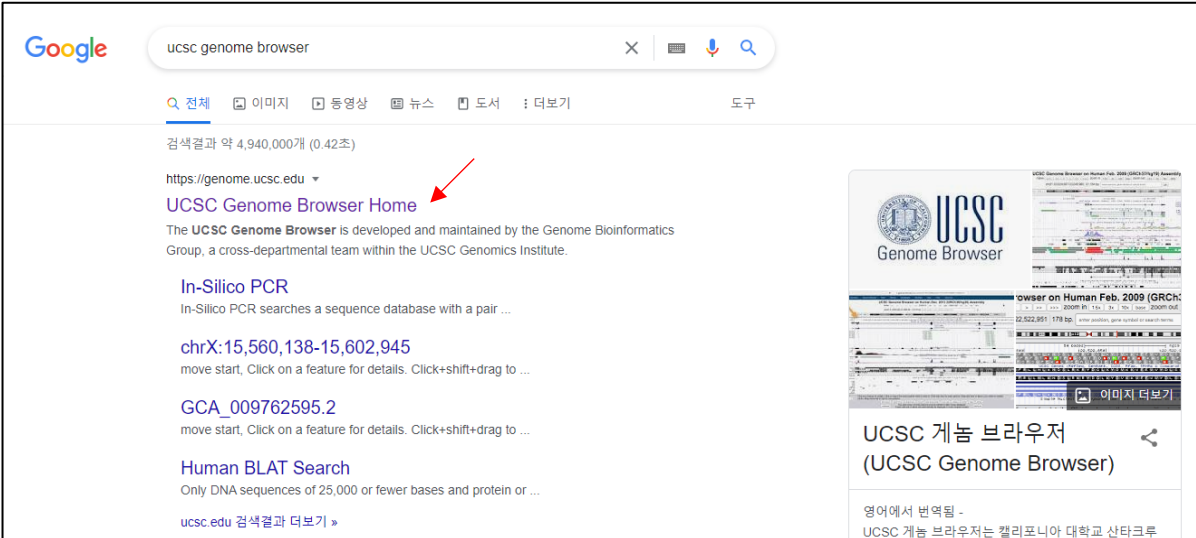# Molecular Biology Laboratory

Bioinformatics and Genomics Lab.

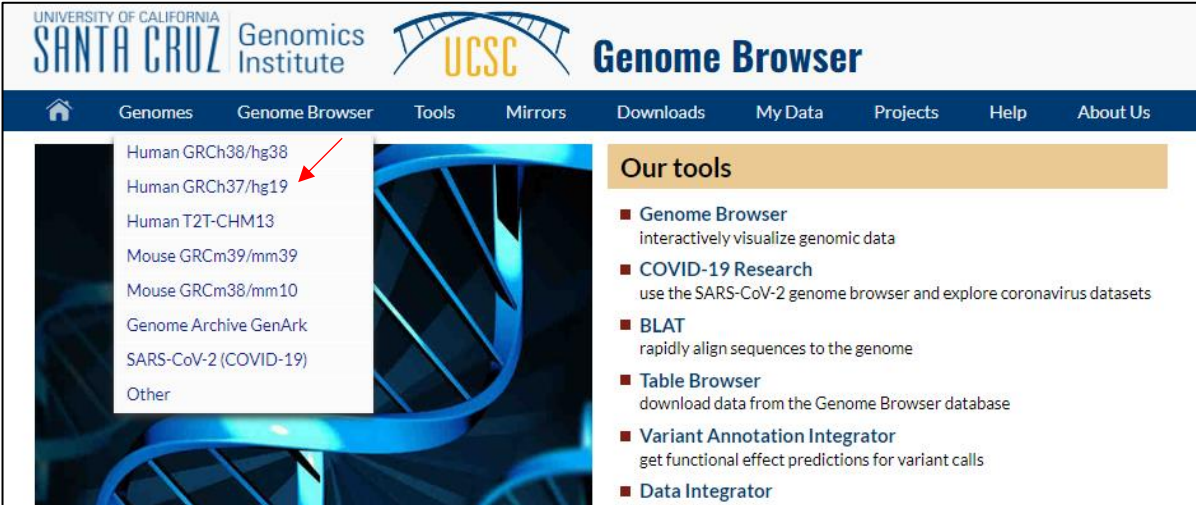## Week1. DNA / RNA Sequence Alignment (BLAST) & Genome Browser

