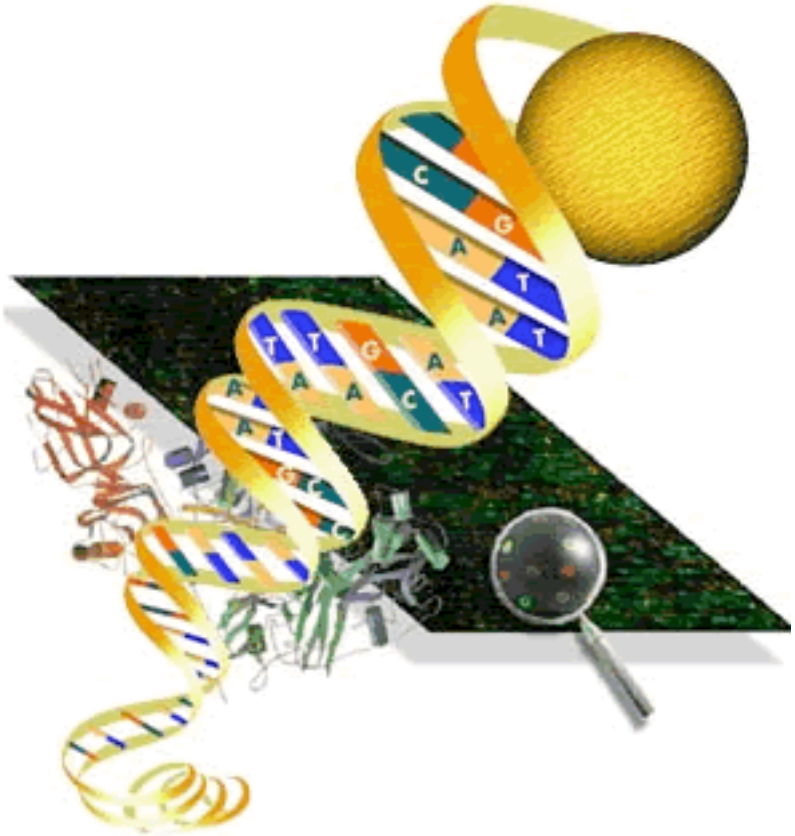


http://www.plantgdb.org/tutorial/annotatemodule/

Making Sense of Genomes

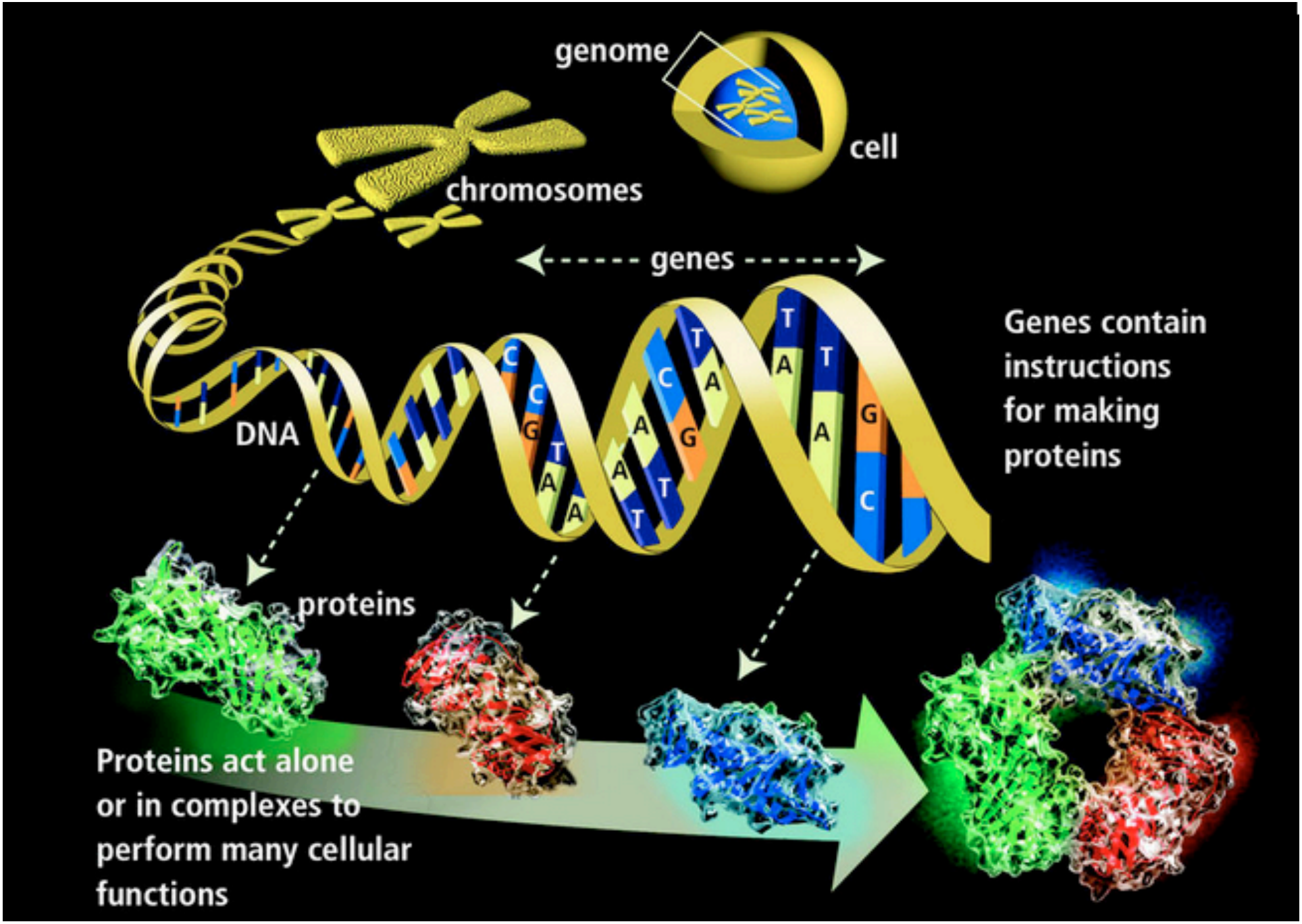
Jane Loveland

Open Door Workshop
11th May 2015
Hinxton

Overview

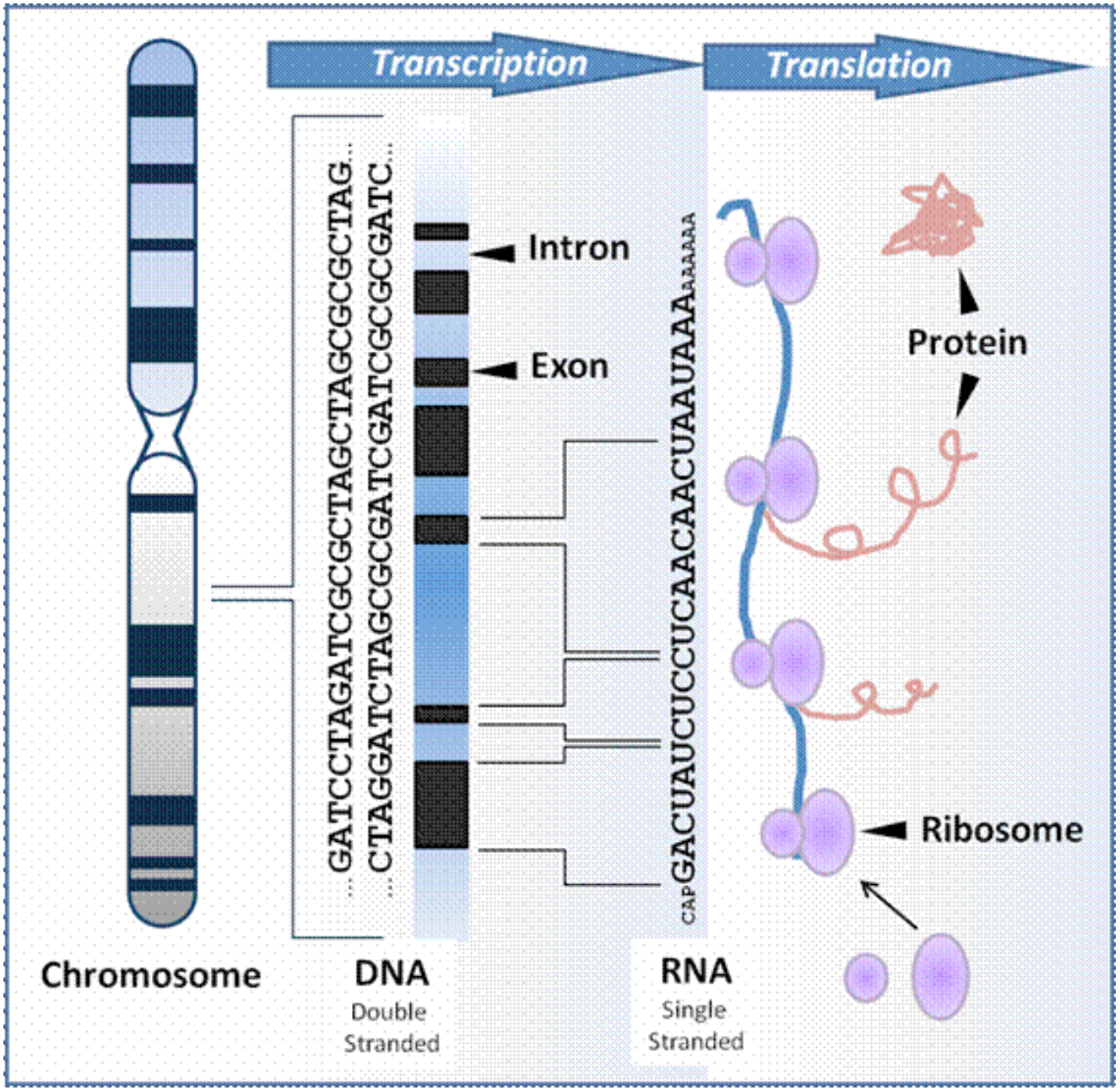
- A bit of background
- Genomes
- Genes
- Some bioinformatics basics

- A bit of background
- Genomes
- Genes
- Some bioinformatics basics



http://chslearns.vail.k12.az.us/pluginfile.php/1738/course/section/616/central%20dogma.png

Central Dogma Of Molecular Biology



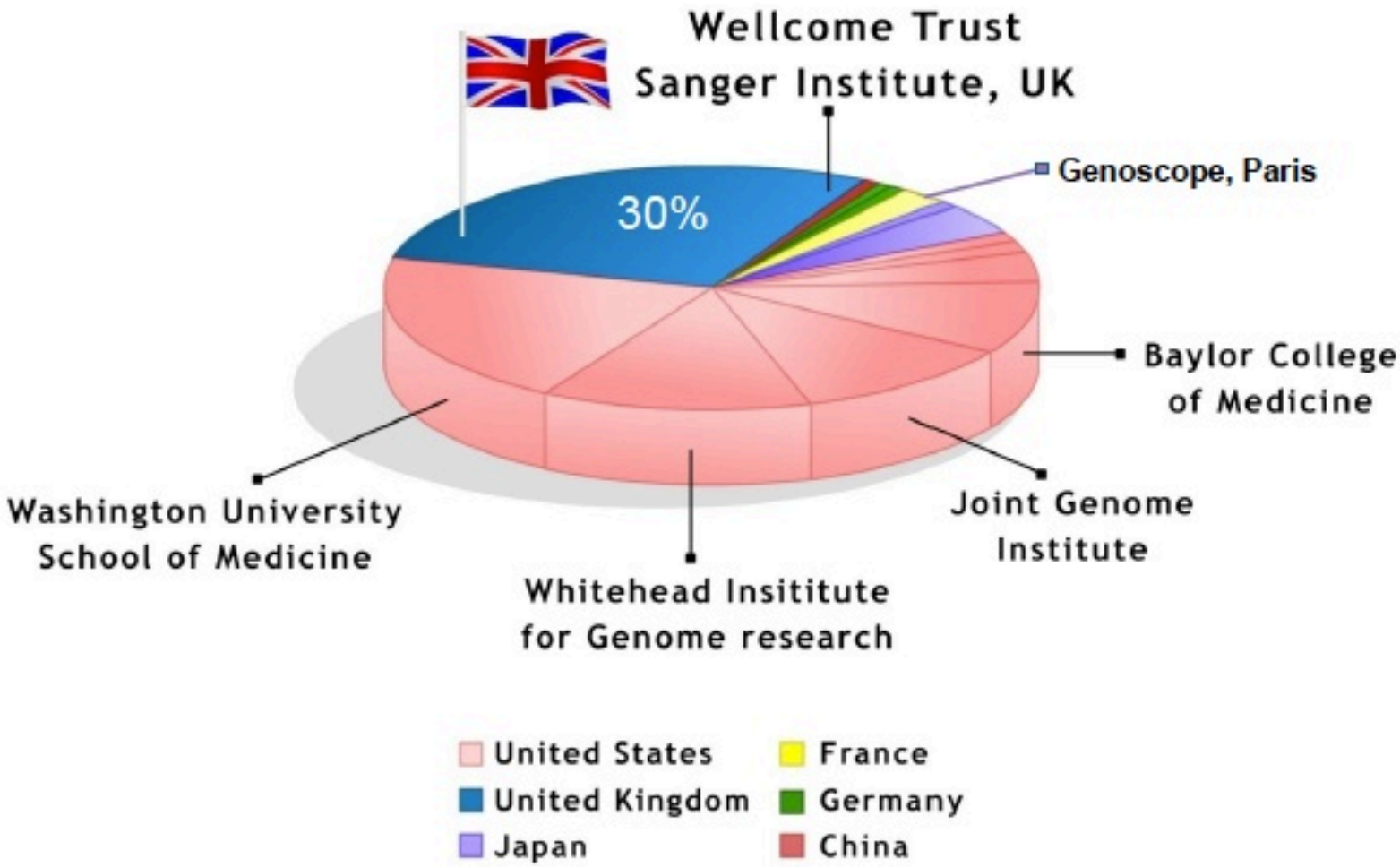
http://www4.ncsu.edu/~bjreadin/

- A bit of background
- Genomes
- Genes
- Some bioinformatics basics

The Human Genome Project

- Possibility first raised in the mid 1980s
- 1990 a plan was created to begin work over 15 years
- Automatic release of draft data (1996 Bermuda Statement)
- Immediate submission of finished sequence

