



Dr. Susan Steinbusch-Coort
susan.coort@maastrichtuniversity.nl

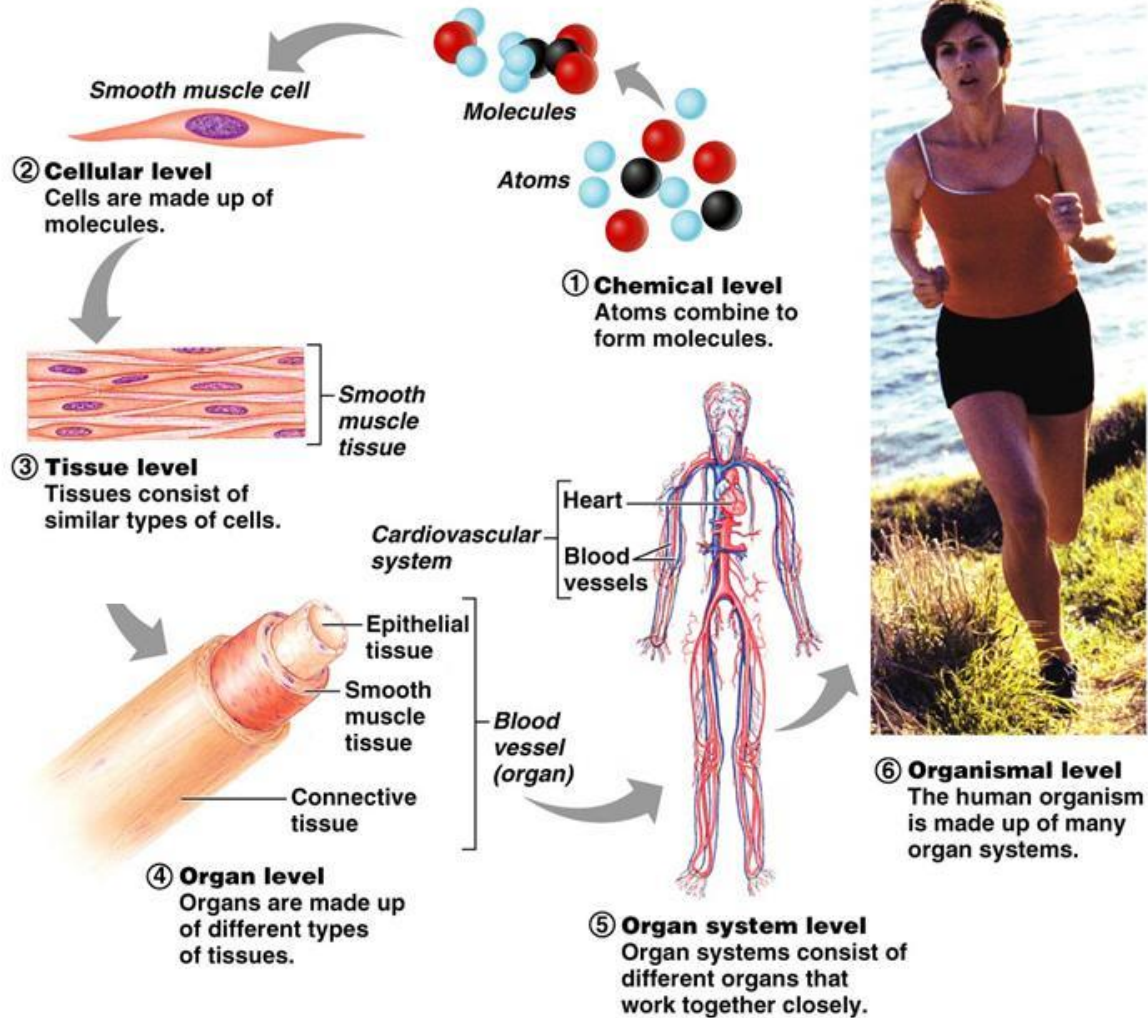
What happens with the human body when you are running?



Organ systems work together

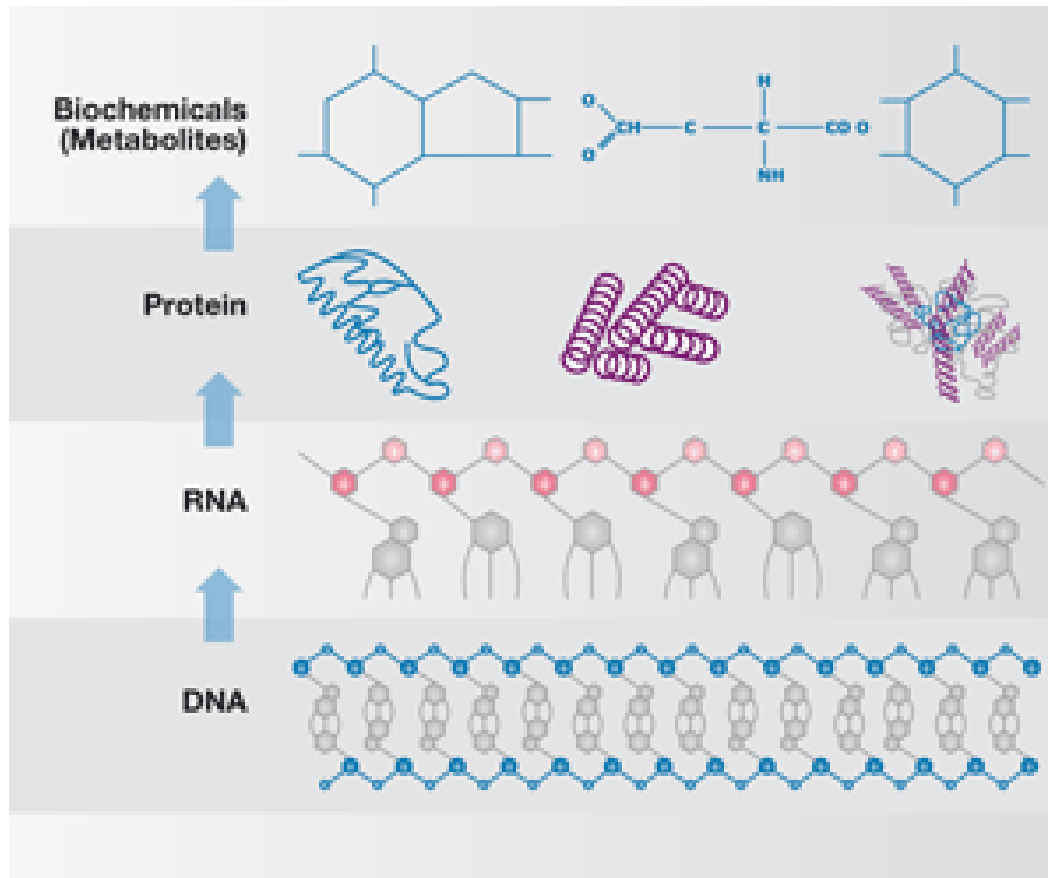
- Skeletal system- supports the skeleton
- Muscular system - pulls on the bones to enable you to move
- Respiratory system - makes sure your muscles have enough oxygen for respiration
- Circulatory system- provides oxygen and glucose to the skeletal muscle cells

