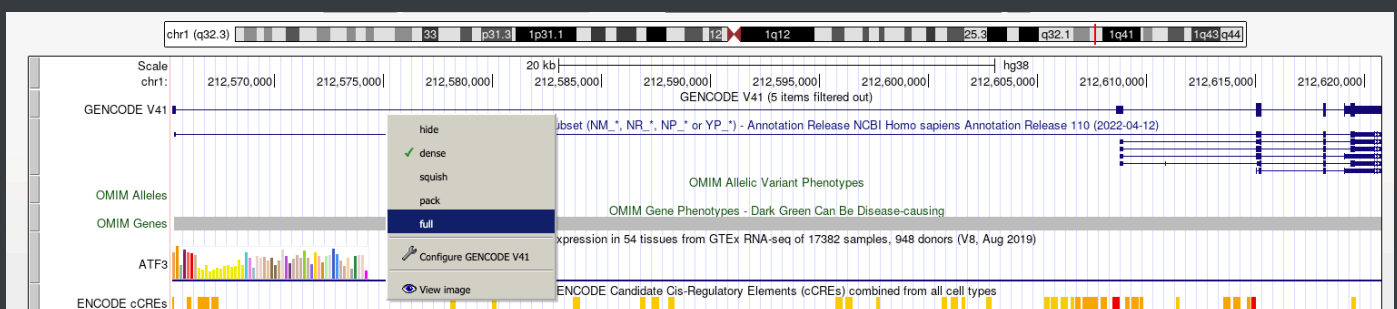
4. In-Silico PCRを実行し、PCR primerを設計します。

1. ターゲット遺伝子をUCSC GBで確認する。

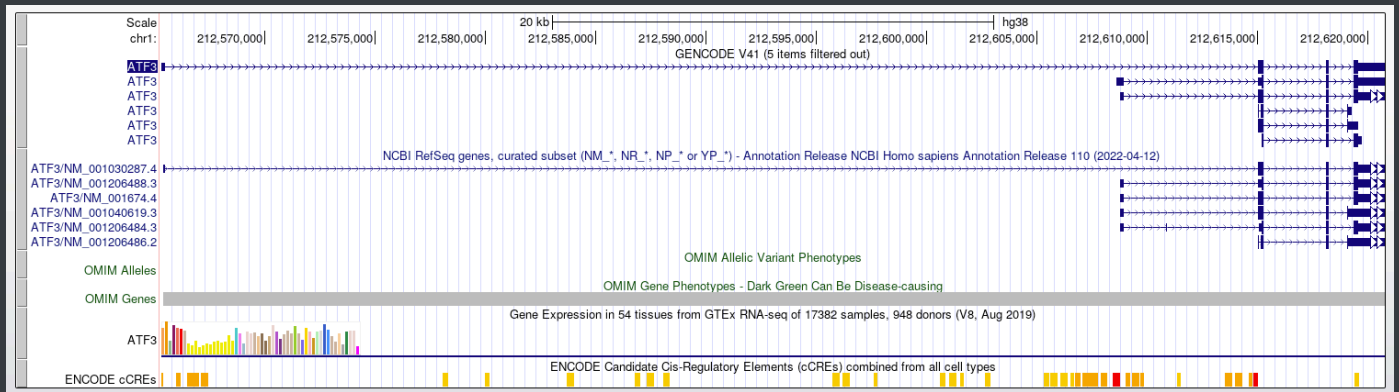
まずATF3をUCSC genome browserで眺めます。GenomesからHuman GRCh38.p13を選択してください。次に、search boxにATF3と入力します。すると、suggestionが現れますので、クリックしてください。



ATF3遺伝子座が現れたら、ATF3遺伝子のsplicing variantも表示したいと思います。GENCODEとNCBI Refseq genesのトラック上で右クリックをして、fullを選択してください。

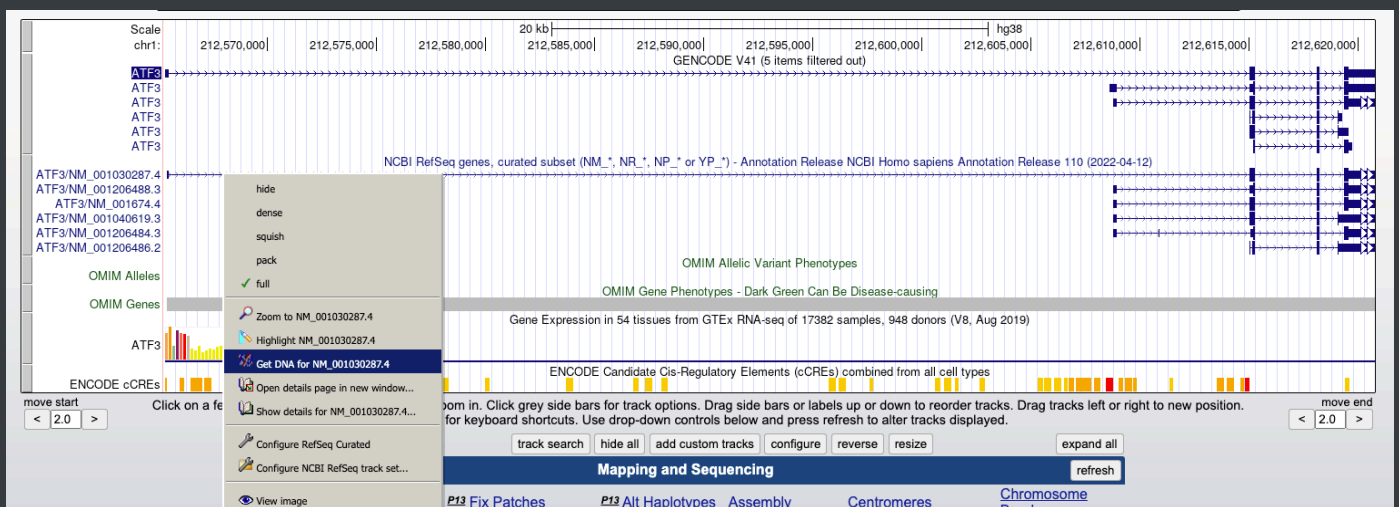


すると遺伝子座にある転写産物が全て表示されます。今回は中でも**最も長いATF3の転写産物**をターゲットにしてみたいと思います。この地図からも、ターゲットは5'に特徴があるので、第1,2 exonをまたぐ配列をPrimerにすることを考えてみたいと思います。