Contributors to finished human sequence



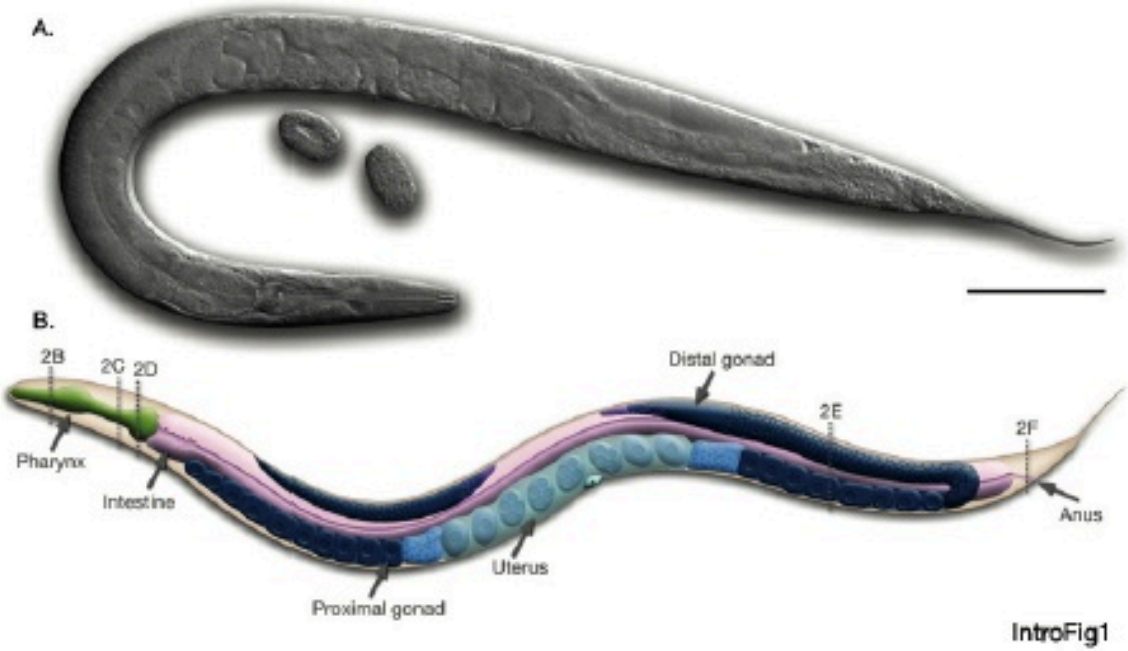
C. elegans

Number of cells – 959

100Mb genome

5 pairs of chromosomes

XX/XO sex chromosomes



~20,500 genes
Published 1998

The race to the finish

- 1998 Celera Genomics began private sector sequencing, led by J. Craig Venter using WGS

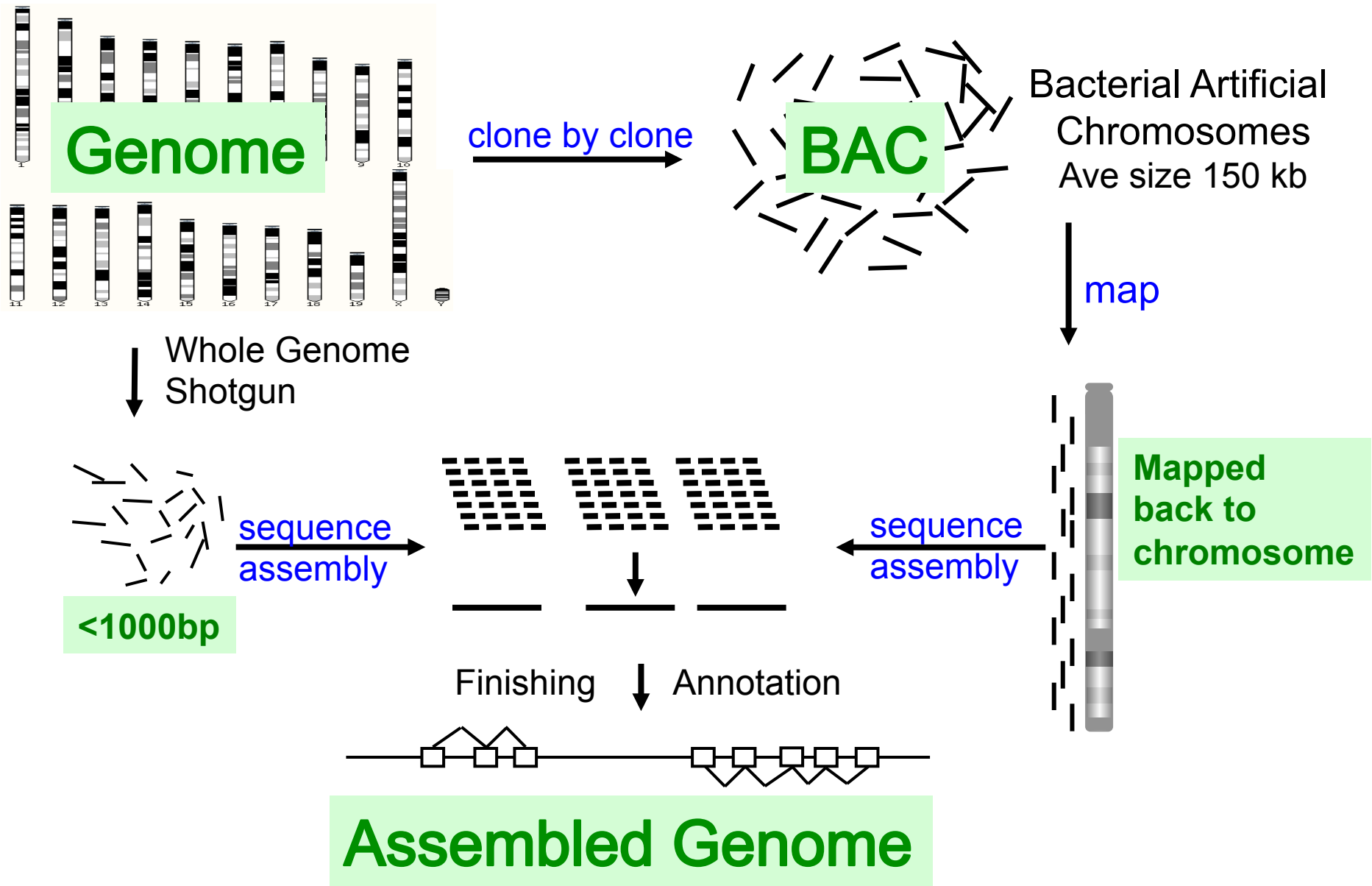
- 2000 President Clinton announced that the genome could not be patented and should be freely available to everyone

- The human genome working draft announced in 2000

- Publicly funded sequence published in 2001 in Nature
- Celera data was published in 2001 in Science

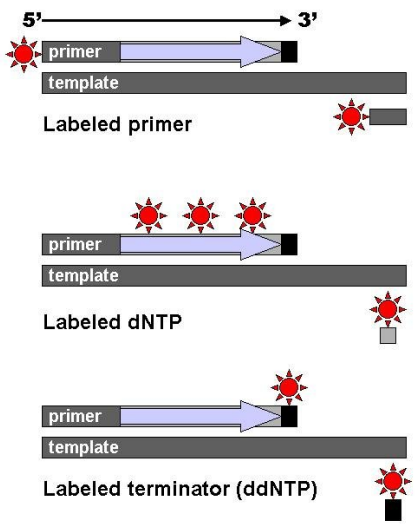
- Essentially complete genome in 2003
- 2006 the last chromosome published (chr 1)

Hybrid Sequencing Strategy



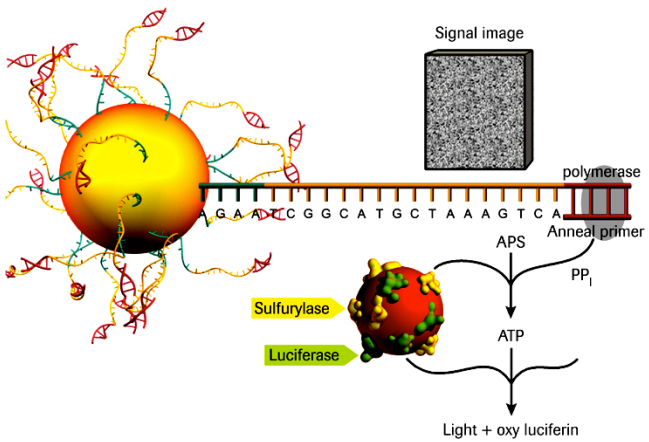
Sequencing chemistry changed in next gen sequencing technologies

Sanger sequencing



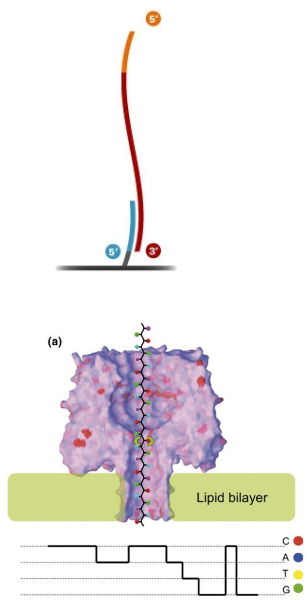
ABI 3730

Sequencing-by-synthesis



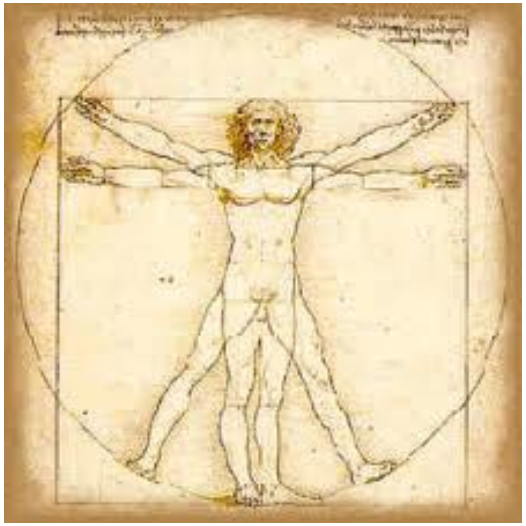
454, Illumina GAIIx and HiSeq2000

Single molecule sequencing



PacBio, Nanopore

Reference genomes:



Human ~3Gb:
22 chromosomes + sex chromosomes
GRCh38

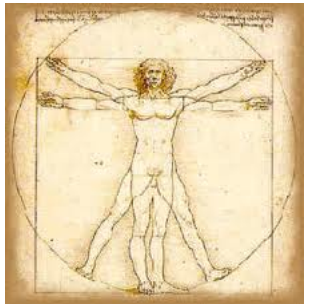
Mouse ~3 Gb:
19 chromosomes + sex chromosomes
GRCm38

Zebrafish ~1.4 Gb:
25 chromosomes, no specific sex chromosomes
Zv10
(GRCz10 coming soon)

ENCODE

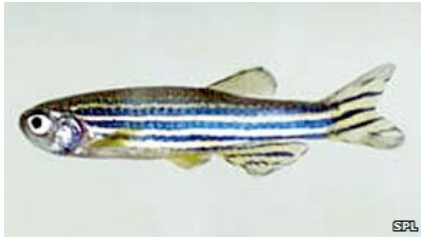
Do we know how many genes there are?

