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

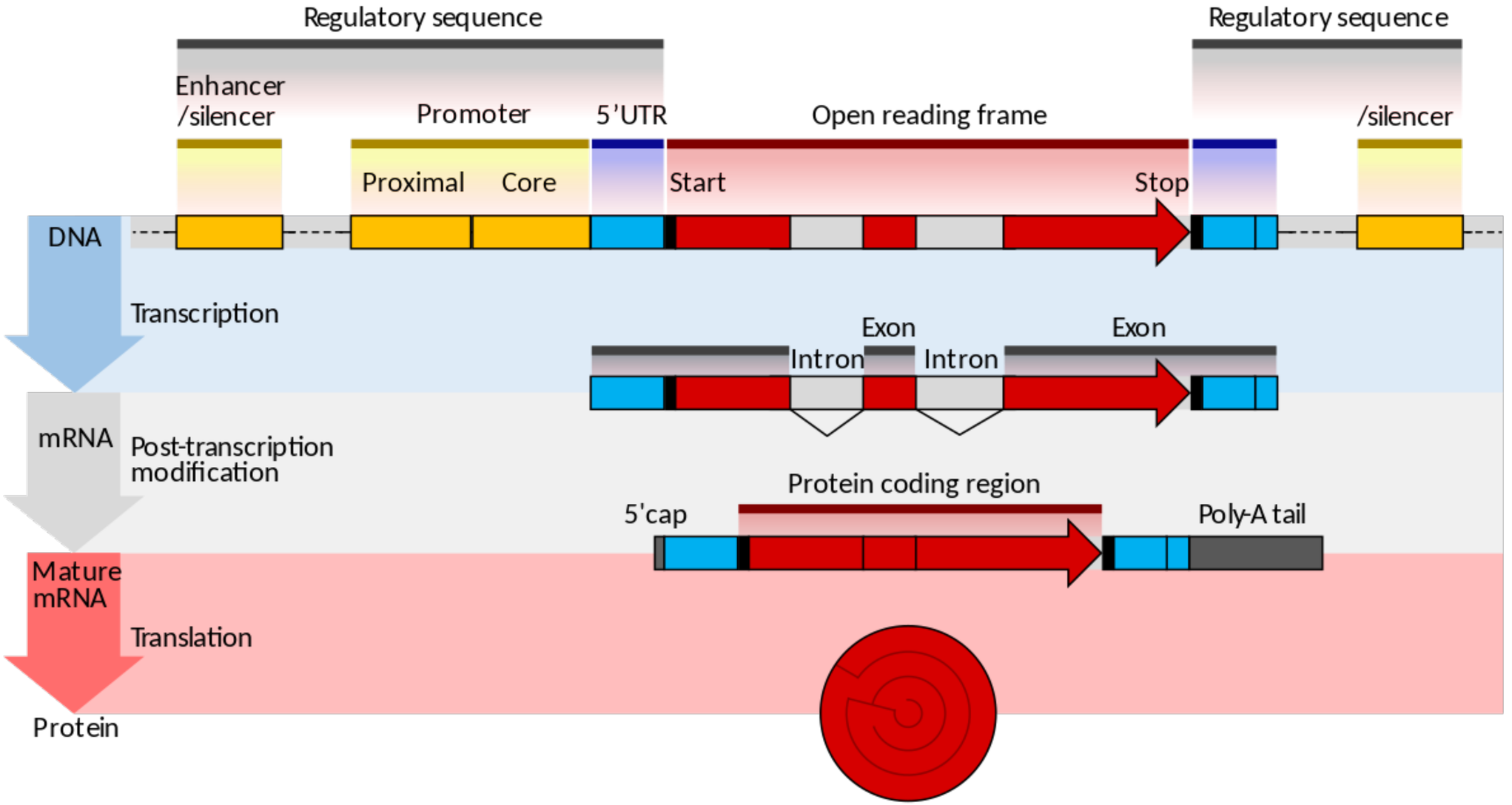
14:35 - 15:05 Part 2 - Accessing & Working with DNA, RNA & Protein Sequences