2. mRNAの配列情報を取得する

NCBI Refseq Genesトラックの一番上の転写産物の配列(NM_001030287.4)を取得したいと思います。カーソルを合わせて右クリックをしてください。



ここで注意点があります。図に示したようにGet DNA for NM...という欄がありますがこれをクリックしてもmRNAの配列を取得ではなく、転写産物の座標領域全体の塩基配列を取得してしまいます。この状況からmRNAの配列を取得するには私が知る限りは2つの方法があります。

- 1. **Show details for NM_00~**から、NCBIのリンクをたどり、FASTAを出力させる。
- 2. Table browserに移動してアクセッションナンバーを入力して配列を取得する。

個人的にはmRNAを1種類だけ調べるときは1のやりの方がおすすめで、2のやり方はtable browserで行うにはやや面倒なやり方のように思います。ここでは2つとも紹介します。

1. NCBI Refseqのリンクに飛び取得する

Show details for NM_00~をクリックしてください。

GenomesGenome BrowserToolsMirrorsDownloadsMy DataProjectsHelpAbout Us

NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - NM_001030287.4

RefSeq Gene ATF3

RefSeq: NM_001030287.4 Status: Reviewed

Description: activating transcription factor 3, transcript variant 3

Molecule type: mRNA

Source: BestRefSeq

Biotype: protein_coding

Other notes: isoform 1 is encoded by transcript variant 3

OMIM: 603148

Protein: NP_001025458.1

HGNC: 785

Entrez Gene: 467

GeneCards: ATF3

AceView: ATF3

Summary of ATF3

This gene encodes a member of the mammalian activation transcription factor/cAMP responsive element-binding (CREB) protein family of transcription factors. This gene is involved in the complex process of cellular stress response. Multiple transcript variants encoding different isoforms have been identified for this gene. [provided by RefSeq, Apr 2011].

mRNA/Genomic Alignments (NM_001030287.4)

BROWSER | SIZE | IDENTITY | CHROMOSOME | STRAND | START | END | QUERY | START | END | TOTAL

browser | 1847 | 100.0% | 1 | + | 212565407 | 212620775 | NM_001030287.4 | 1 | 1847 | 1847

View details of parts of alignment within browser window.

NM_00~をクリックしてください。NCBI Genbankに移動できるはずですが。

NIH

National Library of Medicine

National Center for Biotechnology Information

Log in

Nucleotide

Nucleotide

Search

Advanced

Help

GenBank

Send to

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Show in Genome Data Viewer

Articles about the ATF3 gene

ATF3 in atherosclerosis: a controversial transcription factor. [J Mol Med (Berl). 2022]

ATF3 -activated accelerating effect of LINC00941/lncAPF on fibrobl [Autophagy. 2022]

Nuclear Receptor PXR Confers Irradiation Resistance by Promoting DN [Front Oncol. 2022]

See all...

Reference sequence information

RefSeq alternative splicing

See 11 reference mRNA sequence splice variants for the ATF3 gene.

RefSeq protein product

See the reference protein sequence for ATF3

Homo sapiens activating transcription factor 3 (ATF3), transcript variant 3, mRNA

NCBI Reference Sequence: NM_001030287.4

[FASTA](#) [Graphics](#)

Go to: (v)

LOCUS NM_001030287 1847 bp mRNA linear PRI 29-DEC-2022

DEFINITION Homo sapiens activating transcription factor 3 (ATF3), transcript variant 3, mRNA.

ACCESSION NM_001030287

VERSION NM_001030287.4

KEYWORDS RefSeq.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 1847)

AUTHORS Wang B, Yang X, Sun X, Liu J, Fu Y, Liu B, Qiu J, Lian J and Zhou J.

TITLE ATF3 in atherosclerosis: a controversial transcription factor

JOURNAL J Mol Med (Berl) 100 (11), 1557-1568 (2022)

PUBMED [36207452](#)

REMARK GeneRIF: ATF3 in atherosclerosis: a controversial transcription factor.

Review article

REFERENCE 2 (bases 1 to 1847)

AUTHORS Zhang J, Wang H, Chen H, Li H, Xu P, Liu B, Zhang Q, Lv C and Song X.

TITLE ATF3 -activated accelerating effect of LINC00941/lncAPF on fibroblast-to-myofibroblast differentiation by blocking autophagy depending on ELAVL1/HuR in pulmonary fibrosis

JOURNAL Autophagy 18 (11), 2636-2655 (2022)

PUBMED [35427207](#)

矢印のFASTAをクリックしてください。すると、mRNAの配列がfasta形式で表示されますので、コピーしておいてください。ファイルに保存したい場合は、 **send to**をクリックするとプルダウンメニューが表示され、Fileを選択すると操作できます。

FASTA

Send to

Complete Record

Coding Sequences

Gene Features

Choose Destination

File

Clipboard

Collections

Analysis Tool

Download 1 item.

