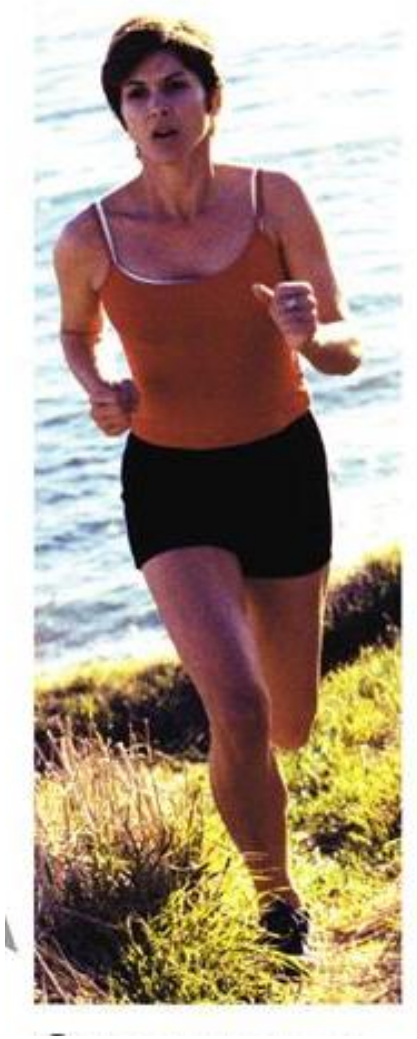
Dr. Susan Steinbusch-Coort

susan.coort@maastrichtuniversity.nl

February 3rd, 2016

What happens with the human body when you are running?



Organ systems work together

- Skeletal system- supports the skeleton
- Muscular system - pulls on the bones to enable you to move
- Respiratory system - makes sure your muscles have enough oxygen for respiration
- Circulatory system- provides oxygen and glucose to the skeletal muscle cells

Human body structure

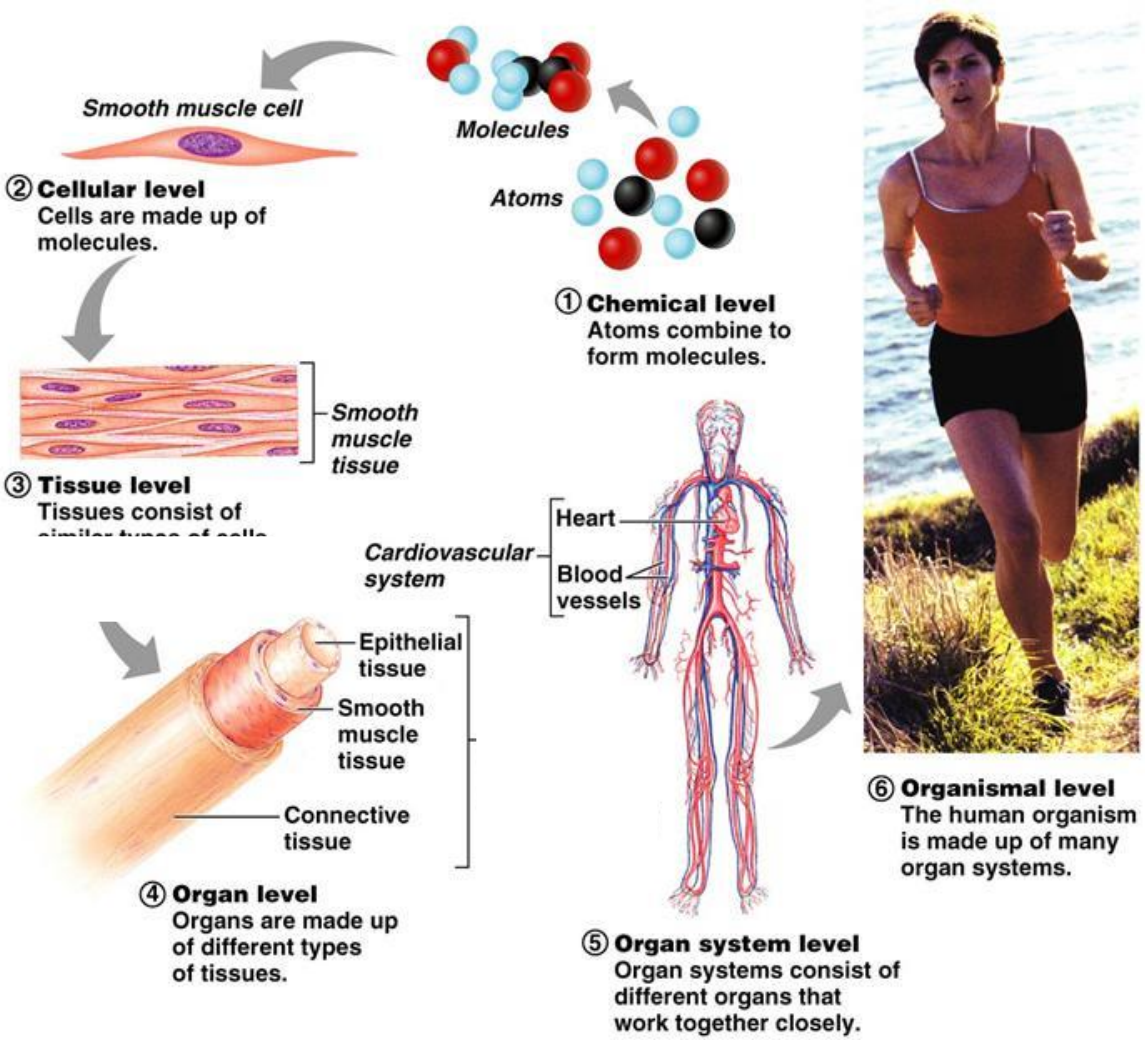
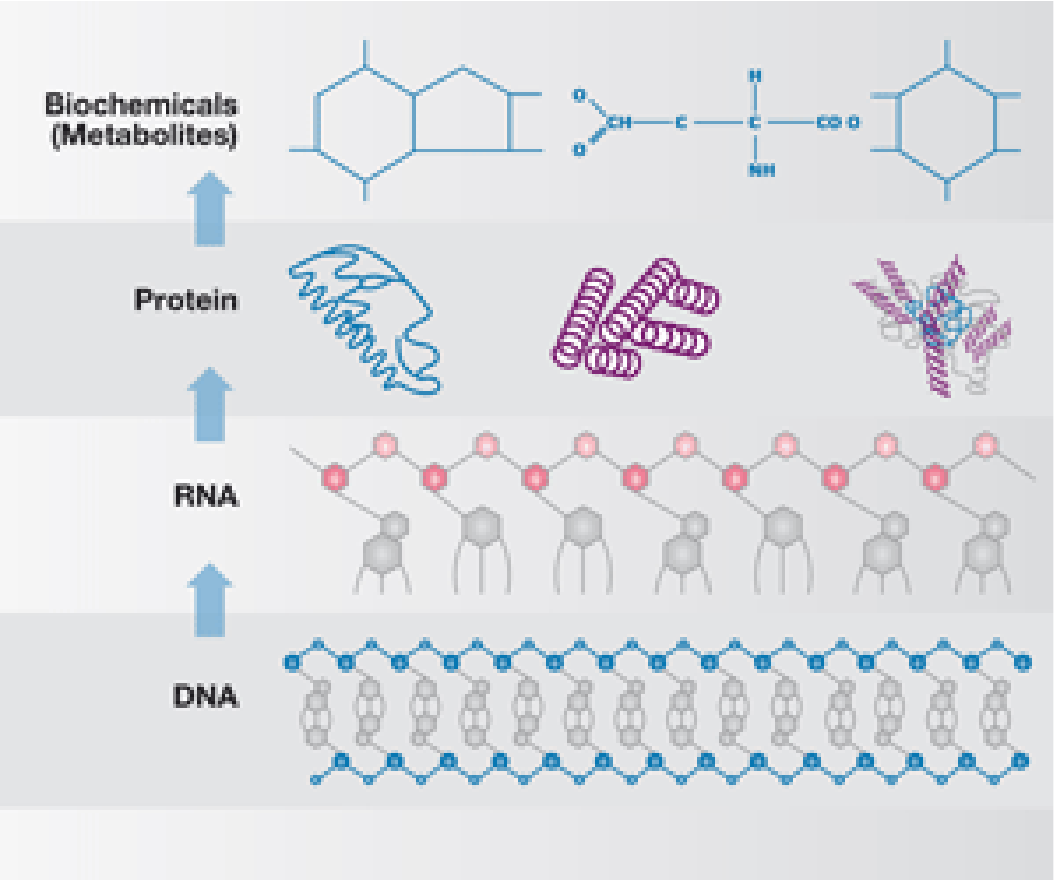
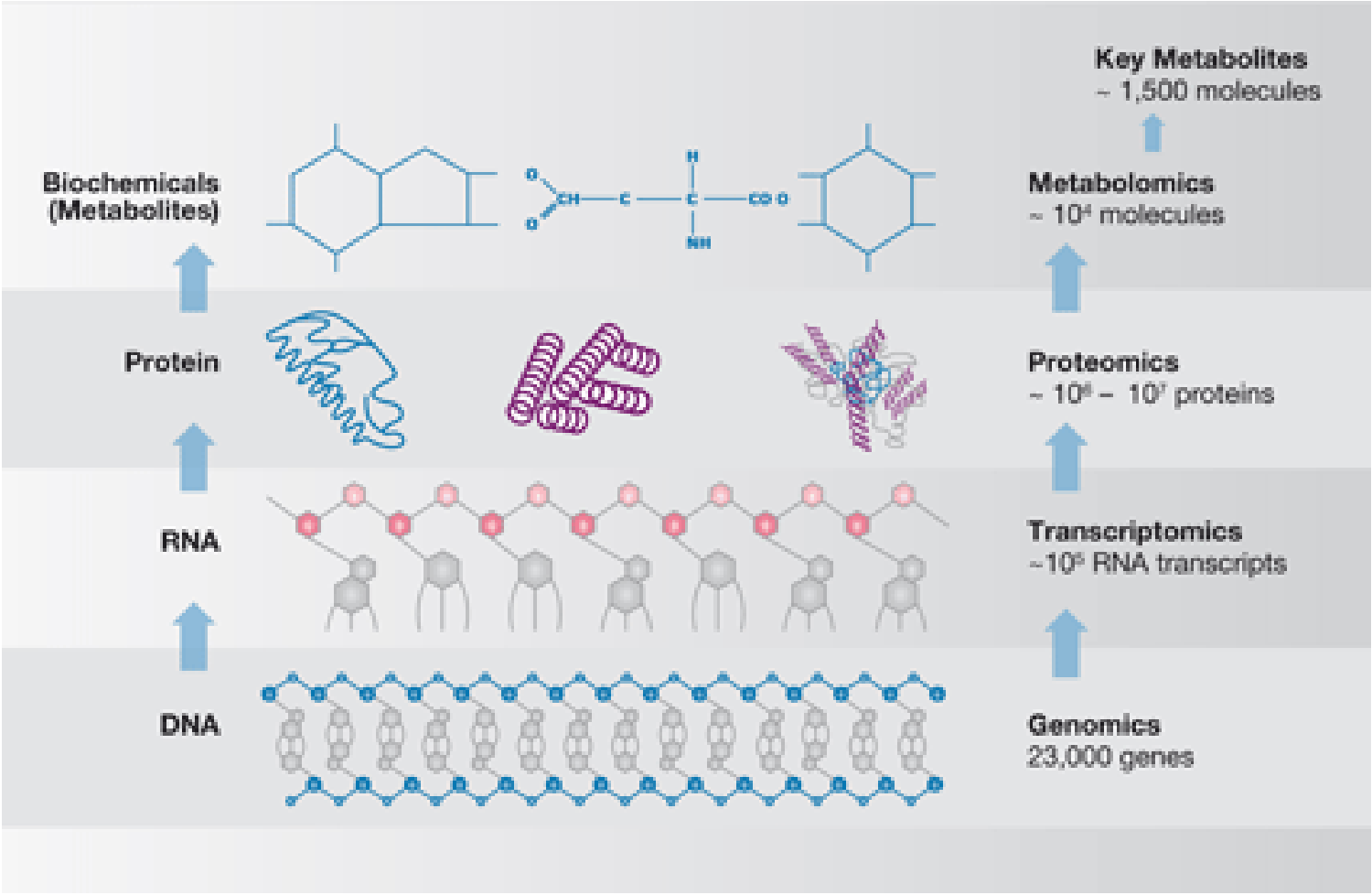