15:05 - 15:15 Introduction to Part 3

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

15:45 - 15:55 Q&A

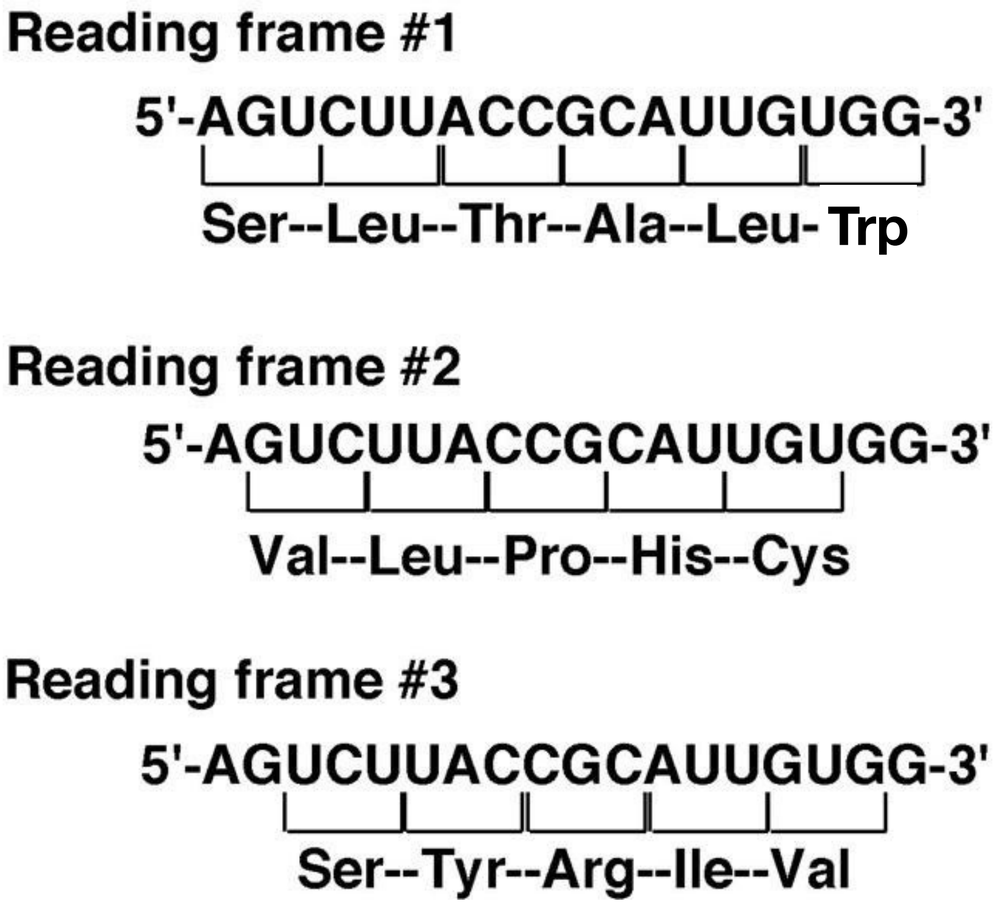
Structure of a Eukaryotic Gene



ian.simpson@ed.ac.uk

How To Read mRNA Sequences

Start: AUG
Stop: UAA, UAG, UGA



ian.simpson@ed.ac.uk

