Molecular Biology Laboratory

Bioinformatics and Genomics Lab.

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

TA
Junseob Han, Hyunseok Song

Contact
Junseob Han
010.2113.6458
hljs502@gmail.com

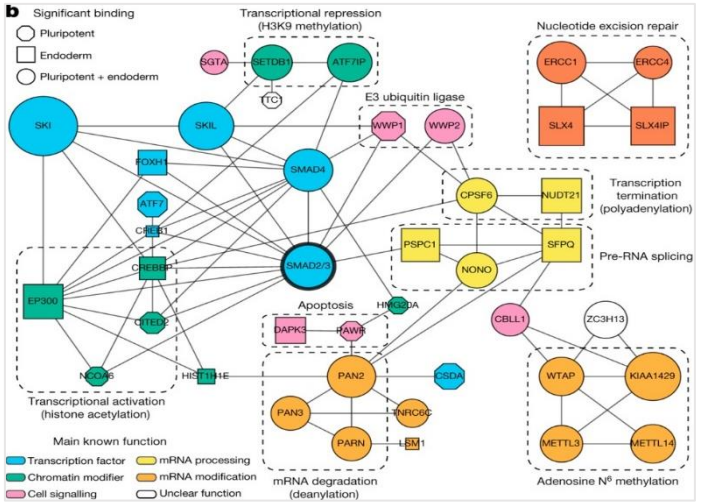
Goal of This Week

1. To know what is bioinformatics
2. To know the format of DNA & RNA data and gene annotation used in bioinformatics
3. To know how to visualize the data (UCSC Genome Browser)
4. To know how to analyze the sequence data of DNA or RNA (BLAST)

What is Bioinformatics?

- Molecular biology covers DNA, RNA, and protein
- Information on these molecules can be changed into mathematical and computerized data
  - DNA & RNA: Nucleotide sequence
  - Protein: Amino acid sequence
  - The interaction of each molecule can be changed to computerized or visualized data

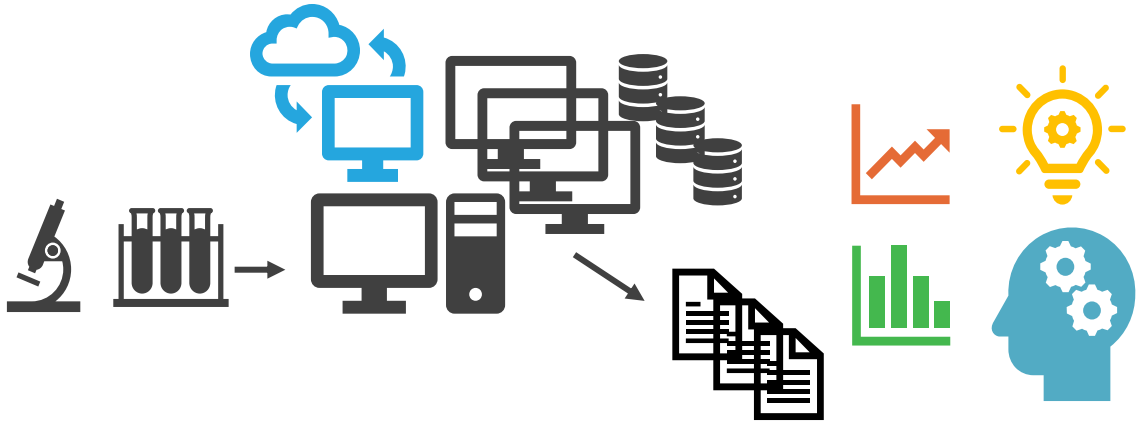
Overview - Bioinformatics Core, Mayo Clinic Research



Alessandro Bertero et al., Nature, 2018

What is Bioinformatics?