Protein coding genes



1980's	100, 000
2000	40, 000
Today	~ 20, 000

~22,000

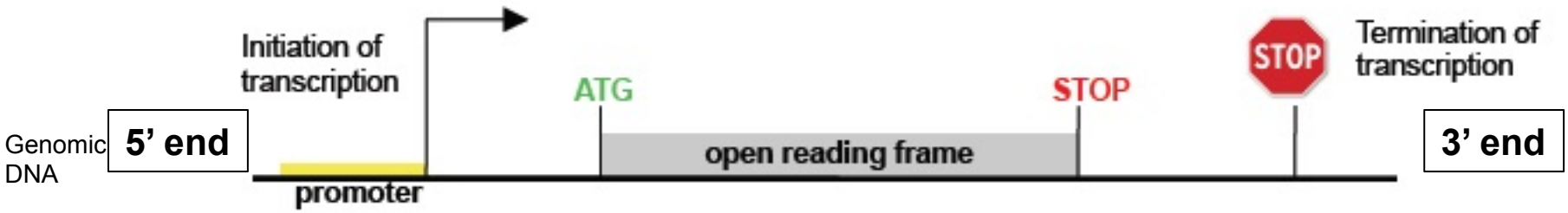


~26,000

- A bit of background
- Genomes
- Genes
- Some bioinformatics basics

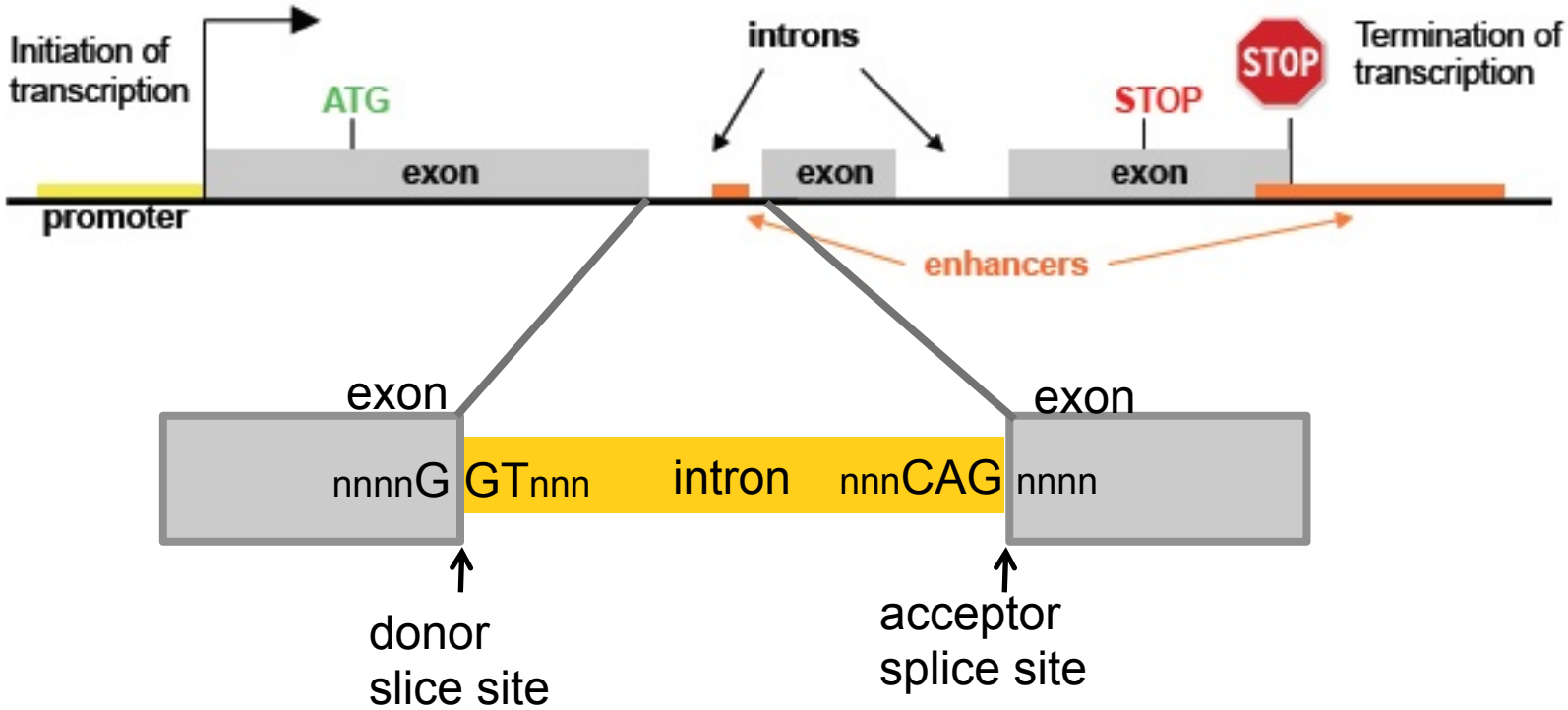
Prokaryotes:

Simple protein-coding gene



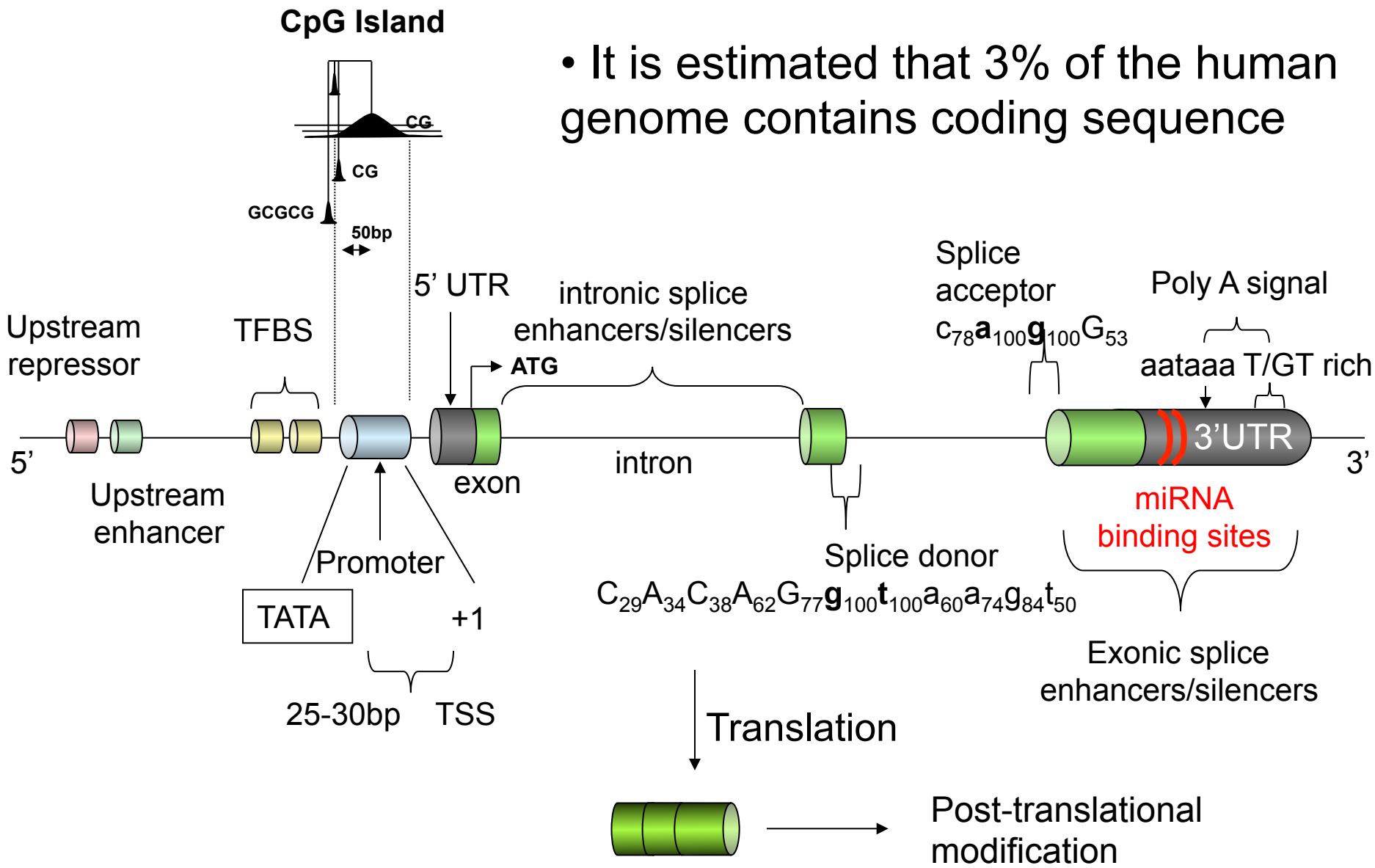
Eukaryotes:

More complex: Introns and Exons



What is in a gene?

- It is estimated that 3% of the human genome contains coding sequence



Evidence for genes: DNA (*in vitro*)

5' end

mRNA extracted from tissue sample

3' end

mRNA AUGCCTG~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~AATAAA~~~~~~~~~AAAAAAAA

Reverse transcriptase transcribes the mRNA into a complimentary strand of DNA (cDNA)

cDNA TACGGAC~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~TTATTT~~~~~~~~~TTTTTTTTT

ATGCCTG~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~AATAAA~~~~~~~~AAAAAAAA

mRNA degraded and a second strand of cDNA made

EST ATGCCTG~~~~~~~~~~~~~

~TTATTT~~~~~~~~~TTTTTTTTT

5' –> 3'

3' <– 5'

Expressed Sequence Tags (ESTs)
shorter stretches of cDNA (`800bp) sequences from both ends