1. Practice how to use UCSC Genome Browser

- Access to the UCSC Genome Browser and find human insulin gene

  ➤ Search "UCSC Genome Browser" in google and access to genome browser.

  ➤ Hover the mouse pointer to "Genomes" and select genome version. We will use "Human GRC38/hg38" in our practical exercise.

> Type "INS" search window and select "Homo sapiens insulin".

> Compare Isoforms of the insulin gene and look through other information.

● Make custom gene annotation and visualize

> Click the marked insulin isoform, "ENST00000397262.5", and look through the information of the isoform.

- ➢ Open notepad and save it as "Custom_INS_Annotation.gtf".

- ➢ Fill the blank with reference to the below information.

chr11	TEST	transcript			.		.	gene_id "INS"; transcript_id "INS";
chr11	TEST	exon			.		.	gene_id "INS"; transcript_id "INS";
chr11	TEST	UTR			.		.	gene_id "INS"; transcript_id "INS";
chr11	TEST	CDS			.		.	gene_id "INS"; transcript_id "INS";
chr11	TEST	exon			.		.	gene_id "INS"; transcript_id "INS";
chr11	TEST	CDS			.		.	gene_id "INS"; transcript_id "INS";
chr11	TEST	UTR			.		.	gene_id "INS"; transcript_id "INS";

A: 2159779; B: 2159855; C: 2159997; D: 2160785; E: 2160972; F: 2161204;
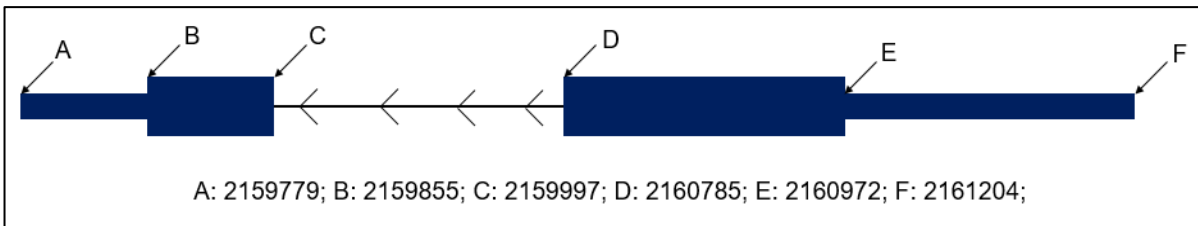
- ➢ Cautions

  - ✓ Each row is arranged in ascending order based on the start position.

  - ✓ There is a difference of 1 between the end position of UTR and the end position of CDS.

  - ✓ Each column is separated by a tab, but gene id and transcript id are written in one column and they are separated by space.