- Bioinformatics is the study that analyzes mathematical and computerized data to solve biological problems
  - Bioinformaticians use applied mathematics, data science, statistics, computer science, AI and et cetera for analyzing data and deducing biological meaning
- There are many tools or programs for analyzing data (BLAST, RNAfold, AlphaFold, etc.)
  - Each program has a different purpose or pros and cons, so we select the program carefully to match the analysis purpose and data type



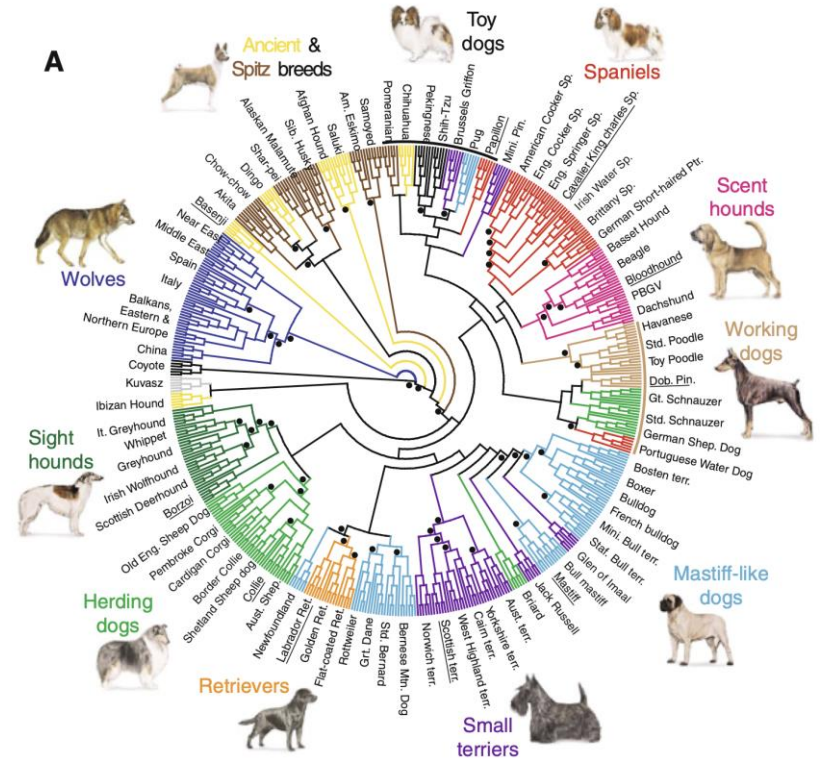
Sequence Data

- DNA and RNA are composed of nucleotides, and proteins are composed of amino acids
  - DNA & RNA: Adenine (A), Thymine (T), Guanine (G), Cytosine (C), Uracil (U)
  - Proteins: Methionine (M), Valine (V), Cysteine (C) …
- We can get the sequence data of these molecules with various sequencing methods
  - DNA sequencing, RNA sequencing, and protein sequencing

RNA-Seq: Basics, Applications and Protocol, Technology Networks, 2018

Sequence Data

- Using these sequence data, we can infer similarities and relations between different species
  - During evolution, genome sequence changing occurs
  - Some species which have common ancestors share a common sequence change
- There are many tools or programs for analyzing sequence data
  - These tools have unique algorithms
  - Each tool uses a different data type or format



Robert K. Wayne and Bridgett M. vonHoldt, Mammalian Genome, 2012

Format of Sequence Data - FASTA

- Sequence data of DNA, RNA, and proteins are written in "FASTA" format generally
- "FASTA" format is consist of two parts: Header part & Sequence part
  - Header part has information on sequences like chromosomes, gene name, protein name, etc.
  - Sequence part has sequence literally
  - Header part and Sequence part is distinct by ">"
- DNA and RNA FASTA files have nucleotide sequences and protein FASTA files have amino acid sequences
  - DNA sequence contains exon and intron sequence, but RNA sequence contains exon sequence only