Format

FASTA

Show GI

Create File

Homo sapiens activating transcription factor 3 (ATF3), transcript variant 3, mRNA

NCBI Reference Sequence: NM_001030287.4

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

>NM_001030287.4 Homo sapiens activating transcription factor 3 (ATF3), transcript variant 3, mRNA

GTGACAAGAGAGAAATCCTCCTCTATAGGATGCTGCTGTTTCCTAAGGATTTTCAGCACCTTGCC

CCAAAATCAAAATGATGCTTCAACACCCAGGCCAGGTCTCTGCTCGGAAGTGAGTGCTTCTGCCATCGT

CCCCGCTGTCCCTCCTGGGTCACTGGTGTGAGGATTTTGCTAACCTGACGCCCTTTGTCAAGGAA

GAGCTGAGTTTGCCATCCAGAACAAGCACCTCTGCCACCGGATGCTCTGCGTGGAATCAGTCACTG

TCAGGCACAGACCCCTCGGGGTGTCATCACAAGGCCAGGTAGCCCTGAAGAAGATGAAAGGAAAAA

GAGGCGACGAGAAAGAAATAGATTGCAGCTGCAAAAGTGCCGAAACAAGAGAAGGAGACGGAGTGC

CTGCAGAAAGAGTCGGAGAAGCTGGAAGGTGTAATGCTGAAGTGAAGGCTCAGATTGAGGAGCTCAAGA

ACGAGAAGCAGCATTTGATATACATGCTCAACCTTCATCGGCCACGTGTATTGTCGGGCTCAGAATGG

GAGGACTCCAGAAGATGAGAGAAACCTCTTATCCAAACAGATAAAAGAAGGAACATTGCAGAGCTAAGCA

GTCTGGTATGGGGCGACTGGGGAGTCTCATTGAATCCTCATTTATACCCAAAACCTGAAGCCATT

GGAGAGCTGTCTCCTGTGTACCTCTAGAATCCCAGCAGCAGAGAACCATCAAGGCGGAGGGCTCGAG

TGATTGAGCAGGCCCTTCCATTGCCCCAGAGTGGGTCTTGACACAGGCAAGTGCACTTTTGCTCA

ACTCCAGGATTTAGGCCCTAACACACTGGCCATTCTTATGTTCCAGATGGCCCCAGCTGGTGTCTGCC

CGCCTTTCATCTGGATTCTACAAAAACAGGATGCCACCGTTAGGATTGAGGAGCAGTGTCTGTACC

TCGGGTGGGAGGATGAGGCCATCTCCTTCACCGTGGTACCATTTGCACTCGTAGGGGATGTGGAGTGA

GAACAGCATTTAGTGAAGTTGTCAACGGCCAGGGTGTGCTTCTAGCAAAATATGCTGTTATGTCCAGA

AATTGTGTGTCAAGAAACTAGGCAATGTACTCTTCGATGTTTGTGTACACAACACTGATGTGACTT

TTATATGCTTTTCTCAGATCTGGTTCTAAGAGTTTGGGGGCGGGGCTGCACCACGTGCAGTATCT

CAAGATATTCAGGTGGCCAGAAGAGCTTGTGAGCAAGAGGAGGACAGAATTTCCACAGCTTAAACACAA

ATCCATGGGAGTATGATGGCAGGTCTCTGTTGCAAACTCAGTTCCAAAGTCACAGGAAGAAGCAGAA

AGTTCAACTTCCAAAGGGTTAGGACTCTCACTCAATGTCTTAGGTCAGGAGTTGTGTCTAGGCTGGAAG

AGCCAAAGAATATTCATTTTCTTTCTTTGTTGTTGAAACACAGTCAGTGGAGAGATGTTTGGAAAC

CACAGTCAGTGGAGCTGGGTGTTACCCAGGCTTTAGCATTATGAGTGTCAATAGCATTGTTTGTGCA

TGTAGCTGTTTAAAGAAATCTGGCCAGGGTGTGTCAGCTGTGAGAAGTCACTACACTGGCCACAAGG

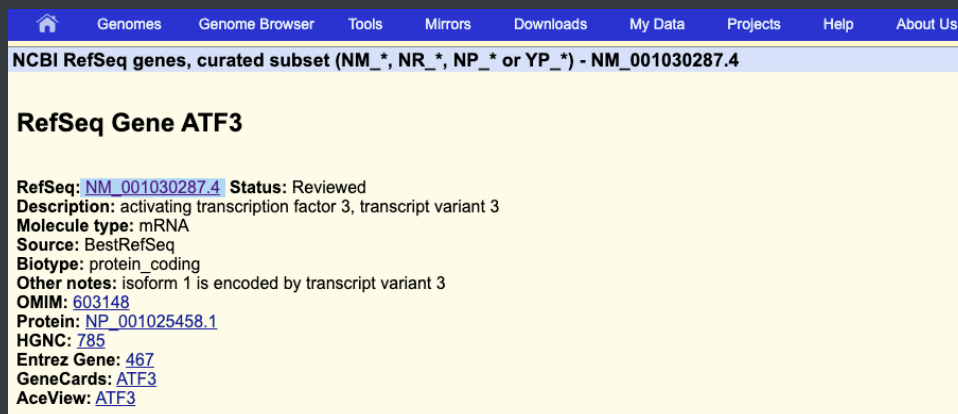
ACGCTGGCTACTGTCTATTAAAAATCTGATGTTTCTGTGAATTCAGAGTGTGTTAATGTACTCAATG

GTATCATTACAATTTCTGTAAGAGAAAAATATTACTTATTATCTAGTATTCCTAACCTGTGAGATAA

TAAATATTGGAACCAAGACATGGTAAA

2. UCSC GBのTable browserを利用する

1.の時と同様にShow details for NM_00~をクリックしてください。

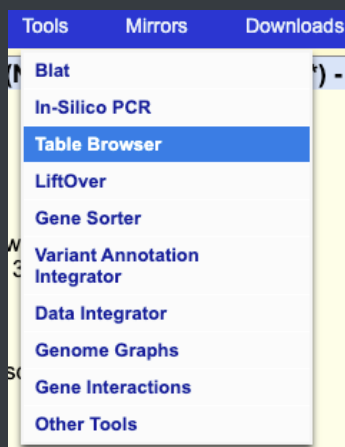


NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - NM_001030287.4

RefSeq Gene ATF3

RefSeq: [NM_001030287.4](#) Status: Reviewed
Description: activating transcription factor 3, transcript variant 3
Molecule type: mRNA
Source: BestRefSeq
Biotype: protein_coding
Other notes: Isoform 1 is encoded by transcript variant 3
OMIM: [603148](#)
Protein: [NP_001025458.1](#)
HGNC: [785](#)
Entrez Gene: [467](#)
GeneCards: [ATF3](#)
AceView: [ATF3](#)