- ➢ For uploading custom an annotation file to the genome browser, hover the mouse on "My Data" and click "Custom Tracks".
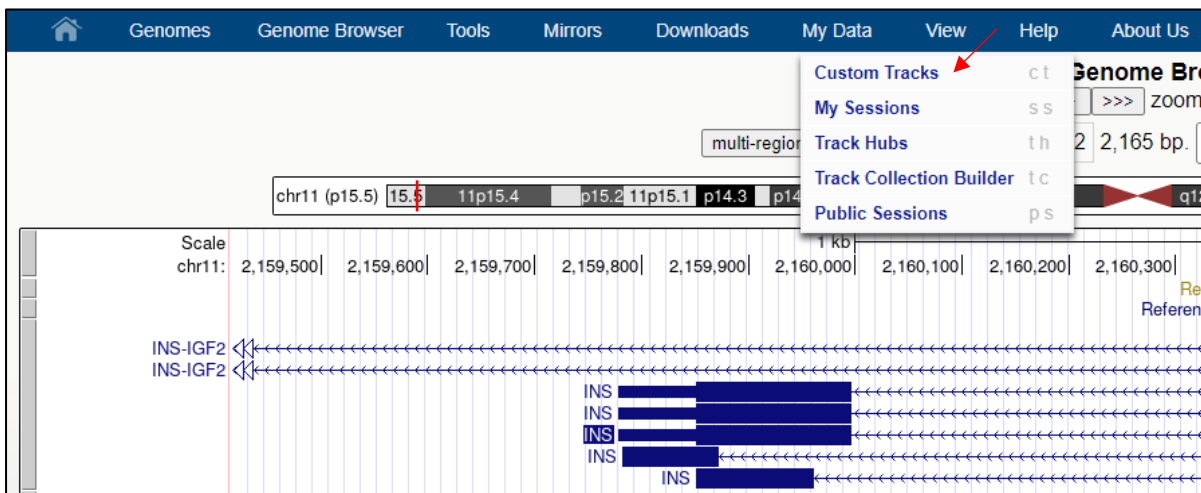
➢ Upload the custom annotation file and click "Submit".

➢ After uploading, click "go" to visualize the custom annotation track.

➢ Compare the custom annotation track and marked Insulin isoform.
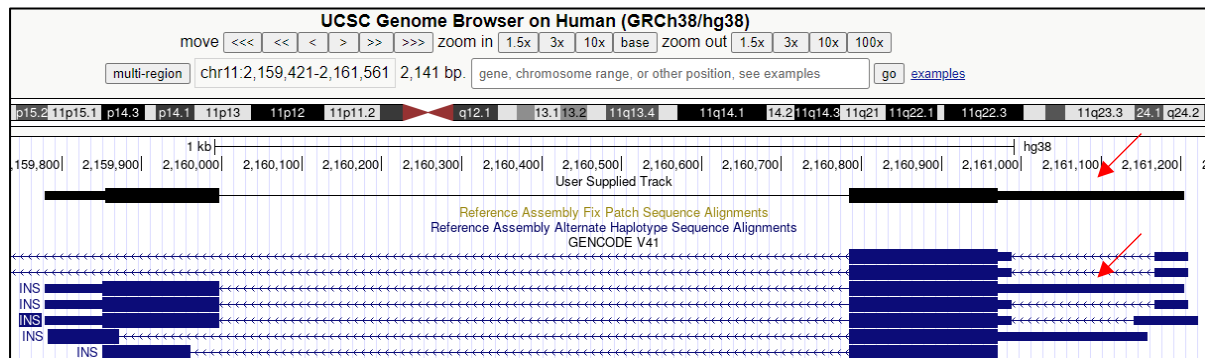
2. Practice how to get sequence data and use BLASTN

- Find GFP gene sequence and align the sequence

  - Search "NCBI" in google and access NCBI.

  - Search "GFP Aequorea victoria" and click "Nucleotide".

➢ Select one from the list. We will use 4ᵗʰ one, "Aequorea victoria green-fluorescent protein (GFP) mRNA, complete cds" with 922 bp linear mRNA in a practical exercise.