Human body structure

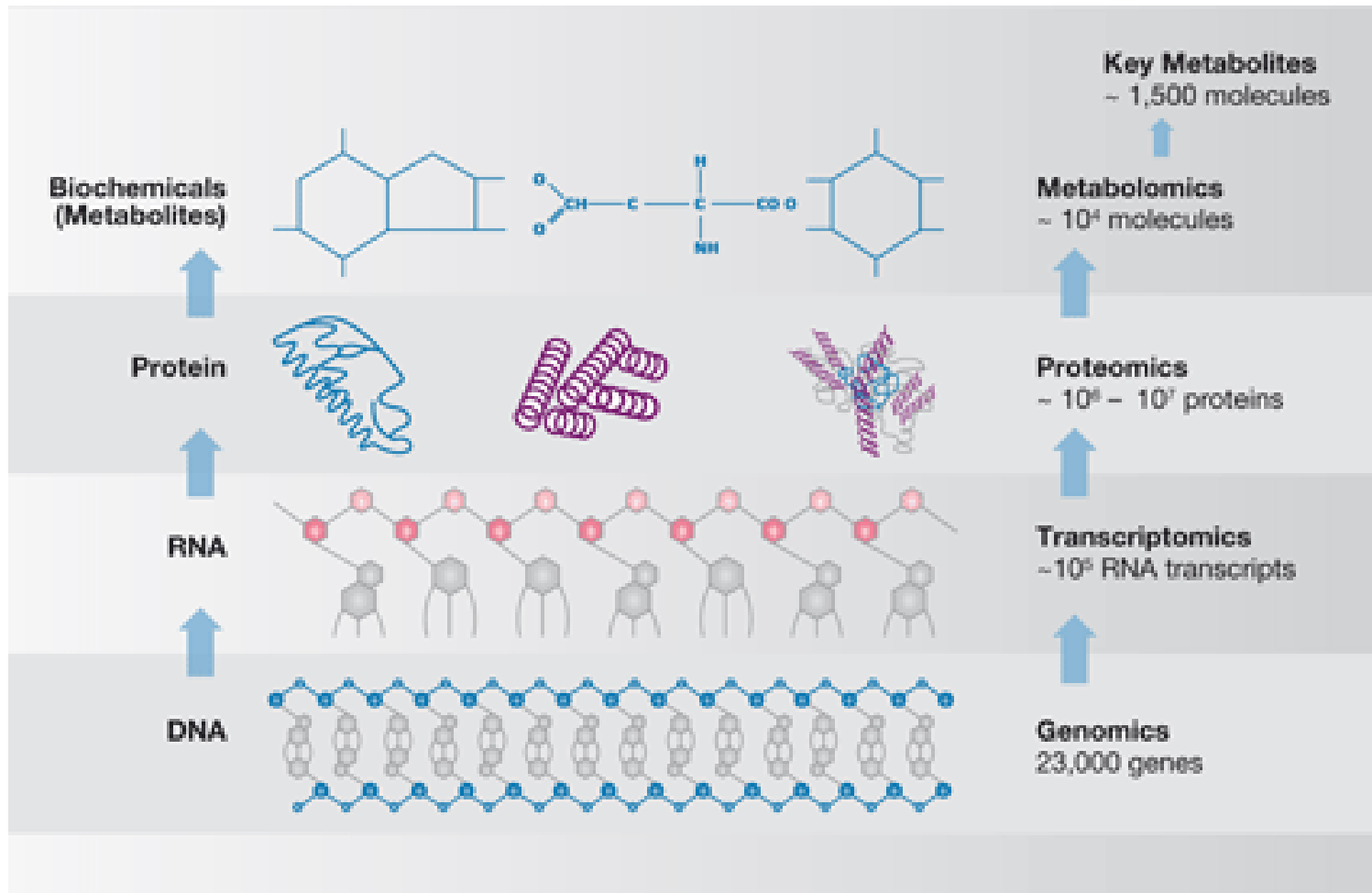


(Bio)Molecules

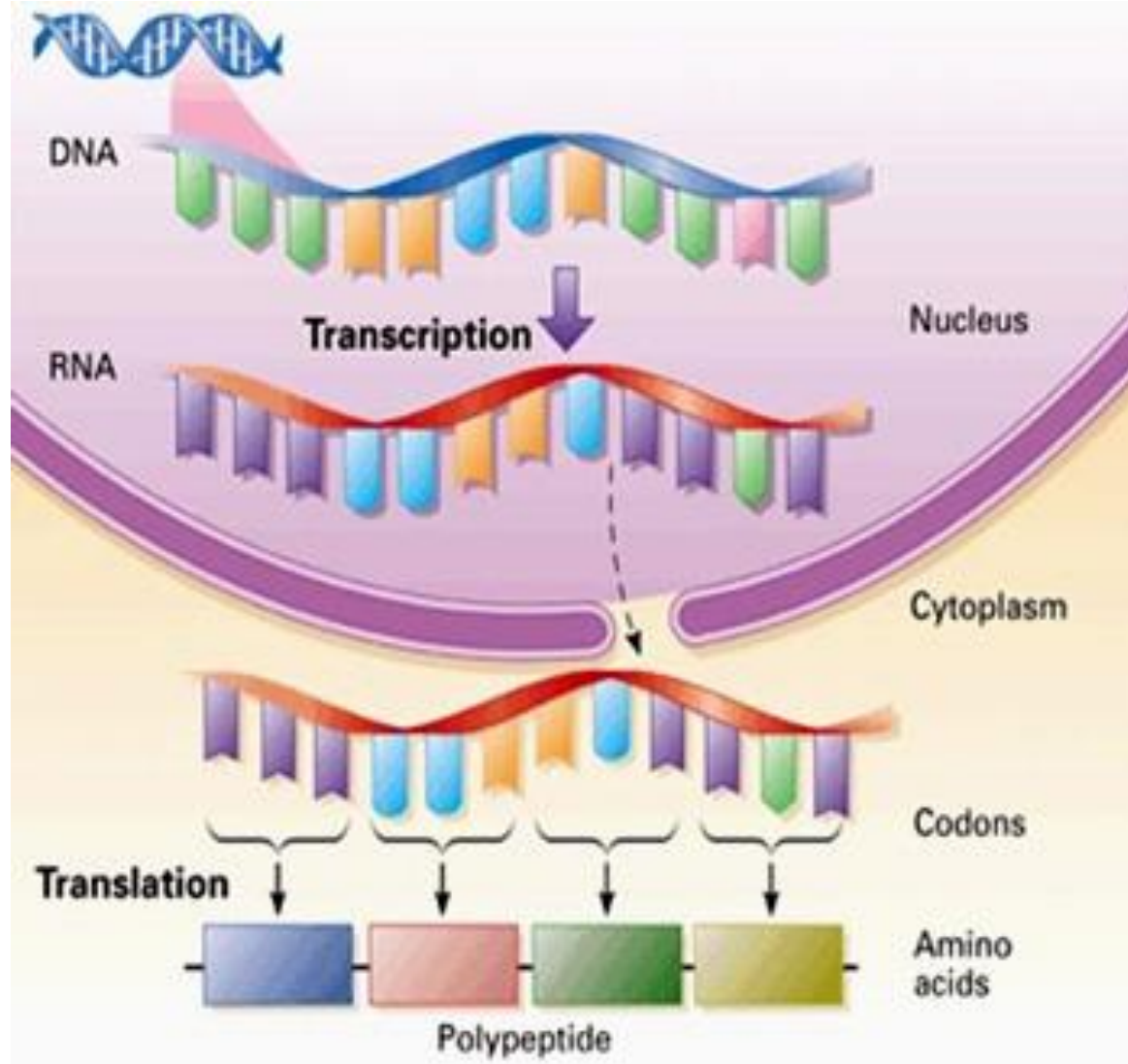
Individual players are important



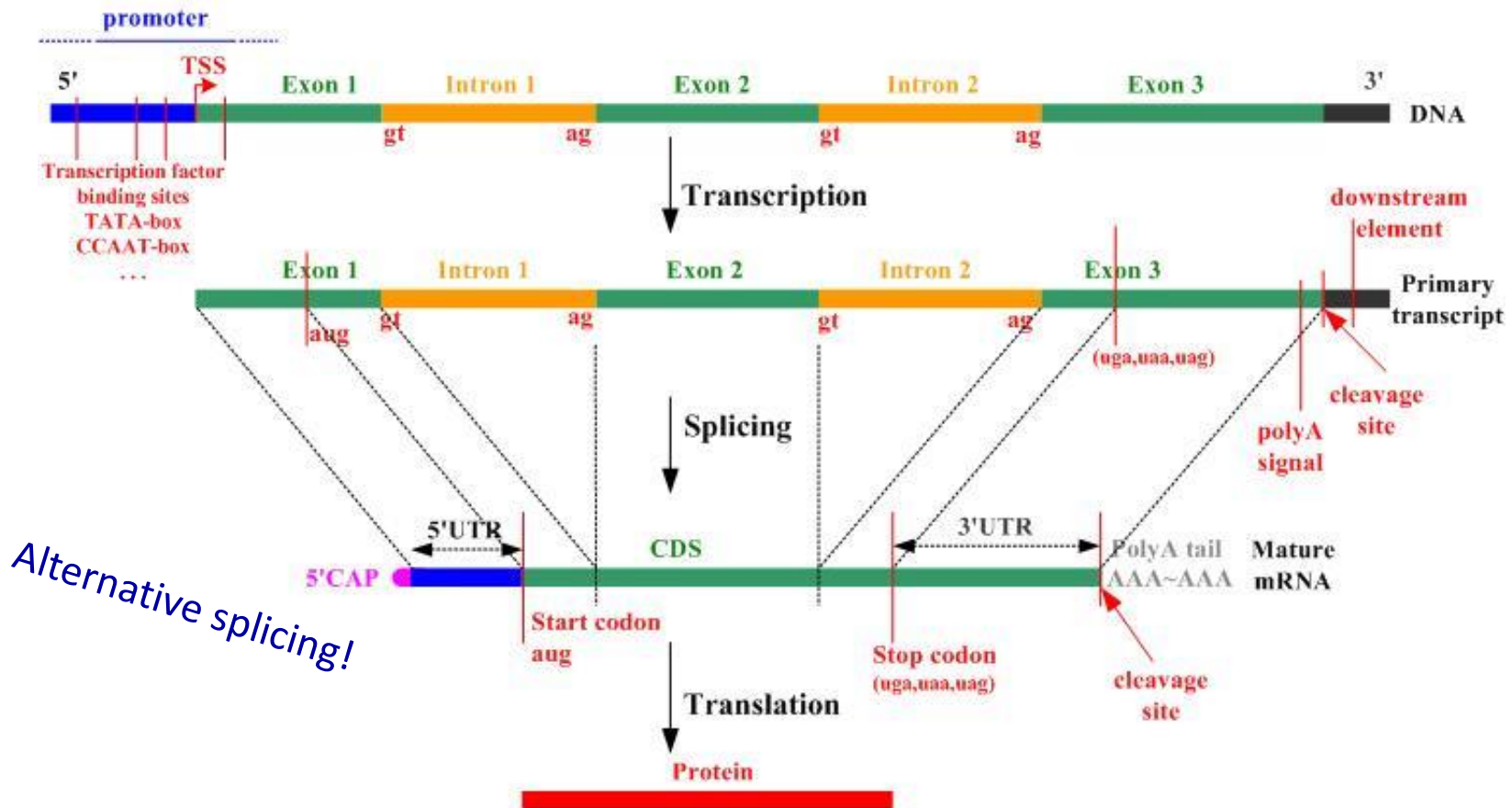
Heaps of knowledge on biomolecules online available.