その後、**NM_001030287.4**の文字列をコピーしてください。次に、上段のメニューにあるToolsから**Table browser**をクリックしてください。



Tools Mirrors Downloads

- Blat
- In-Silico PCR
- Table Browser**
- LiftOver
- Gene Sorter
- Variant Annotation Integrator
- Data Integrator
- Genome Graphs
- Gene Interactions
- Other Tools

すると次のような画面が現れます。Table Browserは、UCSC GBの中でも中核となっているツールで、UCSCが保持しているデータからさまざまなサブセットデータを抽出できる便利なツールです。

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

Table Browser

Use this tool to retrieve and export data from the Genome Browser annotation track database. You can limit retrieval based on data attributes and intersect or merge

Select dataset

clade: genome: assembly:
group: track:
table: [describe table schema](#)

Note: Most dbSNP tables are huge. Trying to download them through the Table Browser usually leads to a timeout. Please see our [Data Access FAQ](#) on how to download dbSNP data.

Define region of interest

region: ☐ genome ☒ position [lookup](#) [define regions](#)
identifiers (names/accessions): [paste list](#) [upload list](#)

Optional: Subset, combine, compare with another track

filter: [create](#)
subtrack merge: [create](#)
intersection: [create](#)

[Explore gene/protein interactions and pathways](#)

Retrieve and display data

output format: Send output to ☐ Galaxy ☐ GREAT
output filename: (add .csv extension if opening in Excel, leave blank to keep output in browser)
output field separator: ☒ tsv (tab-separated) ☐ csv (for excel)
file type returned: ☒ plain text ☐ gzip compressed

[get output](#) [summary/statistics](#)

しかしその反面、多くのデータ抽出の用途に使えるが故に、設定を1つ間違えると目的に沿わないサブセットが得られてしまうので注意が必要です。今回の目的には以下の図の Select datasetのセッティングを行なってください。

- clade: Mammal
- genome: Human
- Group: Genes and Predictions
- Assembly: GRCh38
- track: NCBI RefSeq
- table: RefSeq All

Select dataset

clade: genome: assembly:
group: track:
table: [describe table schema](#)

Define region of interest

region: ☐ genome ☒ position [lookup](#) [define regions](#)
identifiers (names/accessions): [paste list](#) [upload list](#)

その後、Define region of interestを利用してアクセッションナンバーを入力するために、**paste list**をクリックします。

Paste In Identifiers for ncbiRefSeq

Please paste in the identifiers you want to include. The items must be values of the **name** field of the current information about the table fields.) Some example values:
NM_002025.4
NM_031458.3
XM_006715562.5
GUCY2GP
SEMA3C
RHPN1-AS1

すると、IDリストを入力するフォームがありますので、先ほどコピーしておいた **NM ナンバー(NM_001030287.4)**を入力し、submitしてください。

Retrieve and display data

output format: Send output to ☐ [Galaxy](#) ☐ [GREAT](#)
output filename: (leave blank to keep output in browser)
file type returned: ☒ plain text ☐ gzip compressed

すると、Table Browserの画面に自然に戻ります。最後に、Retrive and display dataに出力形式やファイル名を入力して、get outputします。

Select sequence type for RefSeq All

☒ genomic

genomicという選択のみですが、submitしてください。

最後に、どの領域を取得するのか選択する必要があります。今回はmRNA配列部分を取得するので、intronsのcheckを外して**get sequence**してください。

ncbiRefSeq Genomic Sequence

Sequence Retrieval Region Options:

☐ Promoter/Upstream by 1000 bases

☒ 5' UTR Exons

☒ CDS Exons

☒ 3' UTR Exons

☐ Introns

☐ Downstream by 1000 bases

☒ One FASTA record per gene.

☐ One FASTA record per region (exon, intron, etc.) with 0 extra bases upstream (5') and 0 extra downstream (3')

☐ Split UTR and CDS parts of an exon into separate FASTA records

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

Sequence Formatting Options:

☒ Exons in upper case, everything else in lower case.

☐ CDS in upper case, UTR in lower case.

☐ All upper case.

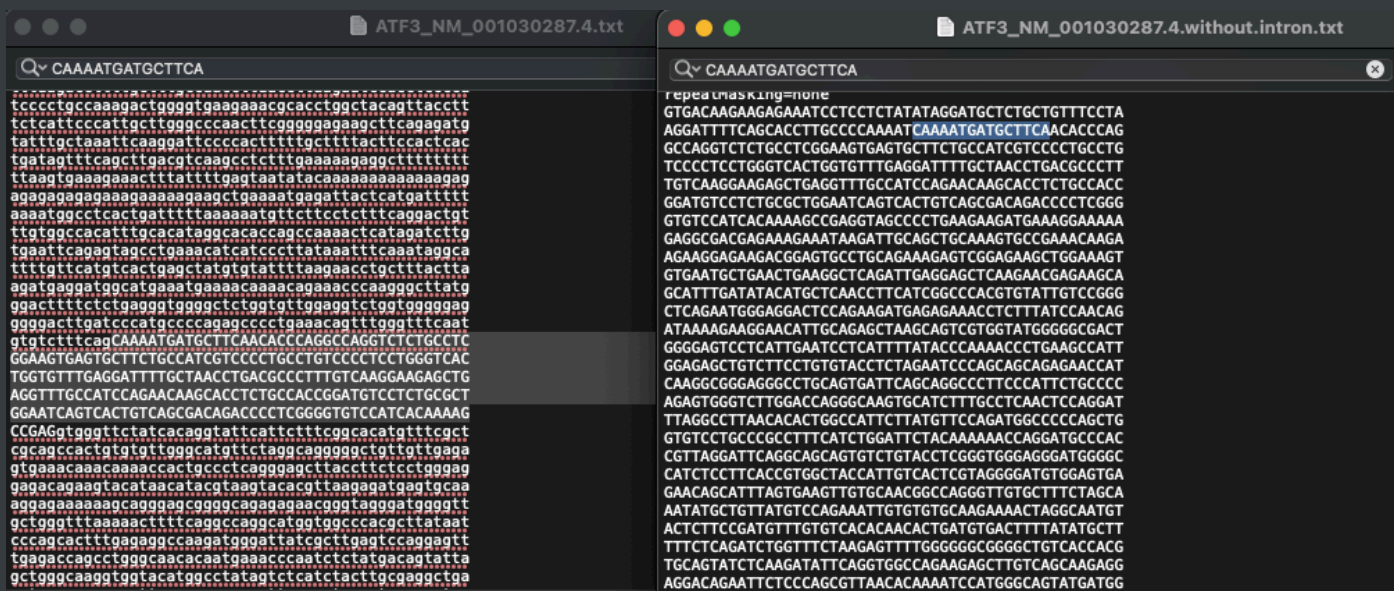
☐ All lower case.