1. ☐ Synthetic construct for **Aequorea victoria** partial **gfp** gene for **GFP**
   735 bp linear other-genetic
   Accession: LN515608.1 GI: 690969141
   PubMed   Taxonomy

   GenBank   FASTA   Graphics

2. ☐ **Aequorea victoria** clone EM1 **GFP** gene, partial cds
   135 bp linear DNA
   Accession: JN800726.1 GI: 353442080
   Protein   Taxonomy

   GenBank   FASTA   Graphics

3. ☐ **Aequorea victoria** isolate sGFP-206 green fluorescent protein (**gfp**) gene, partial cds
   753 bp linear DNA
   Accession: JX472995.1 GI: 506954254
   Protein   PubMed   Taxonomy

   GenBank   FASTA   Graphics   PopSet

4. ☐ **Aequorea victoria** green-fluorescent protein (**GFP**) mRNA, complete cds
   922 bp linear mRNA
   Accession: L29345.1 GI: 606383
   Protein   PubMed   Taxonomy

   GenBank   FASTA   Graphics

➢ Click "FASTA" to get the gene sequence of GFP.

Aequorea victoria green-fluorescent protein (GFP) mRNA, complete cds

GenBank: L29345.1

FASTA   Graphics

Go to: ☑

```
LOCUS       AEVGFP                   922 bp    mRNA    linear   INV 30-DEC-1994
DEFINITION  Aequorea victoria green-fluorescent protein (GFP) mRNA, complete
            cds.
ACCESSION   L29345
VERSION     L29345.1
KEYWORDS    aequorin; bioluminescence; energy acceptor protein;
            green-fluorescent protein.
SOURCE      Aequorea victoria
  ORGANISM  Aequorea victoria
            Eukaryota; Metazoa; Cnidaria; Hydrozoa; Hydroidolina; Leptothecata;
            Aequoreidae; Aequorea.
REFERENCE   1  (bases 1 to 922)
  AUTHORS   Inouye,S. and Tsuji,F.I.
  TITLE     Aequorea green fluorescent protein. Expression of the gene and
            fluorescence characteristics of the recombinant protein
  JOURNAL   FEBS Lett. 341 (2-3), 277-280 (1994)
   PUBMED   8137953
REFERENCE   2  (bases 1 to 922)
  AUTHORS   Inouye,S. and Tsuji,F.I.
  TITLE     Evidence for redox forms of the Aequorea green fluorescent protein
  JOURNAL   FEBS Lett. 351 (2), 211-214 (1994)
   PUBMED   8082767
```

Aequorea victoria green-fluorescent protein (GFP) mRNA, complete cds

GenBank: L29345.1

GenBank   Graphics

```
>L29345.1 Aequorea victoria green-fluorescent protein (GFP) mRNA, complete cds
TACACACGAATAAAAGATAACAAAGATGAGTAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTT
GTTGAATTAGATGGCGATGTTAATGGGCAAAAATTCTCTGTCAGTGGAGAGGGTGAAGGTGATGCAACAT
ACGGAAAACTTACCCTTAAATTTATTTGCACTACTGGGAAGCTACCTGTTCCATGGCCAACACTTGTCAC
TACTTTCTCTTATGGTGTTCAATGCTTTTCAAGATACCCAGATCATATGAAACAGCATGACTTTTTCAAG
AGTGCCATGCCCGAAGGTTATGTACAGGAAAGAACTATATTTTTACAAAGATGACGGGAACTACAAGACAC
GTGCTGAAGTCAAGTTTGAAGGTGATACCCTTGTTAATAGAATCGAGTTAAAAGGTATTGATTTTAAAGA
AGATGGAAACATTCTTGGACACAAAATGGAATACAACTATAACTCACATAATGTATACATCATGGCAGAC
AAACCAAAGAATGGAATCAAAGTTAACTTCAAAATTAGACACAACATTAAAGATGGAAGCGTTCAATTAG
CAGACCATTATCAACAAAATACTCCAATTGGCGATGGCCCTGTCCTTTTACCAGACAACCATTACCTGTC
CACACAATCTGCCCTTTCCAAAGATCCCAACGAAAAGAGAGATCACATGATCCTTCTTGAGTTTGTAACA
GCTGCTGGGATTACACATGGCATGGATGAACTATACAAATAAATGTCCAGACTTCCAATTGACACTAAAG
TGTCCGAACAATTACTAAATTCTCAGGGTTCCTGGTTAAATTCAGGCTGAGACTTTATTTATATATTTAT
AGATTCATTAAAATTTTATGAATAATTTATTGATGTTATTAATAGGGGCTATTTTCTTATTAAATAGGCT
ACTGGAGTGTAT
```

> Search "BLAST" in google and access BLAST.