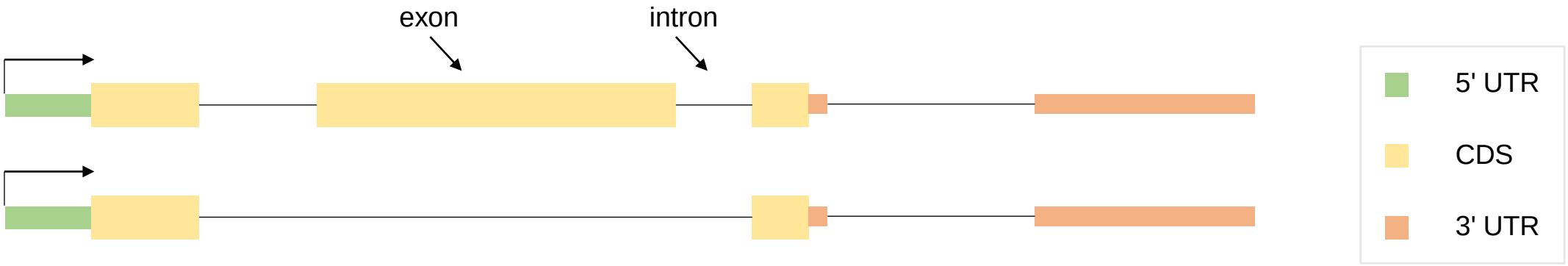
Header →

Sequence

```
>NC_000017.11:c7687490-7668421 Homo sapiens chromosome 17, GRCh38.p14 Primary Assembly
CTCAAAAGTCTAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTGCGTTCGGGC
TGGGAGCGTGCTTTCCACGACGGTGACACGCTTCCCTGGATTGGGTAAGCTCCTGACTGAACTTGATGAG
TCCTCTCTGAGTCACGGGCTCTCGGCTCCGTGTATTTTCAGCTCGGGAAAATCGCTGGGGCTGGGGGTGG
GGCAGTGGGGACTTAGCGAGTTTGGGGGTGAGTGGGATGGAAGCTTGGCTAGAGGGATCATCATAGGAGT
TGCATTGTTGGGAGACCTGGGTGTAGATGATGGGGATGTTAGGACCATCCGAACTCAAAGTTGAACGCCT
AGGCAGAGGAGTGGAGCTTTGGGGAACCTTGAGCCGGCCTAAAGCGTACTTCTTTGCACATCCACCCGGT
GCTGGGCGTAGGGAATCCCTGAAATAAAAGATGCACAAAGCATTGAGGTCTGAGACTTTTGGATCTCGAA
ACATTGAGAACTCATAGCTGTATATTTTAGAGCCCATGGCATCCTAGTGAAAACTGGGGCTCCATTCCGA
AATGATCATTTGGGGGTGATCCGGGGAGCCCAAGCTGCTAAGGTCCCACAACTTCCGGACCTTTGTCCTT
CCTGGAGCGATCTTTCCAGGCAGCCCCCGGCTCCGCTAGATGGAGAAAATCCAATTGAAGGCTGTCAGTC
```

Gene Structure

- Gene structure is divided into exon and intron
  - Exon: Part of a gene that transcripts to RNA
  - Intron: Part of a gene that connects exons
- Some transcripts come from the same gene but have different exon and intron structure
  - We called these transcripts "isoforms"
- Exon of mRNA is divided into coding DNA sequence (CDS) and untranslated region (UTR)
  - UTR is also divided into 5' UTR and 3' UTR depending on the relative position to transcript start site (TSS)



Gene Annotation

- All genes have structure and coordination (position) information, and gene annotation indicates this information
  - Symbol, gene id, coordination information (chromosome, start & end position, strand, etc.)
- Gene annotation is provided from various database
  - RefSeq, UCSC Genome Browser, Ensembl, GENCODE
- There are 3 common formats for gene annotation
  - Gene prediction format (genePred, refFlat)
  - General transfer format (GTF)
  - Browser extensible data (BED)

Gene Annotation - genePred & refFlat

- Gene prediction format (genePred, refFlat)
  - This format is provided from RefSeq database of NCBI
  - "genePred" and "refFlat" are almost the same but "genePred" format doesn't have a gene symbol column
- Example of "refFlat" format
  - Information about each gene or transcript is written in one row