☐ Mask repeats: ☒ to lower case ☐ to N

get sequence

cancel

するとファイルが保存されますので、そのファイルの中身をテキストエディタで開いてください。下の図はIntronsを除いた場合と除いていない場合を並べてご紹介しています。



左はcheckを外す前、右は、Intronを除いたもので、**第二エキソンの位置をハイライト**してみました。確かにmRNA配列を取得できています！

3. mRNAのexon-exon junctionを調べるためのBLATの実行

さて、配列の準備が終わりまりましたので、今度はこの配列をBLATしたいと思います。fastaファイルをコピーしておき（ファイルに保存しても構いません）、**Tools-> BLAT**へ移動してください。移動後、配列をペーストもしくはsubmitしてください。ファイルを保存した場合

Human (hg38) BLAT Results

BLAT Search Results

Go back to [chr1:212,565,334-212,620,777](#) on the Genome Browser.

Custom track name:

Custom track description:

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	hg38_ncbiRefSeq_NM_001030287.4	1844	1	1847	1847	100.0%	chr1	+	212565407	212620775	55369
browser details	hg38_ncbiRefSeq_NM_001030287.4	25	1474	1505	1847	96.3%	chr1	-	71753904	71753938	35
browser details	hg38_ncbiRefSeq_NM_001030287.4	21	1710	1731	1847	100.0%	chr5	-	137642976	137642998	23
browser details	hg38_ncbiRefSeq_NM_001030287.4	20	1483	1502	1847	100.0%	chr1	-	91831502	91831521	20

BLATの結果が表示されます。top hitの結果のdetailを表示してください。

Alignment of hg38_ncbiRefSeq_NM_001030287.4 and chr1:212565407-212620775

Click on links in the frame to the left to navigate through the alignment. Matching bases in cDNA and genomic sequences are colored blue and capitalized. Light blue bases mark the boundaries of gaps in either sequence (often splice sites).

