

## Schedule for Week 2 Practical

14:00 - 14:10 Introduction to Part 1

14:10 - 14:25 Part 1 - Introduction to Python & the Jupyter Notebook

14:25 - 14:35 Introduction to Part 2

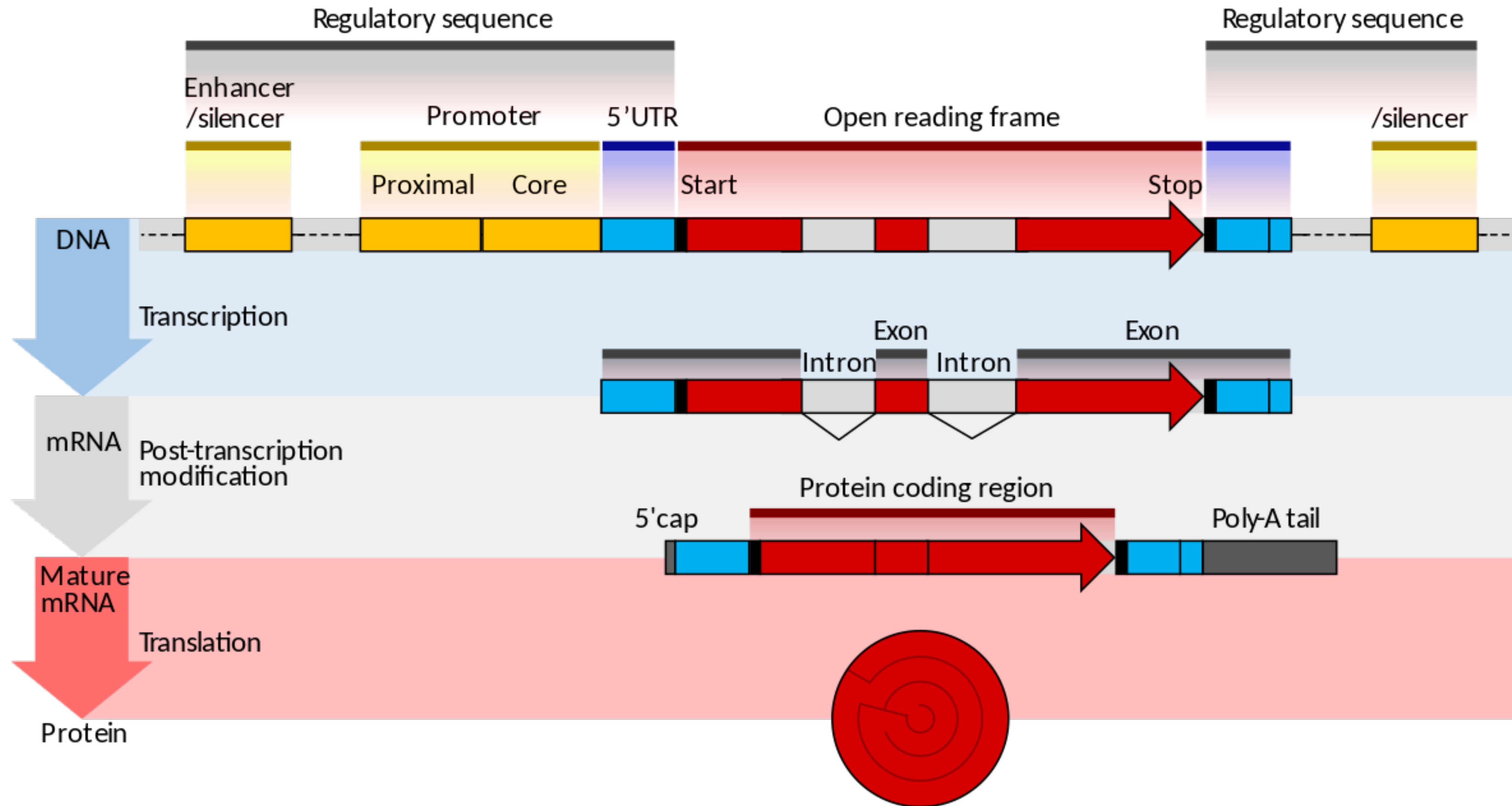
14:35 - 15:05 Part 2 - Accessing & Working with DNA, RNA & Protein Sequences