> Click "Nucleotide BLAST" (BLASTN).

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS
BLAST+ 2.13.0 is here!
Starting with this release, we are including the blastn_vdb and tblastn_vdb executables in the BLAST+ distribution.

Thu, 17 Mar 2022 12:00:00 EST   More BLAST news...

Web BLAST

➢ Copy and paste the GFP sequence to the query sequence box, and set the job title. When pasting the sequence, we must paste the header part too. There are many options, but we will use the default setting in the practical experiment. After then, select the program "BLASTN" and click "BLAST" for alignment.

➢ We can see the alignment result. There are many similar sequences in the database, but the first one is the best-matched sequence. Select it and see "Graphic Summary" and "Alignments".

Job Title	GFP BLAST
RID	HD72483P013 *Search expires on 09-07 09:03 am* Download All ⌄
Program	BLASTN ❓ Citation ⌄
Database	nt See details ⌄
Query ID	lcl|Query_20431
Description	L29345.1 Aequorea victoria green-fluorescent protein (GF ...
Molecule type	dna
Query Length	922
Other reports	Distance tree of results MSA viewer ❓

Filter Results

Organism *only top 20 will appear* ☐ exclude

Type common name, binomial, taxid or group name

✚ Add organism

Percent Identity   E value   Query Coverage

[ ] to [ ]   [ ] to [ ]   [ ] to [ ]

Filter   Reset

Descriptions | **Graphic Summary** | Alignments | Taxonomy

🔍 *hover to see the title* ▶ *click to show alignments*   Alignment Scores ■ < 40 ■ 40 - 50 ■ 50 - 80 ■ 80 - 200 ■ >= 200 ❓

1 sequences selected ❓

Distribution of the top 1 Blast Hits on 1 subject sequences

Query
1   150   300   450   600   750   900

Descriptions | Graphic Summary | **Alignments** | Taxonomy

Alignment view [Pairwise   ⌄] ☐ CDS feature ❓ Restore defaults   Download ⌄

1 sequences selected ❓

⬇ Download ⌄   GenBank Graphics   ▼ Next ▲ Previous ◄Descriptions

Aequorea victoria green-fluorescent protein (GFP) mRNA, complete cds
Sequence ID: L29345.1 Length: 922 Number of Matches: 1

Range 1: 1 to 922 GenBank Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
1663 bits(1844)	0.0	922/922(100%)	0/922(0%)	Plus/Plus

```
Query  1    TACACACGAATAAAAGATAACAAAGATGAGTAAAGGAGAAGAACTTTTCACTGGAGTTGT   60
            |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    TACACACGAATAAAAGATAACAAAGATGAGTAAAGGAGAAGAACTTTTCACTGGAGTTGT   60

Query  61   CCCAATTCTTGTTGAATTAGATGGCGATGTTAATGGGCAAAAATTCTCTGTCAGTGGAGA   120
            |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   CCCAATTCTTGTTGAATTAGATGGCGATGTTAATGGGCAAAAATTCTCTGTCAGTGGAGA   120

Query  121  GGGTGAAGGTGATGCAACATACGGAAAACTTACCCTTAAATTTATTTGCACTACTGGGAA   180
            |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  GGGTGAAGGTGATGCAACATACGGAAAACTTACCCTTAAATTTATTTGCACTACTGGGAA   180

Query  181  GCTACCTGTTCCATGGCCAACACTTGTCACTACTTTCTCTTATGGTGTTCAATGCTTTTC   240
            |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  GCTACCTGTTCCATGGCCAACACTTGTCACTACTTTCTCTTATGGTGTTCAATGCTTTTC   240
```

- Find insulin gene sequences of human and pig, and compare the two sequences
  - Search "human insulin" in NCBI.