cDNA hg38_ncbiRefSeq_NM_001030287.4

50

AGGATTTCACACCTTGGCCCAAAATCAA AATGATGCTT CAACACCCAG

100

GCCAGGCTCT TGCCTCGGAA GTGAGTGCTT CTGCCATCGT CCCCTGCGCTG

150

TCCCCTCTCTG GTGCACTGGT GTTTGAGGAT TTGCTAACC TGACGCCCTT

200

TGTCAAGGAA GAGCTGAGGT TTGCCATCCA GAACAAGCAC CTCTGCCACC

250

GGATGTCTCT TGCCTCGGAA TCAGTCACTG TCAGCGACAC ACCCTCGGG

300

GTGTCCATCA CAAAAGCCCA GTAGCCCTT GAAGAAGATG AAAGGAAAAA

350

GAGCGACGCA GAAAGAAATA AGATTGCAGC TGCAAGGTGC CGAAGCAAGA

400

AGAGGAGAGAA GACCGAATGC CTGCAGAAAG ACTCGGAGAA CCTGGAAAGT

450

GTGAATGCTG AACTGAAGGC TCAGATTGAG GAGCTCAAGA ACGAGAAGCA

500

GCATTTGATA TACATGCTCA ACCTTCATCG GCCACGCTGT ATTGTCGGG

550

CTCAGAATGG GAGGACTCCA GAAGATGAGA GAAACCTCTT TATCCAACAG

600

ATAAAGAAG GAACATTGCA GAGCTAAGCA GTGCTGGTAT GGGGGCGACT

650

GGGGAGTCTC CATTGAATCC TCATTTTATA CCCAAACCC TGAAGCCATT

700

GGAGAGCTGT CTTCCTGTGT ACCTCTAGAA TCCAGCAGC AGAAGACCAT

750

CAAGCGGGGA GGGCTGCAG TGATTCAGCA GGCCTTCCC ATTCTGCCCC

800

AGAGTGGGTC TTGACACAGG GCAAGTGCAT CTTTGCCTCA ACTCCAGGAT

850

TTAGGCTCTA ACACACTGGC CATTCTTATG TTCCAGATGG CCCCGAGCTG

900

GTGCTCGCC CGCCTTTGAT CTGGATTCTA CAAAAGCCA GGATGCCAC

950

CGTTAGGATT CAGGCAGCAG TGCTCTGACC TCGGGTGGGA GGGATGGGG

1000

CATCTCCTTC ACCGTGGCTA CCATTGTGTC TCGTAGGGGA TGTGGAGTGA

1050

GAACAGCATT TAGTGAAGTT GTGCAACGGC CAGGGTTGTG CTTTCTAGCA

1100

AATATGCTGT TATGTCCAGA AATTGTGTGT GCAAGAAAC TAGGCAATGT

1150

ACTCTCCGGA TGTTTGTGTC ACACAACACT GATGTGACTT TTATATGCTT

1200

TTTCTCAGAT CTGGTTTCTA AGAGTTTGG GGGCGGGGC GTGCACCAG

1250

TGCAGTATCT CAAGATATTC AGGTGCCAG AAGAGCTTGT CAGCAAGAGG

1300

AGGACAGAA TCTCCACAGG TTAACACAAA ATCCATGGGC AGTATGATGG

1350

CAGTCTCTCT GTTCAAACCT CAGTCTCAAA GTCCAGAGAA GAAAGCAGAA

1400

AGTCAAGCTT CCAAAGGCTT AGGACTCTCC ACTCAATGTC TTAGTCCAG

1450

AGTTGTGTCT AGGCTGGAAG AGCCAAAGAA TATTCAATT TCCTTCTCT

1500

GTGGTTGAAA ACACAGCTCA GTGGAGAGAT GTTTGGAAAC CACAGTCAGT

1550

GGAGCCTGGG TGGTACCCAG GCTTTAGCAT TATTGGATGT CAATAGCATT

1600

GTTTTTGTC TGTAGCTGTT TTAAGAAATC TGCCCGAGGG TGTTCGACG

1650

TGTGAGAAGT CACTCACACT GGCCACAAGG ACGCTGGCTA CTGTCTATTA

1700

AAATCTGAT GTTCTGTGTA AATTCTCAGA GTGTTAATT GTACTCAATG

1750

GTATCATTAC AATTCTCTGT AAGAGAAAAT ATTACTTATT TATCTAGTA

1800

TTCTAACCT GTCAGAATAA TAAATATTGG AACCAAGACA TGGTAA

このURLで結果に飛ぶことができます。

https://genome.ucsc.edu/cgi-bin/hgc?o=212565406&g=htcUserAli&i=../trash/hgSs/hgSs_genome_34914_4ed890.pslx+.%2Ftrash%2FhgSs%2FhgSs_genome_34914_4ed890.fa+hg38_ncbiRefSeq_NM_001030287.4&c=chr1&l=212565406&r=212620775&db=hg38&hgsid=1545961743_RWsba2vqJjM1GjFppnyhnd6691Z9

BLATしてexon-exon junctionをハイライトすることができ、かつイントロンとの配列の状況も綺麗に可視化できました。では、SYBER GreenでのqPCR primerを設計することを想定して、FRのPrimer位置を決定しましょう。その際にsplice位置をうまく跨ぎたいので、色付けしたわけです。

4. Primerを設計する

PrimerはPracticalには以下のように決定すると思います。

1. 増幅サイズは100-200bp
2. Tm 55-60度（反応効率上、高すぎないように、できればFRを揃える）
3. 3'末端にGまたはCが3個以上連続する配列は避ける（が意識しすぎない）。一方で、3'末端がTになる配列はミスマッチでアニールしやすいので避ける。
4. ダイマーがふえすぎないように、3'末などと一致する配列をさける（が意識しすぎない）

1 回目のトライ

- F: 5- CCTCTATATAGGATGCTCTG-3
- R: complementary処理する前の配列-> 5-GATTTTGCTAACCTGACGC-3
- **Forward:** 49.2 C cctctatataggatgctctg
- **Reverse:** 56.0 C gattttgctaacctgacgc

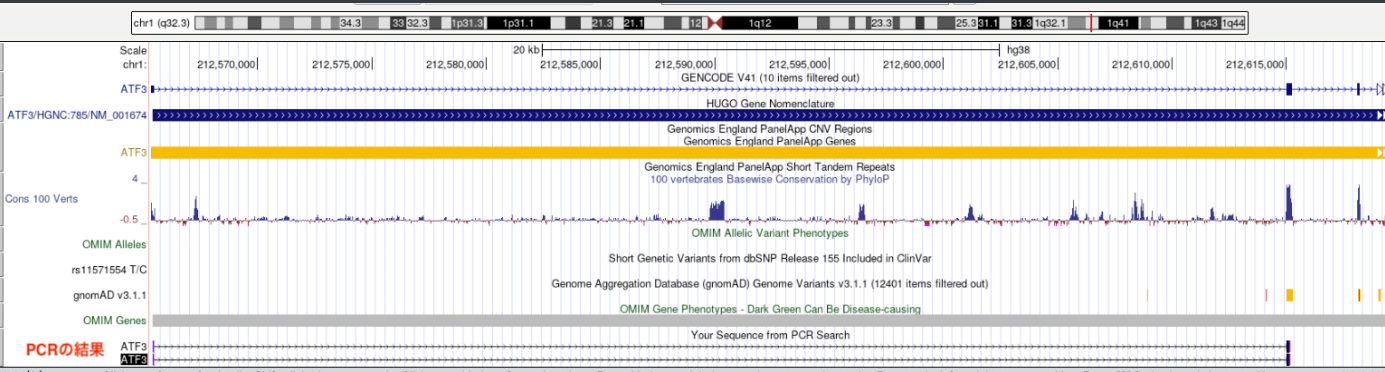
2 回目のトライ

- F: 5-AGGATGCTCTGCTGTTTCC-3
- R: complementary処理する前の配列-> 5-GATTTTGCTAACCTGACGC-3
- **Forward:** 58.0 C aggatgctctgctgtttcc
- **Reverse:** 56.0 C gattttgctaacctgacgc

UCSC In-Silico PCR
<p>The sequences and coordinates shown below are from GENCODE Genes, not from the genome assembly. The links lead to the Genome Browser at the position of the entire target sequence.</p> <p>>ENST00000366981.8_ATF3:103+269 167bp AGGATGCTCTGCTGTTTCC GCGTCAGGTTAGCAAAATC AGGATGCTCTGCTGTTTCCtaaggattttcagcaccttgcccaaatca aaatgatgcttcaacacccaggccaggtctctgcctcggaagtgagtgt tctgccatgctccctgcctgtcccctctgggtcactggtgtttgagGA TTTTGCTAACCTGACGC</p> <p>>ENST00000366987.6_ATF3:103+269 167bp AGGATGCTCTGCTGTTTCC GCGTCAGGTTAGCAAAATC AGGATGCTCTGCTGTTTCCtaaggattttcagcaccttgcccaaatca aaatgatgcttcaacacccaggccaggtctctgcctcggaagtgagtgt tctgccatgctccctgcctgtcccctctgggtcactggtgtttgagGA TTTTGCTAACCTGACGC</p>
Primer Melting Temperatures
<p>Forward: 58.0 C aggatgctctgctgtttcc Reverse: 56.0 C gattttgctaacctgacgc The temperature calculations are done assuming 50 mM salt and 50 nM annealing oligo concentration. The code to calculate the melting temp comes from Primer3.</p>