15:05 - 15:15 Introduction to Part 3

15:15 - 15:45 Part 3 - Pairwise Alignment

15:45 - 15:55 Q&A

# Structure of a Eukaryotic Gene



# How To Read mRNA Sequences

Start: AUG  
Stop: UAA, UAG, UGA

## Reading frame #1

5'-AGUCUUACCGCAUUGUGG-3'  
| | | | | |  
Ser--Leu--Thr--Ala--Leu-Trp

## Reading frame #2

5'-AGUCUUACCGCAUUGUGG-3'  
| | | | |  
Val--Leu--Pro--His--Cys

## Reading frame #3

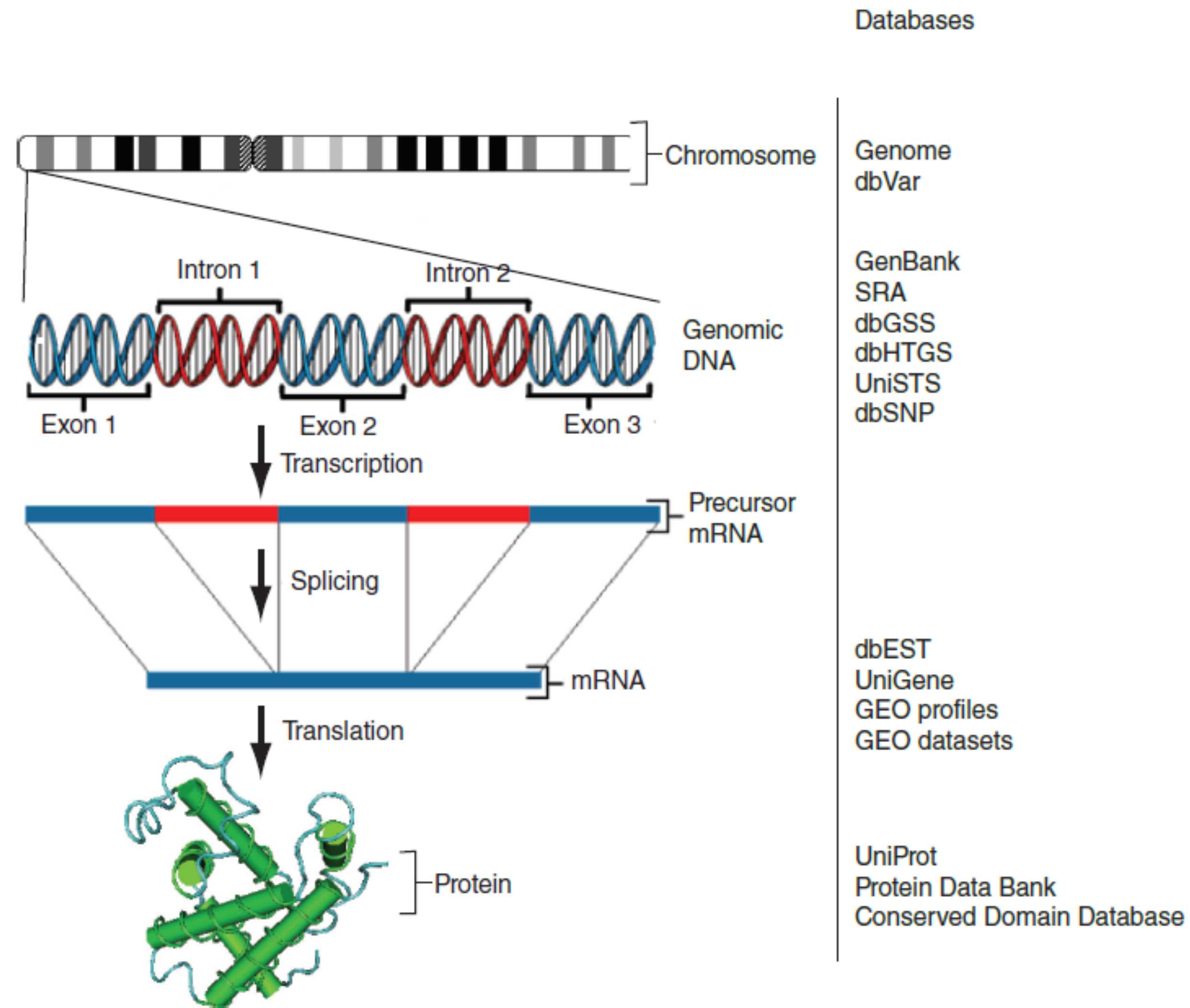
5'-AGUCUUACCGCAUUGUGG-3'  
| | | | |  
Ser--Tyr--Arg--Ile--Val