Figure 1.1

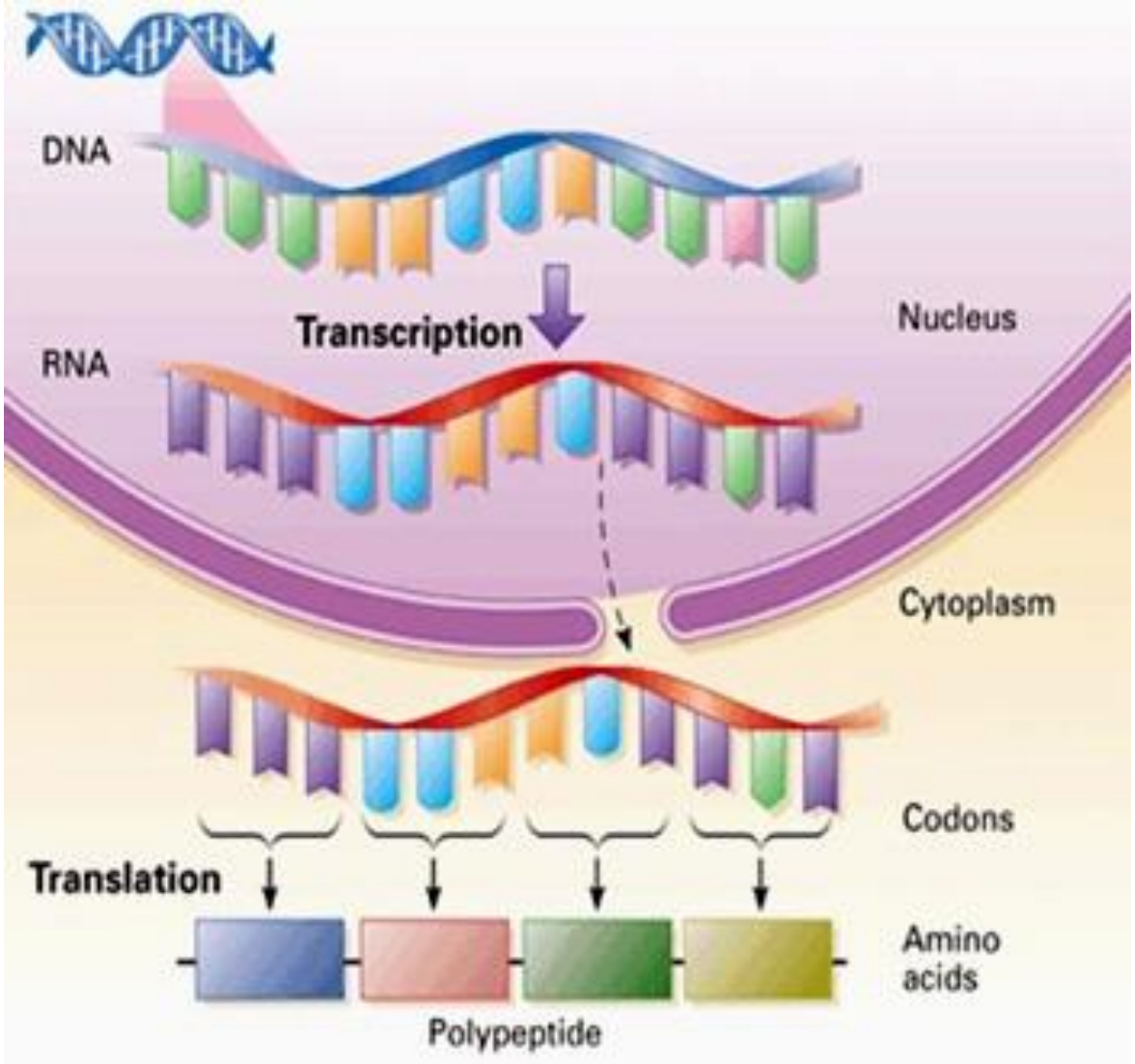
(Bio)Molecules
Individual players are important



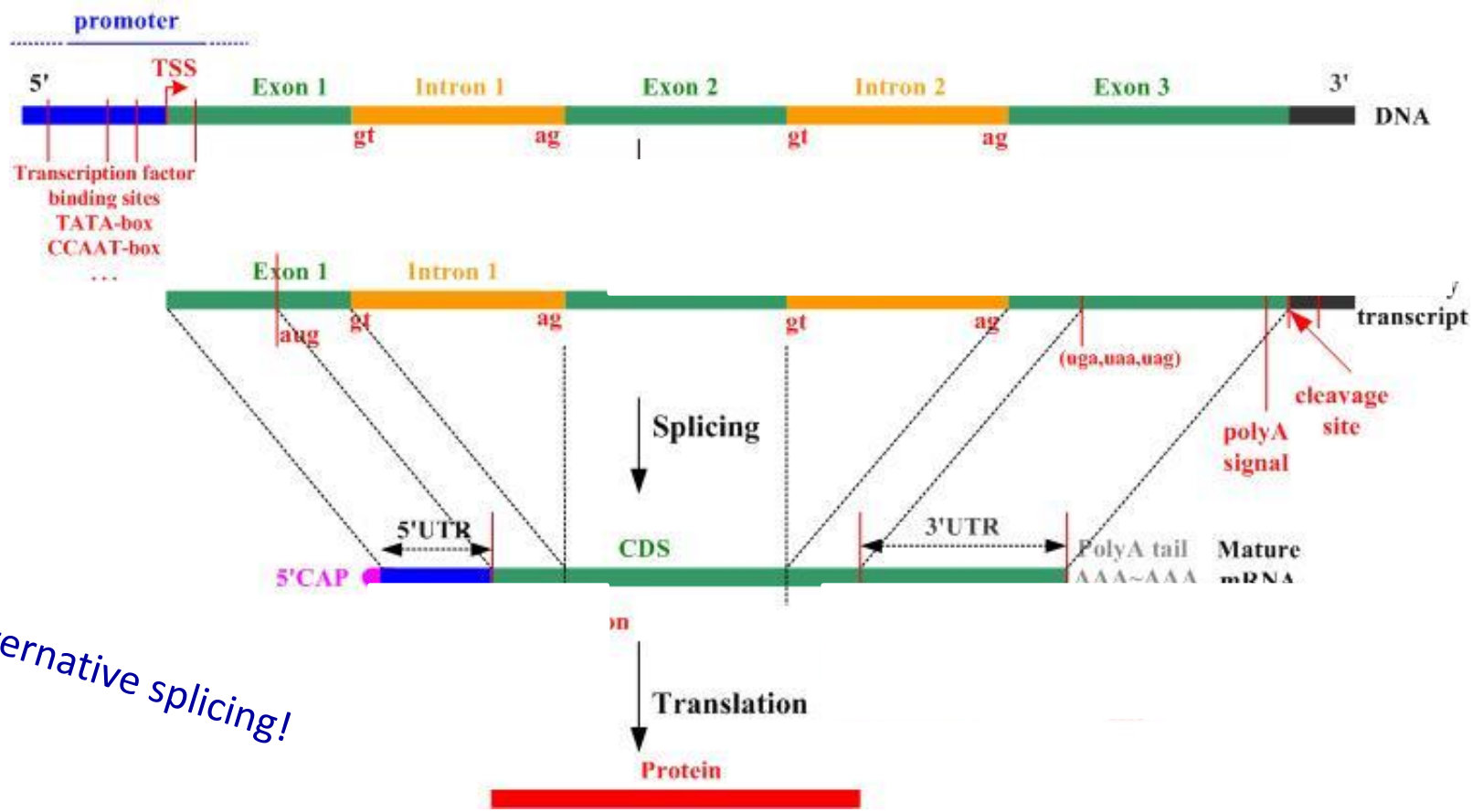
Heaps of knowledge on biomolecules online available.



Protein synthesis



Gene structure



Alternative splicing!

CDS = Coding DNA Sequence
UTR = UnTranslated region

GOAL

To understand biological sequence databases

- Which biological sequence databases are available?
- How can you find information in these databases?
- What is the content of the databases?
- What is Gene Ontology?
- Two projects aimed at deciphering the content of the human genome, the human genome project & ENCODE.

What is a database

https://www.youtube.com/watch?v=gfT7EGibry0

Genes in stead of persons

Name	Identifier	Sequence	Synonyms	Chromosomal location	Disease	Many more
Gene 1	2456	AGTCCCGT	DAH, HSD	4q12	Cancer	.....
Gene2	4333	CGGTAACT	HGR	7p10	Diabetes	.....
Gene 3	6799	AGTCGGCGGG				
etc						

All the available information is stored in databases!

Biological sequence databases

Originally – just a storage place for sequences.

Currently – the databases are bioinformatics work bench which provide many tools for retrieving, comparing and analyzing sequences.

1. Global nucleotide/protein sequence storage databases:
   - GenBank of NCBI (National Center for Biotechnology Information)
   - The European Molecular Biology Laboratory (EMBL) database
   - The DNA Data Bank of Japan (DDBJ)

2. Genome-centered databases
   - NCBI genomes
   - Ensembl Genome Browser
   - UCSC Genome Bioinformatics Site

3. Protein Databases
   - UniProt

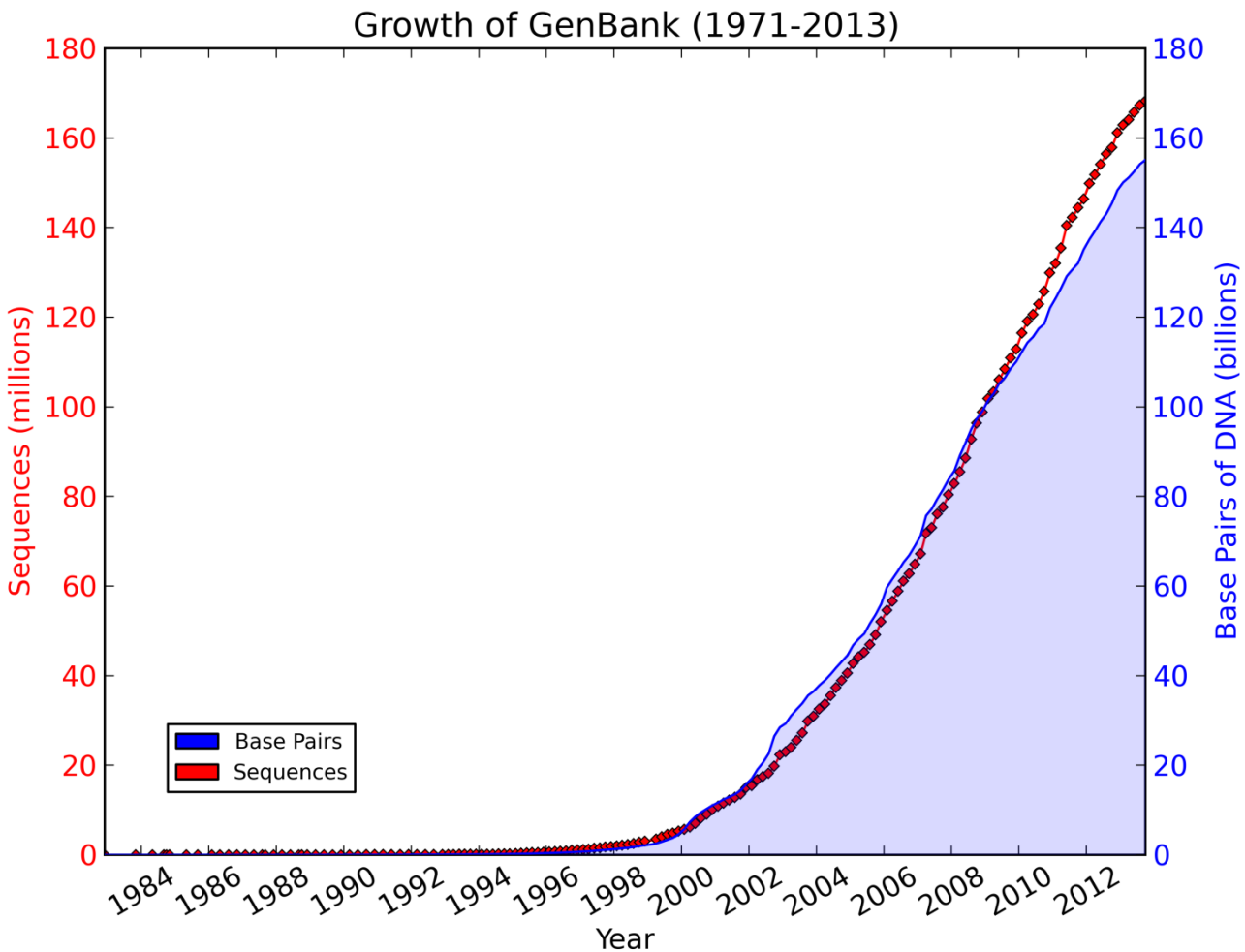
Lecture protein structures

NCBI nucleotide databases

- GenBank
  - Individual submissions (DNA, mRNA, eiwit)
  - Bulk submissions (Genome centers)
    - High throughput sequencing (DNA)
    - Expressed Sequence Tags (mRNA)

- RefSeq
  - Curated subset of GenBank
  - “Reference” sequence
  - Single sequence per locus / molecule