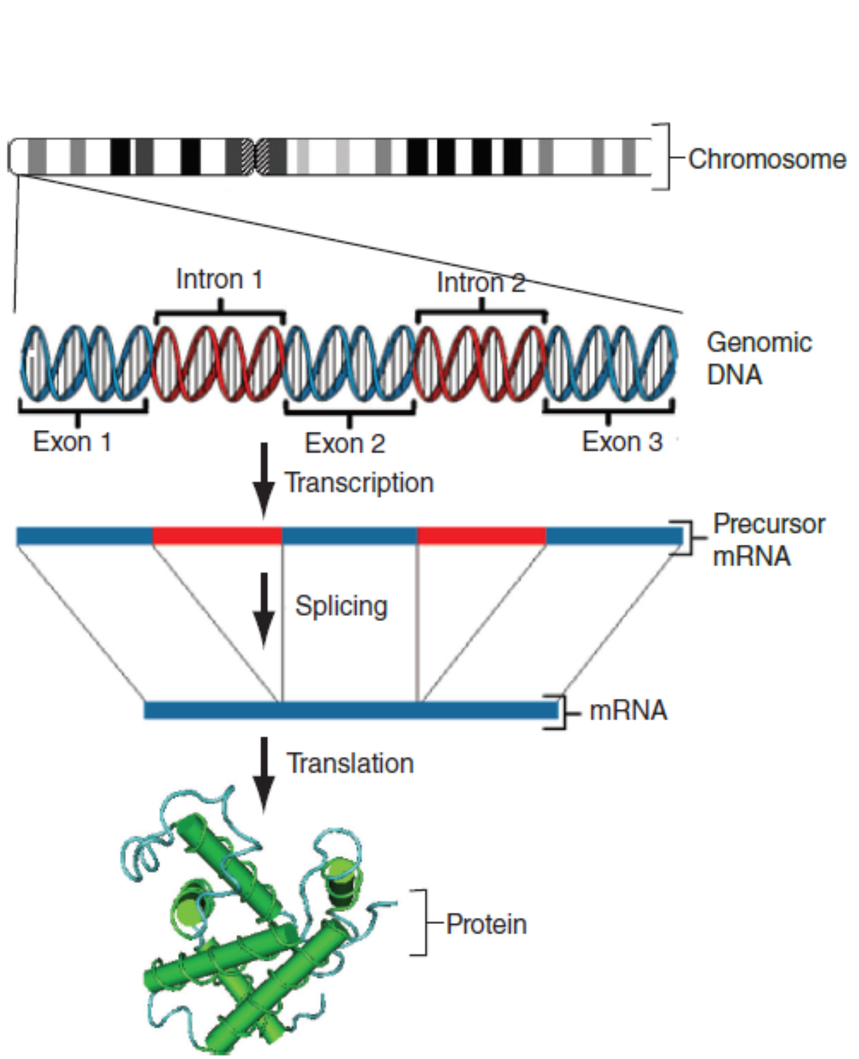
The Genetic Code

Standard genetic code

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

ian.simpson@ed.ac.uk

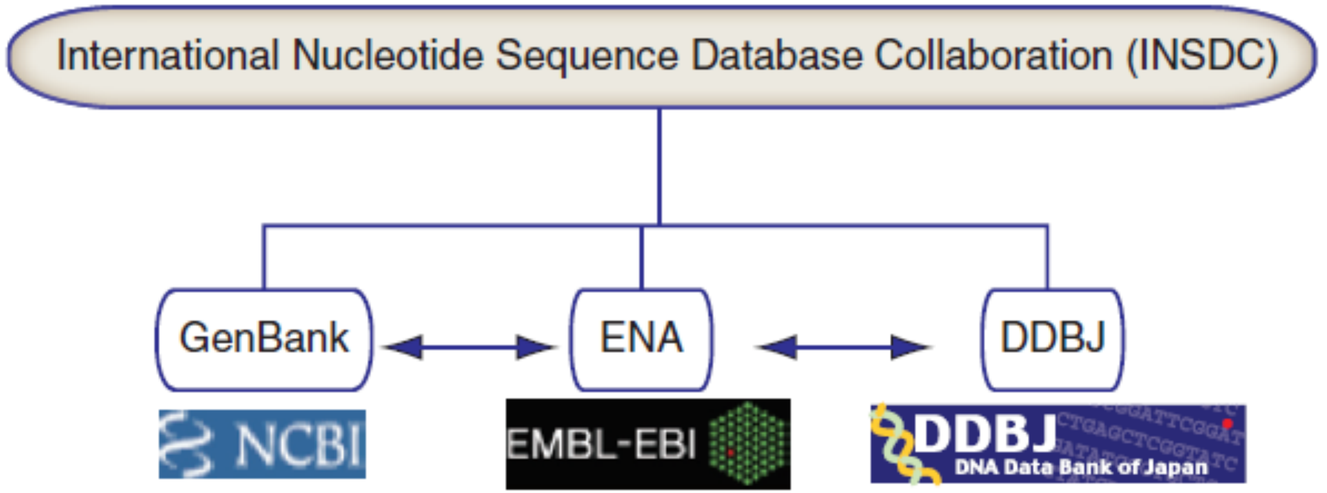
Data Types & Databases



B&FG3e

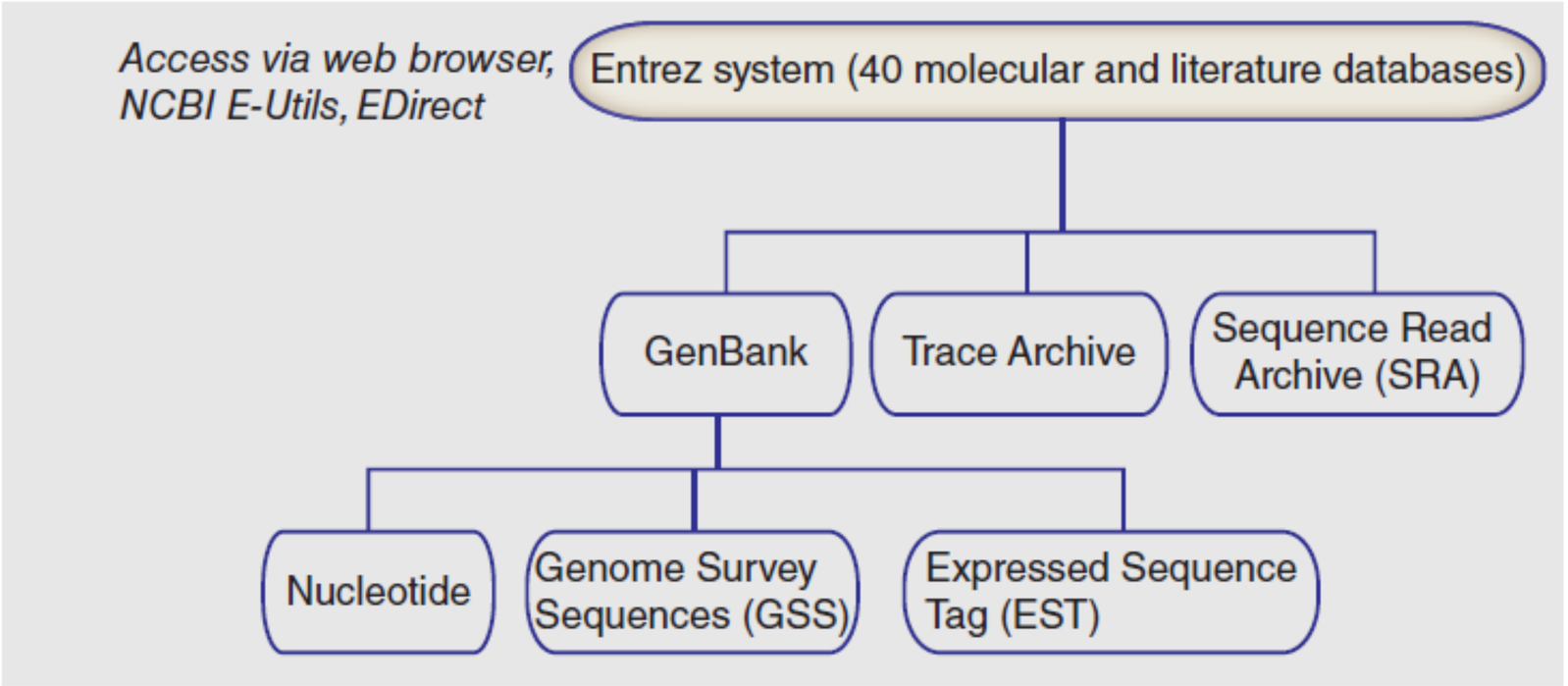
ian.simpson@ed.ac.uk