# The Genetic Code

Standard genetic code								
1st base	2nd base							
	T		C		A		G	
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine
	TTC		TCC		TAC		TGC	
	TTA		TCA		TAA	Stop (Ochre)	TGA	Stop (Opal)
	TTG		TCG		TAG	Stop (Amber)	TGG	(Trp/W) Tryptophan
C	CTT	(Leu/L) Leucine	CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine
	CTC		CCC		CAC		CGC	
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA	
	CTG		CCG		CAG		CGG	
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine
	ATC		ACC		AAC		AGC	
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine
	ATG <sup>[A]</sup>	(Met/M) Methionine	ACG		AAG		AGG	
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA	
	GTG		GCG		GAG		GGG	
								3rd base
								T
								C
								A
								G
								T
								C
								A
								G
								T
								C
								A
								G

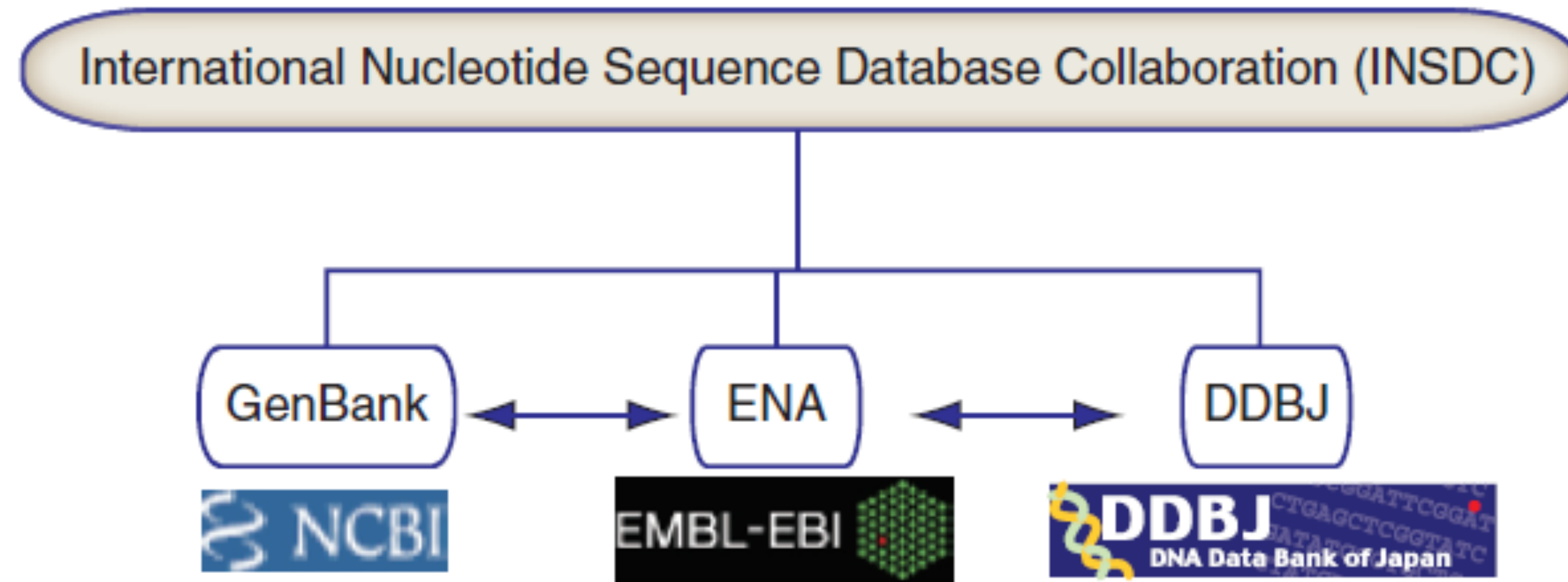


# Data Types & Databases

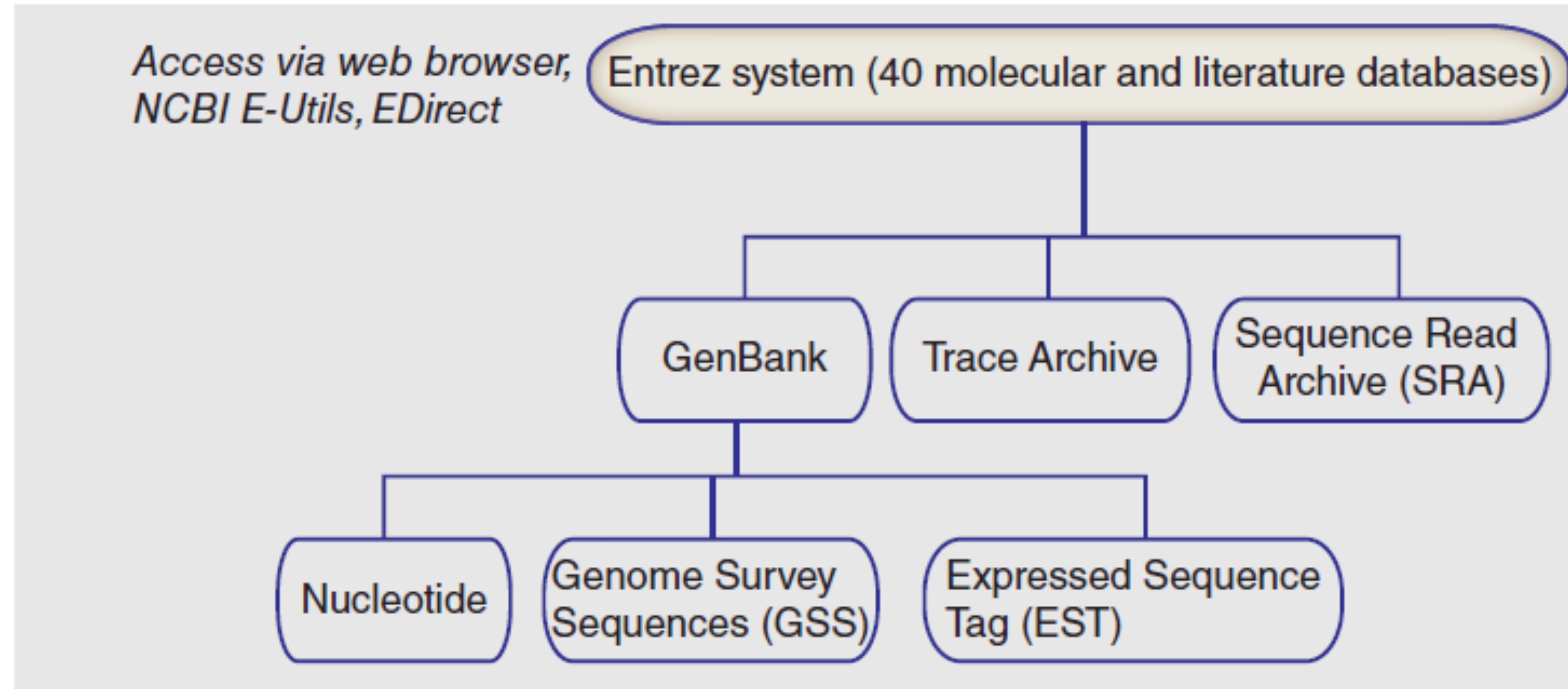


B&FG3e