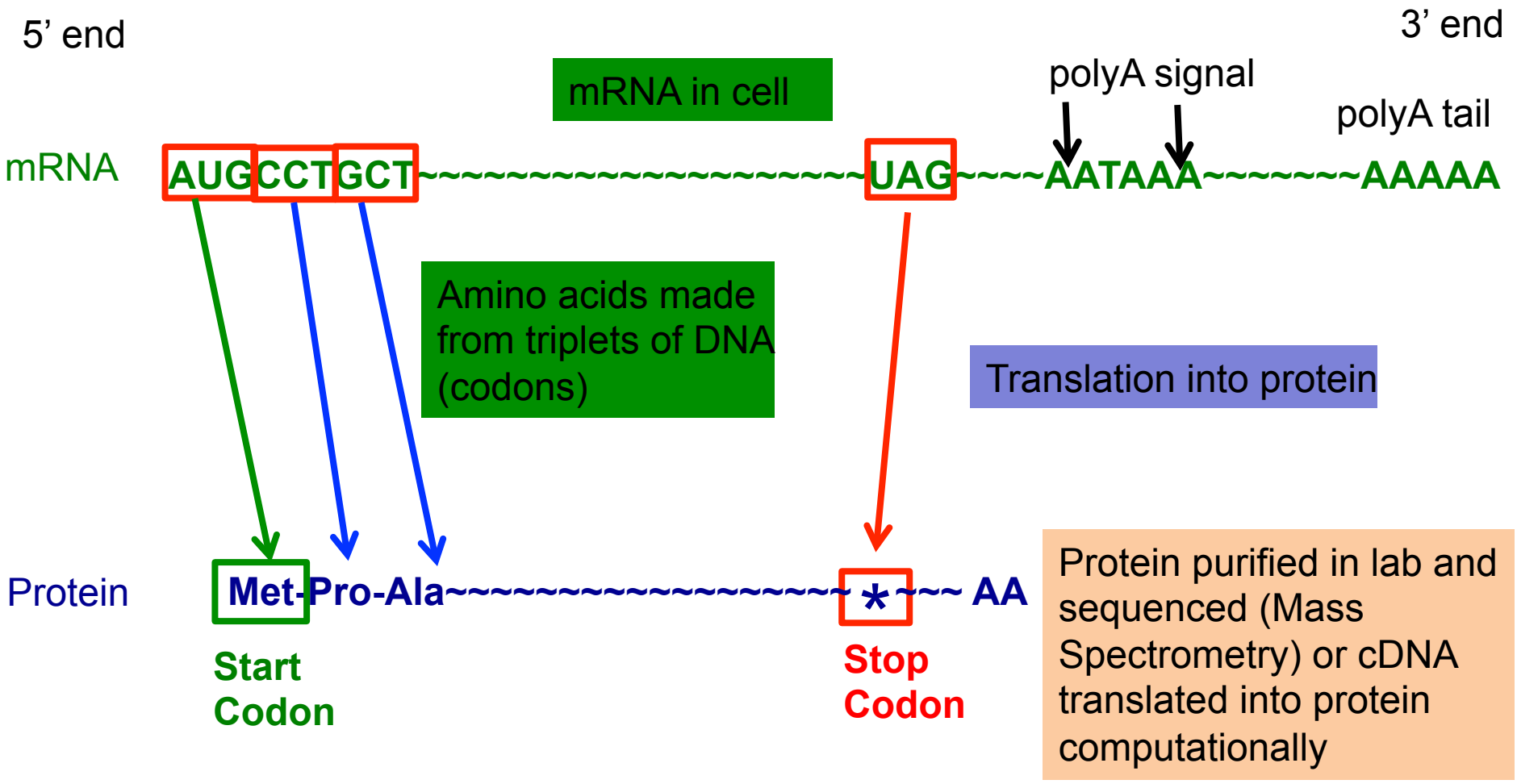
Sequences submitted to databases

ENA European Nucleotide Archive

NCBI GenBank

DDBJ DNA Data Bank of Japan

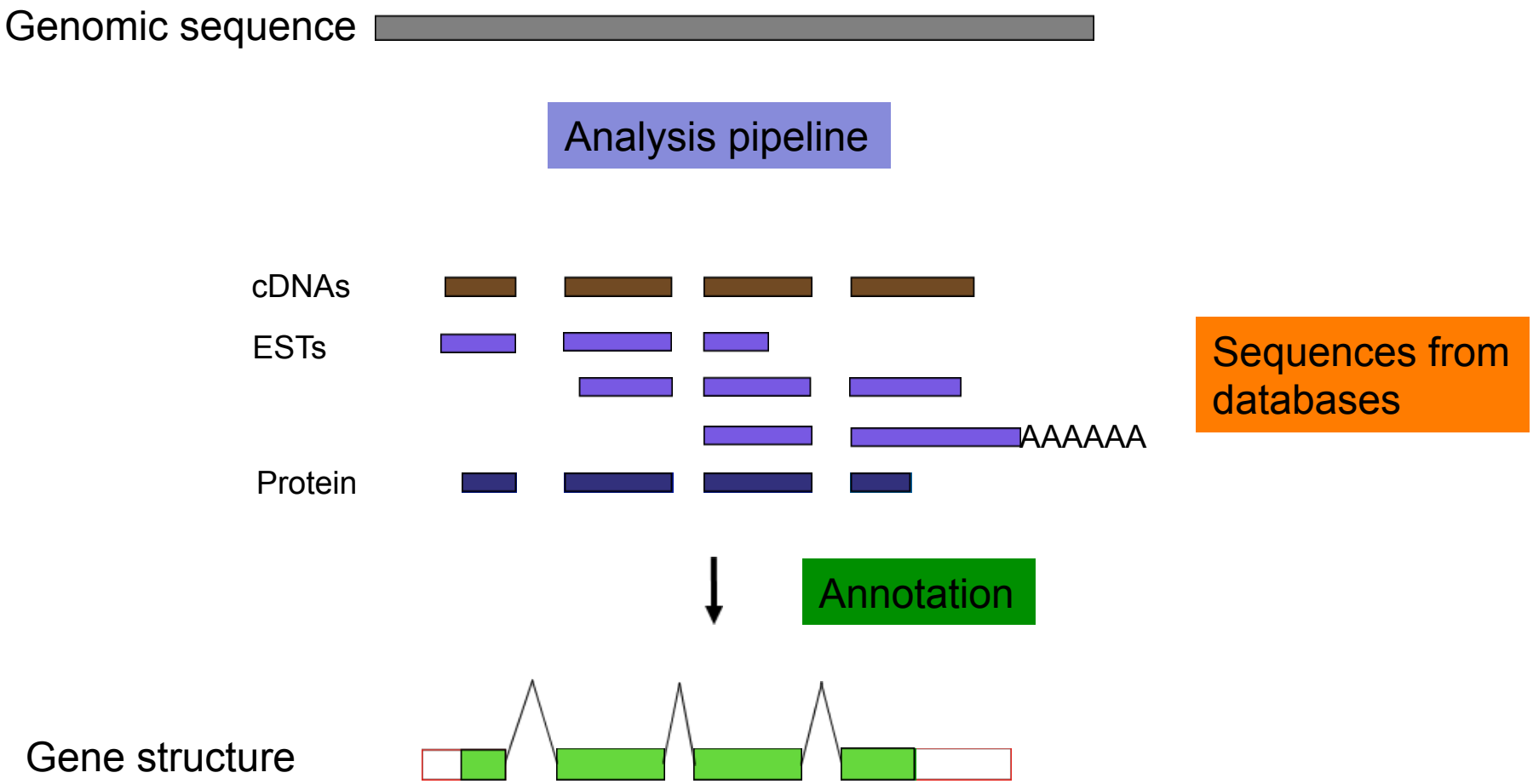
Evidence for genes: Protein



Sequences submitted to database

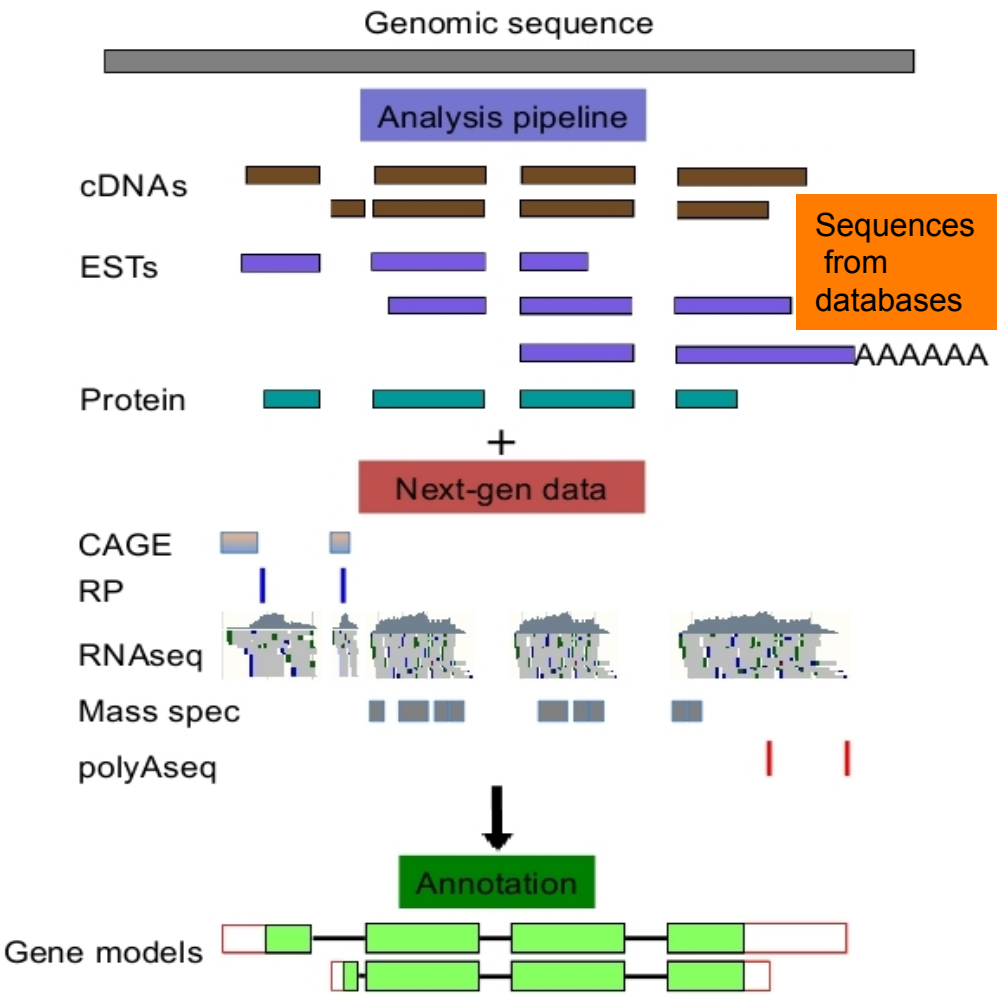
UniProt

Making the transcript from evidence:



Manual Annotation and Biotypes:

Annotation:
based on transcriptional evidence



Biotypes

Protein Coding
- Known_CDS
- Novel_CDS
- Putative_CDS
- Nonsense_mediated_decay

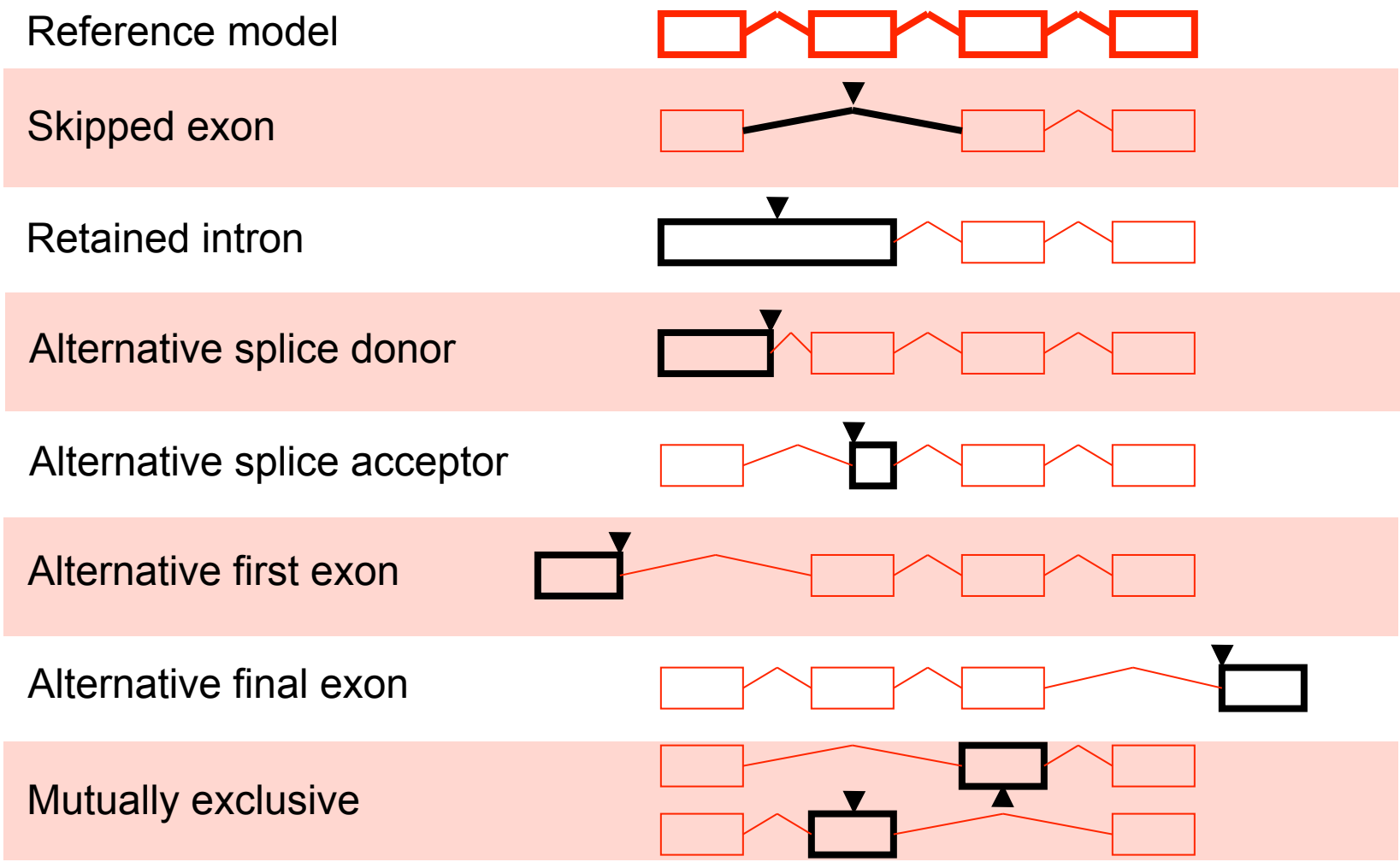
Transcript
- retained intron
- putative

Non-coding
- lincRNA
- Antisense
- Sense_intronic
- Sense_overlapping
- 3'_overlapping_ncRNA

Pseudogene
- Processed
- Unprocessed
- Transcribed
- Translated
- Unitary
- Polymorphic