Co-ordination of Sequence Data



ian.simpson@ed.ac.uk

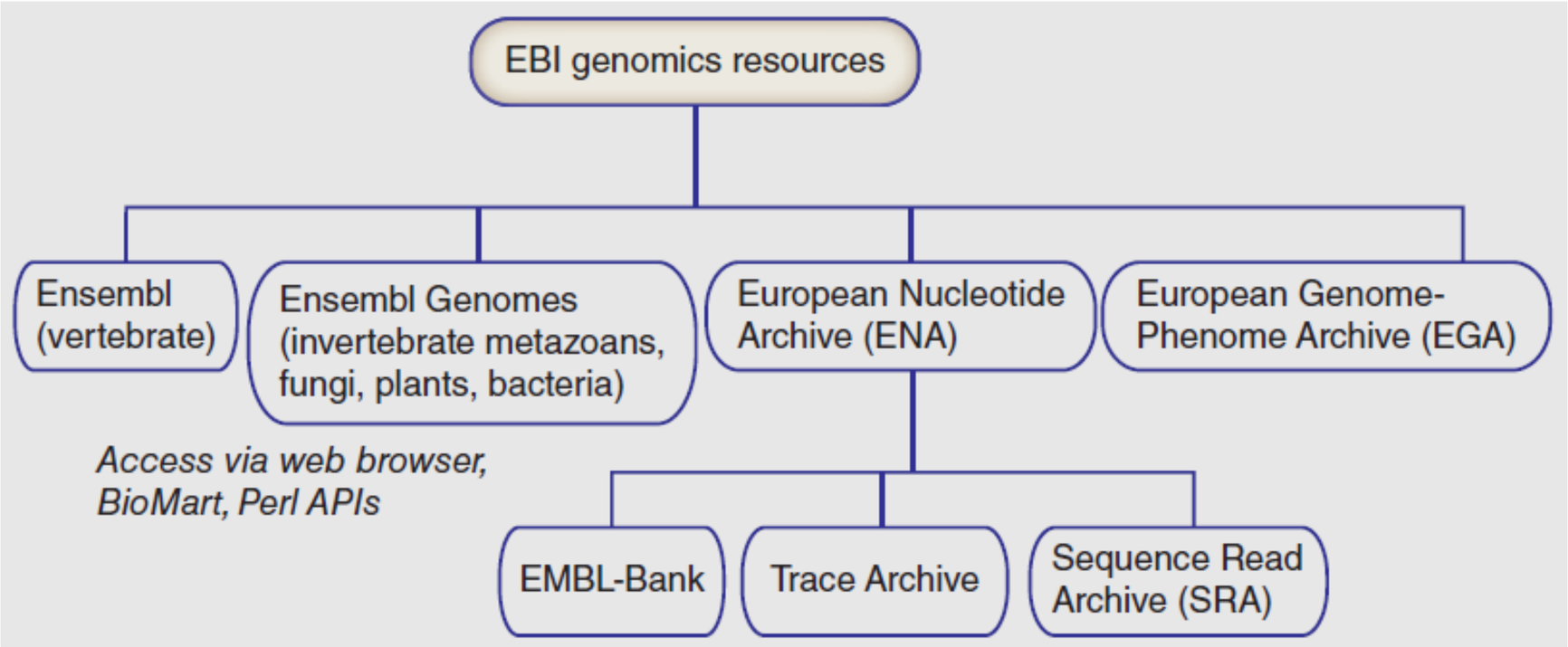
National Center for Biotechnology Information (NCBI)



B&FG3e

ian.simpson@ed.ac.uk

European Bioinformatics Institute (EBI)



B&FG3e

ian.simpson@ed.ac.uk

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.