# Co-ordination of Sequence Data

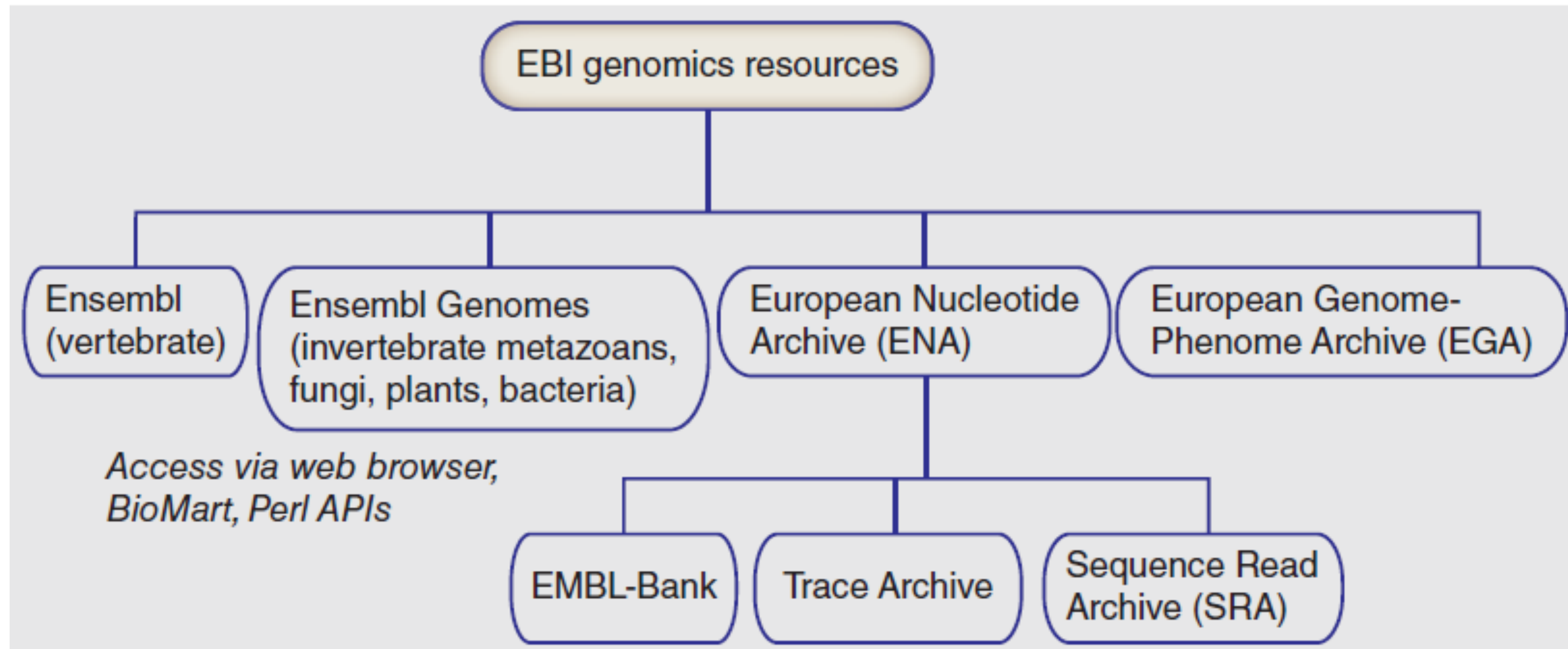


# National Center for Biotechnology Information (NCBI)



B&FG3e

# European Bioinformatics Institute (EBI)



B&FG3e



# Accession Numbers

NCBI includes databases that contain information on DNA, RNA, or protein sequences

You may want to acquire information beginning with a query such as the **name of a protein, gene or organism of interest**, or even **raw nucleotide sequences** comprising a DNA sequence of interest.