Growth of GenBank

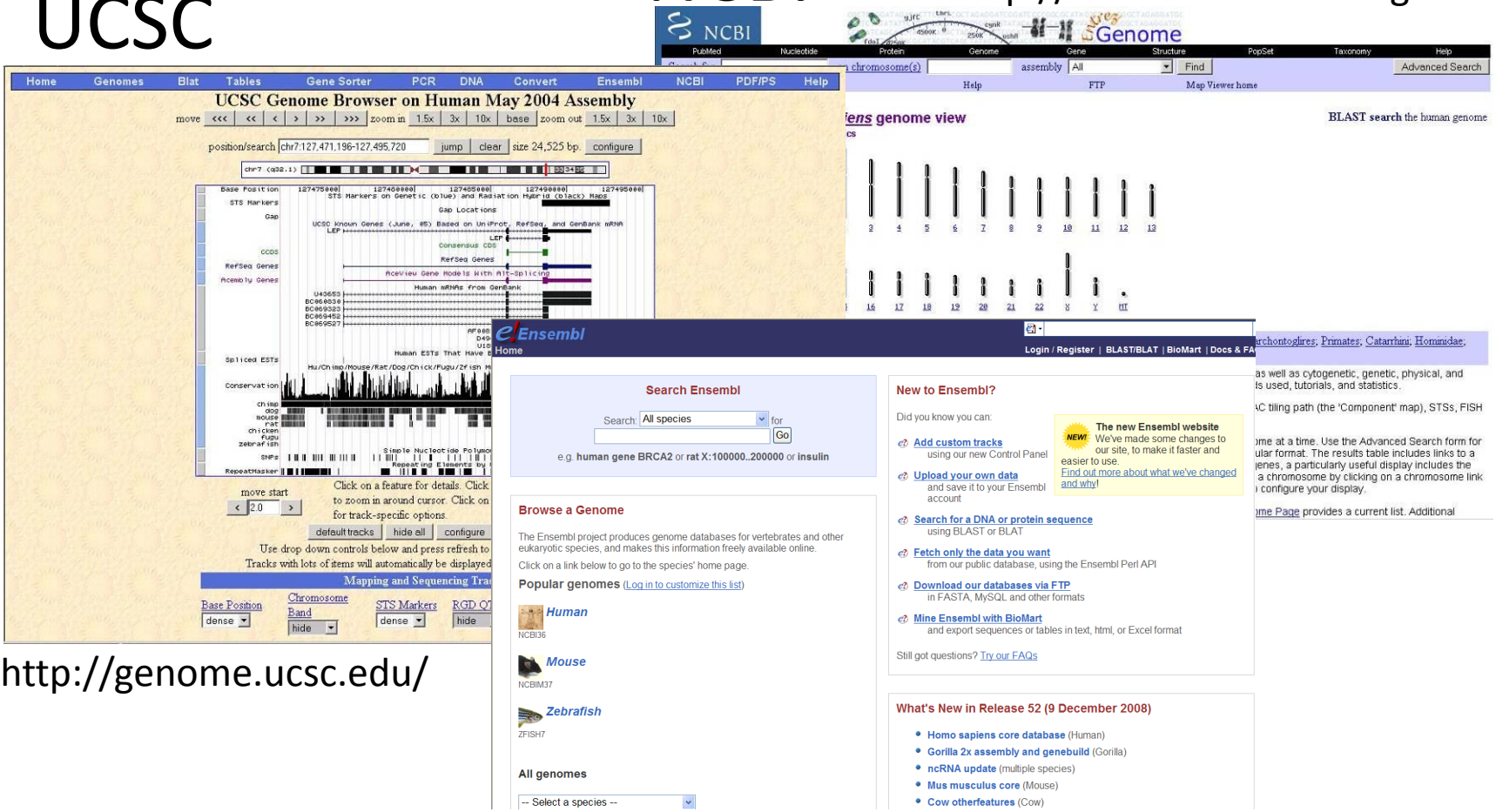


Genome-centered databases

UCSC

NCBI

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

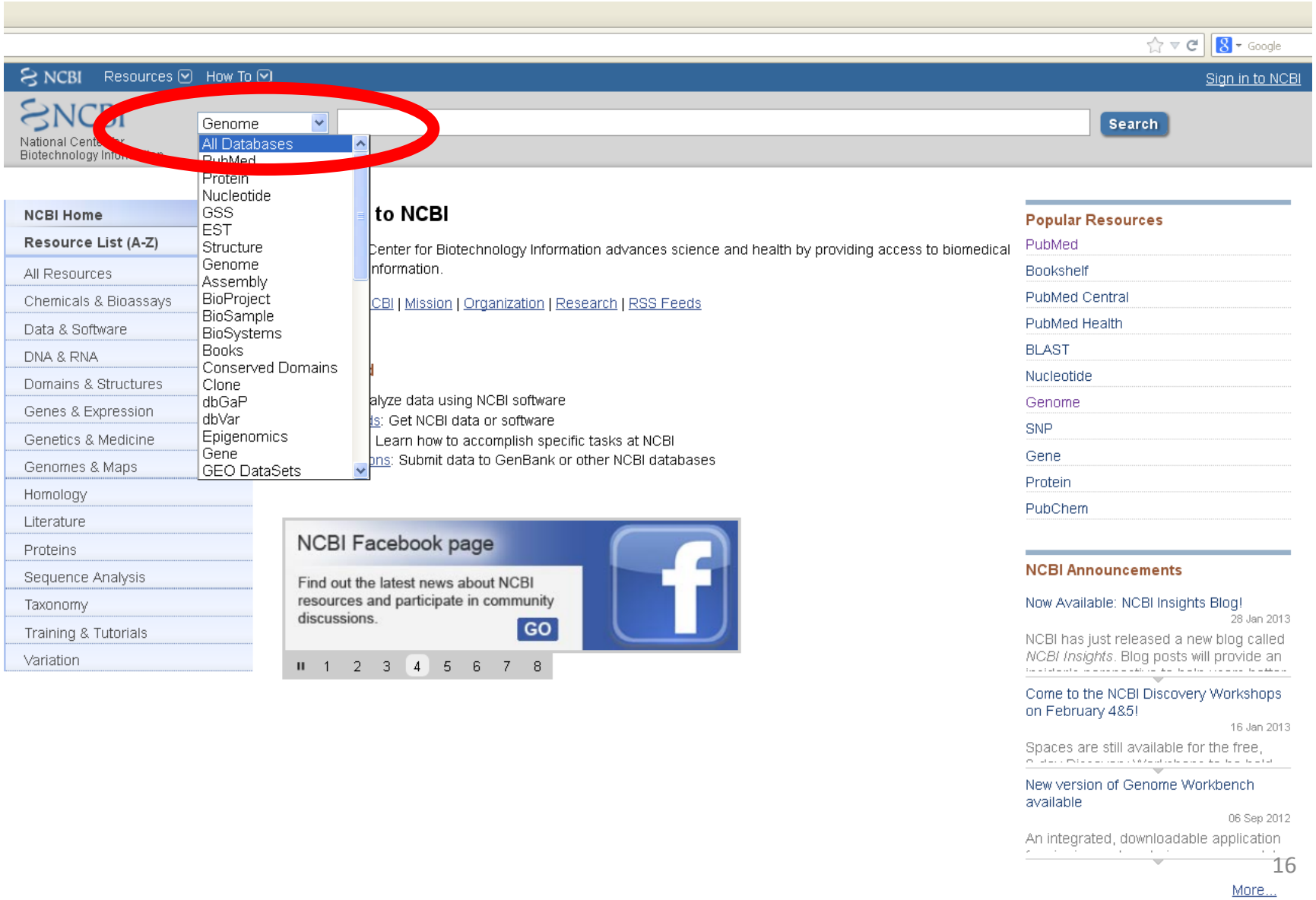


http://genome.ucsc.edu/

Ensembl

http://www.ensembl.org/

NCBI homepage



NCBI Global Cross-database search

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

GQuery

NCBI Global Cross-database Search

Search NCBI databases

Search

Literature

- **PubMed**: scientific & medical abstracts/citations
- **PubMed Central**: full-text journal articles
- **NLM Catalog**: books, journals and more in the NLM Collections
- **MeSH**: ontology used for PubMed indexing
- **Books**: books and reports
- **Site Search**: NCBI web and FTP site index

Health

- **PubMed Health**: clinical effectiveness, disease and drug reports
- **MedGen**: medical genetics literature and links
- **GTR**: genetic testing registry
- **dbGaP**: genotype/phenotype interaction studies
- **ClinVar**: human variations of clinical significance
- **OMIM**: online mendelian inheritance in man
- **OMIA**: online mendelian inheritance in animals

Organisms

- **Taxonomy**: taxonomic classification and nomenclature catalog

Nucleotide Sequences

- **Nucleotide**: DNA and RNA sequences
- **GSS**: genome survey sequences
- **EST**: expressed sequence tag sequences
- **SRA**: high-throughput DNA and RNA sequence read archive
- **PopSet**: sequence sets from phylogenetic and population studies
- **Probe**: sequence-based probes and primers

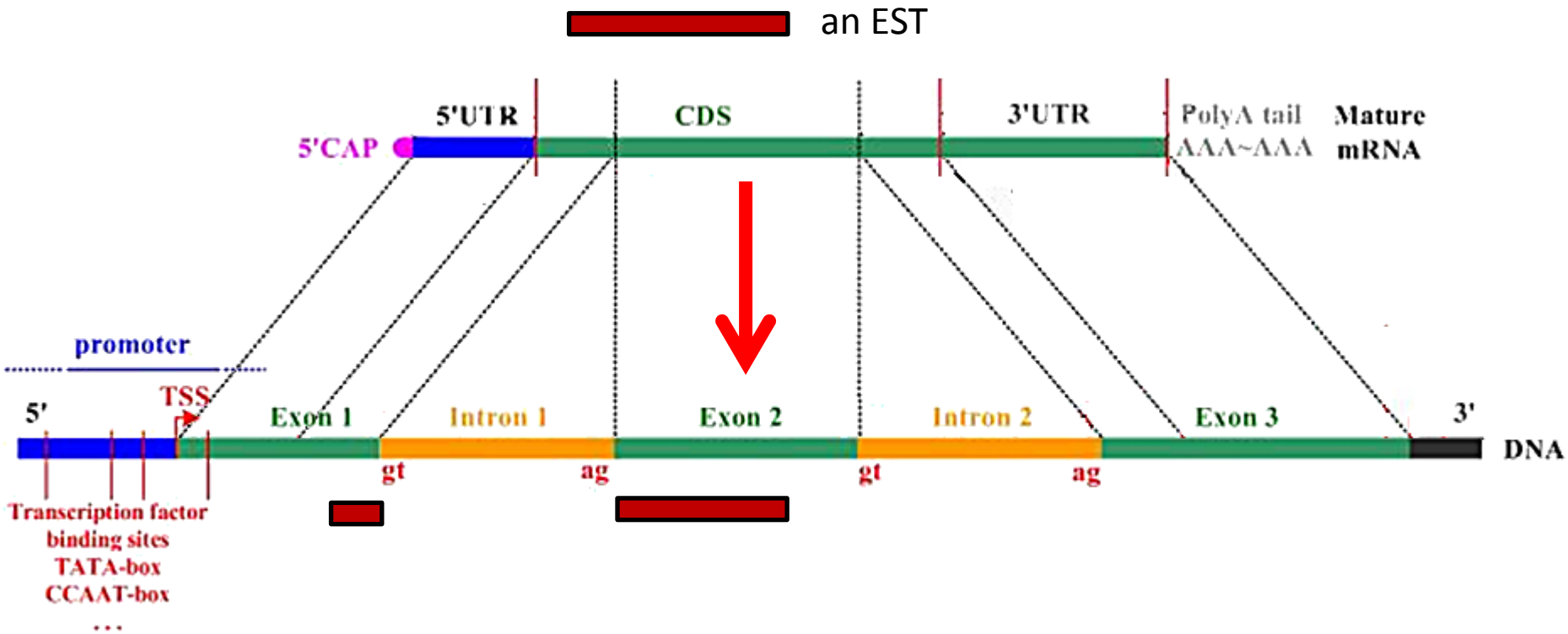
Genomes

- **Genome**: genome sequencing projects by organism
- **Assembly**: genomic assembly information
- **Epigenomics**: epigenomic studies and display tools
- **UniSTS**: sequence-tagged sites for genome mapping
- **SNP**: short genetic variations
- **dbVar**: genome structural variation studies
- **BioProject**: biological projects providing data to NCBI
- **BioSample**: descriptions of biological source materials
- **Clone**: genomic and cDNA clones

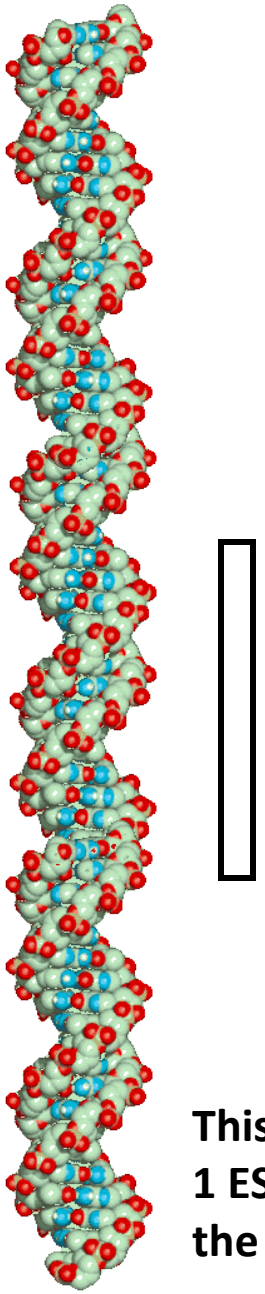
UniGene

- EST:
  - DNA sequence corresponding to mRNA from expressed gene
  - ~500 base pairs long
  - Sequenced from a cDNA library
- Predict genes based on ESTs (expressed sequence tags)
- Cluster ESTs from many cDNA libraries to predict distinct genes

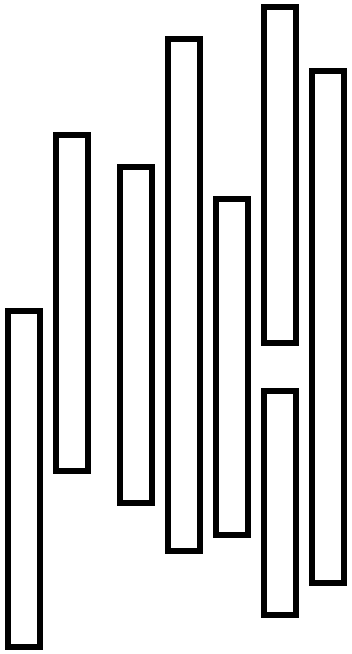
Map mRNA (EST) back to DNA

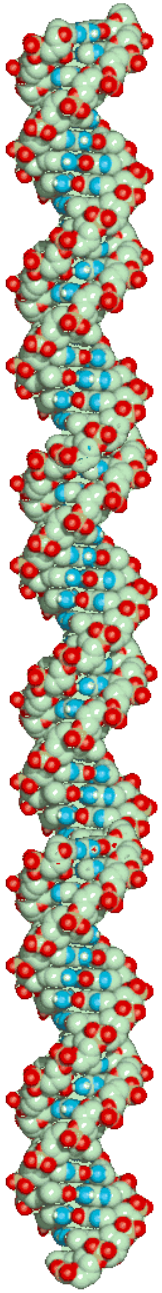


EST clusters



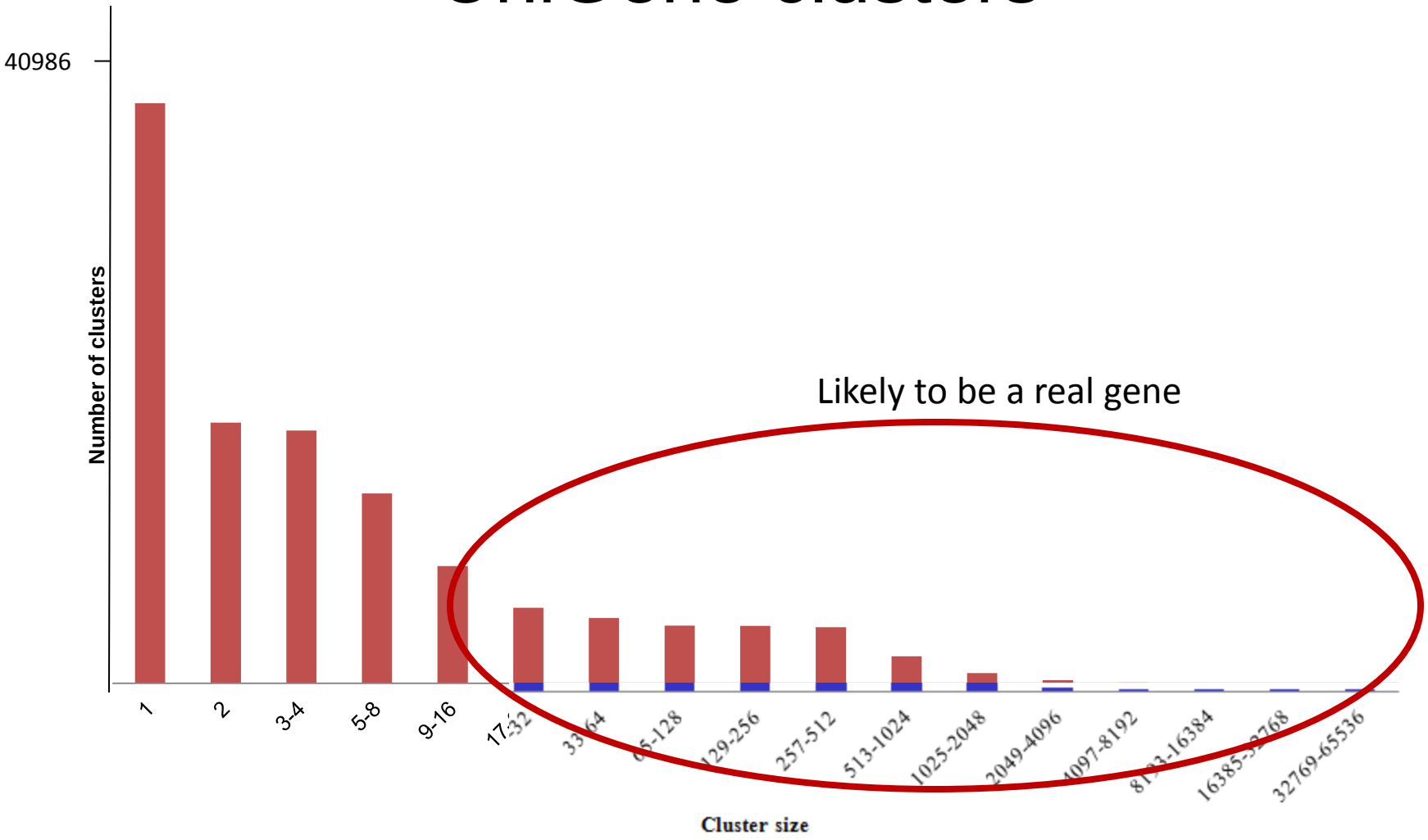
This is a gene with 1 EST associated; the cluster size is 1