ian.simpson@ed.ac.uk

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

ian.simpson@ed.ac.uk

RefSeq

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

ian.simpson@ed.ac.uk

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

ian.simpson@ed.ac.uk

Find a gene

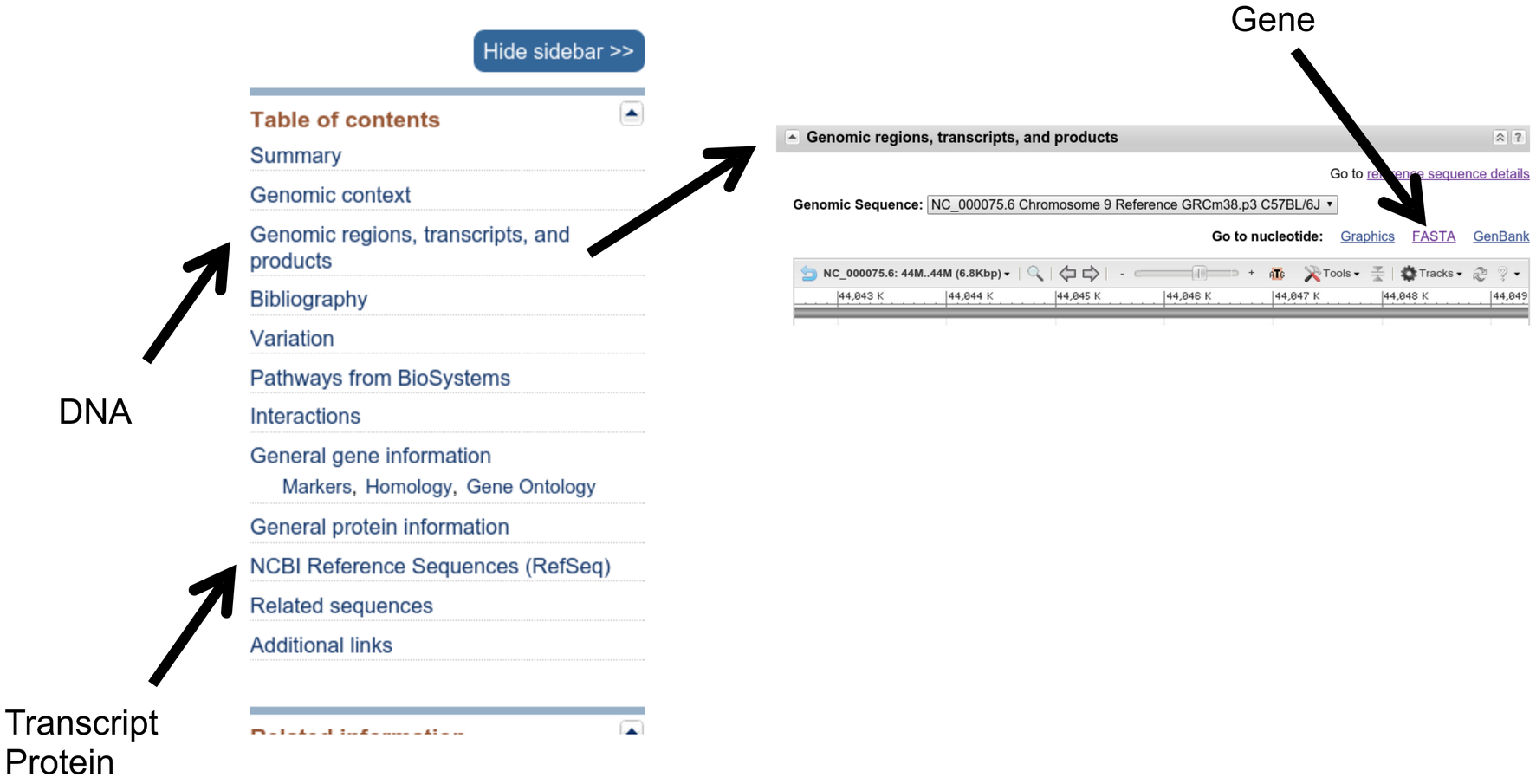
Go to: http://www.ncbi.nlm.nih.gov/

Select 'Gene'