DNA sequences and other molecular data are tagged with **accession numbers** that are used to identify a sequence or other record relevant to molecular data.

# Accession Numbers are Diverse

## DNA

**CH471100.2** GenBank genomic DNA sequence

**NC\_000001.10** Genomic contig

**rs121434231** dbSNP (single nucleotide polymorphism)

## RNA

**AI687828.1** An expressed sequence tag (1 of 184)

**NM\_001206696** RefSeq DNA sequence (from a transcript)

## Protein

**NP\_006138.1** RefSeq protein

**CAA18545.1** GenBank protein

**O14896** SwissProt protein

**1KT7** Protein Data Bank structure record

<https://www.ncbi.nlm.nih.gov/refseq/>

RefSeq provides an expertly curated accession number that corresponds to the most stable, agreed-upon “reference” version of a sequence.

RefSeq identifiers include the following formats:

- Complete genome NC\_#####
- Complete chromosome NC\_#####
- Genomic contig NT\_#####
- mRNA (DNA format) NM\_##### e.g. NM\_006744
- Protein NP\_##### e.g. NP\_006735
-

# Biological Sequence Databases

- ♦ Standardised data formats
- ♦ Agreed ways of annotation
- ♦ Comprehensive meta-data
- ♦ Consistent and efficient cross-referencing
- ♦ Curated data sets for referencing



<https://www.ncbi.nlm.nih.gov>



<https://www.ensembl.org/index.html>



# Find a gene

Go to: <http://www.ncbi.nlm.nih.gov/>



# Obtain the genomic sequence

**DNA** →

**Transcript Protein** →

**Gene** →

Hide sidebar >>

**Table of contents**

- Summary
- Genomic context
- Genomic regions, transcripts, and products**
- Bibliography
- Variation
- Pathways from BioSystems
- Interactions
- General gene information
  - Markers, Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

**Genomic regions, transcripts, and products**

Go to [reference sequence details](#)