結果的に2つの転写産物のprimerになっていますが、どうやらNCBIのサイトで調べたところ、この2つの違いはより3'側にあるようです。

Transcripts for ATF3: 1 to 11 of 11							Filter...
Transcript	Location	Size	Type	Protein	Exons		
ENST00000341491.9 - NM_001674.4 MANE Select	1 212608761 212620775	12.02 kb	Protein coding	ENSP00000344352.4 181aa	4		
ENST00000366983.5 NM_001040619.3	1 212615022 212619555	4.53 kb	Protein coding	ENSP00000355950.1 135aa	3		
ENST00000366987.6 NM_001030287.4	1 212655334 212620772	55.44 kb	Protein coding	ENSP00000355954.2 181aa	4		
ENST00000613954.4 NM_001206488.3 NM_001206484.3	1 212608628 212620772	12.14 kb	Protein coding	ENSP00000483576.1 124aa	5		
ENST00000336937.8 NM_001206486.2	1 212615022 212619266	4.25 kb	Protein coding	ENSP00000336908.4 106aa	4		
ENST00000366981.8	1 212655334 212619535	54.20 kb	Protein coding	ENSP00000355948.4 175aa	4		
ENST00000366985.5	1 212608915 212619611	10.70 kb	Protein coding	ENSP00000355952.2 153aa	5		
ENST00000464547.5	1 212615022 212619555	4.53 kb	Nonsense mediated decay	ENSP00000432208.1 135aa	4		
ENST00000613104.1	1 212615178 212619746	4.57 kb	Protein coding	ENSP00000480606.1 124aa	3		
ENST00000465155.5	1 212608670 212619247	10.58 kb	Retained Intron		3		
ENST00000492118.2	1 212613436 212620777	7.34 kb	Protein coding CDS not defined		2		



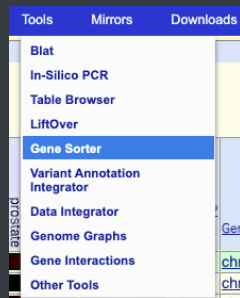
結果のリンクをクリックすると、Track情報にIn Silico PCRで設計したプライマーの構造的な関係が描かれていると思います。このようにPCR設計の位置をゲノム地図上にダイレクトに図示することができます。

GeneSorterを使った発現解析

こちらは時間が余った時のおまけに近いのですが、Gene Sorterを使った発現解析の流れを説明します。

- 1. Gene SorterでHBB遺伝子を検索
 - 1. 表示対象のセッティング
 - 2. データの抽出
- 2. GenomeGraphで、あるLD領域（SNPsマーカーで囲まれている）の遺伝子を抽出し、発現データを得る。
 - 1. SNPマーカーを入力する。
 - 2. Gene SorterでLD領域の遺伝子抽出。

1. Gene SorterでHBB遺伝子を検索



上段のメニューから、**Tools -> Gene Sorter**を選択してください。

Gene Sorterでは、検索した遺伝子と発現プロファイルの似ている遺伝子を各種データベースから取得することができます。今回は**HBB**遺伝子を検索してみたいと思います。下の図のようにHBBを検索してみてください。その時に、データは**GTEx**を利用してみましょう。

Known Gene Names

Known Gene Descriptions

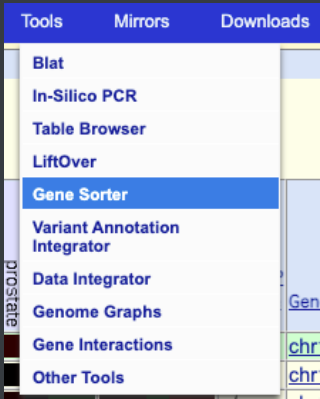
検索結果として、さまざまな遺伝子がヒットしますが、HBB遺伝子を選択してください。

HBB遺伝子とGTExデータの中で発現プロファイルが似ている遺伝子が上から表示されます。GTExは、約53の健常組織のデータから発現データを取得したものです。しかし、GTExのデータは、13列ほどしか表示されていないので、**Configure**で設定を変えてみましょう。

GenomeGraphで、（SNPsマーカーで囲まれている）あるLD領域の遺伝子を抽出する。

これは特殊な事例なので、使う人はそうそういないと思いますが、ゲノムの座標から遺伝子抽出することができます。用途を考えてみましたが、連鎖不平衡（LD）の領域がわかった場合、その中の遺伝子が気になることがあるかもしれません。LD領域の遺伝子をごっそり取得して、そしてGTExの発現テーブルを作成してみます。

今回用意した連鎖不平衡の領域は、染色体一番の約65kbの領域（chr1: 2556224-2622185）になります。次のマーカーで囲まれた位置になります。このLSに乗っているSNPsは、GWASの結果、Eosinophil percentage of white cells, Chronic inflammatory diseases, Chronic inflammatory diseases などの炎症に関連しそうな形質に関わっている可能性が示唆されています。どんな遺伝子が乗っているのでしょうか？駆け足で説明します。



Gene Sorterを開いてください。

GenomesGenome BrowserToolsMirrorsDownloadsMy DataView

Upload Data to Genome Graphs

name of data set:LD_test

description:LD region

file format:best guess

markers are:dbSNP rsID

column labels:best guess

display min value:-2max value:2

label values:rs2227312-rs6671426

draw connecting lines between markers separated by up to25000000bases.