Search NCBI   [Human insulin   ✖] Search

Results found in 29 databases

GENE

INS – insulin
Homo sapiens (human)
Also known as: IDDM, IDDM1, IDDM2, ILPR, IRDN, MODY10, PNDM4
Gene ID: 3630

RefSeq transcripts (4) RefSeq proteins (4) RefSeqGene (2) PubMed (970)

[Orthologs] [Genome Data Viewer] [BLAST] [Download]

Was this helpful? 👍 👎

RefSeq transcripts

INS – 3 of 4 transcripts

Transcript	Isoform	Len (nt)
NM_000207.3		465
NM_001185097.2		491
NM_001291897.2		525

[View full table]   NCBI Datasets

RefSeq Sequences   ✚

> Go below and click "FASTA" to get the gene sequence of INS.

> Click "all" to see the orthologs of the INS gene.

➤ Click "INS" of pig and get gene sequence in the same way.

Species	Gene	Architecture	aa	
Homo sapiens human	INS insulin		110	⌄
Mus musculus house mouse	Ins2 insulin II		110	⌄
Rattus norvegicus Norway rat	Ins2 insulin 2		119	⌄
Bos taurus cattle	INS insulin		221	⌄
Sus scrofa pig	INS insulin		175	⌄
Gallus gallus chicken	INS insulin		107	⌄

Sus scrofa isolate TJ Tabasco breed Duroc chromosome 2, Sscrofa11.1, whole genome shotgun sequence

NCBI Reference Sequence: NC_010444.4

GenBank   Graphics

```
>NC_010444.4:c1498052-1496842 Sus scrofa isolate TJ Tabasco breed Duroc chromosome 2,
Sscrofa11.1, whole genome shotgun sequence
TGGCCCTCAGGCCGCTCGTTAAGACTCTAATGACCTCAAGGCCCCCAGAGGCGCTGATGACCCACGGAGA
TGATCCCGCAGGCCTGGCAGCAGGGAAATGATCCAGAAAGTGCCACCTCAGCCCCCAGCCATCTGCCACC
CACCTGGAGGCCCTCAGGGGCCGGGCGCCGGGGGGCAGGCGCTATAAAGCCGGCCGGGCCCAGCCGCCCC
CAGCCCTCTGGGACCAGCTGCGTTCCCAGGCCGCCGGCAAGCAGGTCTGTCCCCCTGGGCTCCCGTCAGC
TGGGTCTGGGCTGTCCTGCTGGGGCCAGGGCATCTCGGCAGGAGGACGTGGGCTCCTCTCTCGGAGCCCT
TGGGGGGTGAGGCTGGTGGGGGCTGCAGGTGCCCCTGGCTGGCCTCAACGCCGCCCGTCCCCCAGGTCCT
CACCCCCCGCCATGGCCCTGTGGACGCGCCTCCTGCCCCTGCTGGCCCTGCTGGCCCTCTGGGCGCCCGC
CCCGGCCCAGGCCTTCGTGAACCAGCACCTGTGCGGCTCCCACCTGGTGGAGGCGCTGTACCTGGTGTGC
GGGGAGCGCGGCTTCTTCTACACGCCCAAGGCCCGTCGGGAGGCGGAGAACCCTCAGGGTGAGCCGAGGG
GGCGTCCCGGGAGCGGTCGGGGGAGTTTTTAAGGAGGAAATTGGTAAAAGTGACCAACTCCCTGGGAGCT
GAGCCCAGAGACACCCCTCCCACGCCCCGGTCCCGCTCGAGAAGCCCCCCTTCCCTCCCCTCCTCCCGAG
GCGGCTCCAGGGAGGAATCTTACGGAGTCAAGGCCCGGGTGCCGCTGGTCTCCGAGTGACATGGCCGTGG
TGTCCCGTCTGCCGGCCCACATGCCCGTGGGAGATGCCCCATCCCCCTGGGAGGGGGCCCCGTGCCGGGC
AGGCAGCGGGAGGCCCAGGACCGGCGGCTGCTGCGGCTTCCACTCCAGGGTGGGCGGGGTGGGGGGTGGC
TGTCTCTGTGTGACCGGCTCTCCCCGCAGCAGGTGCCGTGGAGCTGGGCGGAGGCCTGGGCGGCCTGCAG
GCCCTGGCGCTGGAGGGGCCCCCGCAGAAGCGTGGCATCGTGGAGCAGTGCTGCACCAGCATCTGTTCCC
TCTACCAGCTGGAGAACTACTGCAACTAGGCCGCCCCTGAGGGCGCCTGCTGCTCCCCGCACCCCAAAAC
CCAATAAAGTCCTGAATGAGC
```