Protein synthesis



Gene structure



CDS = Coding DNA Sequence
UTR = UnTranslated region

GOAL

To understand biological sequence databases

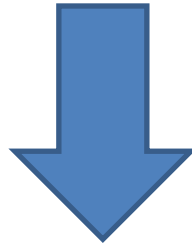
- Which biological sequence databases are available?
- How can you find information in these databases?
- What is the content of the databases?
- What is Gene Ontology?
- Two projects aimed at deciphering the content of the human genome, the human genome project & ENCODE.

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

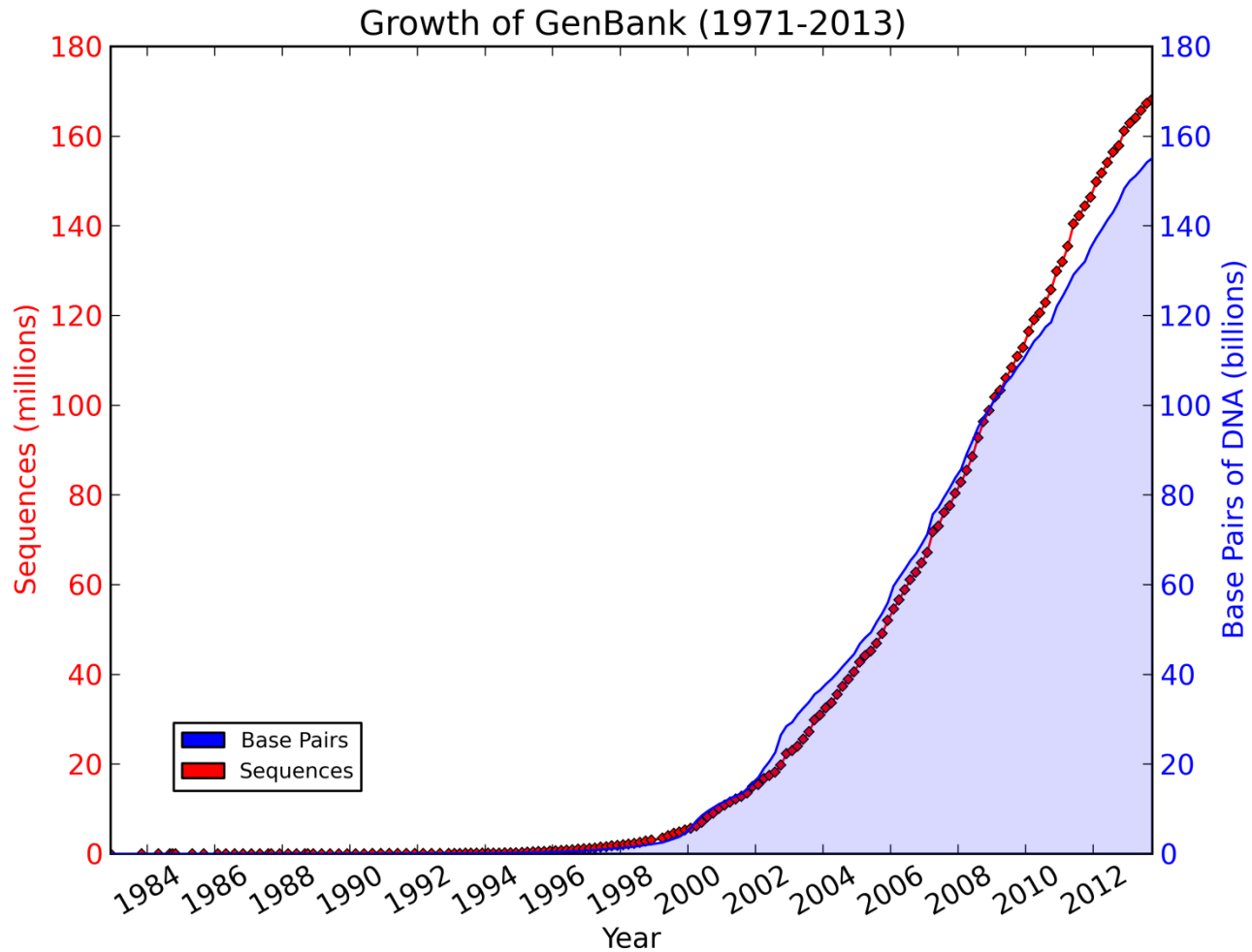
- UniProt

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>

UCSC Genome Browser on Human May 2004 Assembly

position/search chr7:127,471,196-127,495,720

chr7 (chr7:1)

Base Position 127470000 127480000 127490000 127500000

STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps

UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA

RefSeq Genes

Consensus CDS

Accession Numbers

Human mRNAs from GenBank

Spliced ESTs

Conservation

Click on a feature for details. Click to zoom in around cursor. Click on for track-specific options

Use drop-down controls below and press refresh to update tracks. Tracks with items will automatically be displayed

Mapping and Sequencing Tracks

Base Position Chromosome STS Markers RGID

dense hide dense hide

<http://genome.ucsc.edu/>

NCBI Genome

chromosome(s) assembly All Find

Help FTP Map Viewer home

Genomes genome view

BLAST search the human genome

2 4 5 6 7 8 9 10 11 12 13

16 17 18 19 20 21 22 X Y MT

Ensembl

Search Ensembl

Search: All species for

Go

e.g. human gene BRCA2 or rat X:100000..200000 or insulin

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes (Log in to customize this list)

Human NCBI36

Mouse NCBI37

Zebrafish ZFISH7

All genomes

-- Select a species --

New to Ensembl?

Did you know you can:

- Add custom tracks
- Upload your own data
- Search for a DNA or protein sequence
- Fetch only the data you want
- Download our databases via FTP
- Mine Ensembl with BioMart

The new Ensembl website

We've made some changes to our site, to make it faster and easier to use. Find out more about what we've changed and why!

What's New in Release 52 (9 December 2008)

- Homo sapiens core database (Human)
- Gorilla 2x assembly and genebuild (Gorilla)
- ncRNA update (multiple species)
- Mus musculus core (Mouse)
- Cow otherfeatures (Cow)

Ensembl

<http://www.ensembl.org/>

NCBI homepage

The screenshot shows the NCBI homepage with a search bar at the top. A dropdown menu is open, listing various databases and resources. The 'All Databases' option is highlighted with a red circle. Below the search bar, there is a navigation menu on the left with categories like 'NCBI Home', 'Resource List (A-Z)', and 'All Resources'. The main content area features a 'Welcome to NCBI' message and a list of 'Popular Resources' including PubMed, Bookshelf, and BLAST. A 'NCBI Facebook page' banner is also visible, along with a 'NCBI Announcements' section.

NCBI Resources How To

Sign in to NCBI

Genome

All Databases

PubMed

Protein

Nucleotide

GSS

EST

Structure

Genome

Assembly

BioProject

BioSample

BioSystems

Books

Conserved Domains

Clone

dbGaP

dbVar

Epigenomics

Gene

GEO DataSets

Welcome to NCBI

Center for Biotechnology Information advances science and health by providing access to biomedical information.

[NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

[Analyze data using NCBI software](#)

[Get NCBI data or software](#)

[Learn how to accomplish specific tasks at NCBI](#)

[Submit data to GenBank or other NCBI databases](#)

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Popular Resources

[PubMed](#)

[Bookshelf](#)

[PubMed Central](#)

[PubMed Health](#)

[BLAST](#)

[Nucleotide](#)

[Genome](#)

[SNP](#)

[Gene](#)

[Protein](#)

[PubChem](#)

NCBI Announcements

Now Available: NCBI Insights Blog!