Enter search term

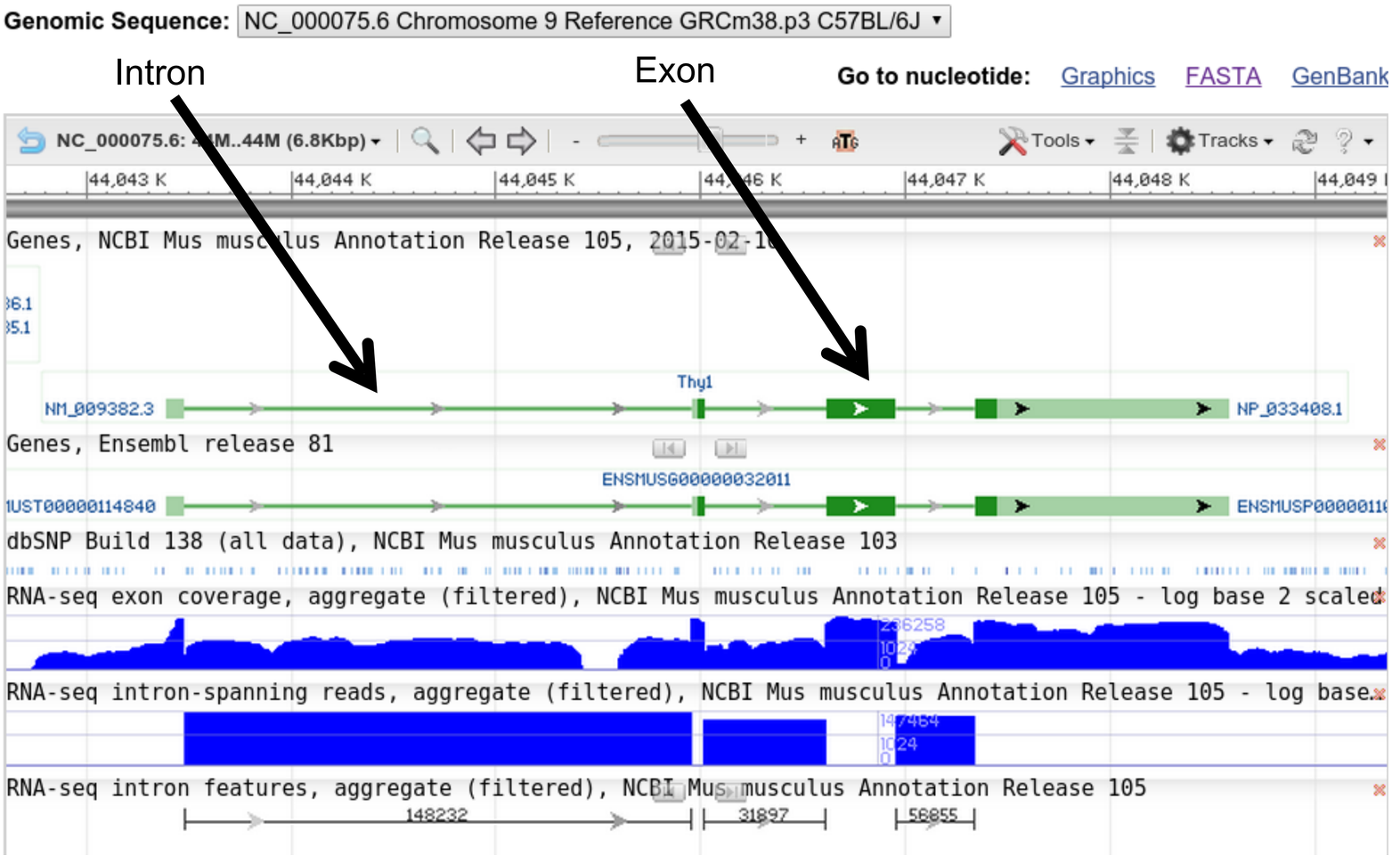
ian.simpson@ed.ac.uk

Obtain the genomic sequence



ian.simpson@ed.ac.uk