1. Gene Symbol	2. ID	3. Chromosome	4. Strand	5. Transcription Start Position	6. Transcription End Position	7. CDS Start Position	8. CDS End Position	9. Number of Exons	10. Exon Start Position	11. Exon End Position
OR4F29	NM_001005221	chr1	+	367658	368597	367658	368597	1	367658,	368597,
OR4F3	NM_001005224	chr1	+	367658	368597	367658	368597	1	367658,	368597,
OR4F16	NM_001005277	chr1	+	367658	368597	367658	368597	1	367658,	368597,
OR4F29	NM_001005221	chr1	-	621095	622034	621095	622034	1	621095,	622034,

  - If the strand is "-" like OR4F29, the start position (5, 7, 10 columns) indicates the end position and the end position (6, 8, 11 column) indicate the start position

Gene Annotation - GTF

- Gene transfer format (GTF)
  - This format is provided from Ensembl, GENCODE database and this format is commonly used
- Example of "GTF" format
  - Information of each gene or transcript is written in multi rows

1. Chromosome 3. Feature 5. End Position 7. Strand 9. Attribute

chr1	HAVANA	gene	65419	71585	.	+	.	gene_id "ENSG00000186092.6_5"; ge
chr1	HAVANA	transcript	65419	71585	.	+	.	gene_id "ENSG00000186092.
chr1	HAVANA	exon	65419	65433	.	+	.	gene_id "ENSG00000186092.6_5"; tr
chr1	HAVANA	UTR	65419	65433	.	+	.	gene_id "ENSG00000186092.6_5"; tr

2. Source 4. Start Position 6. Score 8. Frame

  - Attribute column has various information like gene id, transcript id, symbol, gene type etc.
  - Generally, Score and Frame columns are not used
    - Score: this column mean the probability that the information of row is real
    - Frame: If feature is CDS, this column mean start position's codon frame

Gene Annotation - BED

- Browser extensible data (BED)
  - This format is used for showing gene information but it can be used for showing sequencing read information too
- Example of "BED" format
  - Information about each gene or transcript is written in one row

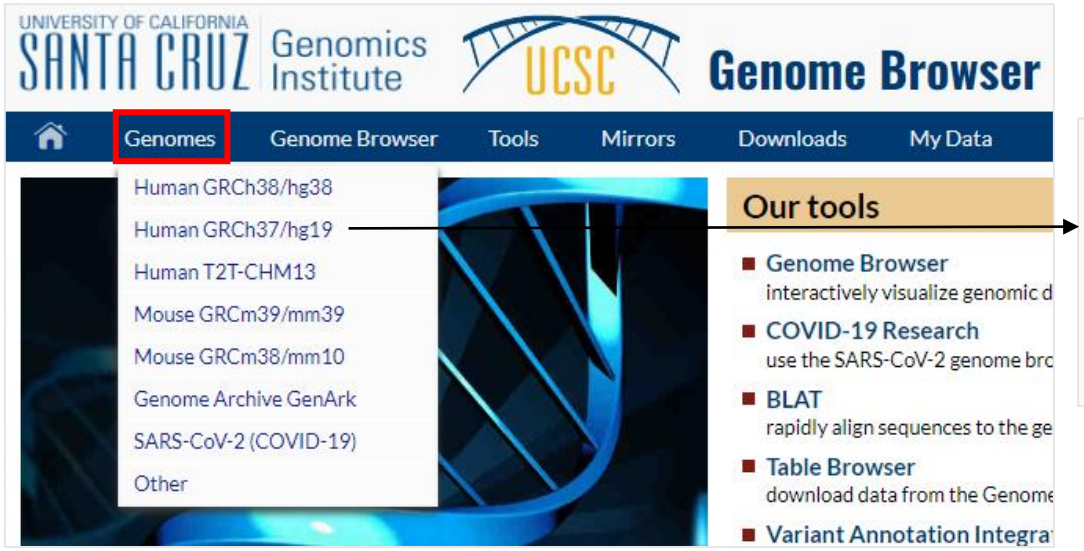
1. Chromosome
2. Start Position
3. End Position
4. ID
5. Score
6. Strand
7. CDS Start Position
8. CDS End Position
9. RGB Code for Visualization
10. Number of Exons
11. Size of each Exons
12. Start Position of each Exons