This is a gene with 10 ESTs associated; the cluster size is 10

UniGene clusters



Gene (NCBI)
DHH as example

DHH desert hedgehog [Homo sapiens] - Gene - NCBI - Mozilla Firefox

File Edit View History Bookmarks Tools Help

DHH desert hedgehog [Homo sapiens] - Gen...

www.ncbi.nlm.nih.gov/gene/50846

NCBI Resources How To Sign in to NCBI

Gene Gene Search

Limits Advanced Help

Display Settings: Full Report Send to:

DHH desert hedgehog [*Homo sapiens*]

Gene ID: 50846, updated on 6-Jan-2013

Summary

Official Symbol DHH provided by HGNC
Official Full Name desert hedgehog provided by HGNC
Primary source HGNC:2865
See related Ensembl:ENSG00000139549; HPRD:05664; MIM:605423; Vega:OTTHUMG00000170408
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 GDXYM; HHG-3; SRXY7
Summary This gene encodes a member of the hedgehog family. The hedgehog gene family encodes signaling molecules that play an important role in regulating morphogenesis. This protein is predicted to be made as a precursor that is autocatalytically cleaved; the N-terminal portion is soluble and contains the signalling activity while the C-terminal portion is involved in precursor processing. More importantly, the C-terminal product covalently attaches a cholesterol moiety to the N-terminal product, restricting the N-terminal product to the cell surface and preventing it from freely diffusing throughout the organism. Defects in this protein have been associated with partial gonadal dysgenesis (PGD) accompanied by minifascicular polyneuropathy. This protein may be involved in both male gonadal differentiation and perineurial development. [provided by RefSeq, May 2010]

Genomic context

Location: 12q13.1
Sequence: Chromosome: 12; NC_000012.11 (49483204..49488602, complement)

See DHH in Epigenomics, MapViewer

Chromosome 12 - NC_000012.11

[49412758 ▶ [49525384 ▶
MLL2 RHEBL1 DHH LMBR1L TUBA1B

Genomic regions, transcripts, and products

Genomic Sequence NC_000012 chromosome 12 reference GRCh37.p10 Primary Assembly

Go to reference sequence details

Go to nucleotide Graphics FASTA GenBank

NC_000012.11: 49M..49M (7.0Kbp) C Find on Sequence: Tools Configure

489,500 | 49,489 K | 49,488,500 | 49,488 K | 49,487,500 | 49,487 K | 49,486,500 | 49,486 K | 49,485,500 | 49,485 K | 49,484,500 | 49,484 K | 49,483,500 | 49,483 K

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Interactions
- General gene info
- General protein info
- Reference sequences
- Related sequences
- Additional links

Related information

- Order cDNA clone
- 3D structures
- BioAssay
- BioProjects
- BioSystems
- Books
- CCDS
- Conserved Domains
- dbVar
- Full text in PMC
- Genome
- GEO Profiles
- GTR
- HomoloGene
- Map Viewer
- MedGen
- Nucleotide
- OMIM
- Probe
- Protein
- PubChem Compound

OMIM (NCBI)

Home | About | Statistics | Downloads/API | Help | External Links | Terms of Use | Contact Us | MIMmatch

Select Language

dhh Search Sort by: Relevance Date updated

Advanced Search: OMIM, Clinical Synopses, OMIM Gene Map **Toggle:** search terms highlighted
Search History: View, Clear

*605423

DESERT HEDGEHOG; DHH

HGNC Approved Gene Symbol: *DHH*

Cytogenetic location: *12q13.12* ***Genomic coordinates (GRCh37):*** *12:49,483,203 - 49,488,601* (from NCBI)

Gene Phenotype Relationships

Location	Phenotype	Phenotype MIM number
12q13.12	46XY partial gonadal dysgenesis, with minifascicular neuropathy	607080
	46XY sex reversal 7	233420

- Table of Contents - *605423

External Links:

- Genome
- DNA
- Protein
- Gene Info
- Clinical Resources
- Variation
- Animal Models
- Cellular Pathways

TEXT

Description

The hedgehog gene family encodes signaling molecules that play an important role in regulating morphogenesis. Mammalian hedgehog genes share striking homology to the Drosophila segment polarity gene hedgehog, a key regulator of pattern formation in the embryonic and adult fly.

Cloning

Tate et al. (2000) found that the human DHH gene encodes a 396-amino acid polypeptide (GenBank AB010994).

Bitgood and McMahon (1995) and Parmantier et al. (1999) showed that during development in the mouse, Dhh mRNA shows a very restricted distribution, being expressed primarily in Sertoli cells of developing testes and in Schwann cells of peripheral nerves.

Gene Structure

Homologene

NCBI Resources How To

HomoloGene HomoloGene

Limits Advanced

Display Settings: HomoloGene

HomoloGene:22431. Gene conserved in Eutheria

Genes

Genes identified as putative homologs of one another during the construction of HomoloGene.

DHH, *H.sapiens*
desert hedgehog

DHH, *C.lupus*
desert hedgehog

DHH, *B.taurus*
desert hedgehog

Dhh, *M.musculus*
desert hedgehog

Dhh, *R.norvegicus*
desert hedgehog

Proteins

Proteins used in sequence comparisons and their conserved domain architectures.

NP_066382.1
396 aa

XP_003640009.1
391 aa

XP_002687352.1
396 aa

NP_031883.1
396 aa

NP_445819.1
396 aa

Protein Alignments

Protein multiple alignment, pairwise similarity scores and evolutionary distances.

Show Multiple Alignment

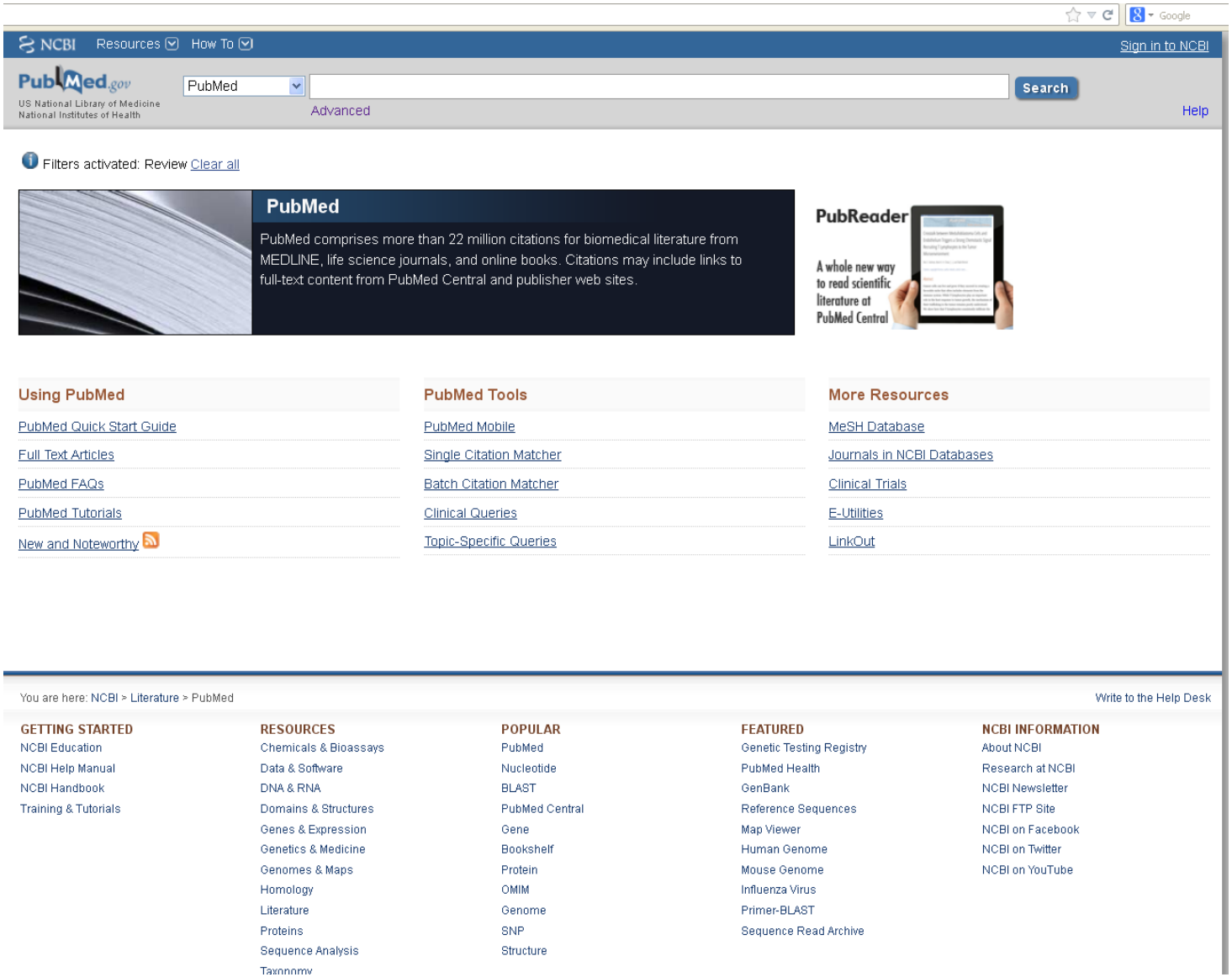
Conserved Domains

Conserved Domains from CDD found in protein sequences by rpsblast searching.

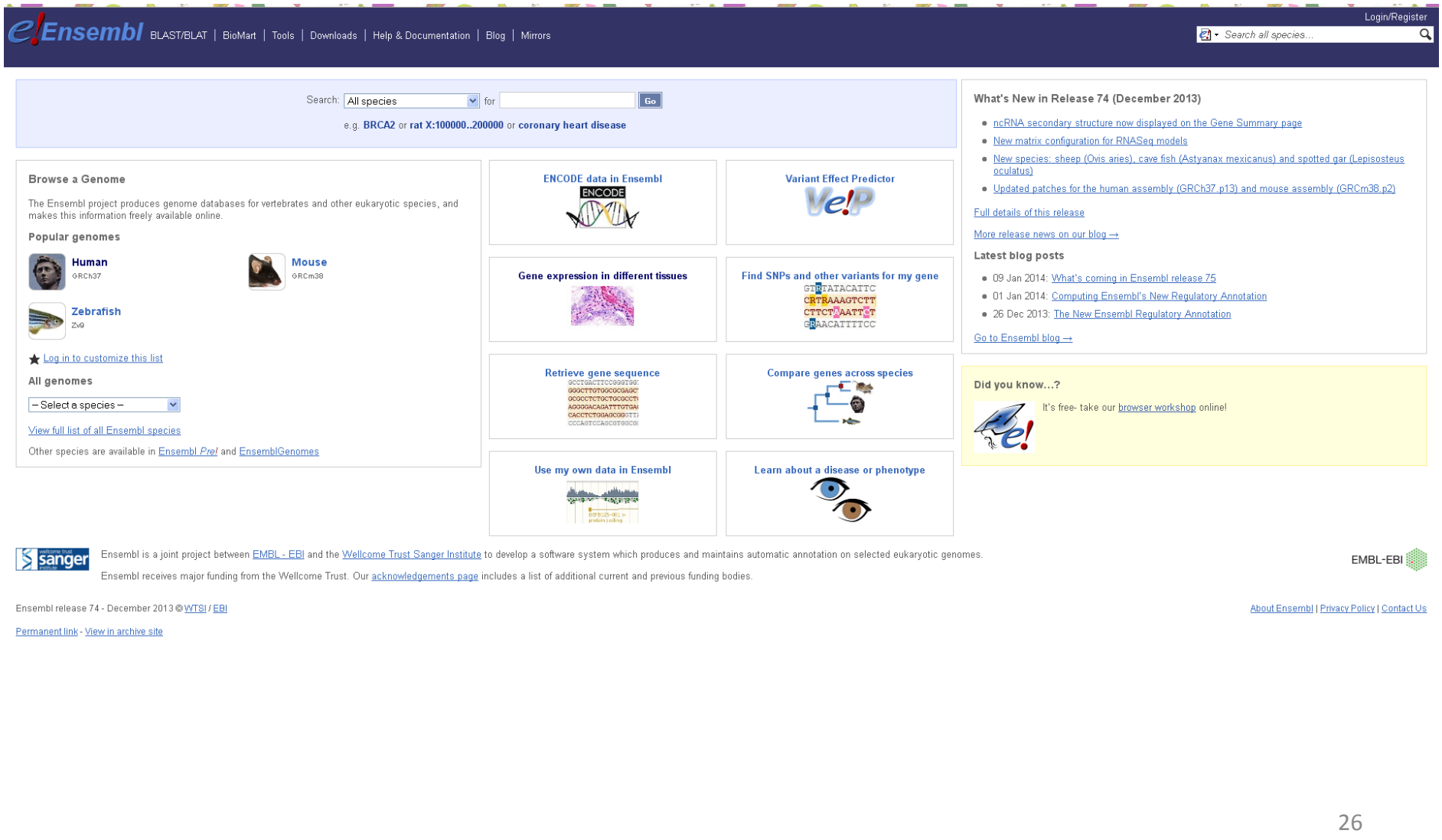
Hint (pfam01079)
Hint module.