file name:ファイルを選択 選択されていません

or

Paste URLs or data:

rs22273121

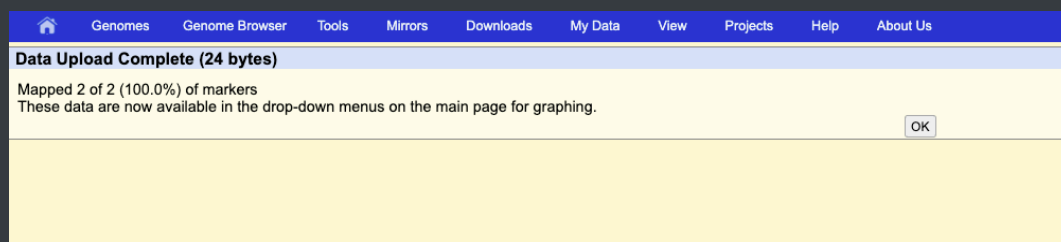
rs66714261

submit

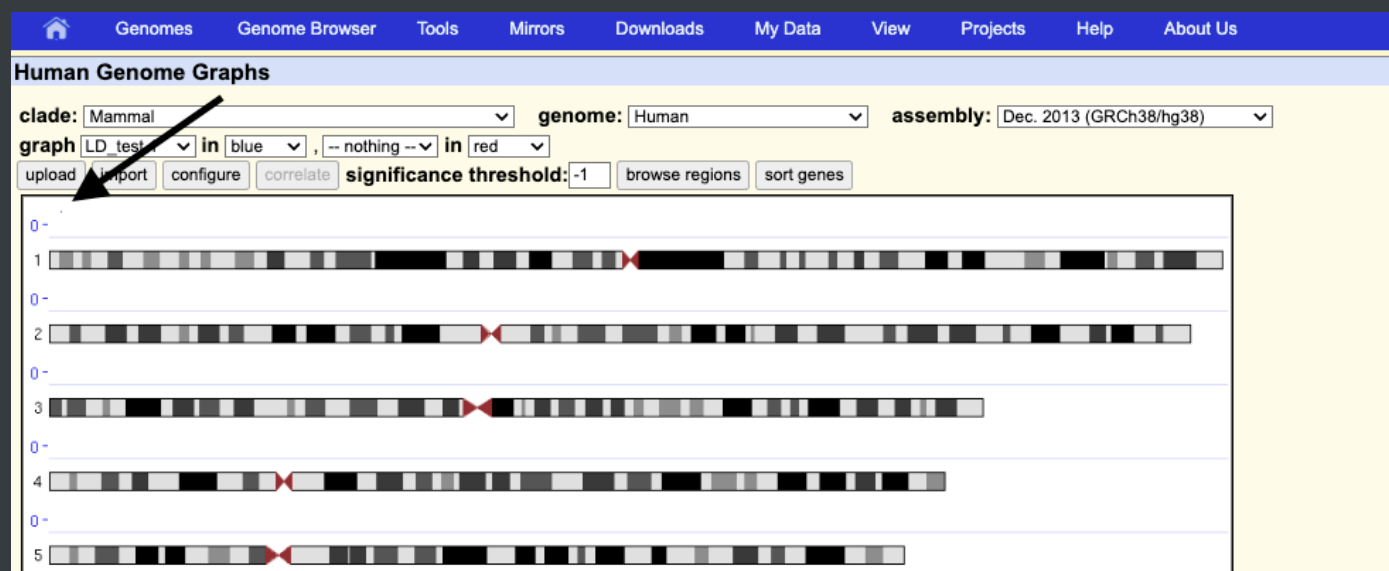
rs2227312 1

rs6671426 1

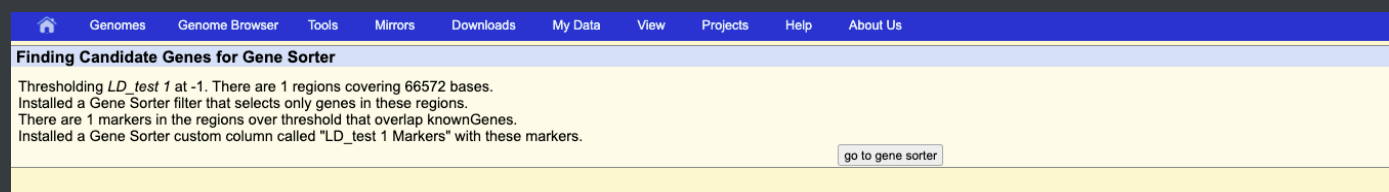
上記のSNPs（LDブロックに使われていた終端マーカーs）を適当なスコアをつけてタブ区切りでsubmitしてみます。名前や詳細については適当に入力しました。みやすいように、スコアを-2から2にしてみました。submit!



OK



何も写ってないじゃないか？と思うかもしれませんが、領域が60kbほどの領域は目に見えないのですよ。実は、矢印のところになります。ゲノム全長からすると小さな領域に見えますね。さて、**sort genes!**



go to gene sorter!

UCSC Human Gene Sorter

genome Human assembly Dec. 2013 (GRCh38/hg38) search ENST00000373020.9 Go!									
sort by Expression (GTEx) configure filter (now on) display 50 output sequence text									
#	Name	VisGene	brain/amygdala brain/cerebellum brain/cortex	spleen	adipose/visceral	pancreas	heart/atrial appendage	lung	BLASTP E-Value
1	MMEL1	n/a							n/a
2	PRXL2B	n/a	n/a	n/a	n/a	n/a	n/a	n/a	chr1 2,611,827
3	TNFRSF14	n/a	n/a	n/a	n/a	n/a	n/a	n/a	chr1 2,589,122
4	ENSG00000225931	n/a							chr1 2,560,100
5	ENSG00000228037	n/a							chr1 2,568,149
6	TNFRSF14-AS1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	chr1 2,583,046
7	ENSG00000272449	n/a							chr1 2,553,475
8	ENSG00000289610	n/a	n/a	n/a	n/a	n/a	n/a	n/a	chr1 2,538,762
LD_test 1 Markers over threshold									
n/a									
n/a									
n/a									
n/a									
n/a									
rs2227312									
n/a									
n/a									
n/a									

というわけで、LD領域の遺伝子たちが抽出できました。先ほどと同じ要領でデータを増やしたりしてみてください。ただし、GTExは健常組織なので炎症性の遺伝子が発現しているようすは厳しいかも。そして、ごらんいただくとTNFのファミリーに属する遺伝子がいたことがわかります。LDの形質情報と遺伝子がつながりました！（終）