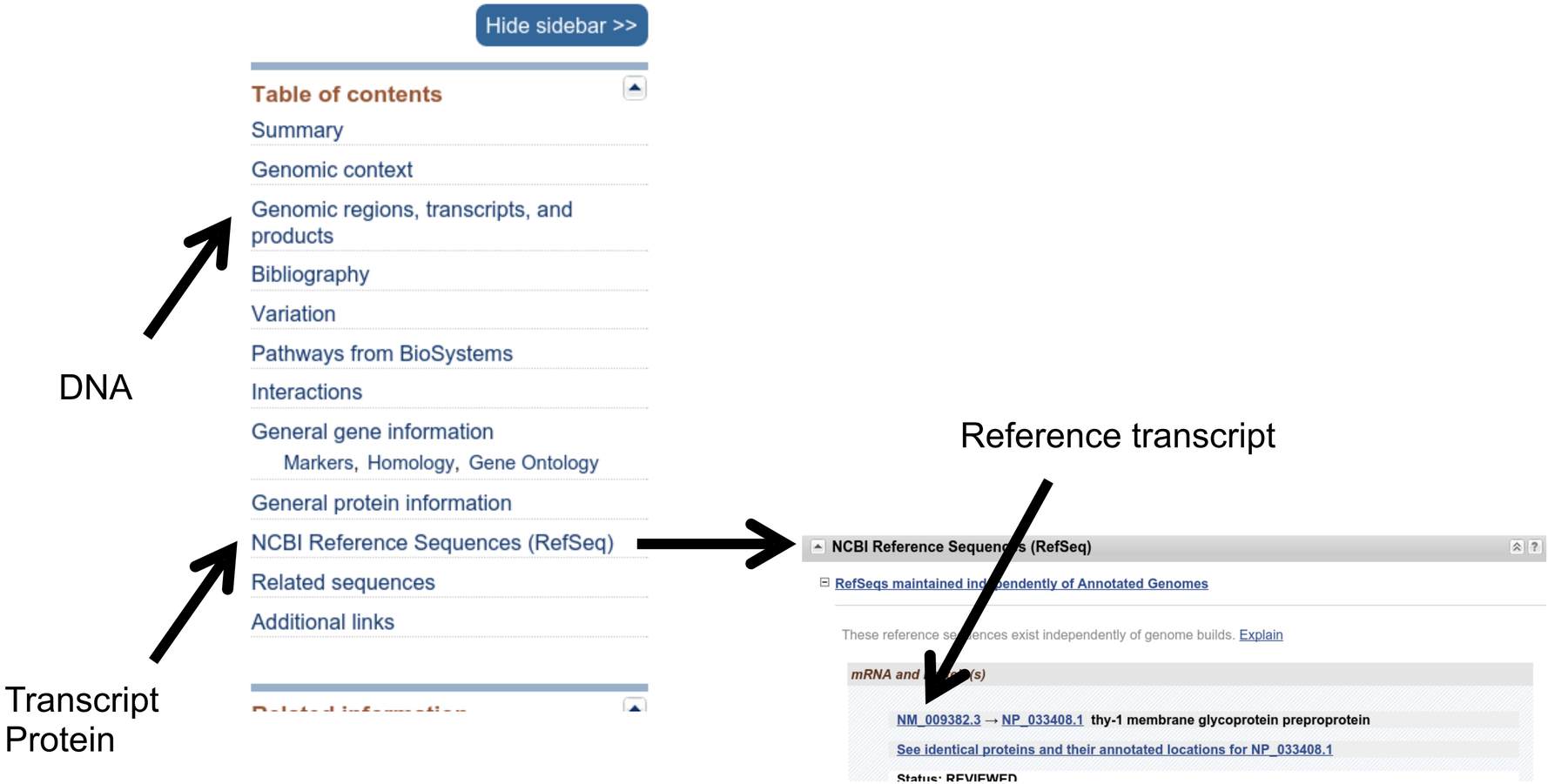
Genomic regions



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