Homologue = One of a group of similar DNA sequences that share a common ancestry.

PubMed (NCBI)


NCBI
Resources
How To
Sign in to NCBI
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Search
Advanced
Help
Filters activated: Review Clear all
PubMed
PubMed comprises more than 22 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.
PubReader
A whole new way to read scientific literature at PubMed Central
Using PubMed
PubMed Quick Start Guide
Full Text Articles
PubMed FAQs
PubMed Tutorials
New and Noteworthy
PubMed Tools
PubMed Mobile
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Topic-Specific Queries
More Resources
MeSH Database
Journals in NCBI Databases
Clinical Trials
E-Utilities
LinkOut
You are here: NCBI > Literature > PubMed
Write to the Help Desk
GETTING STARTED
NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials
RESOURCES
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
POPULAR
PubMed
Nucleotide
BLAST
PubMed Central
Gene
Bookshelf
Protein
OMIM
Genome
SNP
Structure
FEATURED
Genetic Testing Registry
PubMed Health
GenBank
Reference Sequences
Map Viewer
Human Genome
Mouse Genome
Influenza Virus
Primer-BLAST
Sequence Read Archive
NCBI INFORMATION
About NCBI
Research at NCBI
NCBI Newsletter
NCBI FTP Site
NCBI on Facebook
NCBI on Twitter
NCBI on YouTube

Ensembl homepage



Ensembl
example DHH (human)

Ensembl genome browser 70: Homo sapiens ...

www.ensembl.org/Homo_sapiens/Location/View?db=core;g=ENSG00000139549;r=12:49483204-49488602;t=ENST00000266991

Ensembl | BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors | Login · Register | Search Human...

Human (GRCh37) | Location: 12:49,483,204-49,488,602 | Gene: DHH | Transcript: DHH-001

Transcript-based displays
- Transcript summary
- Supporting evidence (6)
- Sequence
  - Exons (3)
  - cDNA
  - Protein
- External References
  - General identifiers (26)
  - Oligo probes (21)
- Ontology
  - Ontology graph (18)
  - Ontology table (18)
- Genetic Variation
  - Variation table
  - Variation image
  - Population comparison
  - Comparison image
- Protein Information
  - Protein summary
  - Domains & features (21)
  - Variations (45)
- External data
  - Personal annotation
- ID History
  - Transcript history
  - Protein history

Configure this page
Add your data
Export data
Bookmark this page
Share this page

Transcript: DHH-001 ENST00000266991

Description	desert hedgehog [Source:HGNC Symbol;Acc:2865]
Location	Chromosome 12: 49,483,204-49,488,602 reverse strand.
Gene	This transcript is a product of gene ENSG00000139549 - This gene has 1 transcript

Show/hide columns | Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
DHH-001	ENST00000266991	1936	ENSP00000266991	396	Protein coding	CCDS8779

Transcript and Gene level displays

Views in Ensembl are separated into gene based views and transcript based views according to which level the information is more appropriately associated with. This view is a transcript level view. To flip between the two sets of views you can click on the Gene and Transcript tabs in the menu bar at the top of the page.

Transcript summary

Reverse strand — 5.40 Kb

Statistics	**Exons:** 3 **Coding exons:** 3 **Transcript length:** 1,936 bps **Translation length:** 396 residues
CCDS	This transcript is a member of the Human CCDS set: CCDS8779
Ensembl version	ENST00000266991.2
Type	Known protein coding
Prediction Method	Transcript where the Ensembl genebuild transcript and the Vega manual annotation have the same sequence, for every base pair. See article.
Alternative transcripts	**This transcript corresponds to the following database identifiers:** **Transcript having exact match between ENSEMBL and HAVANA:** OTTHUMT00000408973 (version 1)

Ensembl release 70 - January 2013 © WTSI / EBI

About Ensembl | Privacy Policy | Contact Us

Permanent link - View in archive site

antisense

RP11-386G11.8-002 >
antisense

UCSC homepage

UCSC Genome Browser Home - Mozilla Firefox

File Edit View History Bookmarks Tools Help

UCSC Genome Browser Home

genome.ucsc.edu

UCSC Genome Bioinformatics

Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - FAQ - Help

Genome Browser
ENCODE
Neandertal
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
VisiGene
Utilities
Downloads
Release Log
Custom Tracks
Microbial Genomes
Mirrors
Archives
Training
Credits
Publications
Cite Us
Licenses
Jobs
Staff
Contact Us

About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the ENCODE and Neandertal projects.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of *in situ* mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.

News

News Archives ▶

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.

25 January 2013 - Southern White Rhinoceros Genome Browser Released

A Genome Browser is now available for the Southern White Rhinoceros (*Ceratotherium simum simum*) assembly released by the Broad Institute in May 2012 (Broad version cerSimSim1.0, UCSC version cerSim1). This genome was sequenced and assembled at the Broad Institute using samples provided by Dr. Oliver Ryder at the San Diego Zoo Institute for Conservation Research. For more information and statistics about this assembly, see the NCBI assembly record for CerSimSim1.0.

Bulk downloads of the sequence and annotation data may be obtained from the Genome Browser FTP server or the Downloads page. Please observe the conditions for use when accessing and using these data sets. The annotation tracks for this browser were generated by UCSC and collaborators worldwide. See the Credits page for a detailed list of the organizations and individuals who contributed to this release.

---

22 January 2013 - New Baboon (papAnu2) Assembly Now Available in the Genome Browser: We are pleased to announce the release of a Genome Browser for the March 2012 assembly of the Olive Baboon, *Papio anubis* (Baylor Panu_2.0, UCSC version papAnu2). Read more.

15 January 2013 - New Lamprey (petMar2) Assembly Now Available in the Genome Browser: We are pleased to announce the release of a Genome Browser for the September 2010 assembly of the Lamprey, *Petromyzon marinus* (WUGSC 7.0, UCSC version petMar2). Read more.

==> News Archives

Conditions of Use

The sequence and annotation data displayed in the Genome Browser are freely available for any use with the following conditions:

- Genome sequence data use restrictions are noted within the species sections on the Credits page.
- Some annotation tracks contributed by external collaborators contain proprietary data that have specific use restrictions. To check for restrictions associated with a particular genome assembly, review the *database/README.txt* file in the assembly's downloads directory.

The UCSC, Ensembl, and NCBI browser and annotation groups have established a common set of minimum requirements for public display of genome data made available after Spring 2009, described here.

The Genome Browser and Blat software are free for academic, nonprofit, and personal use. A license is required for commercial use. See the Licenses page for more information.

Program-driven use of this software is limited to a maximum of one hit every 15 seconds and no more than 5,000 hits per day.

For assistance with questions or problems regarding the UCSC Genome Browser software, database, genome assemblies, or release cycles, see the FAQ.

UCSC: Search Gene (DHH)

Human (Homo sapiens) Genome Browser Gateway - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Human (Homo sapiens) Genome Browser Gat...

genome.ucsc.edu/cgi-bin/hgGateway

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | About Us | Help

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

group	genome	assembly	position	search term	
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr21:33,031,597-33,041,570	DHH	submit

DHH (Homo sapiens desert hedgehog (DHH), mRNA)

Click here to reset the browser user interface settings to their defaults.

track search | add custom tracks | track hubs | configure tracks and display

Human Genome Browser – hg19 assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

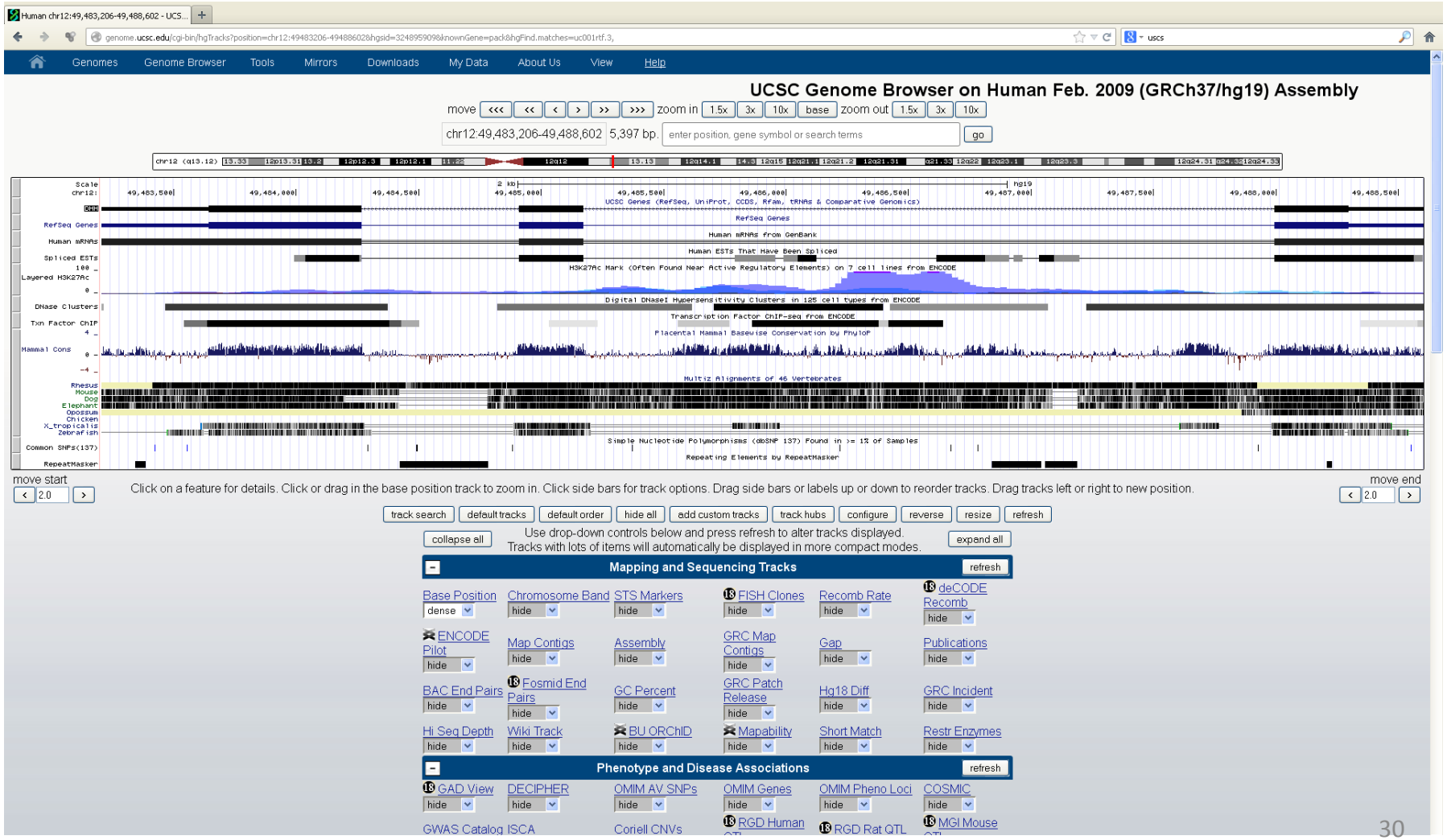
Homo sapiens
(Graphic courtesy of CBSE)

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061;RH80175 15q11;15q13 rs1042522;rs1800370	Displays region between genome landmarks, such as the STS markers RH18061 and RH80175, or chromosome bands 15q11 to 15q13, or SNPs rs1042522 and rs1800370. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, refSeqs, etc.
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps. Includes 100,000 bases on each side as well.
AA205474	Displays region of EST with GenBank accession AA205474 in BRCA1 cancer gene on chr 17
AC008101	Displays region of clone with GenBank accession AC008101
AF083811	Displays region of mRNA with GenBank accession number AF083811
PRNP	Displays region of genome with HUGO Gene Nomenclature Committee identifier PRNP
NM_017414	Displays the region of genome with RefSeq identifier NM_017414
NP_059110	Displays the region of genome with protein accession number NP_059110
pseudogene mRNA	Lists transcribed pseudogenes, but not cDNAs
homeobox caudal	Lists mRNAs for caudal homeobox genes
zinc finger	Lists many zinc finger mRNAs
kruppel zinc finger	Lists only kruppel-like zinc fingers
huntington	Lists candidate genes associated with Huntington's disease
zahler	Lists mRNAs deposited by scientist named Zahler
Evans,J.E.	Lists mRNAs deposited by co-author J.E. Evans

UCSC: Entry page (DHH)



Search for genomic information using identifiers

How can you store genes with a unique name?

- Regular gene names are not suited

- Structured identifiers
- These are different for different databases

NCBI identifiers

- RefSeq:
  - Chromosome: NC_
  - mRNA: NM_
  - Protein: NP_
- Genbank:
  - Many types of IDs
- NCBI gene ID:
  - Number
- OMIM ID:
  - Number
- Pubmed ID:
  - Number

Ensembl identifiers

- ENSG### Ensembl Gene ID
- ENST### Ensembl Transcript ID
- ENSP### Ensembl Peptide ID
- ENSE### Ensembl Exon ID

- For other species than human a suffix is added:

MUS (*Mus musculus*) for mouse: ENSMUSG###
DAR (*Danio rerio*) for zebrafish: ENSDARG###, etc.

Where does all this information come from?