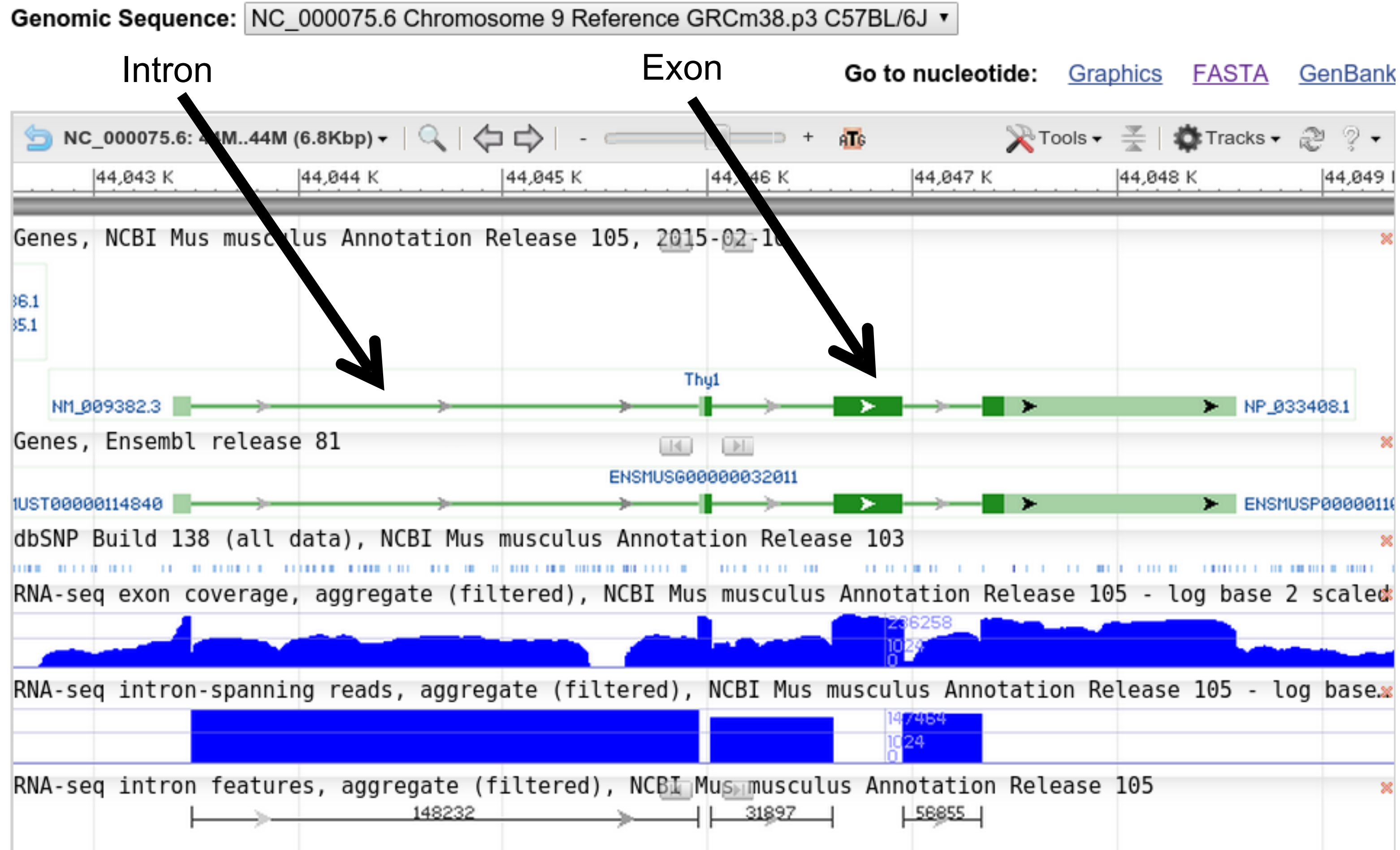
Genomic Sequence: **NC\_000075.6 Chromosome 9 Reference GRCh38.p3 C57BL/6J**

Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)

NC\_000075.6: 44M..44M (6.8Kbp)