28 Jan 2013

NCBI has just released a new blog called *NCBI Insights*. Blog posts will provide an

16 Jan 2013

Spaces are still available for the free,

06 Sep 2012

New version of Genome Workbench available

An integrated, downloadable application

16

[More...](#)

NCBI Global Cross-database search

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

Genes

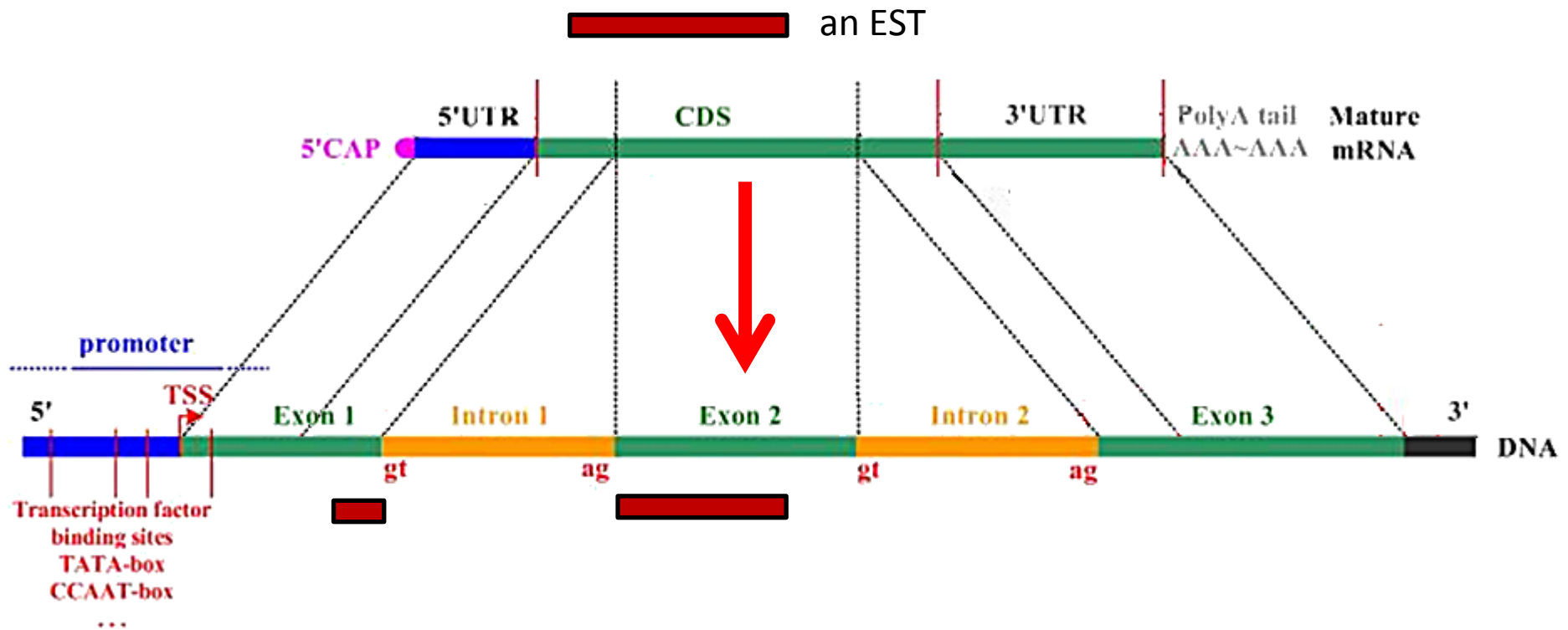
EST	expressed sequence tag sequences
Gene	collected information about gene loci
GEO Data Sets	functional genomics studies
GEO Profiles	gene expression and molecular abundance profiles
HomoloGene	homologous gene sets for selected organisms
PopSet	sequence sets from phylogenetic and population studies
UniGene	clusters of expressed transcripts

Genome	genome sequencing projects by organism
GSS	genome survey sequences
Nucleotide	DNA and RNA sequences
Probe	sequence-based probes and primers
SNP	short genetic variations
SRA	high-throughput DNA and RNA sequence read archive
Taxonomy	taxonomic classification and nomenclature catalog

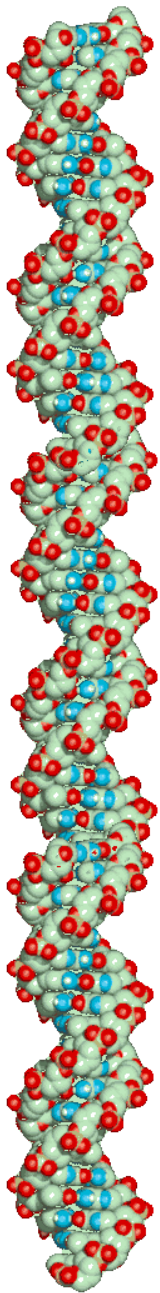
UniGene

- 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

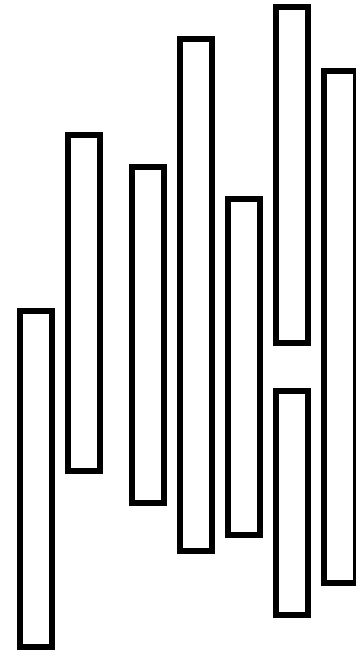
Map mRNA (EST) back to DNA



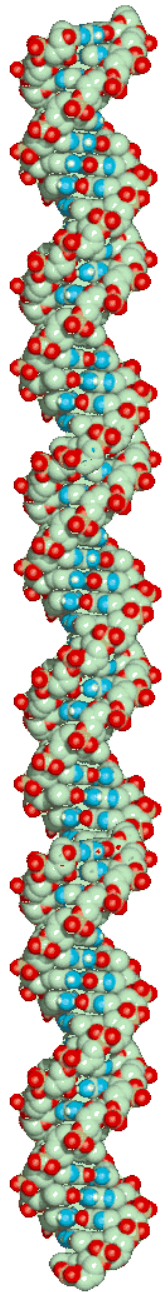
EST clusters



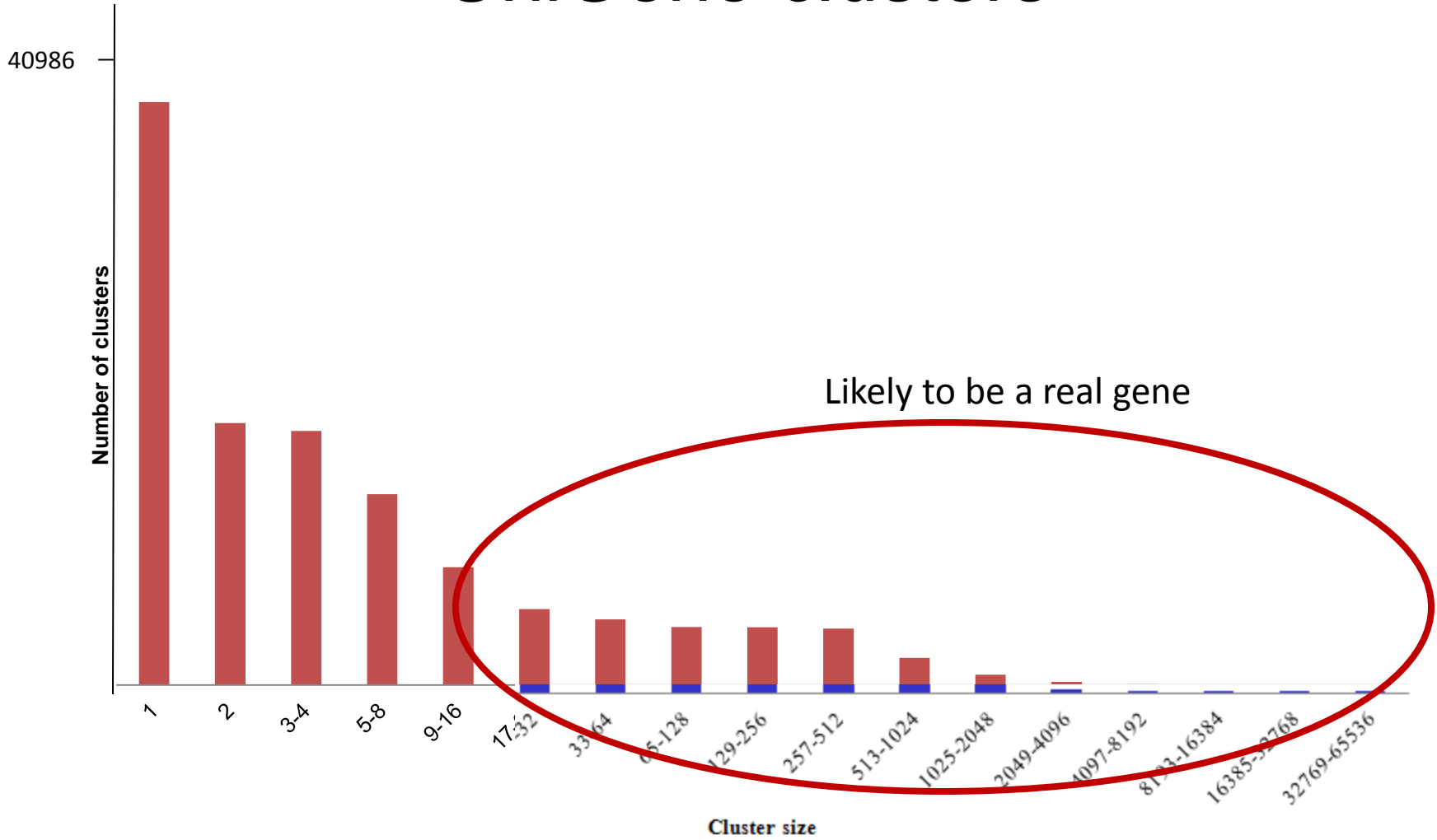
**This is a gene with
1 EST associated;
the cluster size is 1**



**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

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](#)

Genomic regions, transcripts, and products

Genomic Sequence: [NC_000012.11:49M..49M \(7.0Kbp\) C](#) Find on Sequence:

Go to [reference sequence details](#)

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

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)

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

Protein Alignments

Protein multiple alignment, pairwise similarity scores and evolutionary distances.