➤ Go to BLASTN and check the "Align two or more sequences". Paste the human sequence in the query sequence and paste the pig's sequence in the subject sequence. After then, select "blastn" for the program, and click "BLAST".

Align Sequences Nucleotide BLAST

| blastn | blastp | blastx | tblastn | tblastx |

BLASTN programs search nucleotide subjects using a nucleotide query.

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ❓ Clear

>NC_000011.10:c2161209-2159779 Homo sapiens chromosome 11, GRCh38.p14 Primary Assembly
AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGGTCTGTT
CCAAGGGCCTTTGCGTCAGGT

Query subrange ❓
From
To

Or, upload file   파일 선택 | 선택된 파일 없음   ❓

Job Title   Human INS vs. Pig INS
Enter a descriptive title for your BLAST search ❓

☑ Align two or more sequences ❓

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ❓   Clear

>NC_010444.4:c1498052-1496842 Sus scrofa isolate TJ Tabasco breed Duroc chromosome 2, Sscrofa11.1, whole genome shotgun sequence
TGGCCCTCAGGCCGCTCGTTAAGACTCTAATGACCTCAAGGCCCCCAGA
GGCGCTGATGACCCACGGAGA

Subject subrange ❓
From
To

Or, upload file   파일 선택 | 선택된 파일 없음   ❓

Program Selection

Optimize for
- ○ Highly similar sequences (megablast)
- ○ More dissimilar sequences (discontiguous megablast)
- ◉ Somewhat similar sequences (blastn)

Choose a BLAST algorithm ❓

BLAST   Search nucleotide sequence using Blastn (Optimize for somewhat similar sequences)
☐ Show results in a new window

- We can see how similar human's and pig's insulin gene sequence from the result.

Job Title	Human INS vs. Pig INS
RID	HD8JZJPW114 Search expires on 09-07 09:29 am Download All ⌄
Program	Blast 2 sequences Citation ⌄
Query ID	lcl|Query_43979 (dna)
Query Descr	NC_000011.10:c2161209-2159779 Homo sapiens chromo ...
Query Length	1431
Subject ID	lcl|Query_43981 (dna)
Subject Descr	NC_010444.4:c1498052-1496842 Sus scrofa isolate TJ Ta ...
Subject Length	1211
Other reports	MSA viewer ❓

Filter Results

Percent Identity [ ] to [ ]
E value [ ] to [ ]
Query Coverage [ ] to [ ]
[Filter] [Reset]

Tabs: **Descriptions** | Graphic Summary | Alignments | Dot Plot

Sequences producing significant alignments   Download ⌄   Select columns ⌄   Show 100 ⌄   ❓

☑ select all 1 sequences selected   Graphics   MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑ NC_010444.4:c1498052-1496842 Sus scrofa isolate TJ Tabasco breed Duroc chromosome 2, Sscrofa11.1, whole genom...		250	476	31%	8e-70	83.83%	1211	Query_43981