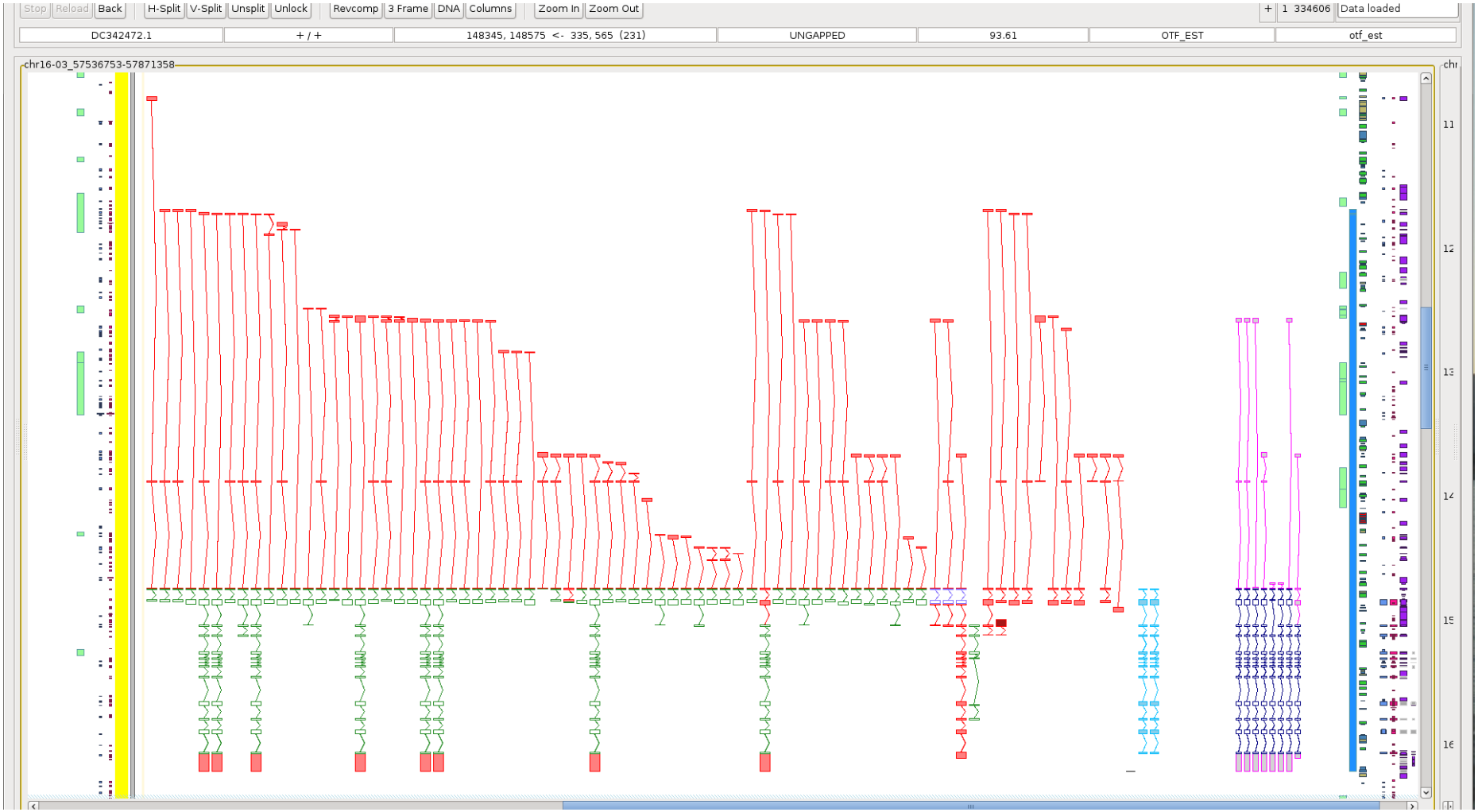
Immunoglobulin
- IG_pseudogene
- IG_Gene
- TR_Gene

Alternative Splicing

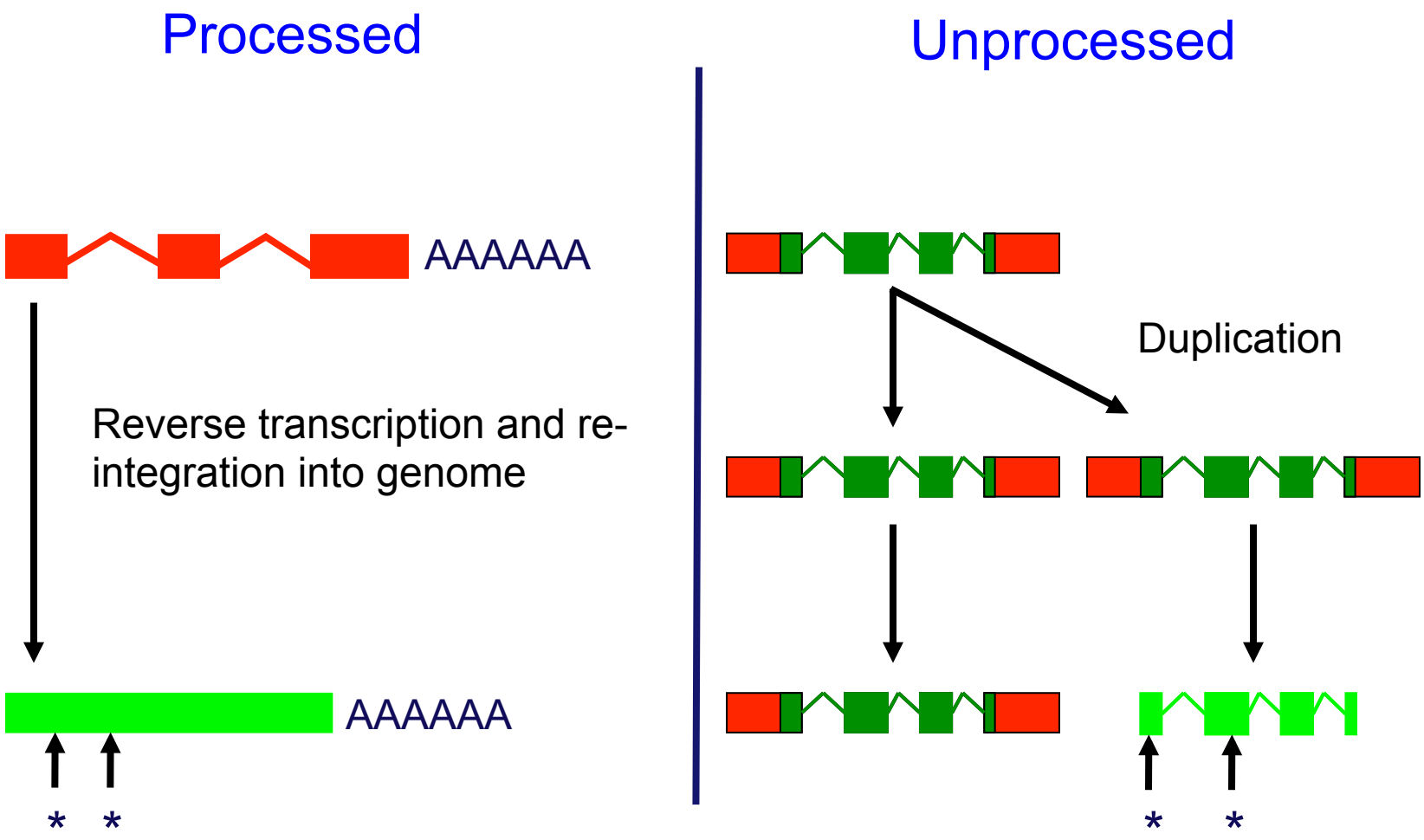


GPR56:

Human G protein-coupled receptor 56 gene



Havana pseudogenes



Disruption to coding sequence and stop codons

Non-coding genes:

microRNAs:

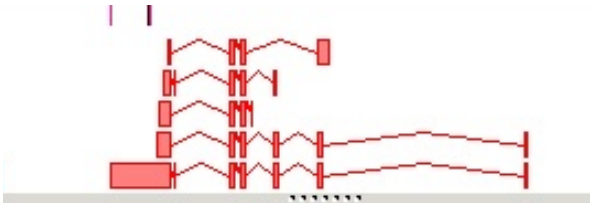
Under 200 residues
Highly conserved signature
Imported from Rfam database

Family: *mir-30* (RF00131)
Description: *mir-30 microRNA precursor*

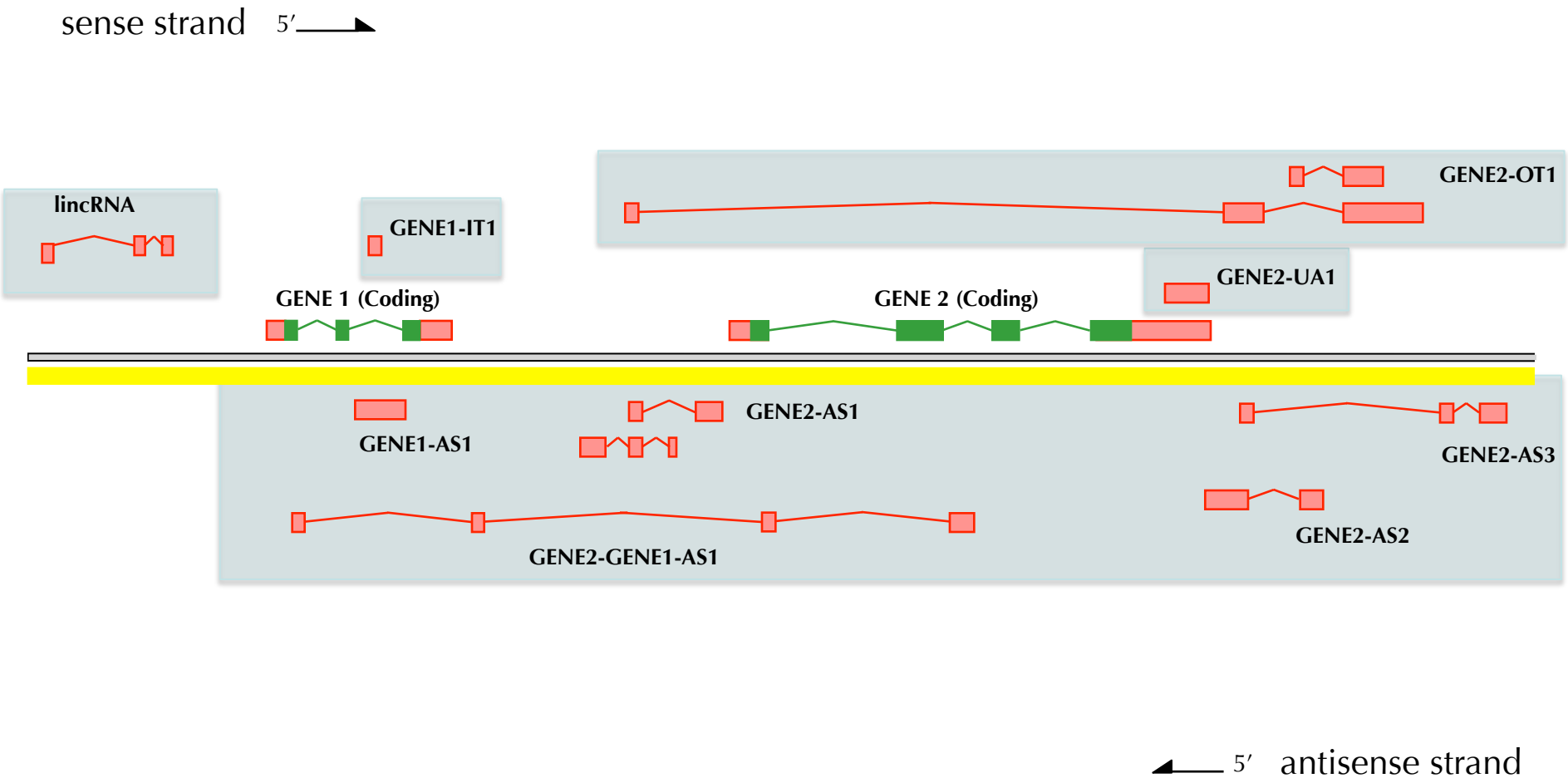
```
Mus musculus (house mouse)          UGUAAACAUCCCCGACUGGAAGCUGUAAGCCAC....AGCCAAGCUUUCAGUCAGAUGUUUGCU
Spermophilus tridecemlineatus (thUGUAAACAUCCCCGACUGGAAGCUGUAGGACAC....AGCUGAGCUUUCAGUCAGAUGUUUGCU
Macaca mulatta (Rhesus monkey)      UGUAAACAUCCUA..CACUCAGCUGUAAUACAU....GGAUUGGCUGGGAGGUGGAUGUUUACU
Macaca nemestrina (pig-tailed macUGUAAACAUCCUACACUCUCAGCUGUGGAAAGU....AAGAAAGCUGGGAGAAGGCUGUUUACU
Macaca nemestrina (pig-tailed macUGUAAACAUCCUA..CACUCAGCUGUAAUACAU....GGAUUGGCUGGGAGGUGGAUGUUUACU
Gorilla gorilla (western gorilla)UGUAAACAUCCUA..CACUCAGCUGUAAUACAU....GGAUUGGCUGGGAGGUGGAUGUUUACU
Homo sapiens (human)                UGUAAACAUCCCCGACUGGAAGCUGUAAGACAC....AGCUAAGCUUUCAGUCAGAUGUUUGCU
```

Long non-coding RNAs:
(lncRNAs)