- Submissions (e.g. Sequences)
- Literature
- Curators and contributors
- Automated generation by computer tools
- High-throughput lab screenings
- Individual contributions and large scale contributions

Functional genomics

Single biomolecules		High throughput
DNA	*Sequencing and gene identification*	GENOME
⇩		⇩
RNA	*Sequencing and gene expression*	TRANSCRIPTOME
⇩		⇩
PROTEIN	*Identification and structure determination*	PROTEOME

nu.nl – Sept. 6th 2012

Gezondheid

Gepubliceerd: 6 september 2012 18:42
Laatste update: 6 september 2012 18:59

Deel:

'Wegenkaart' menselijk DNA gepubliceerd

AMSTERDAM – Een gecoördineerde massapublicatie van 30 wetenschappelijke artikelen, waarvan zes in Nature, doet deze week vrijwel alle functies van het menselijk DNA uit de doeken.

Foto: ANP

Elk van onze cellen bevat bijna drie meter aan minutieus opgevouwen DNA. Slechts één procent daarvan doet dienst als gen. Lange tijd was dan ook de vraag: wat is het nut van al het overige, zogenaamde junk-DNA?

Het antwoord daarop wordt deze week gegeven door ENCODE (Encyclopedia of DNA Elements), een internationaal samenwerkingsverband tussen 440 onderzoekers uit 32 laboratoria.

Junk-DNA

De belangrijkste vondst is dat in het menselijk 'junk-DNA' maar liefst vier miljoen genetische schakelaars liggen besloten. Deze schakelaars bepalen of een gen meer of minder actief wordt, zoals de dimmer op een schemerlamp. Het systeem van genetische schakelaars blijkt extreem complex. De computerberekeningen om de data te analyseren duurden bij elkaar opgeteld meer dan 300 jaar.

Human Genome Project

ENCODE is een vervolg op het Human Genome Project, één van de

dimmer op een schemerlamp. Het systeem van genetische schakelaars blijkt extreem complex. De computerberekeningen om de data te analyseren duurden bij elkaar opgeteld meer dan 300 jaar.

Human Genome Project

ENCODE is een vervolg op het Human Genome Project, één van de
grootste wetenschappelijke projecten uit de geschiedenis. Hiermee werd in 2003 het bijna volledige menselijke DNA uitgelezen. ENCODE ging vervolgens op zoek naar alle functionele elementen daarin. Ze vonden dat ten minste 80 procent van ons DNA een biologische functie vervult.

De resultaten vormen een doorbraak in de biologie en wellicht ook de geneeskunde. Experts vergelijken het met de wegenkaart van het menselijk DNA. Het schept enorme potentie voor de ontwikkeling van nieuwe medicatie voor een veelvoud aan ziektes. Al moet daar, gezien de complexiteit, nog wel een slag om de arm worden gehouden.

Door: NU.nl/Kevin Janssen

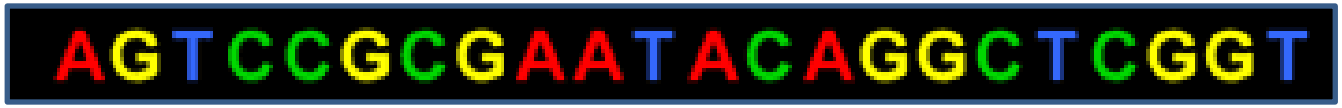
HGP and ENCODE

- We will now discuss these two major projects that contributed a lot of data

- The Humane Genome Project (1990-2003)
  - Sequencing of the human genome
  - Characterizing the genes on the DNA sequence

- The ENCODE project (2003-2012)
  - Focuses on regulatory elements on the DNA

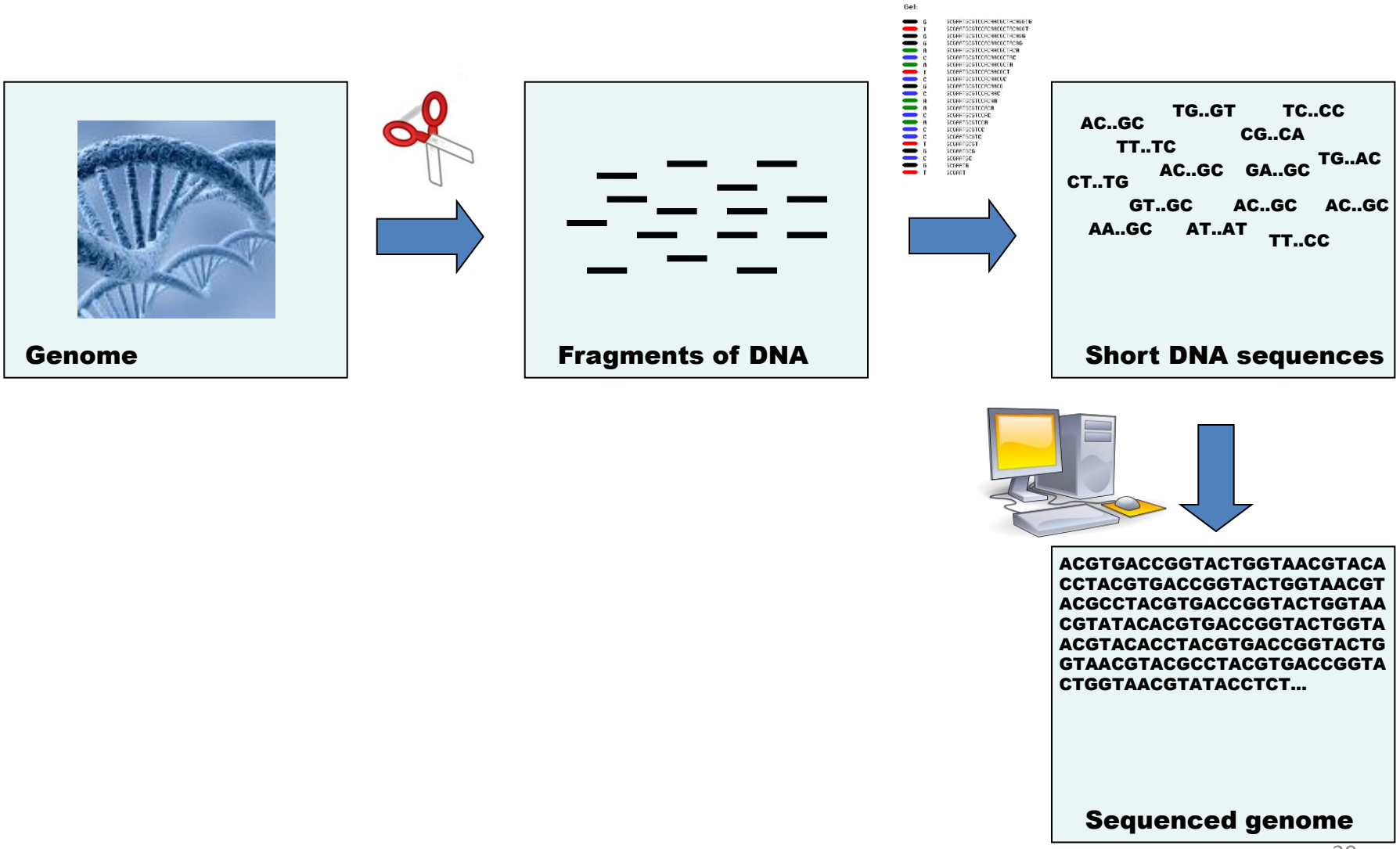
the Human Genome Project



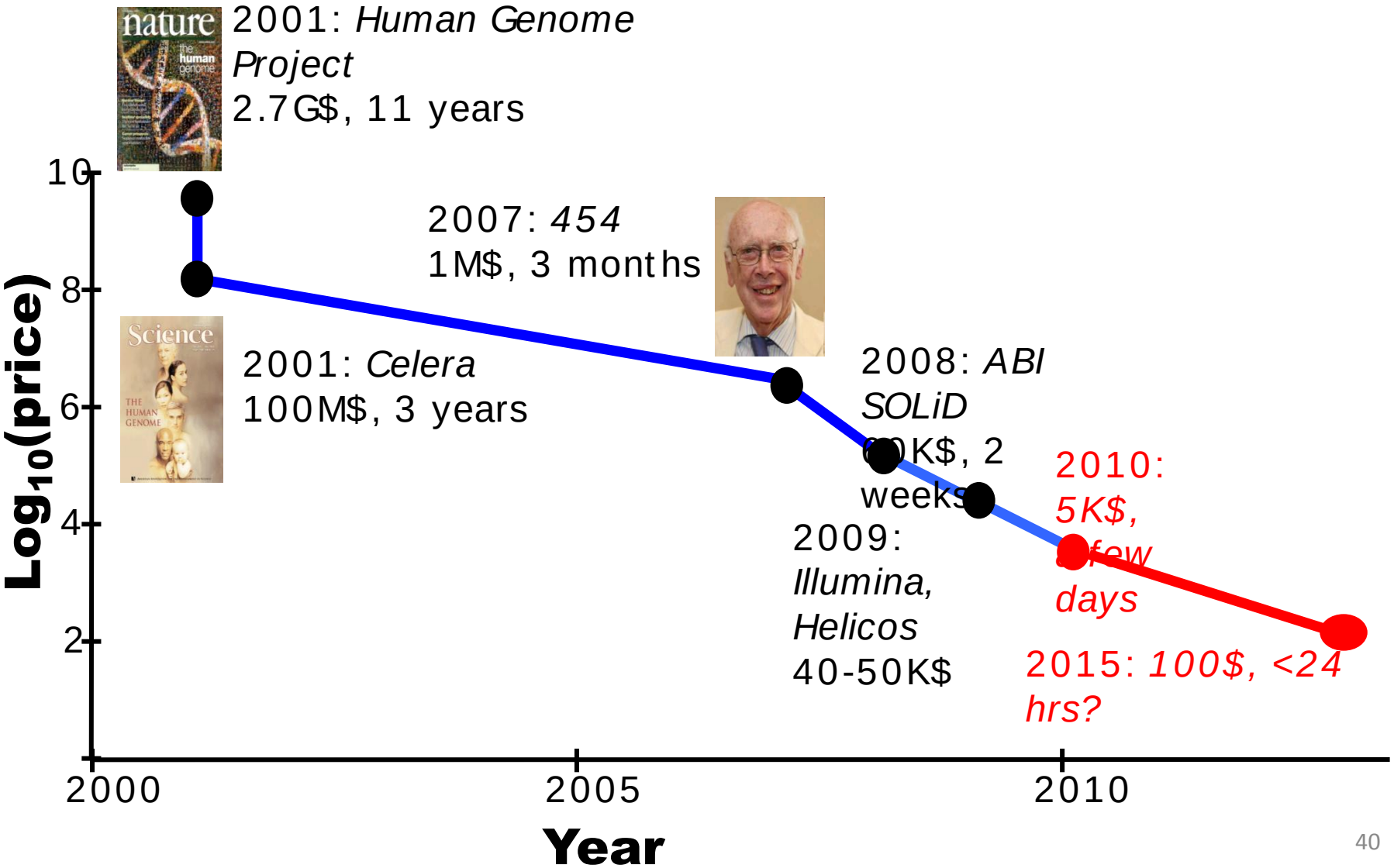
movie

International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. Nature 431, 931-945 (21 October 2004).

Genome sequencing: general principle



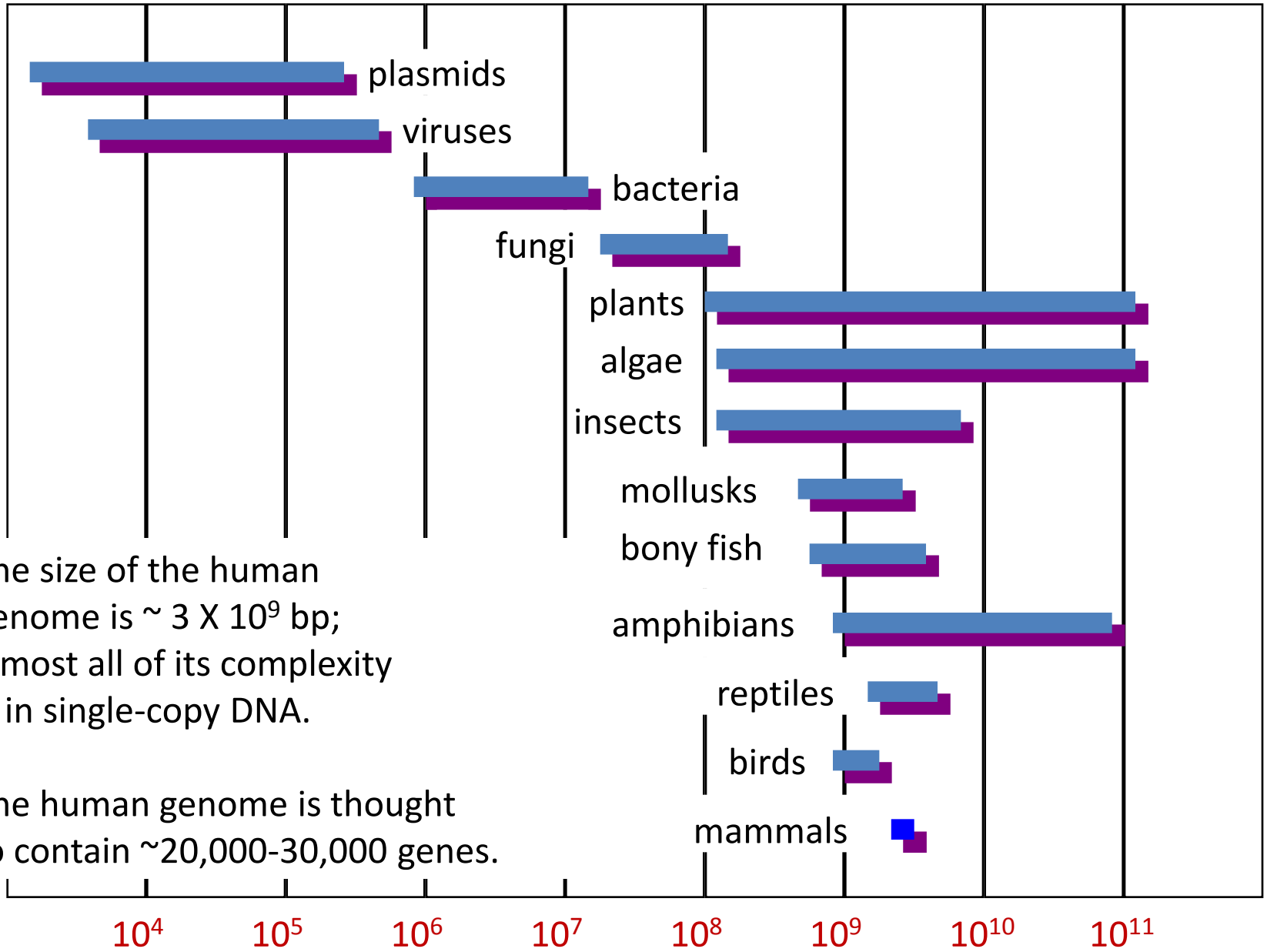
Sequencing the Human Genome



When has a genome been fully sequenced?