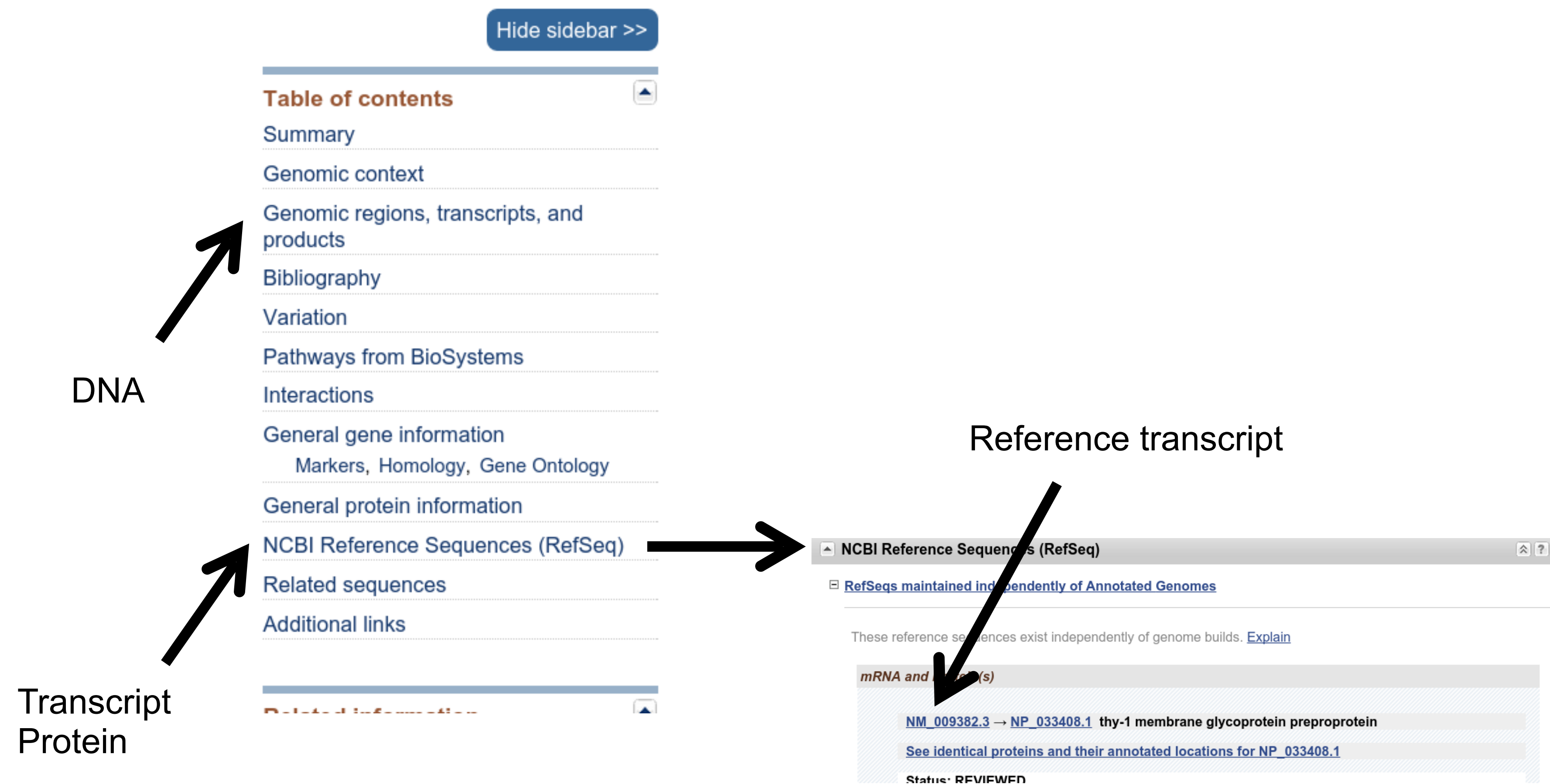
44,043 K 44,044 K 44,045 K 44,046 K 44,047 K 44,048 K 44,049

# Genomic regions



<http://www.ncbi.nlm.nih.gov/tools/sviewer/legends/>

# Obtain the transcript





# Gene Sequences

GenBank

Send

## Homo sapiens Thy-1 cell surface antigen (THY1), transcript variant 2

NCBI Reference Sequence: NM\_001311160.1

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

Go to:

LOCUS NM\_001311160 2944 bp mRNA linear PRI 11-SEP-2016  
DEFINITION Homo sapiens Thy-1 cell surface antigen (THY1), transcript variant 2, mRNA.  
ACCESSION NM\_001311160  
VERSION NM\_001311160.1 GI:902967470  
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 2944)  
AUTHORS Zhu GC, Gao L, He J, Long Y, Liao S, Wang H, Li X, Yi W, Pei Z, Wu M, Xiang J, Peng S, Ma J, Zhou M, Zeng Z, Xiang B, Xiong W, Tang K, Cao L, Li X, Li G and Zhou Y.  
TITLE CD90 is upregulated in gastric cancer tissues and inhibits gastric

- ☐ Complete Record  
☒ Coding Sequences  
☐ Gene Features

Download features.

Format

FASTA Nucleotide

Create File

### Articles about the THY1

CD90 is upregulated in  
inhibits gastric cancer c

CD90+ liver cancer cell  
phenotype through the

The stromal cell-surface  
activation protein- $\alpha$  I [B