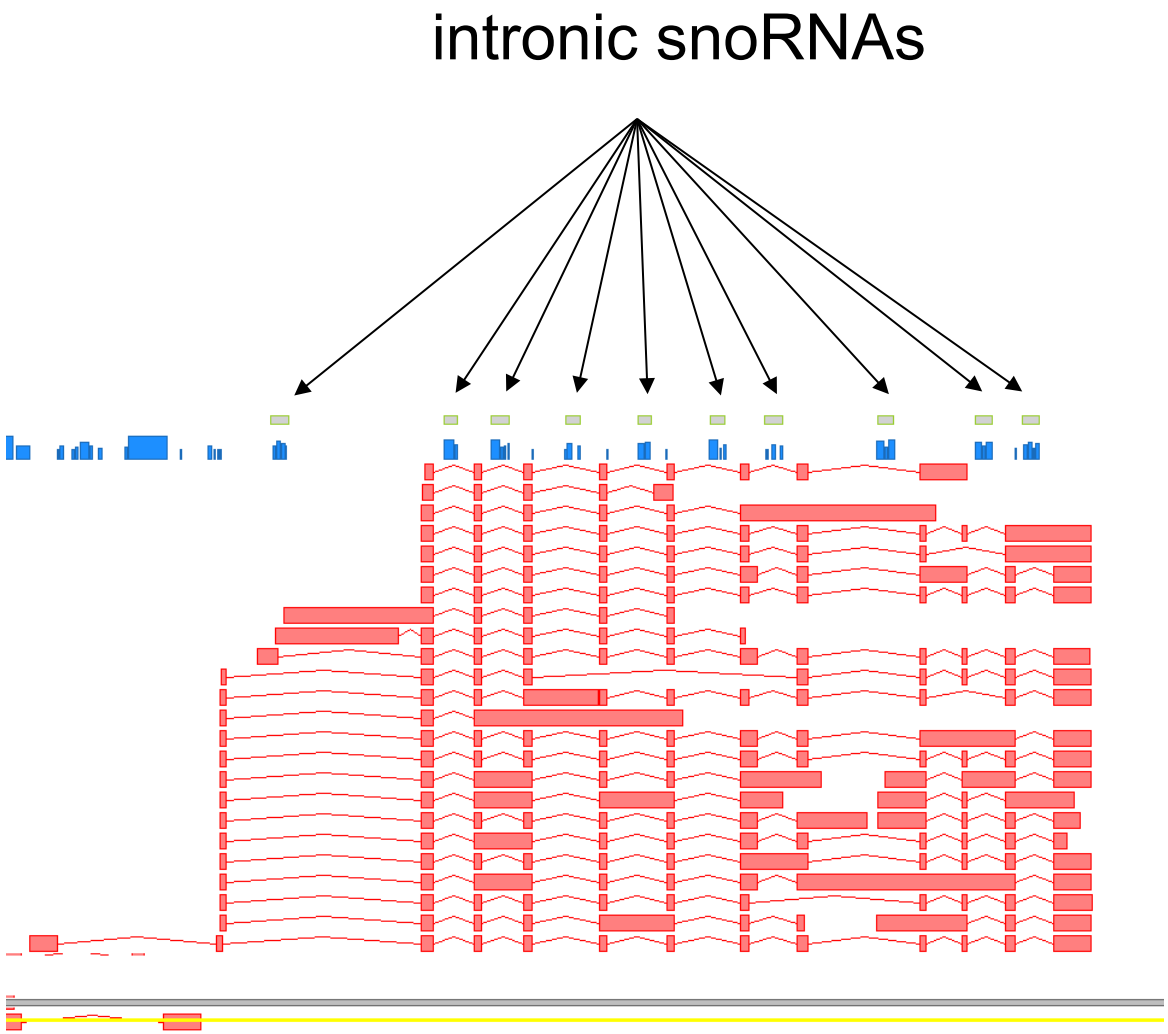
Over ~ 200 residues
Not highly conserved between species
Manually annotated
Some very well-known e.g. Hotair (HOX antisense intergenic RNA)
Many others not yet characterised



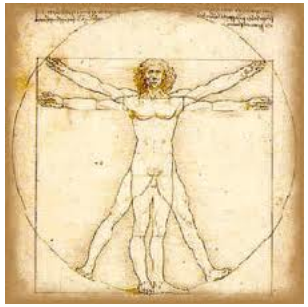
lncRNAs

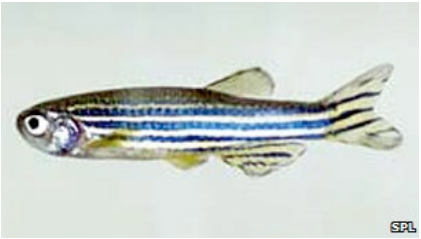

sense strand 5′
lincRNA
GENE1-IT1
GENE2-OT1
GENE2-UA1
GENE 1 (Coding)
GENE 2 (Coding)
GENE1-AS1
GENE2-AS1
GENE2-AS3
GENE2-AS2
GENE2-GENE1-AS1
5′ antisense strand

GAS5: growth arrest-specific 5 (non-protein coding)



Do we know how many genes there are?





Coding	Non-coding	Pseudogene
20,300	24,885	14,424
22,606	11,662	8,015
26,459	7,014	264

Who provides gene sets?

http://www.ncbi.nlm.nih.gov/RefSeq/

http://genome.ucsc.edu/

http://www.ensembl.org

http://vega.sanger.ac.uk

http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi (CDS only)

Differences in the gene sets

Automatic annotation

- Fast
- Unfinished sequence or shotgun sequence
- Consistent
- Under/Over-prediction
- Limited functional annotation
- Predicts ~75% loci

Manual annotation

- Slow
- Prefer finished sequence
- Flexible – can deal with inconsistencies
- Consult publications
- Extensive biotypes
- Excellent functional annotation

Automated annotation alone is not sufficient for researchers needs

GENCODE gene set

Ensembl NCBI UCSC | Vega

NCBI - RefSeq

- Non-redundant gene set
- Accessed via browsers or Entrez Gene
- Accessions for genomic DNA, transcripts and proteins
- Primarily protein-coding
- Semi-curated

	Automated	Curated
Genomic	NC_12345	
mRNA	XM_12345	NM_12345
ncRNA	XR_12345	NR_12345
Protein	XP_12345	NP_12345

UCSC gene set

- Non-redundant gene set
- Automatic annotation based on BLAT alignments
- Transcripts require Genbank accession plus one other supporting feature (eg. Uniprot)
- Includes RefSeq models (require no additional support)
- Both protein-coding and non-coding
- Data hub for ENCODE data, displays GENCODE geneset (for human and mouse)

http://genome.ucsc.edu/

Ensembl gene set

- Multiple biotypes (Known, Novel, ESTgenes, Pseudogenes)
- Automatic annotation based on pair-wise alignment using exonerate
- Transcripts require supporting mRNA and protein evidence
- Both protein-coding and non-coding transcripts
- Includes merged data from VEGA and CCDS and is called the GENCODE geneset

VEGA gene set

- Manually annotated using Otterlace/ZMAP annotation software
- Based on direct pairwise alignment of mRNA, EST and protein evidence (including cross-species)
- Multiple biotypes, reflect confidence levels
- Includes additional data sources as DAS tracks (eg. CAGE tags, RNAseq)

- Consensus CoDing Sequence project

The Consensus CDS (CCDS) project is a collaborative effort to identify a core set of human and mouse protein coding regions that are consistently annotated and of high quality. The long term goal is to support convergence towards a standard set of gene annotations.

- Havana, Ensembl, RefSeq, HGNC and MGI

- Produce reference CDS: set ATG-STOP on human and mouse genome - must agree. No UTRs.

CCDS website: GATA3 gene ATG->STOP

NCBI

Consensus CDS protein set

CCDS Database

EBI • NCBI • UCSC • WTSI

PubMed | Entrez | Gene | BLAST | OMIM

Search CCDS ID for CCDS15674.1 in All Organisms and Current Builds Go Clear

CCDS
- Home
- FTP
- Process
- Statistics
- AUG-guidelines

Collaborators
- EBI
- NCBI
- UCSC
- WTSI

Contact Us
- GenComp eMail

Genome Displays
- E Ensembl
- U Genome Browser
- N Map Viewer
- V VEGA

Related Resources
- Entrez Gene
- HomoloGene
- RefSeq
- UniGene

Report for CCDS ID CCDS15674.1

CCDS	Status	Species	Chrom.	Gene	NCBI Builds	Links
15674.1	Public	*Mus musculus*	2	Gata3	36.1 - 37.1	H G G

Sequence IDs included in CCDS 15674.1

Original	Current	Source	Nucleotide ID	Protein ID	Status in CCDS	Seq. Status	Links
✔	✔	EBI,WTSI	ENSMUST00000102976	ENSMUSP00000100041	Accepted	alive	N P N P
✔	✔	EBI,WTSI	OTTMUST00000026063	OTTMUSP00000011932	Accepted	alive	N P N P
✔		NCBI	NM_008091.2	NP_032117.1	Updated	not alive	N P N P
	✔	NCBI	NM_008091.3	NP_032117.1	Accepted	alive	N P N P B

Chromosomal Locations for CCDS 15674.1

On '-' strand of Chromosome 2 (NC_000068.6)

Genome Browser links: N U E V

Chromosome	Start	Stop	Links
2	9779997	9780281	N N U E V
2	9784722	9784847	N N U E V
2	9790388	9790533	N N U E V
2	9796016	9796552	N N U E V
2	9798979	9799216	N N U E V

CCDS Sequence Data

Blue highlighting indicates alternate exons.

Red highlighting indicates amino acids encoded across a splice junction.

Mouse over the nucleotide or protein sequence below and click on the highlighted codon or residue to select the pair.

Nucleotide Sequence (1332 nt):

```
ATGGAGGTGACTGCGGACCAGCCGCGCTGGGTGAGCCACCATCACCCCGCGGTCCTCAACGGTCAGCACC
CAGACACGCACCACCCGGGCCTCGGCCATTCGTACATGGAAGCTCAGTATCCGCTGACGGAAGAGGTGGA
CGTACTTTTTAACATCGATGGTCAAGGCAACCACGTCCCGTCCTACTACGGAAACTCCGTCAGGGCTACG
GTGCAGAGGTATCCTCCGACCCACCACGGGAGCCAGGTATGCCGCCCGCCTCTGCTGCACGGATCTCTGC
CCTGGCTGGATGGCGGCAAAGCCCTGAGCAGCCACCACACCGCCTCGCCCTGGAACCTCAGCCCCTTCTC
CAAGACGTCCATCCACCACGGCTCTCCGGGGCCTCTGTCCGTTTACCCTCCGGCTTCATCCTCTTCTCTG
GCGGCCGGCCACTCCAGTCCTCATCTCTTCACCTTCCCGCCCACCCCGCCGAAAGACGTCTCCCCAGACC
CGTCGCTGTCCACCCCGGGATCCGCCGGGTCGGCCAGGCAAGATGAGAAAGAGTGCCTCAAGTATCAGGT
GCAGCTGCCAGATAGCATGAAGCTGGAGACGTCTCACTCTCGAGGCAGCATGACCACCCTGGGTGGGGCC
TCATCCTCAGCCCACCACCCCATTACCACCTATCCGCCCTATGTGCCCGAGTACAGCTCTGGACTCTTCC
CACCCAGCAGCCTGCTGGGAGGATCCCCTACCGGGTTCGGATGTAAGTCGAGGCCCAAGGCACGATCCAG
CACAGAAGGCAGGGAGTGTGTGAACTGCGGGGCAACCTCTACCCCACTGTGGCGGCGAGATGGTACCGGG
CACTACCTTTGCAATGCCTGCGGACTCTACCATAAAATGAATGGGCAGAACCGGCCCCTTATCAAGCCCA
AGCGAAGGCTGTCGGCAGCAAGGAGAGCAGGGACATCCTGCGCGAACTGTCAGACCACCACCACCACCCT
CTGGAGGAGGAACGCTAATGGGGACCCGGTCTGCAATGCCTGTGGGCTGTACTACAAGCTTCATAATATT
AACAGACCCCTGACTATGAAGAAAGAAGGCATCCAGACCCGAAACCGGAAGATGTCTAGCAAATCGAAAA
AGTGCAAAAAGGTGCATGACGCGCTGGAGGACTTCCCCAAGAGCAGCTCCTTCAACCCGGCCGCTCTCTC
CAGACACATGTCATCCCTGAGCCACATCTCTCCCTTCAGCCACTCCAGCCACATGCTGACCACACCGACG
CCCATGCATCCGCCCTCCGGCCTCTCCTTCGGACCTCACCACCCTTCCAGCATGGTCACCGCCATGGGTT
AG
```

Translation (443 aa):

```
MEVTADQPRWVSHHHPAVLNGQHPDTHHPGLGHSYMEAQYPLTEEVDVLFNIDGQGNHVPSYYGNSVRAT
VQRYPPTHHGSQVCRPPLLHGSLPWLDGGKALSSHHTASPWNLSPFSKTSIHHGSPGPLSVYPPASSSSL
AAGHSSPHLFTFPPTPPKDVSPDPSLSTPGSAGSARQDEKECLKYQVQLPDSMKLETSHSRGSMTTLGGA
SSSAHHPITTYPPYVPEYSSGLFPPSSLLGGSPTGFGCKSRPKARSSTEGRECVNCGATSTPLWRRDGTG
HYLCNACGLYHKMNGQNRPLIKPKRRLSAARRAGTSCANCQTTTTTLWRRNANGDPVCNACGLYYKLHNI
NRPLTMKKEGIQTRNRKMSSKSKKCKKVHDALEDFPKSSSFNPAALSRHMSSLSHISPFSHSSHMLTTPT
PMHPPSGLSFGPHHPSSMVTAMG
```