- *N*-fold coverage
  - A typical goal is to obtain five to ten-fold coverage.
  - With next-generation sequencing typically even more, like 30-fold coverage
  - Mostly both strands are sequenced
- Finished sequence
  - Usually no gaps in the sequence
  - High quality standard; error rate <0.01%.

Genome sizes in nucleotide base pairs (log scale)



The size of the human genome is ~ 3 X 10^9 bp; almost all of its complexity is in single-copy DNA.

The human genome is thought to contain ~20,000-30,000 genes.

Which genomes are sequenced?

Selection of genomes for sequencing is based on criteria such as:

- genome size (some plants are >>> human genome)
- cost
- relevance to human disease (or other disease)
- relevance to basic biological questions
- relevance to agriculture or other food production

Species List - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Species List

www.ensembl.org/info/about/species.html

Most Visited | Aan de slag | Laatste nieuws | http://ftp.bigcat.unim... | https://webmail.maas... | Log in | Kääntäjä

Collared flycatcher (preview - assembly only)
Ficedula albicollis
FicAlb_1.4

Cow
Bos taurus
UMD3.1

Dog
Canis lupus familiaris
CanFam3.1

Dolphin
Tursiops truncatus
turTru1

Duck (preview - assembly only)
Anas platyrhynchos
duck1

Elephant
Loxodonta africana
loxAfr3

Ferret
Mustela putorius furo
MusPutFur1.0

Fruitfly
Drosophila melanogaster
BDGP5

Mouse
Mus musculus
GRCm38

Mouse Lemur
Microcebus murinus
micMur1

Opossum
Monodelphis domestica
BROAD05

Orangutan
Pongo abelii
PPYG2

Painted Turtle (preview - assembly only)
Chrysemys picta bellii
ChrPicBel3.0.1

Panda
Ailuropoda melanoleuca
ailMel1

Pig
Sus scrofa
Sscrofa10.2

Pig FPC_map (preview - assembly only)
Sus scrofa map
MAP

Tilapia
Oreochromis niloticus
Oreni1.0

Tree Shrew
Tupaia belangeri
TREESHREW

Turkey
Meleagris gallopavo
UMD2

Wallaby
Macropus eugenii
Meug_1.0

Xenopus
Xenopus tropicalis
JGI_4.2

Zebra Finch
Taeniopygia guttata
taeGut3.2.4

Zebrafish
Danio rerio
Zv9

Credits page for species images

Other Metazoa

Additional metazoan genomes (initially insect vectors and nematodes) are available from EnsemblMetazoa

Plants and Fungi

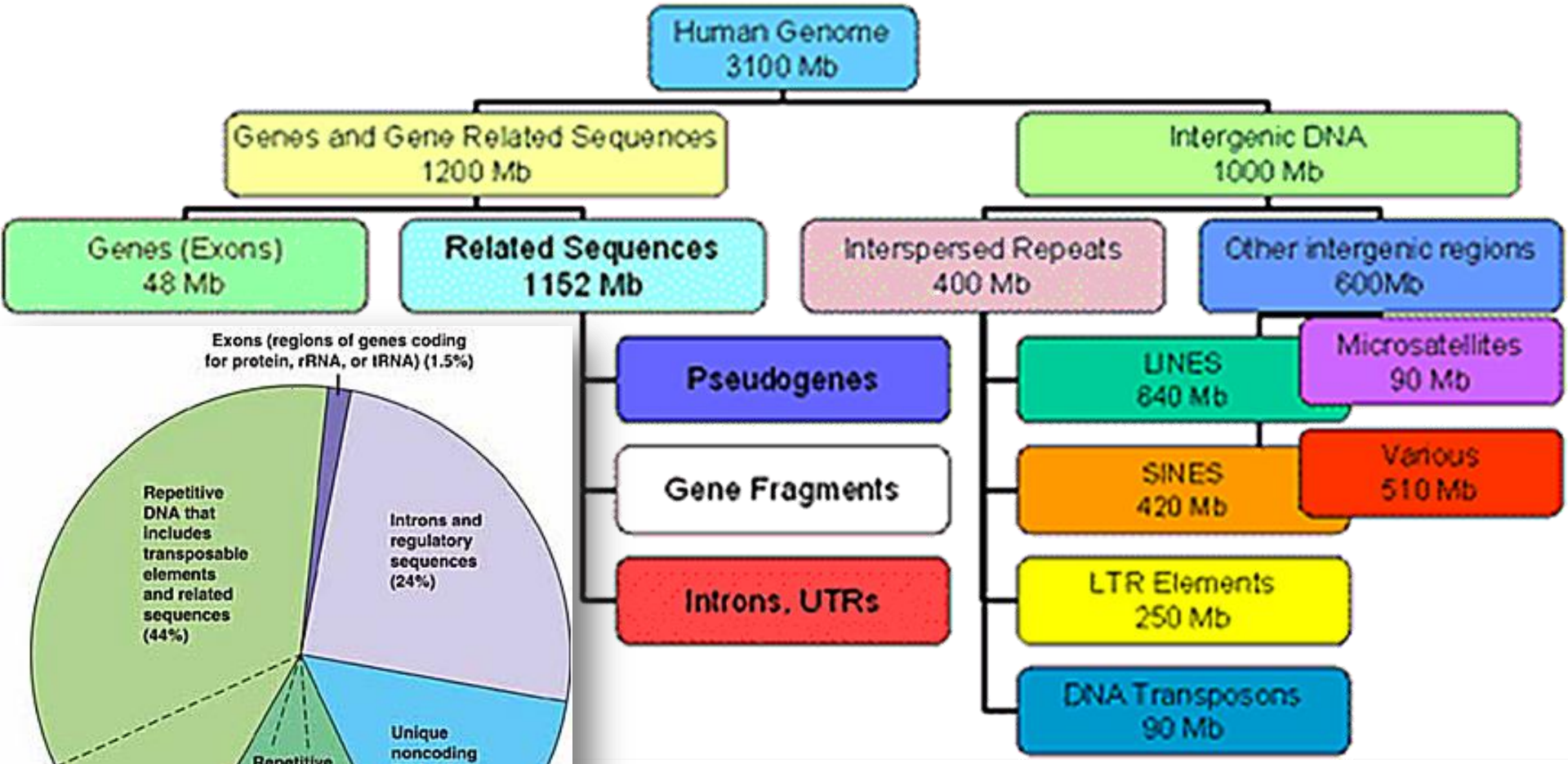
Number of genes

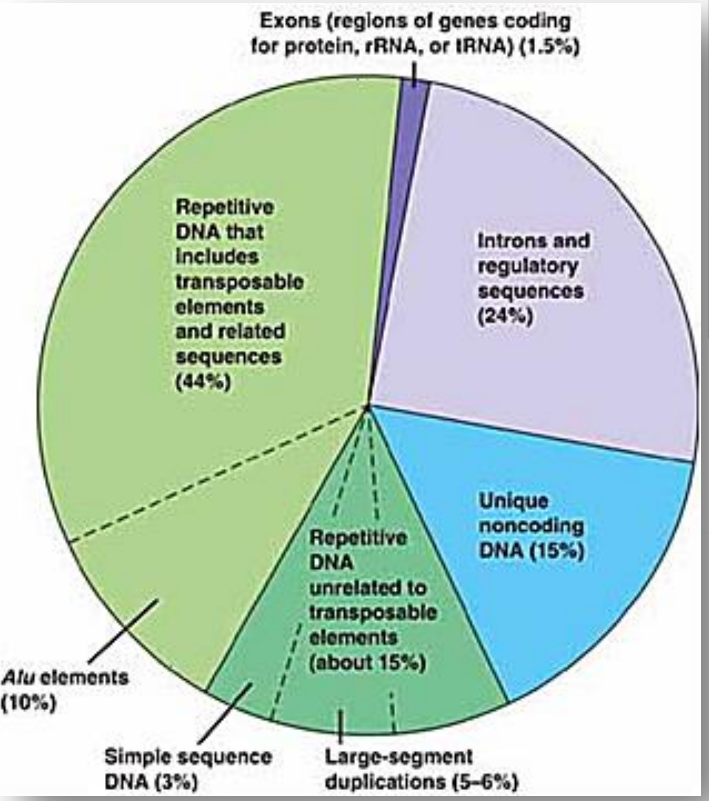
Species and Common Name	Estimated Total Size of Genome (bp)*	Estimated Number of Protein-Encoding Genes*
Saccharomyces cerevisiae (unicellular budding yeast)	12 million	**6,000**
Trichomonas vaginalis	160 million	**60,000**
Plasmodium falciparum (unicellular malaria parasite)	23 million	**5,000**
Caenorhabditis elegans (worm)	95.5 million	**18,000**
Drosophila melanogaster (fruit fly)	170 million	**14,000**
Arabidopsis thaliana (mustard; thale cress)	125 million	**25,000**
Oryza sativa (rice)	470 million	**51,000**
Gallus gallus (chicken)	1 billion	**20,000-23,000**
Canis familiaris (domestic dog)	2.4 billion	**19,000**
Mus musculus (laboratory mouse)	2.5 billion	**30,000**
Homo sapiens (human)	2.9 billion	**20,000-25,000**

Plants and amphibians with huge genomes (not in table) do not have huge amounts of genes

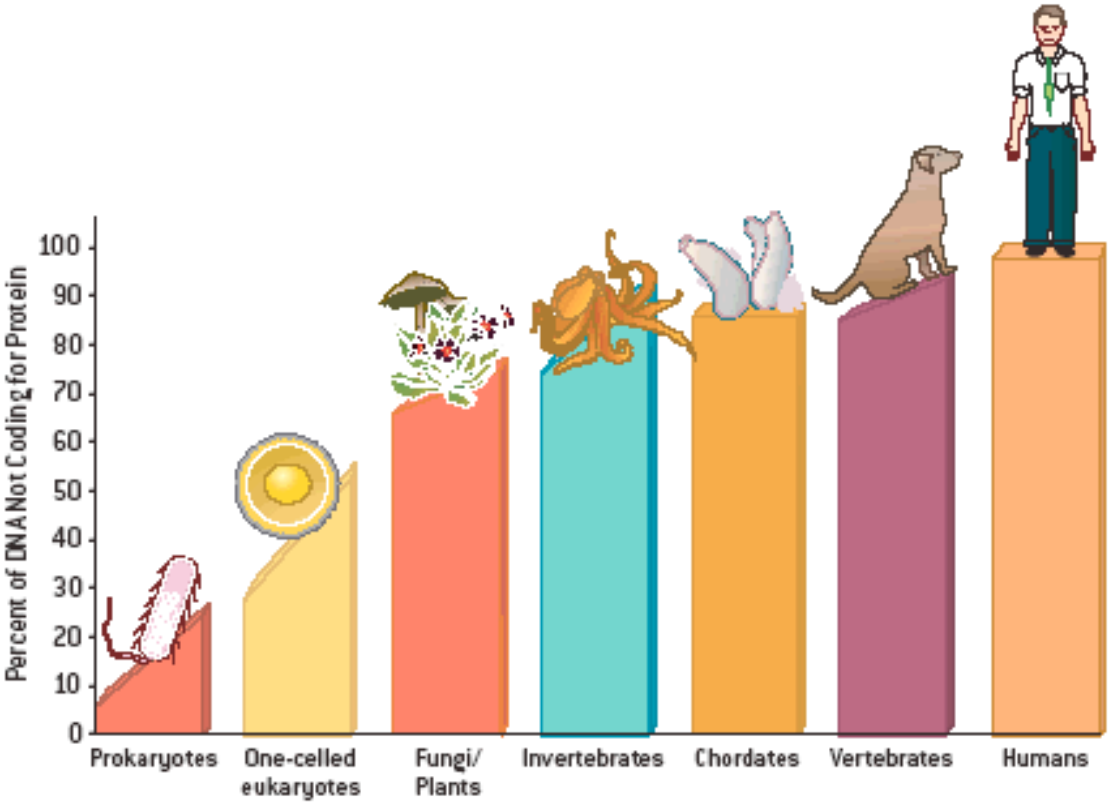
Pray, L. (2008) Eukaryotic genome complexity. Nature Education 1(1)