Show Multiple Alignment

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

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 PubMed Search

US National Library of Medicine National Institutes of Health 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



BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Login/Register

Search all species...

Search: All species for

e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes



Human
GRCh37



Mouse
GRCm38



Zebrafish
Zv9

★ [Log in to customize this list](#)

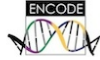
All genomes

— Select a species —

[View full list of all Ensembl species](#)

Other species are available in [Ensembl Pre!](#) and [EnsemblGenomes](#)

ENCODE data in Ensembl



Variant Effect Predictor



Gene expression in different tissues



Find SNPs and other variants for my gene

```
GTTATACATTC  
CCTRAAAGTCTT  
CTTCTAAATTC  
GACACATTTCC
```

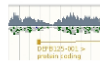
Retrieve gene sequence

```
GGCGGATTCGAGTGG  
GGGCTTGTGGCCGAGC  
GGGCTGTGCTGGGCTT  
AGGGGAGATTTGTGG  
CACCTCTGAGAGGTTT  
CCCACTCCAGCGGCG
```

Compare genes across species



Use my own data in Ensembl



Learn about a disease or phenotype



What's New in Release 74 (December 2013)

- [ncRNA secondary structure now displayed on the Gene Summary page](#)
- [New matrix configuration for RNASeq models](#)
- [New species: sheep \(*Ovis aries*\), cave fish \(*Astyanax mexicanus*\) and spotted gar \(*Lepisosteus oculatus*\)](#)
- [Updated patches for the human assembly \(GRCh37.p13\) and mouse assembly \(GRCm38.p2\)](#)

[Full details of this release](#)

[More release news on our blog →](#)

Latest blog posts

- 09 Jan 2014: [What's coming in Ensembl release 75](#)
- 01 Jan 2014: [Computing Ensembl's New Regulatory Annotation](#)
- 26 Dec 2013: [The New Ensembl Regulatory Annotation](#)

[Go to Ensembl blog →](#)

Did you know...?



It's free- take our [browser workshop](#) online!



Ensembl is a joint project between [EMBL - EBI](#) and the [Wellcome Trust Sanger Institute](#) to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

Ensembl receives major funding from the Wellcome Trust. Our [acknowledgements page](#) includes a list of additional current and previous funding bodies.



Ensembl release 74 - December 2013 © [WTSI](#) / [EBI](#)

[About Ensembl](#) | [Privacy Policy](#) | [Contact Us](#)

[Permanent link](#) - [View in archive site](#)

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

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

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

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.

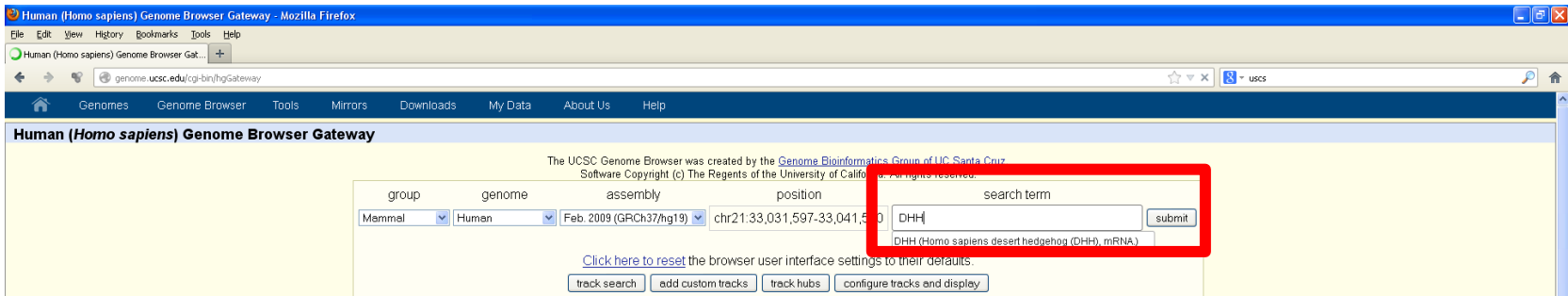
The UCSC, Ensembl, and NCBI browser and annotation groups have established a common set of minimum requirements for public display of genome data made available after Spring 2009, described [here](#).

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 Gateway

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) chr21:33,031,597-33,041,500 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 end 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.

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 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.

15q11;15q13

rs1042522;rs1800370

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



Homo sapiens
(Graphic courtesy of [CSDB](#))

UCSC: Entry page (DHH)

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

chr12:49,483,206-49,488,602 enter position, gene symbol or search terms go

chr12 (410.12) hg19

Scale chr12: 49,483,500 49,484,000 49,484,500 49,485,000 49,485,500 49,486,000 49,486,500 49,487,000 49,487,500 49,488,000 49,488,500

RefSeq Genes UCSC Genes (RefSeq, Unifrot, CCDS, Rfam, tRNAs & Comparative Genomics) RefSeq Genes

Human mRNAs Human mRNAs from GenBank

Spliced ESTs Human ESTs That Have Been Spliced

Layered H3K27Ac H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

DNase Clusters Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE

Tbx Factor ChIP Transcription Factor ChIP-seq from ENCODE

Mammal Cons Placental Mammal Baseuse Conservation by PhyloP

PheSus Mouse Dog Elephant Opossum Chimpanzee X_Tropid Callis Zebrafish Multi Alignments of 46 Vertebrates

Common SNPs (137) Single Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples

RepeatMasker Repeating Elements by RepeatMasker

move start < 2.0 > Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. move end < 2.0 >

track search default tracks default order hide all add custom tracks track hubs configure reverse resize refresh

collapse all Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be shown in more compact modes. expand all

Mapping and Sequencing Tracks refresh

Base Position	Chromosome Band	STS Markers	FISH Clones	Recomb Rate	deCODE
dense	hide	hide	hide	hide	Recomb
hide					hide
ENCODE	Map Contigs	Assembly	GRC Map Contigs	Gap	Publications
Pilot	hide	hide	hide	hide	hide
hide					
BAC End Pairs	Fosmid End Pairs	GC Percent	GRC Patch Release	Hg18 Diff	GRC Incident
hide	hide	hide	hide	hide	hide
Hi Seq Depth	Wiki Track	BU ORCID	Mapability	Short Match	Restr Enzymes
hide	hide	hide	hide	hide	hide

Phenotype and Disease Associations refresh

GAD View	DECIPHER	OMIM AV SNPs	OMIM Genes	OMIM Pheno Loci	COSMIC
hide	hide	hide	hide	hide	hide
GWAS Catalog ISCA		Coriell CNVs	RGD Human	RGD Rat QTL	MGI Mouse

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

- ENS**G**### Ensembl **Gene** ID
 - EN**S**T### Ensembl **Transcript** ID
 - ENS**P**### Ensembl **Peptide** ID
 - ENS**E**### Ensembl **Exon** ID
-
- For other species than human a suffix is added:

MUS (*Mus musculus*) for mouse: ENS**MUS**G###

DAR (*Danio rerio*) for zebrafish: ENS**DAR**G###, 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

DNA



RNA



PROTEIN

*Sequencing and gene
identification*

*Sequencing and gene
expression*

*Identification and
structure determination*

High throughput

GENOME



TRANSCRIPTOME



PROTEOME

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

nu.nl – Sept. 6th 2012

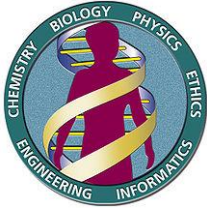
dimmer op een schemerlamp. Het systeem van genetische schakelaars blijkt extreem complex. De computerberekeningen om de data te analyseren duurden bij elkaar opgeteld meer dan 300 jaar.

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

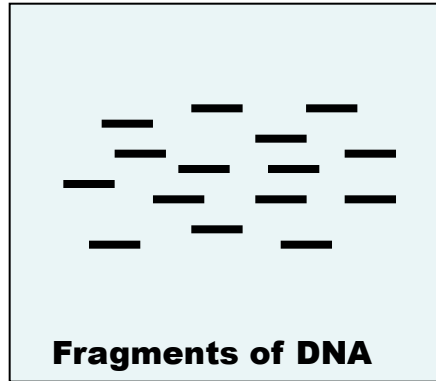
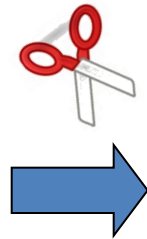
the Human Genome Project

AGTCCGCGAATACAGGCTCGGT

movie

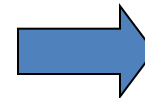
International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. Nature 431, 931-945 (21 October 2004).