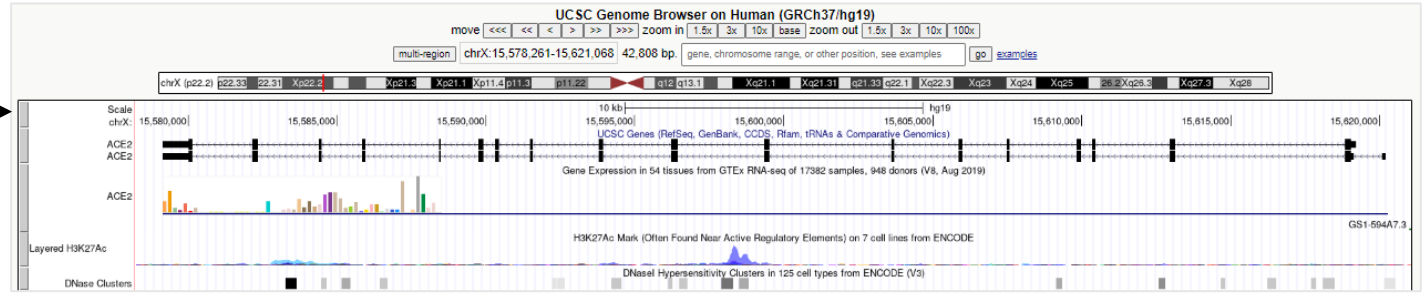
chr1	320161	321056	ENST00000432964.1	0	+	320161	320161	0	3	492,58,25,	0,719,870,
chr1	320161	324461	ENST00000423728.1	0	+	320161	320161	0	3	492,58,23,	0,4126,4277,
chr1	320334	322097	ENST00000601486.1	0	+	320334	320334	0	4	319,58,259,60,	0,546,697,1703,
chr1	320880	322203	ENST00000599771.2	0	+	320880	320880	0	3	58,233,166,	0,151,1157,
chr1	322077	342806	ENST00000455464.2	0	+	322077	322077	0	3	151,169,415,	0,12051,20314,
chr1	322671	324955	ENST00000419160.3	0	+	322671	322671	0	2	402,205,	0,2079,
chr1	323860	334505	ENST00000601814.1	0	+	323860	323860	0	3	200,58,377,	0,427,10268,

- "BED" format doesn't always have 12 columns
  - The minimum column number is 4, and the file format is named by column number (BED4 ~ BED12)

UCSC Genome Browser

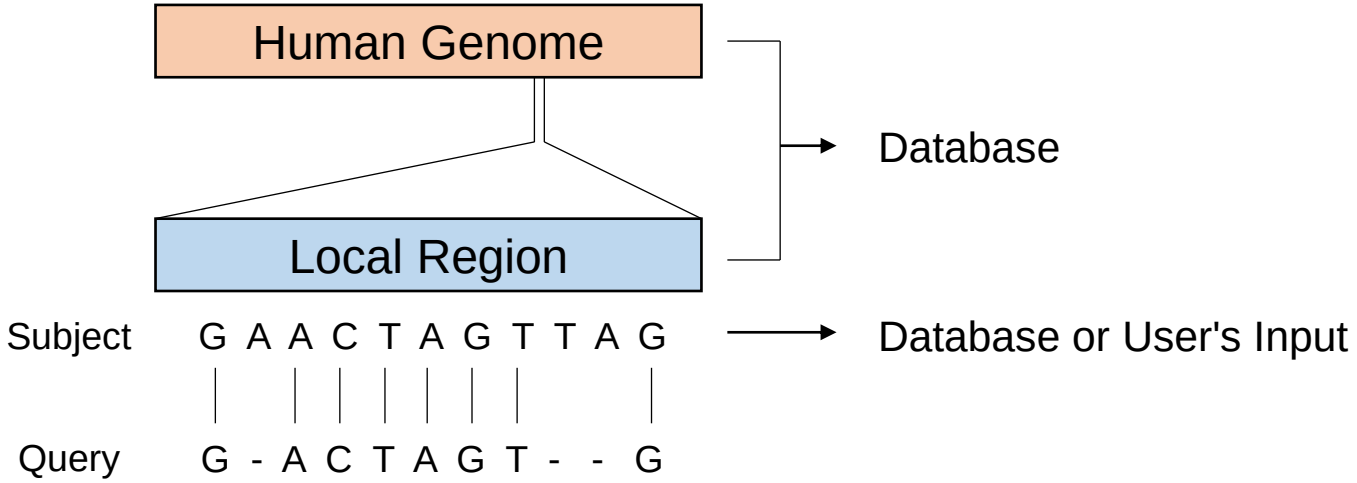
- UCSC genome browser is a web-based tool used for visualizing genome data easily
  - Search "UCSC genome browser" in google or use the hyperlink https://genome.ucsc.edu
- We can find specific regions, genome structure, expression patterns, chromatin accessible regions, and other information from this tool
- Hover a mouse pointer on "Genomes" and select species, and then we can show visualized information