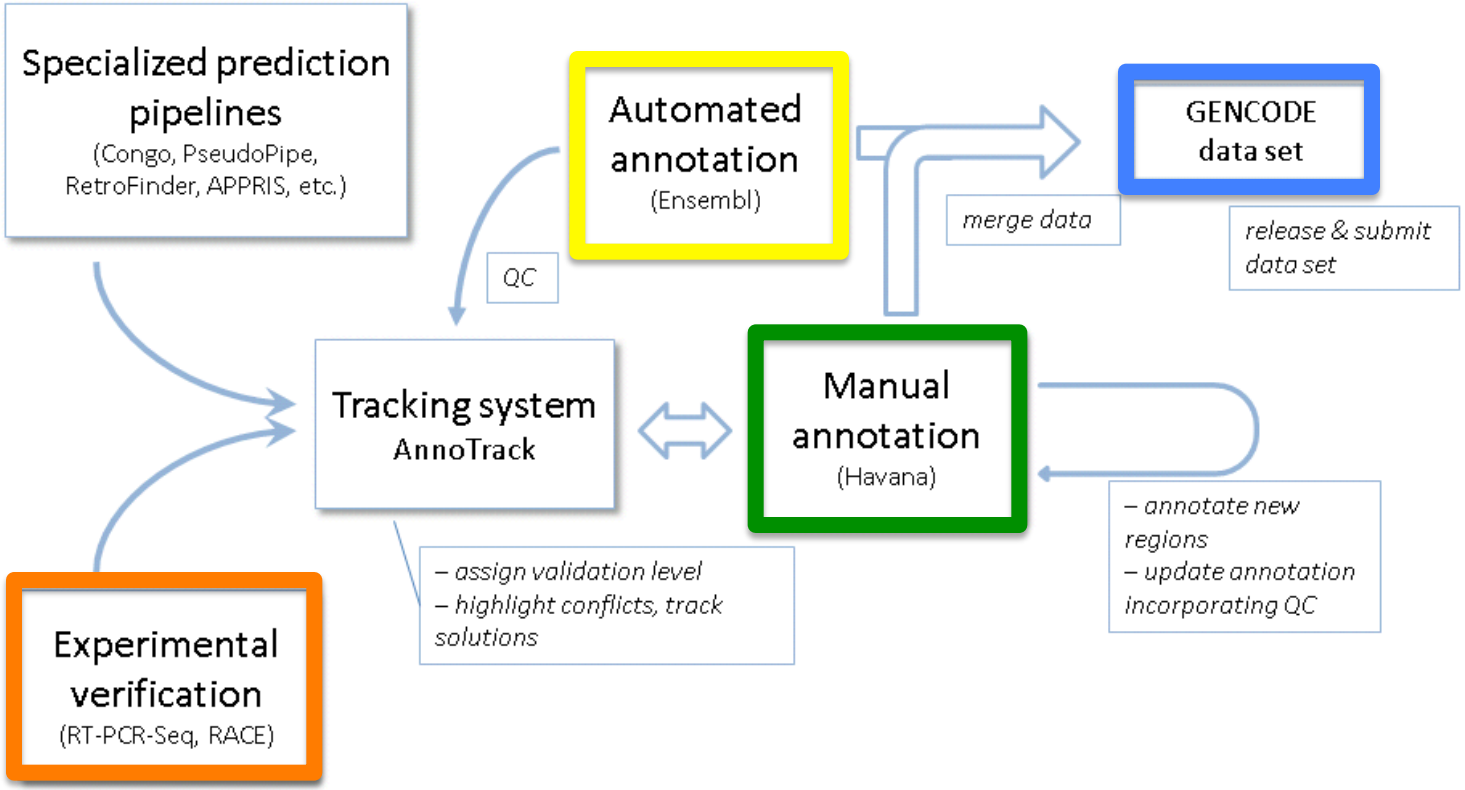
Comparing Gene Sets

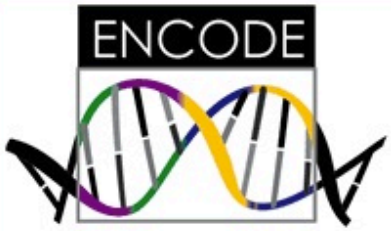
How many genes are there?

	Human	Mouse
RefSeq (2013)		
Coding genes	19,119	20,553
Total transcripts	35,539	27,113
UCSC		
Coding genes	21,520	21,181
Total transcripts	82,960	55,121
Ensembl (Gencode)		
Known coding	20,805 (e75)	23,871
Total transcripts	196,501	94,647
VEGA (unfinished)		
Coding genes	19,520 (v59)	16,359
Total transcripts	181,669	73,869
CCDS	18,800 (v106) (30,499)	20,080 (v104) (23,880)

GENCODE pipeline

Aim of GENCODE :annotate all evidence based gene features in the human genome





The GENCODE Project:

Encyclopædia of genes and gene variants

Current GENCODE version

The current version in **Human** is **Gencode 21**, released on the 2nd October 2014.

For more information about the human releases please see the **README.txt** file.

The current version in **Mouse** is **Gencode M4**, released on the 3rd December 2014.

Introduction

The National Human Genome Research Institute (NHGRI) launched a public research consortium named **ENCODE**, the Encyclopedia Of DNA Elements, in September 2003, to carry out a project to identify all functional elements in the human genome sequence. After a successful pilot phase on 1% of the genome, the scale-up to the entire genome is now underway. The Wellcome Trust Sanger Institute was **awarded a grant** to carry out a scale-up of the GENCODE project for integrated annotation of gene features.

Having been involved in successfully delivering the definitive annotation of functional elements in the human genome, the GENCODE group were **awarded a second grant** in 2013 in order to continue their human genome annotation work and expand GENCODE to include annotation of the mouse genome.

The **international team** working in the GENCODE project is headed by **Tim Hubbard** at the **Wellcome Trust Sanger Institute**, and includes members from **Centre de Regulació Genòmica**, **Spanish National Cancer Research Centre**, **The University of Lausanne**, **Massachusetts Institute of Technology**, **Yale University** and **The University of California, Santa Cruz** .

The GENCODE gene sets are used by the entire ENCODE consortium and by many other projects (eg. 1000 Genomes) as reference gene sets.

Acknowledgements

The GENCODE project is funded through an NHGRI ENCODE grant with additional funding from the Wellcome Trust.

When referencing, please use "Harrow J, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project" (**PubMed**).

Genome Reference Consortium

Goal:

- Correct regions in the genome that are currently misrepresented
- To close as many gaps as possible
- To produce alternative assemblies of structurally variant loci where necessary
- Scientific community can report loci in need of review
- Human, mouse and zebrafish

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

wellcome trust sanger institute

THE Genome CENTER AT WASHINGTON UNIVERSITY

EMBL-EBI

NCBI

gEVAL

Tools | Help & Documentation | Blog | About us

Search: Choose an assembly for Go

gEVAL: Genome Evaluation Browser

The **gEVAL Browser** allows the evaluation of genome assemblies through its tools and pre-computed analyses.

The strength of this browser is the ability to navigate an up to date assembly and identify problematic regions and assist in strategizing potential solutions for these issues.

This facilitates the improvement of overall assemblies to a "gold" standard for release as reference genomes.

Commonly viewed genomes

Zebrafish_GRCz10
CURRENT

Mouse
GRCm38.p3

Human_GRCh38
CURRENT

About the Project

gEVAL utilizes the Ensembl framework and is maintained by the *Genome Reference Informatics Team* at the Wellcome Trust Sanger Institute.

The team is part of the *Genome Reference Consortium (GRC)*, a multi-centre collaboaration tasked with providing improved reference assemblies that better represent complex diversity.

Browse a Genome

Browse Human Genome Assemblies

Browse Mouse Genome Assemblies

Browse Zebrafish Genome Assemblies

Browse Pig Assemblies

Browse Rat Assemblies

Browse Parasitic Helminth Assemblies

Browse Chicken Assemblies

- A bit of background
- Genomes
- Genes
- Some bioinformatics basics

EMBL-EBI

About us

Bioinformatics services

We maintain the world's most comprehensive range of **freely available** and up-to-date molecular databases. Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically.

DNA & RNA

genes, genomes & variation

Gene expression

RNA, protein & metabolite expression

Proteins

sequences, families & motifs

Structures

Molecular & cellular structures

Systems

reactions, interactions & pathways

Chemical biology

chemogenomics & metabolomics

Ontologies

taxonomies & controlled vocabularies

Literature

Scientific publications & patents

Other software

cross-domain tools & resources

ENA
European Nucleotide Archive

Search
Examples: BN000065, histone
Advanced
Sequence

Home | Search & Browse | Submit & Update | Software | About ENA | Support

Sequence: BN000065.1

TPA: Homo sapiens SMP1 gene, RHD gene and RHCE gene

Send Feedback

View: TEXT FASTA XML

Download: XML FASTA TEXT

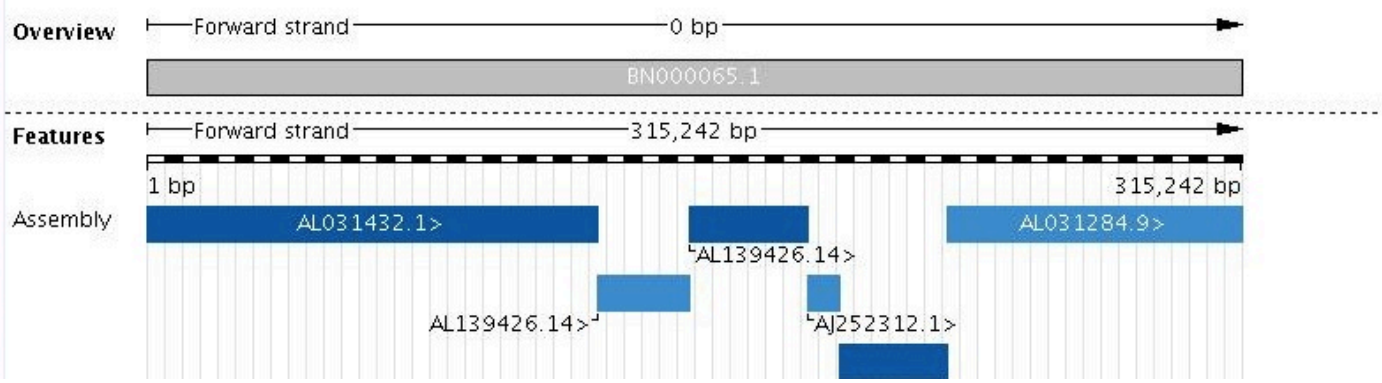
Organism	Molecule type	Topology	Data class	Taxonomic Division
Homo sapiens	genomic DNA	linear	STD	HUM
Sequence length	**Sequence Version**	**First public**	**Last updated**	**Show Version History**
315,242	1	23-APR-2002	14-NOV-2006	BN000065

Keywords
RHCE gene, RhCE protein, RHD gene, RhD protein, small membrane protein 1, SMP1 gene, Third Party Data, TPA, TPA:inferential.

Lineage
Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Euarchontoglires, Primates, Haplorrhini, Catarrhini, Hominidae, Homo

Navigation | Overview | Source Feature(s) | Sequence | Assembly | Other Feature(s) | Publications

Base range: 1 - 315242 Apply



UniProt

UniProtKB | Advanced

BLAST Align Retrieve/ID Mapping Help Contact

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB

Swiss-Prot (547,357)

Manually annotated and reviewed.

TrEMBL (89,451,166)

Automatically annotated and not reviewed.

UniRef

Sequence clusters

UniParc

Sequence archive

Proteomes

Supporting data

Literature citations	Taxonomy	Subcellular locations
Cross-ref. databases	Diseases	Keywords

News

Thalidomide, the pharmacological version of yin and yang | Cross-references to DEPOD, MoonProt and Proteomes

UniProt release 2015_01

Higher and higher | New mouse and zebrafish variation files | Structuring of 'cofactor' annotations

UniProt release 2014_11

News archive

Q9HD64 - XAGE1_HUMAN

Protein	**X antigen family member 1**
Gene	**XAGE1A** more
Organism	*Homo sapiens (Human)*
Status	Reviewed - ●●●○○ - Experimental evidence at transcript level [i]

Names & Taxonomy [i]

Protein names [i]	*Recommended name:* **X antigen family member 1** ▪ *Short name:*XAGE-1 *Alternative name(s):* • Cancer/testis antigen 12.1 ▪ *Short name:*CT12.1 • G antigen family D member 2
Gene names [i]	*Name:*XAGE1A Synonyms:GAGED2, XAGE1 AND *Name:*XAGE1B AND *Name:*XAGE1C AND *Name:*XAGE1D AND *Name:*XAGE1E
Organism [i]	Homo sapiens (Human)
Taxonomic identifier [i]	9606 [NCBI]
Taxonomic lineage [i]	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › H Hominidae › Homo »
Proteomes [i]	UP000005640: Chromosome X

NCBI Resources How To Sign in to N

NCBI National Center for Biotechnology Information All D Search

How To

- Save text searches and set up automated searches with E-mailed results
- Find bioassays in which a given drug is active
- Find bioassays that test a particular disease or protein target
- Submit data to NCBI
- Download NCBI Software
- Submit sequence data to NCBI
- Retrieve all sequences for an organism or taxon
- Find the function of a gene or gene product
- View all SNPs associated with a gene
- Find genes associated with a phenotype or disease
- Find expression patterns
- Obtain genomic sequence for/near a gene, marker, transcript or protein
- Find human variations associated with a phenotype or disease (clinical association)
- Convert feature coordinates between genomic assemblies
- View/download features around an object or between two objects on a chromosome
- Compare protein homologs between two microbial genomes
- Find sequenced genomes, including those in progress, for a taxonomic group
- Download the complete genome for an organism
- Display genomic annotation graphically
- Determine conserved synteny between the genomes of two organisms
- Find a homolog for a gene in another organism
- Find articles about a topic similar to that in a given article
- Obtain the full text of an article
- View the 3D structure of a protein
- Find a curated version of a sequence record (NCBI Reference Sequence)
- Align two or more 3D structures to a given structure
- Find published information on a gene or sequence
- Find transcript sequences for a gene
- Link from an object on a map to another resource
- Run BLAST software on a local computer

Submissions

BioProject Submission
An online form that provide
genomic and genetic data

ClinVar Submissions
Guidelines and instructions
level/aggregate data); sup

Database of Genotype and
Guidelines and requiremen

Database of Major Histoco
Guidelines and template fo

GenBank: BankIt
A web-based sequence su

GenBank: Barcode
Tool for submission to the

GenBank: Sequin
A stand-alone software too
simple submissions that co
sequences from phylogene

GenBank: tbl2asn
A command-line program t
submission of complete ge

Gene Expression Omnibus
Submit expression data. su

t for the submission of

data about a variant (variant

and easy.

n.