Genome sequencing: general principle

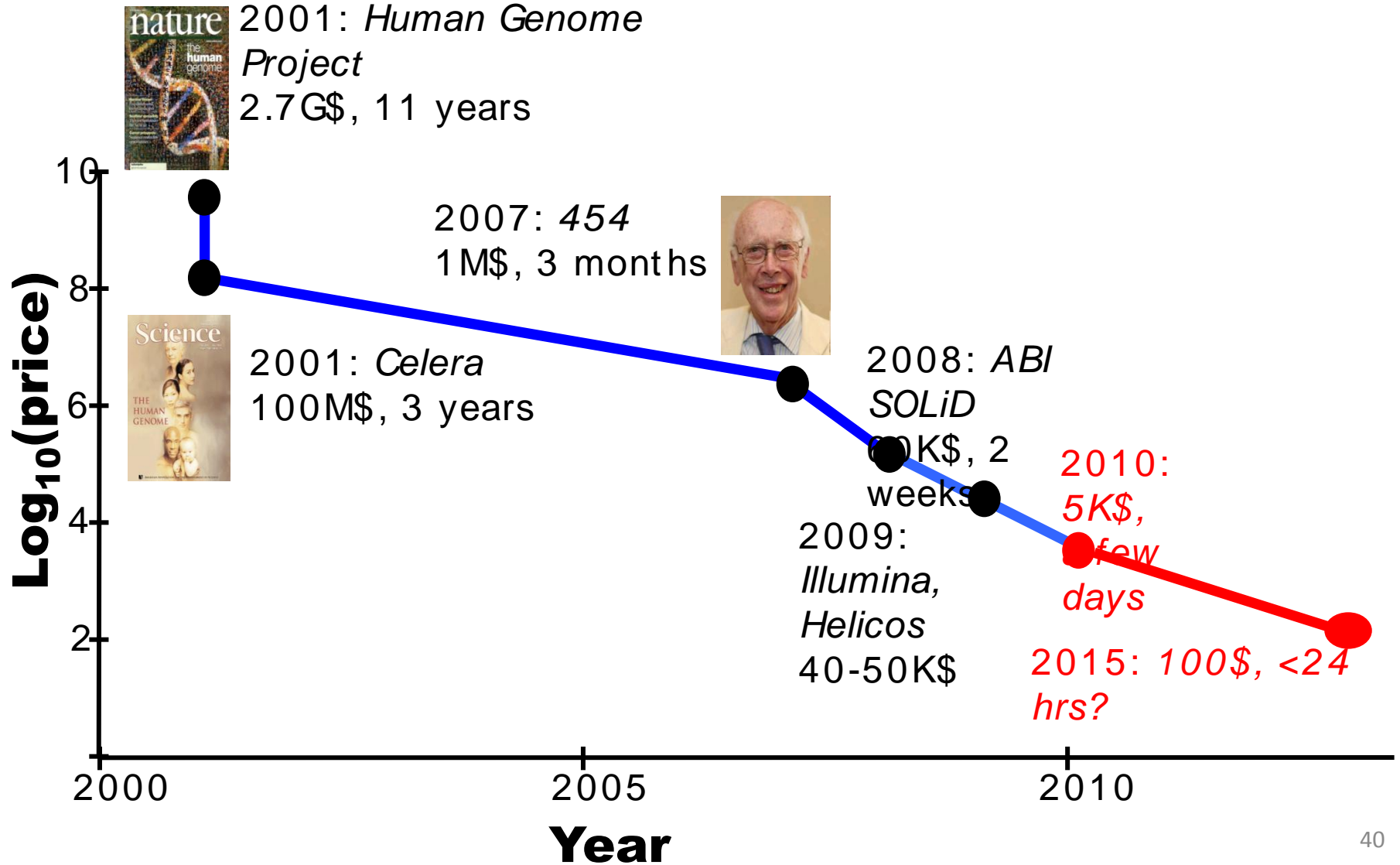


Gel:

- G GGGGTTCCCGCCAGGCTTCCGGGTTT
- T TTTTCCCGCCAGGCTTCCGGGTTT
- G GGGGTTCCCGCCAGGCTTCCGGGTTT
- G GGGGTTCCCGCCAGGCTTCCGGGTTT
- A AAAAAA
- C CCCCCC
- A AAAAAA
- T TTTTCCCGCCAGGCTTCCGGGTTT
- A AAAAAA
- C CCCCCC
- G GGGGTTCCCGCCAGGCTTCCGGGTTT
- C CCCCCC
- A AAAAAA
- T TTTTCCCGCCAGGCTTCCGGGTTT
- G GGGGTTCCCGCCAGGCTTCCGGGTTT
- C CCCCCC
- A AAAAAA
- T TTTTCCCGCCAGGCTTCCGGGTTT



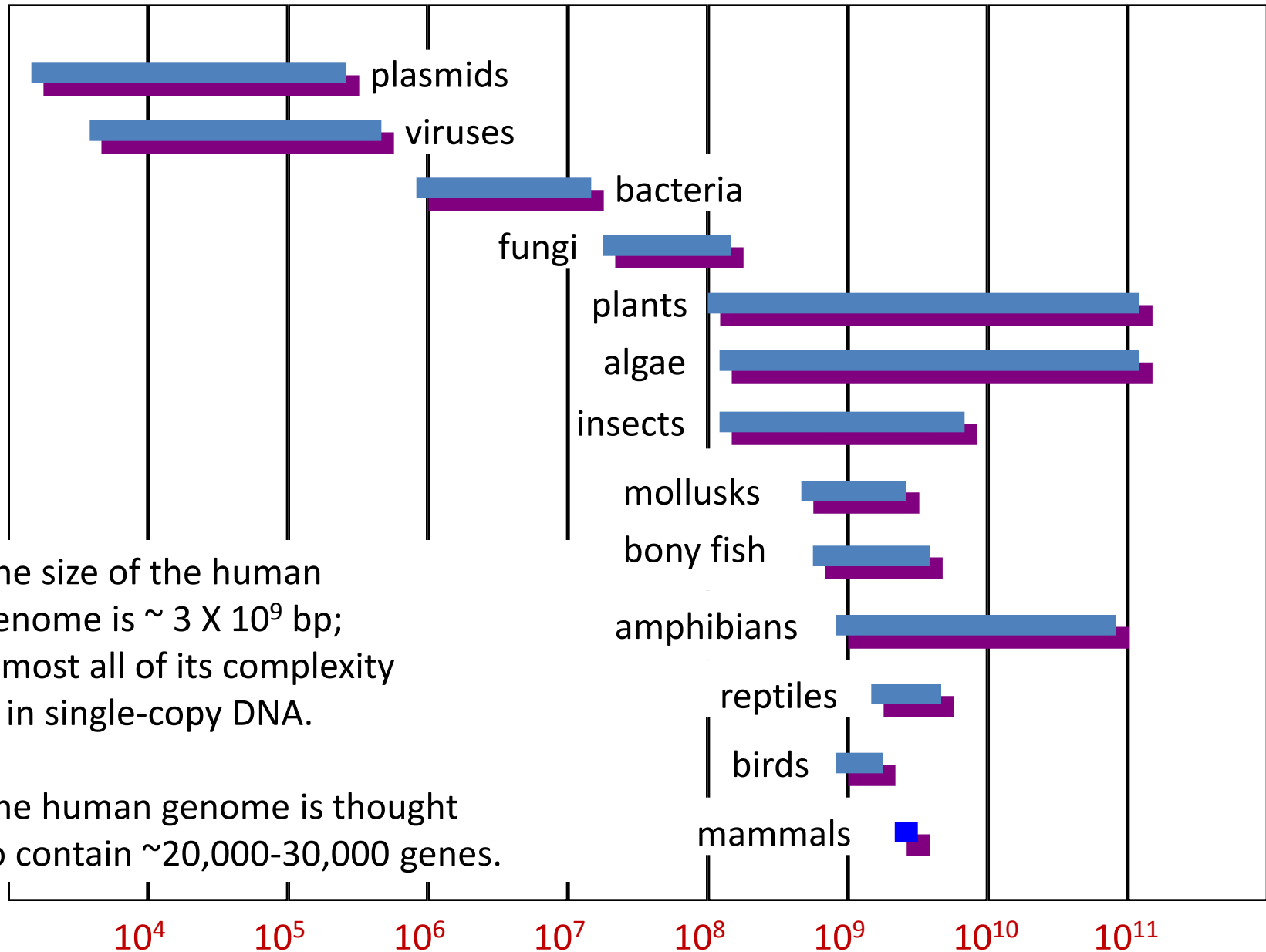
Sequencing the Human Genome



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)