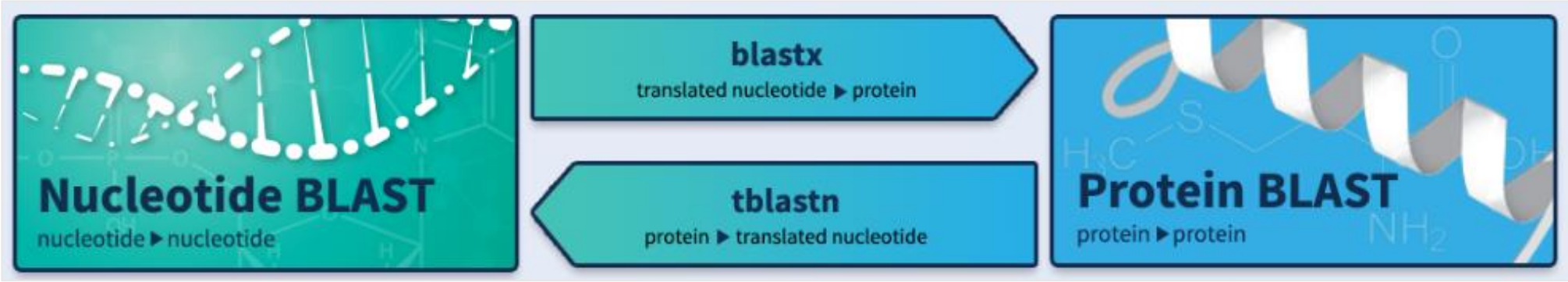
Basic Local Alignment Search Tool - BLAST

- "BLAST" is the program used for comparing multiple sequences and finding local alignment region
  - Local alignment region: Similar sequence region in two compared sequence
  - Search "BLAST" in google or use the hyperlink https://blast.ncbi.nlm.nih.gov/Blast.cgi
- We can use "BLAST" to search sequences from the genome database or to compare two sequences
  - Query: Input sequence for comparing
  - Subject: Matched sequence with query sequence in database or input sequence



Basic Local Alignment Search Tool - BLAST

- There are 4 tools in BLAST
  - BLASTN: Compare nucleotide sequence (query) and nucleotide sequence (subject)
  - BLASTP: Compare amino acid sequence (query) and amino acid sequence (subject)
  - BLASTX: Compare nucleotide sequence (query) and amino acid sequence (subject)
  - TBLASTN: Compare amino acid sequence (query) and nucleotide sequence (subject)



Practical Exercise

1. Practice how to use UCSC Genome Browser
    - Access the UCSC Genome Browser and find the human insulin gene
    - Make custom gene annotation and visualize
2. Practice how to get sequence data and how to use BLASTN
    - Find GFP gene sequence in NCBI database and align the sequence
    - Find human and pig insulin gene sequences in NCBI database and compare two sequence