. It is capable of handling
ed sets of DNA, as well as

in. It is used primarily for

http://www.ncbi.nlm.nih.gov/gene

NCBI Resources How To Sign in to NCBI

Gene Gene Search
Advanced Help

Display Settings: Full Report Send to: Hide sidebar >>

XAGE1B X antigen family, member 1B [*Homo sapiens* (human)]

Gene ID: 653220, updated on 7-Dec-2014

Summary

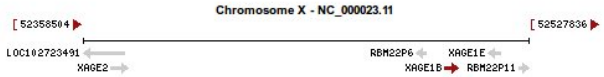
Official Symbol XAGE1B provided by HGNC
Official Full Name X antigen family, member 1B provided by HGNC
Primary source HGNC:HGNC:25400
Locus tag RP11-485B17.2
See related Ensembl:ENSG00000204379; MIM:300742; MIM:300743; Vega:OTTHUMG00000021557; Vega:OTTHUMG00000163577
Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as CTP9; XAGE1; CT12.1; GAGED2; XAGE-1; XAGE1A; CT12.1A; CT12.1B
Summary This gene is a member of the XAGE subfamily, which belongs to the GAGE family. The GAGE genes are expressed in a variety of tumors and in some fetal and reproductive tissues. This gene is strongly expressed in Ewing's sarcoma, alveolar rhabdomyosarcoma and normal testis. The protein encoded by this gene contains a nuclear localization signal and shares a sequence similarity with other GAGE/PAGE proteins. Because of the expression pattern and the sequence similarity, this protein also belongs to a family of CT (cancer-testis) antigens. Alternative splicing of this gene, in addition to alternative transcription start sites, results in multiple transcript variants. [provided by RefSeq, Jan 2010]

Genomic context

Location: Xp11.22 See XAGE1B in Epigenomics, MapViewer

Exon count: 4

Annotation release	Status	Assembly	Chr	Location
106	current	GRCh38 (GCF_000001405.26)	X	NC_000023.11 (52495668..52500812)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	X	NC_000023.10 (52238810..52243954)



Genomic regions, transcripts, and products

Go to reference sequence details

Genomic Sequence: NC_000023.11 chromosome X reference GRCh38 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank

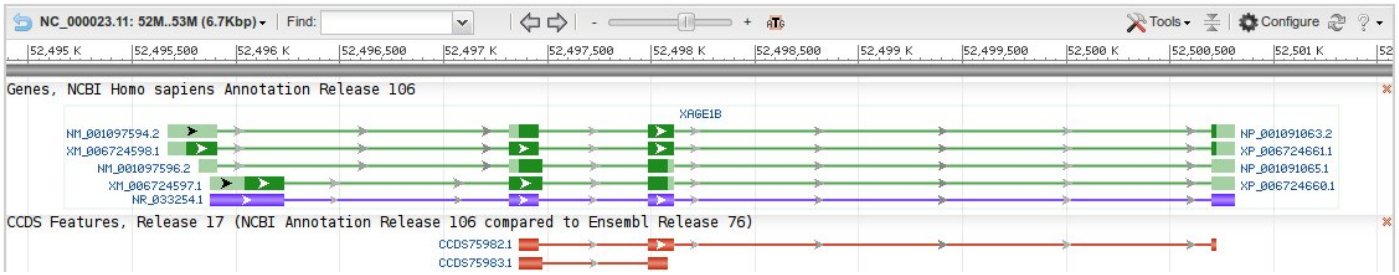


Table of contents
- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Variation
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

Related information
- BioProjects
- CCDS
- ClinVar
- Conserved Domains
- dbVar
- Full text in PMC
- Full text in PMC_nucleotide
- Gene neighbors
- Genome
- GEO Profiles
- HomoloGene
- Map Viewer
- Nucleotide
- OMIM
- Probe
- Protein
- PubChem Compound
- PubChem Substance
- PubMed
- PubMed (GeneRIF)
- PubMed (OMIM)
- PubMed(nucleotide/PMC)
- RefSeq Proteins
- RefSeq RNAs
- SNP
- SNP: GeneView
- Taxonomy
- UniGene

Links to other resources
- HGNC
- Ensembl
- Vega

Searching the databases: How to find gene location

BLAST

(Basic Local Alignment Search Tool)

Genomic DNA as query against the databases

ATGCTAGGATCCGATTGCAAG
CCTGAATCCGGCCTAATTTAC
G[illegible]CC
A[illegible]AG
A[illegible]AG
ATAGCAGATAGACAGTAAGAC
ATGATAGACGATAGATACAGA

Pattern matching to millions of sequences in the databases

```
>ref|NM_000059.3| UEGMD Homo sapiens breast cancer 2, early onset (BRCA2), mRNA
Length=11386

 GENE ID: 675 BRCA2 | breast cancer 2, early onset [Homo sapiens]
(Over 100 PubMed links)


 Score =   348 bits (188),  Expect = 5e-93
 Identities = 188/188 (100%), Gaps = 0/188 (0%)
 Strand=Plus/Plus

Query  7    GTGGCGCGAGCTTCTGAAACTAGGCGGCAGAGGCGGAGCCGCTGTGGCACTGCTGCGCCT  66
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    GTGGCGCGAGCTTCTGAAACTAGGCGGCAGAGGCGGAGCCGCTGTGGCACTGCTGCGCCT  60

Query  67   CTGCTGCGCCTCGGGTGTCTTTTGCGGCGGTGGGTCGCCGCCGGGAGAAGCGTGAGGGGA  126
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   CTGCTGCGCCTCGGGTGTCTTTTGCGGCGGTGGGTCGCCGCCGGGAGAAGCGTGAGGGGA  120

Query  127  CAGATTTGTGACCGGCGCGGTTTTTGTCAGCTTACTCCGGCCAAAAAAGAACTGCACCTC  186
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  CAGATTTGTGACCGGCGCGGTTTTTGTCAGCTTACTCCGGCCAAAAAAGAACTGCACCTC  180

Query  187  TGGAGCGG  194
            ||||||||
Sbjct  181  TGGAGCGG  188
```

BLAST results and alignment

BLAST®

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New DELTA-BLAST, a more sensitive protein-protein search Go

BLAST Assembled Gen

Find Genomic BLAST pages:

Enter organism name o

- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice

Basic BLAST

Choose a BLAST program to

nucleotide blast	Search a Algo
protein blast	Search p Algo
blastx	Search p
tblastn	Search t
tblastx	Search t

Specialized BLAST

Choose a type of specialized

- Make specific p
- Cluster multiple
- Find conserved
- Find sequences
- Search sequen
- Search immuno
- Screen sequen
- Align two (or m
- Search protein
- Search SRA by
- Constraint Base
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene
- Search trace archives

Human BLAT Search

BLAT Search Genome

Genome:

Human | Dec. 2

Paste in a query sequence to fi
separated by lines starting with

File Upload: Rather than pastir
sequence.
Upload sequence: Browse... N

Only DNA sequences of 25,000
letters will be processed. Up to
multiple sequence submissions

For locating PCR primers, use I

HMMER

biosequence analysis using profile hidden Markov models

Home | Search | Results | Software | Help | About

phmmer | hmmscan | hmmsearch | jackhmmer

protein alignment/profile-HMM vs protein sequence database

Paste a Sequence | Upload a File | Accession Search

Paste in your alignment/hmm or use the example

Submit | Reset

▼ **Sequence Database**

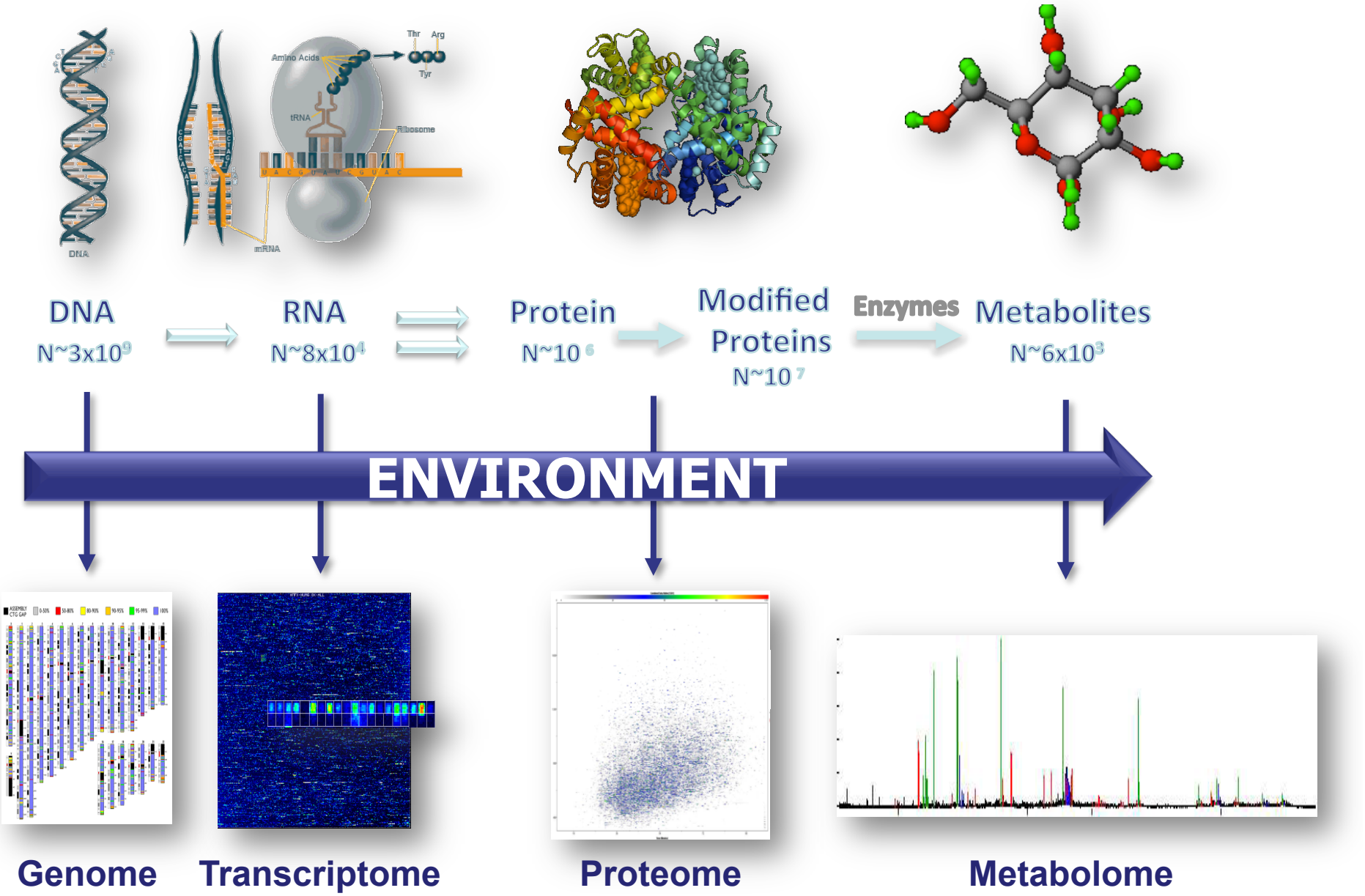
Large collections: NR | RefSeq | UniProtKB | Pfamseq

Annotated and Structures: SwissProt | PDB

Metagenomics and Environmental: UniMes | env NR

Representative Sets (UniProt): rp75 | rp55 | rp35 | rp15 | Reference Proteomes

▶ **Restrict by Taxonomy**



Modified from GD Lewis, R Gerszten et. al. JACC 2008;52;118

Information | Topics | Course manuals | Course urls | Timetable

Introduction

- EBI homepage
- EBI bioinformatics tools
- UniProt
- NCBI homepage
- Gquery
- NCBIGene
- NCBI BLAST server
- ORF finder
- Splign
- MUSCLE
- ClustalOmega

Module 1

- Ensembl

Module 2

- UCSC genome browser
- VEGA
- NCBI Map Viewer

Module 3

- ECR browser
- VISTA Enhancer Browser
- BLINK
- Galaxy
- Homologene
- NCBI Trace Archives
- PHYLIP
- PhyML
- Pipmaker
- Promo
- RepeatMasker
- rVISTA
- TreeBeST