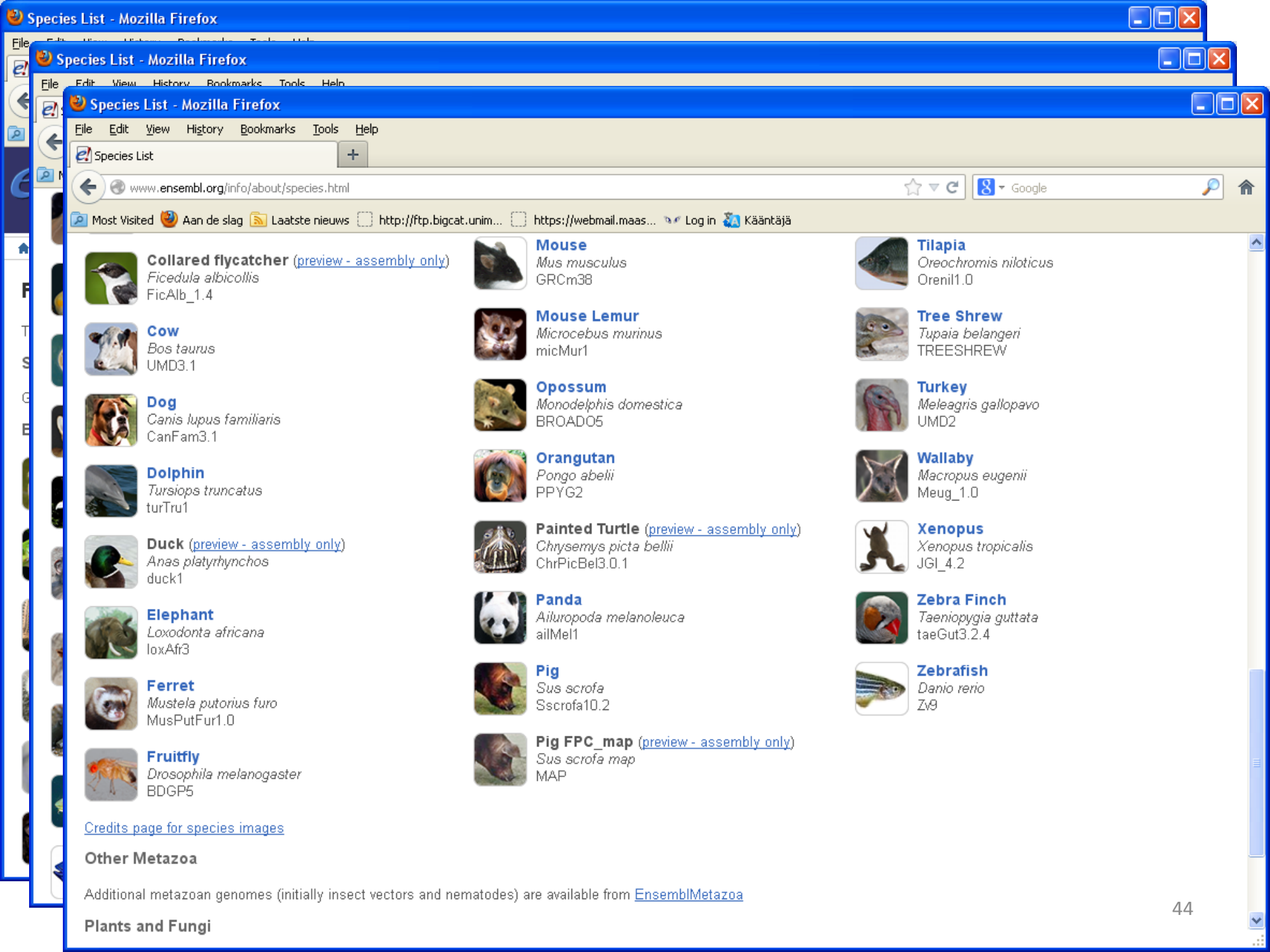
The size of the human genome is $\sim 3 \times 10^9$ bp; almost all of its complexity is in single-copy DNA.

The human genome is thought to contain $\sim 20,000$ - $30,000$ 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



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

[Credits page for species images](#)

Other Metazoa

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

Plants and Fungi



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
Orenil1.0



Tree Shrew
Tupaia belangeri
TREETSHREW



Turkey
Meleagris gallopavo
UMD2



Wallaby
Macropus eugenii
Meug_1.0



Xenopus
Xenopus tropicalis
JGI_4.2



Zebra Finch
Taeniopygia guttata
taeGut3.2.4



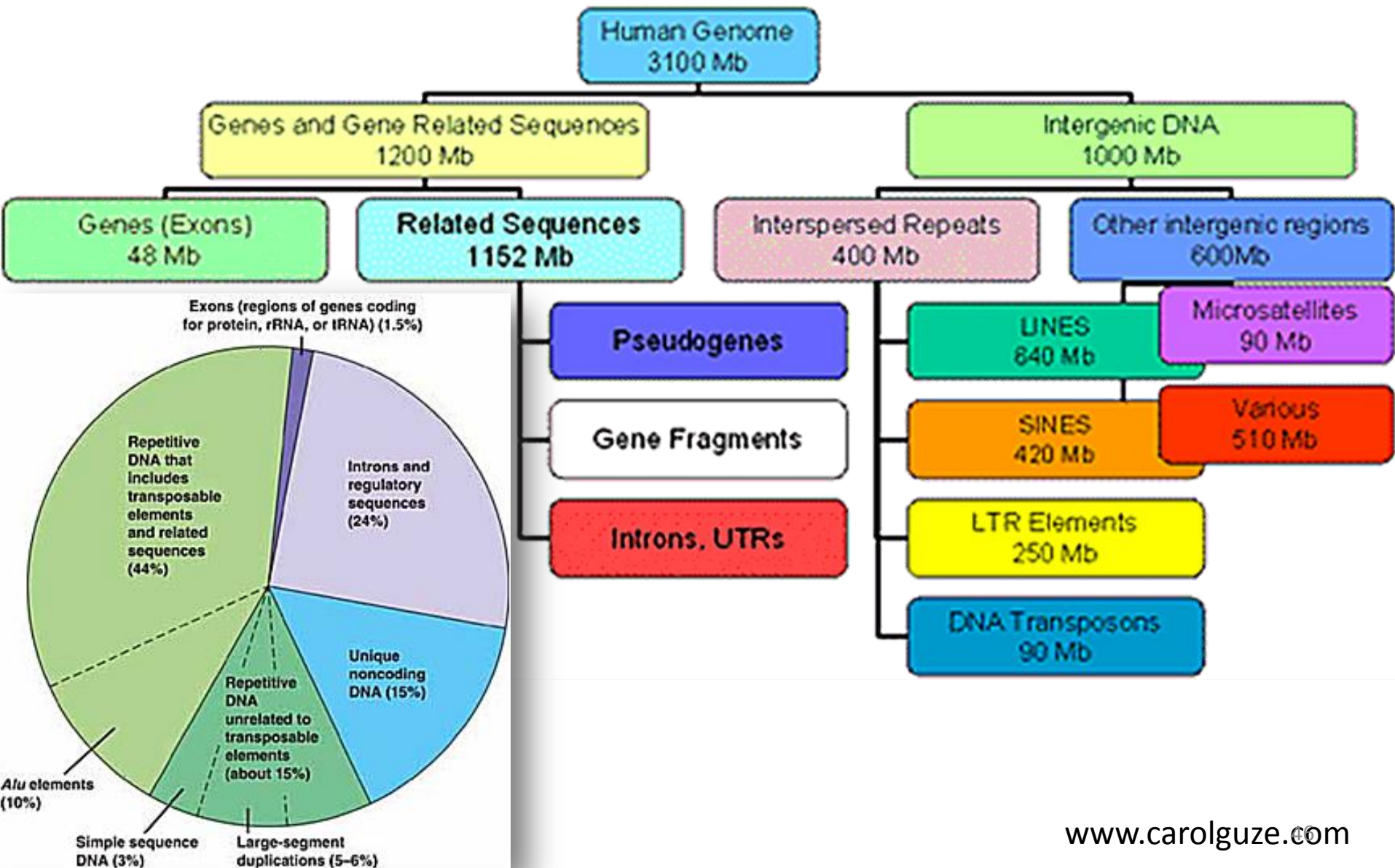
Zebrafish
Danio rerio
Zf9

Number of genes

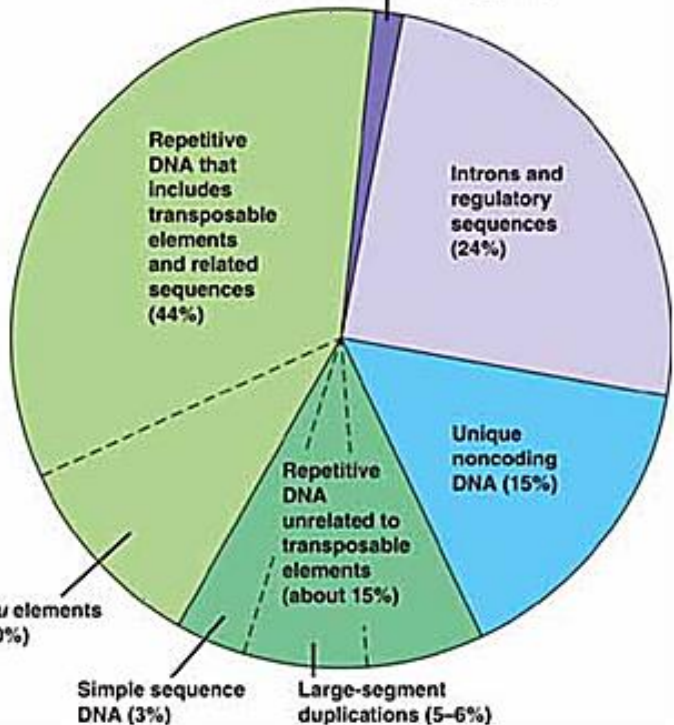
Species and Common Name	Estimated Total Size of Genome (bp)*	Estimated Number of Protein-Encoding Genes*
<i>Saccharomyces cerevisiae</i> (unicellular budding yeast)	12 million	6,000
<i>Trichomonas vaginalis</i>	160 million	60,000
<i>Plasmodium falciparum</i> (unicellular malaria parasite)	23 million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