Organization of the human genome





Non-Protein coding DNA



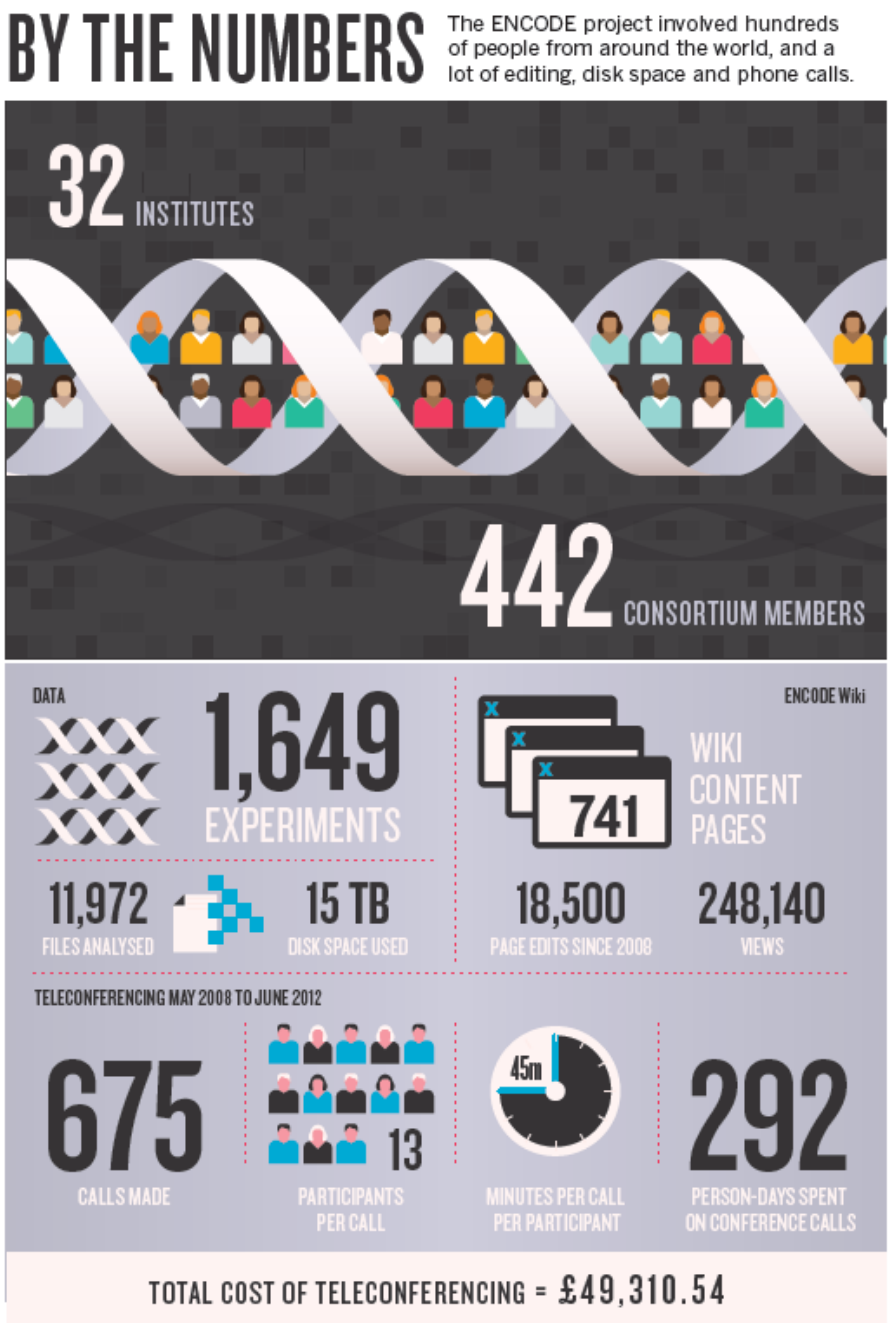
NONPROTEIN-CODING SEQUENCES make up only a small fraction of the DNA of prokaryotes. Among eukaryotes, as their complexity increases, generally so, too, does the proportion of their DNA that does not code for protein. The noncoding sequences have been considered junk, but perhaps it actually helps to explain organisms' complexity.

The ENCODE Project: ENCyclopedia Of DNA Elements

A public research consortium

Launched: September 2003, upgraded to the entire genome September 2007.

Goal: to carry out a project to identify all the functional elements in the human genome sequence.

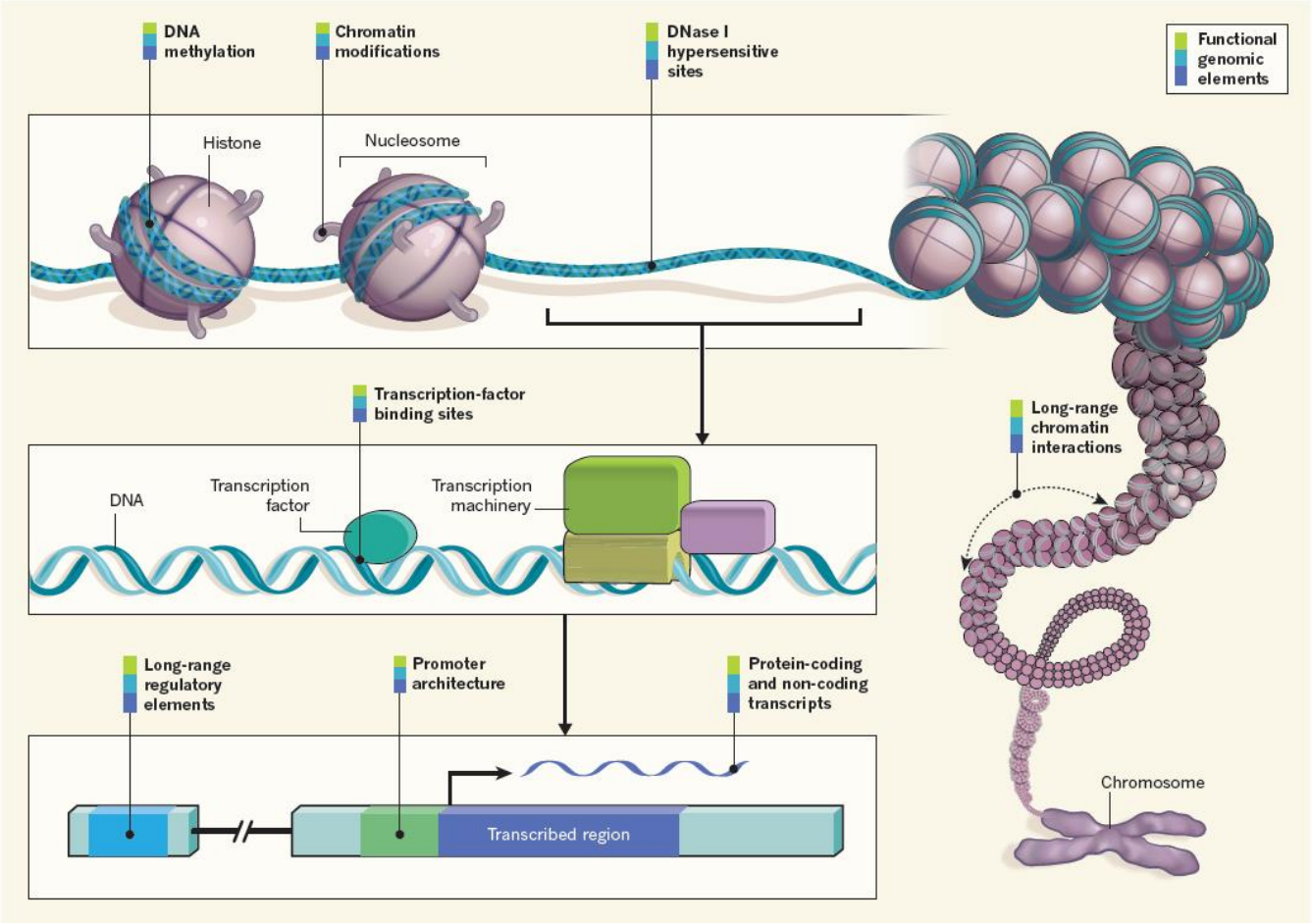


Understanding of the human genome is far from complete. We are missing knowledge on:

1. non-coding RNA
2. Alternatively spliced transcripts
3. Regulatory sequences

The making of ENCODE: Lessons for big-data projects. Birney E.
Nature. 2012 Sep 6;489(7414):49-51

Data retrieved from ENCODE project



Genomics: ENCODE explained. Ecker JR, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, Segal E Nature. 2012 Sep 6;489(7414):52-5.

ENCODE data in Ensembl

Gene Ontology

- Built for a very specific purpose:

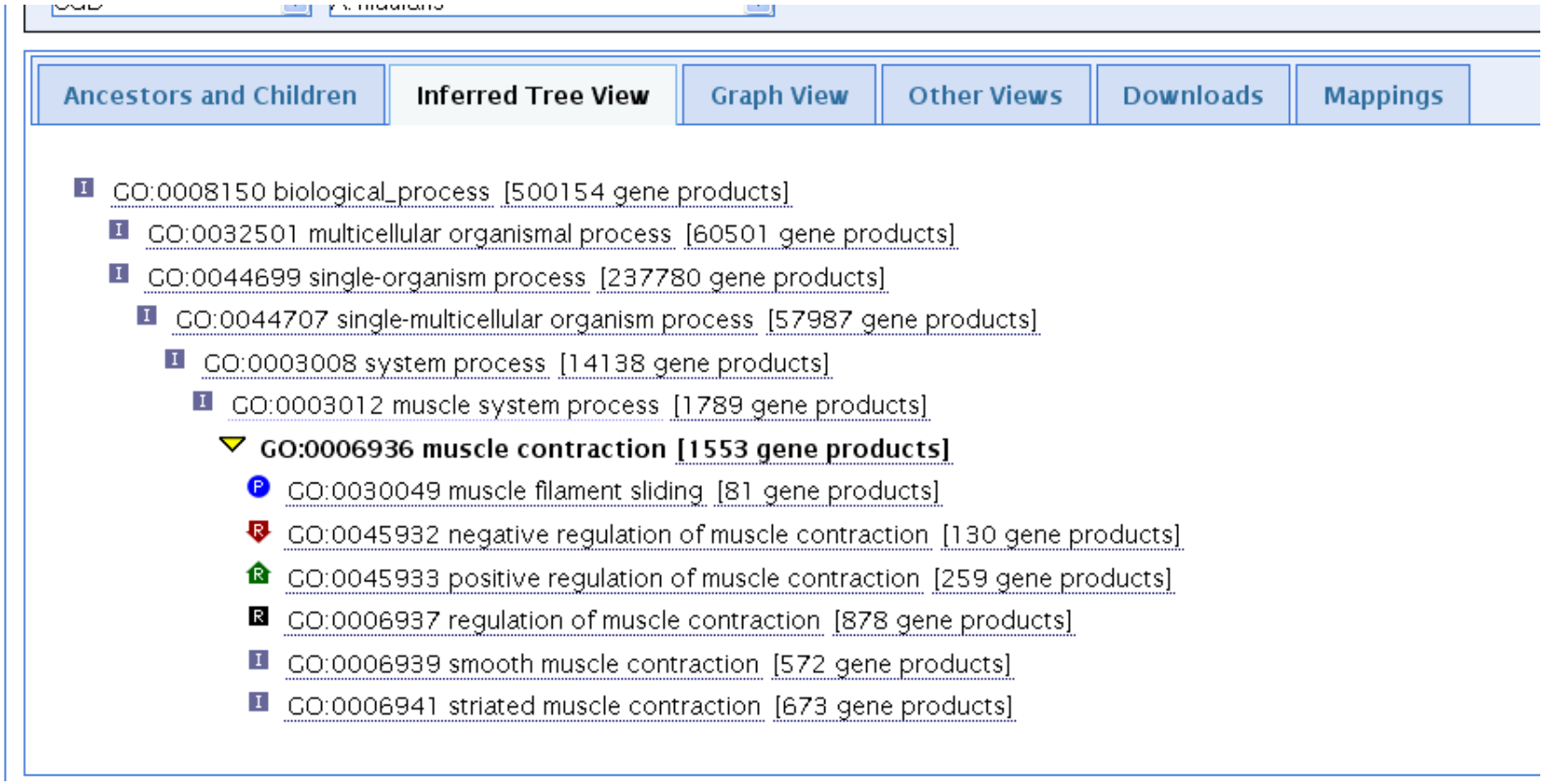
"annotation of genes and proteins in genomic and protein databases"

- Applicable to all species

The 3 Gene Ontologies

- **Molecular Function** = elemental activity/task
  - the tasks performed by individual gene products; examples are ***carbohydrate binding*** and ***ATPase activity***
- **Biological Process** = biological goal or objective
  - broad biological goals, such as ***mitosis*** or ***purine metabolism***, that are accomplished by ordered assemblies of molecular functions
- **Cellular Component** = location or complex
  - subcellular structures, locations, and macromolecular complexes; examples include ***nucleus***, ***telomere***, and ***RNA polymerase II holoenzyme***

GO muscle contraction – tree view



Gene products - Striated muscle contraction (GO:0006941)

striated muscle contraction

Term associations ↓ Term information → Term lineage → External references →

Gene Product Associations to striated muscle contraction ; GO:0006941 and children

Download all association information in: gene association format RDF-XML

Filter associations displayed

Filter by Gene Product — Gene Product Type: All, complex, gene, gene product; Data source: All, ASAP, AspGD, CGD; Species: All, Arabidopsis thaliana, Aspergillus fumig..., Aspergillus fumig...

Filter by Association — Evidence Code: All, IBA, IKR, IRD

View associations: All / Direct associations

Set filters

Remove all filters

1 2 3 4 5 6 7 8 9 ... 17 View all results

striated muscle contraction ; GO:0006941 [show def] [view in tree]

Symbol, full name		Information	Qualifier	Evidence	Reference	Assigned by
Aldoa aldolase A, fructose-bisphosphate	15 associations	**protein** from *Mus musculus*		ISO With UniProtKB:P04075	MGI:MGI:4834177	MGI
Aldoa aldolase A, fructose-bisphosphate	27 associations BLAST	**gene** from *Rattus norvegicus*		ISO With RGD:735815	RGD:1624291	RGD
ALDOA Fructose-bisphosphate aldolase	12 associations BLAST	**protein** from *Bos taurus*		IEA With Ensembl:ENSP00000378669	GO REF:0000019	Ensembl (via UniProtKB)
ALDOA Fructose-bisphosphate aldolase A	29 associations BLAST	**protein** from *Homo sapiens*		IMP	PMID:14615364	BHF-UCL (via UniProtKB)
Arg2 arginase 2	35 associations BLAST	**gene** from *Rattus norvegicus*		IEA With Ensembl:ENSMUSP00000021550 ISO With RGD:736823	RGD:1600115 RGD:1624291	Ensembl (via RGD) RGD
Arg2 arginase type II	13 associations BLAST	**protein** from *Mus musculus*		IMP	PMID:16537391	MGI

Anatomy of a GO term

id: GO:0006094	unique GO ID
name: gluconeogenesis	term name
namespace: process	ontology
def: The formation of glucose from noncarbohydrate precursors, such as pyruvate, amino acids and glycerol. [http://cancerweb.ncl.ac.uk/omd/index.html]	definition
exact_synonym: glucose biosynthesis	synonym
xref_analog: MetaCyc:GLUCONEO-PWY	database ref
is_a: GO:0006006 is_a: GO:0006092	parentage

No GO Areas

- GO covers ‘normal’ functions and processes
  - No pathological processes
  - No experimental conditions
- NO evolutionary relationships
- NOT a system of nomenclature

Searching and Browsing GO

- AmiGO
  - http://www.godatabase.org
- Downloads
  - http://www.godatabase.org/dev/database/
  - XML or as a MySQL database dump
- http://www.geneontology.org/GO.tools.annotation.shtml
  - Annotate gene by sequence similarity.

Practical session

- Ensembl tutorials
- Ensembl genome browser

- Several NCBI databases
  - Gene
  - OMIM

- Gene Ontology