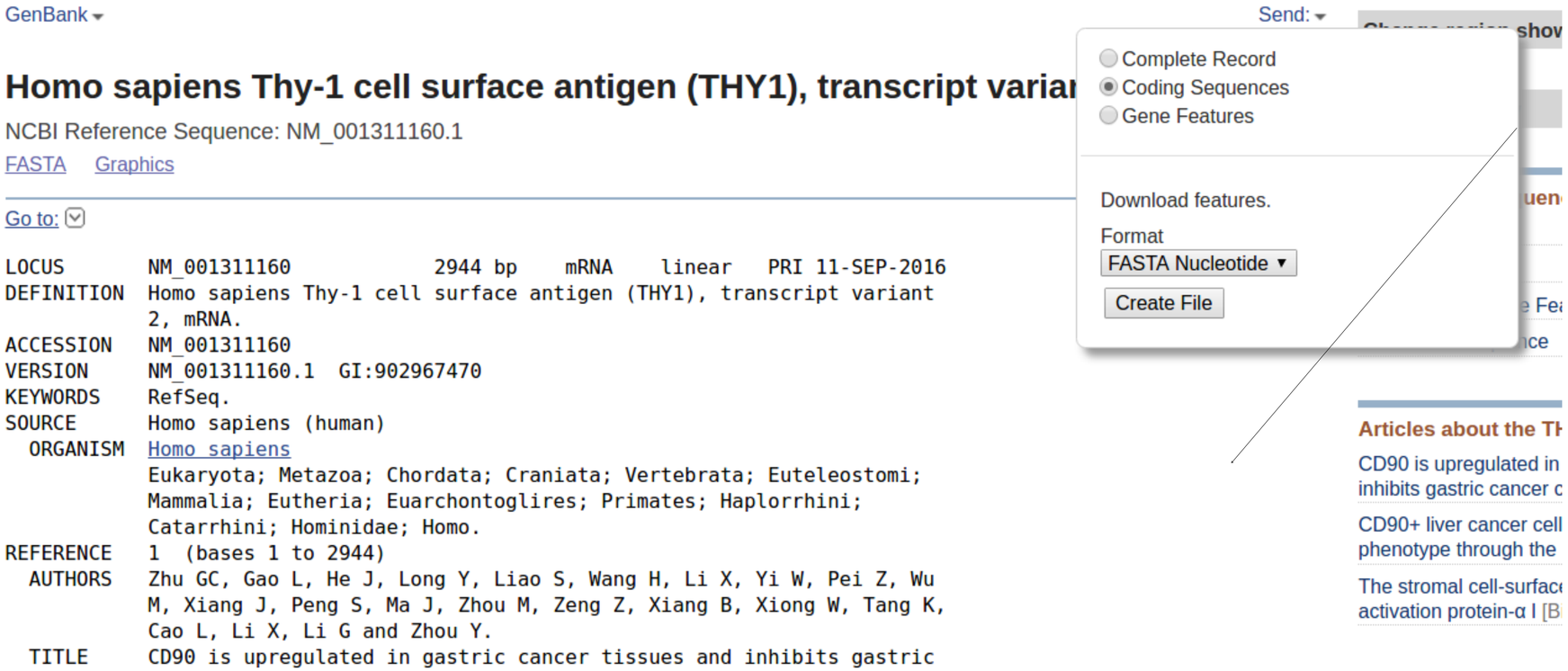
ian.simpson@ed.ac.uk

Obtain the transcript



ian.simpson@ed.ac.uk

Gene Sequences



ian.simpson@ed.ac.uk