Descriptions	**Graphic Summary**	Alignments	Dot Plot

🖑 *hover to see the title* ➤ *click to show alignments*

Alignment Scores ■ < 40 ■ 40 - 50 ■ 50 - 80 ■ 80 - 200 ■ >= 200 ❓

1 sequences selected ❓

Distribution of the top 4 Blast Hits on 1 subject sequences

Query
1   250   500   750   1000   1250
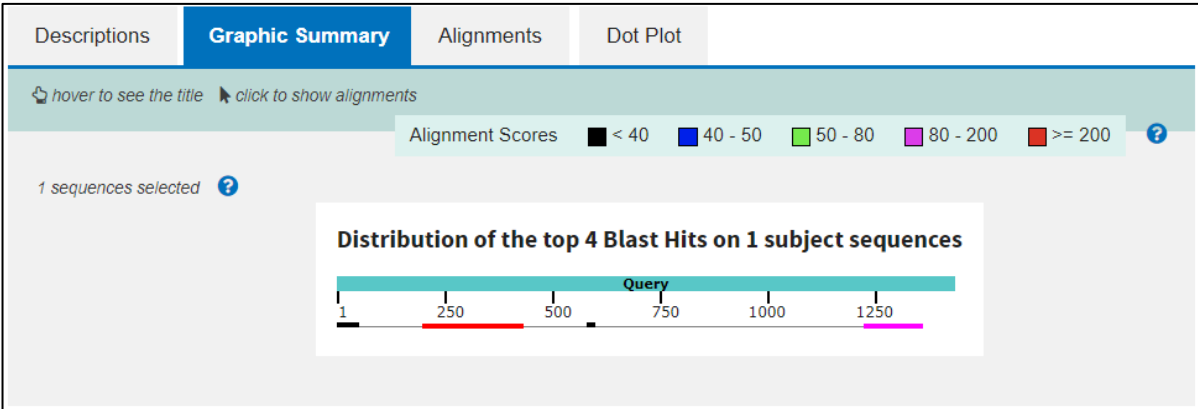
⬇ Download ⌄   Graphics   Sort by: [E value ⌄]   ▽ Next △ Previous ◀ Descriptions

NC_010444.4:c1498052-1496842 Sus scrofa isolate TJ Tabasco breed Duroc chromosome 2, Sscrofa11.1, whole genome shotgun sequence

Sequence ID: **Query_43981** Length: **1211** Number of Matches: **4**

Range 1: 392 to 625 Graphics   ▼ Next Match △ Previous Match

Score	Expect	Identities	Gaps	Strand
250 bits(276)	8e-70	197/235(84%)	1/235(0%)	Plus/Plus

```
Query  198  GCCTCAGCCCTGCCTGTCTCCCAGATCACTGTCCTTCTGCCATGGCCCTGTGGATGCGCC  25
            |||||| | | ||| ||| ||||| || ||   ||  | |||||||||||||| |||||
Sbjct  392  GCCTCAACGCCGCCCGTCCCCCAGGTC-CTCACCCCCCGCCATGGCCCTGTGGACGCGCC  45

Query  258  TCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGA  31
            |||||||||||||| ||| ||||||||||||||||| || | ||   ||||| |||| ||||
Sbjct  451  TCCTGCCCCTGCTGGCCCTGCTGGCCCTCTGGGCGCCCGCCCCGGCCCAGGCCTTCGTGA  51

Query  318  ACCAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAG  37
            |||| ||||||||||||| |||||||||| |||  ||||||| || | |||||||  || |
Sbjct  511  ACCAGCACCTGTGCGGCTCCCACCTGGTGGAGGCGCTGTACCTGGTGTGCGGGGAGCGCG  57

Query  378  GCTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGGTGAGCC  432
            ||||||||||||| |||| ||||| ||| ||||||||| ||| ||| ||||||||||||
Sbjct  571  GCTTCTTCTACACGCCCAAGGCCCGTCGGGAGGCGGAGAACCCTCAGGGTGAGCC  625
```