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

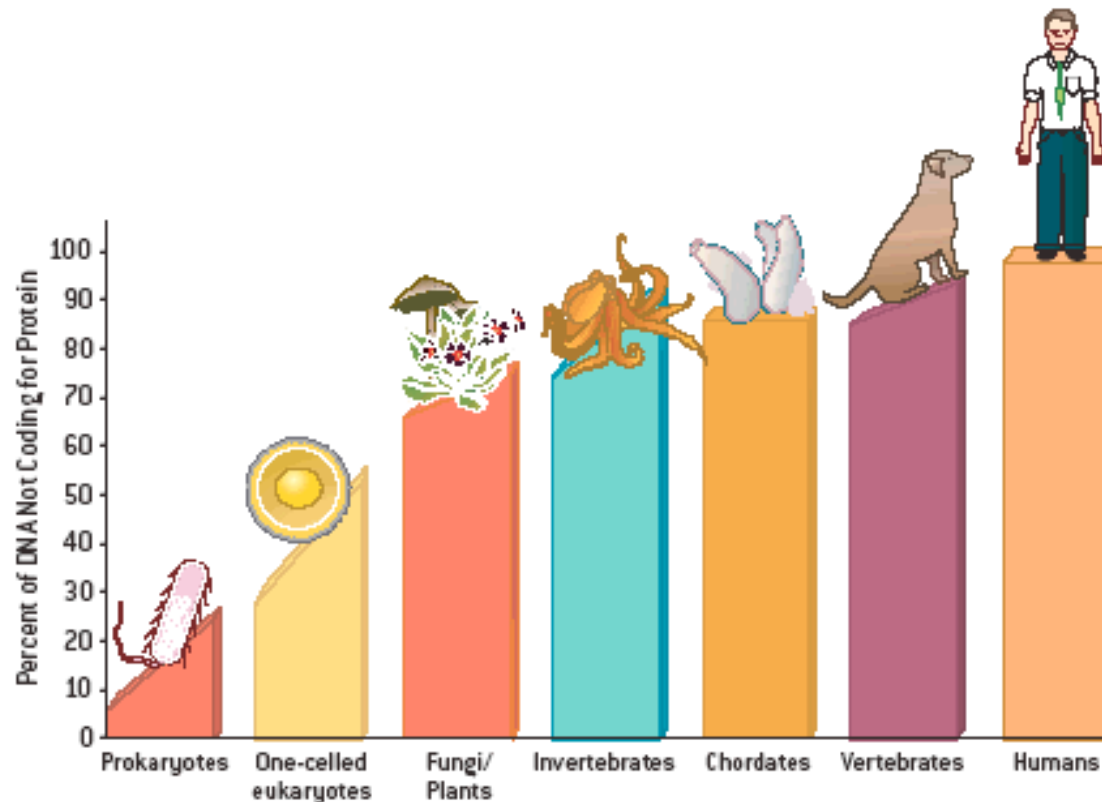
Organization of the human genome



Exons (regions of genes coding for protein, rRNA, or tRNA) (1.5%)



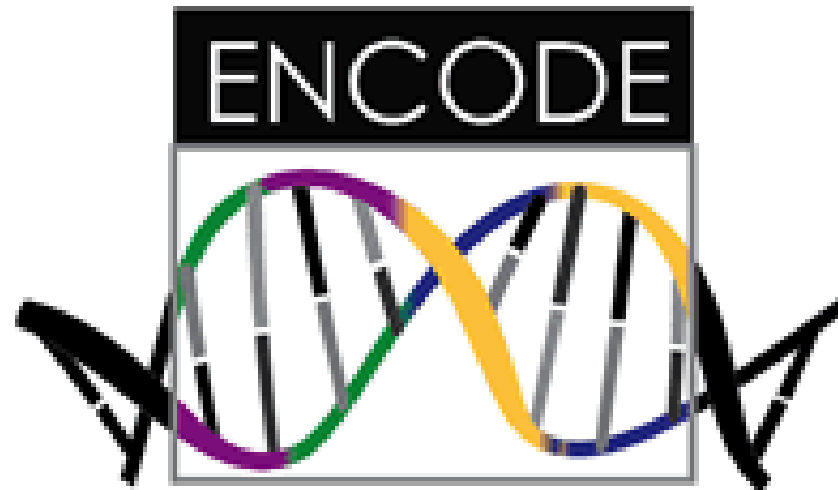
Non-Protein coding DNA



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

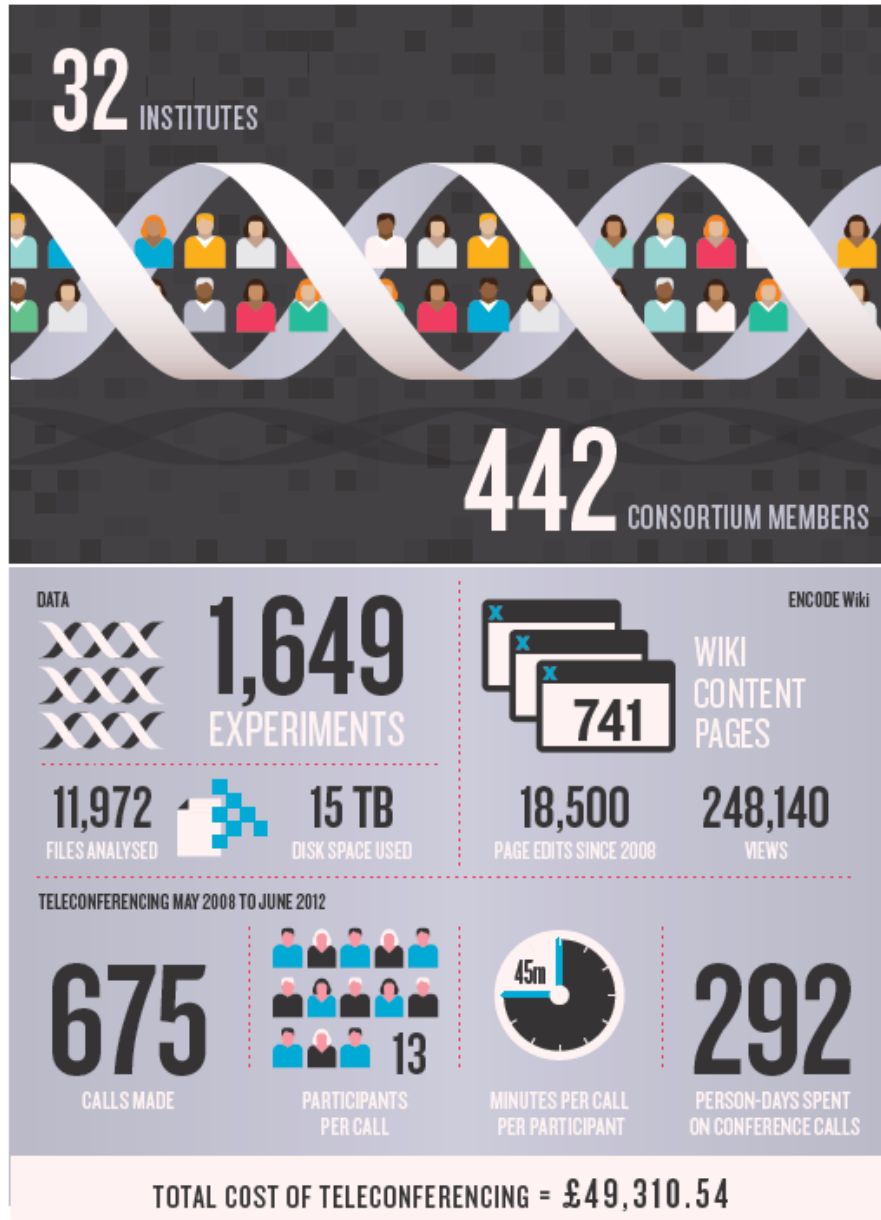


Launched: September 2003, upgraded to the entire genome September 2007.

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

BY THE NUMBERS

The ENCODE project involved hundreds of people from around the world, and a lot of editing, disk space and phone calls.



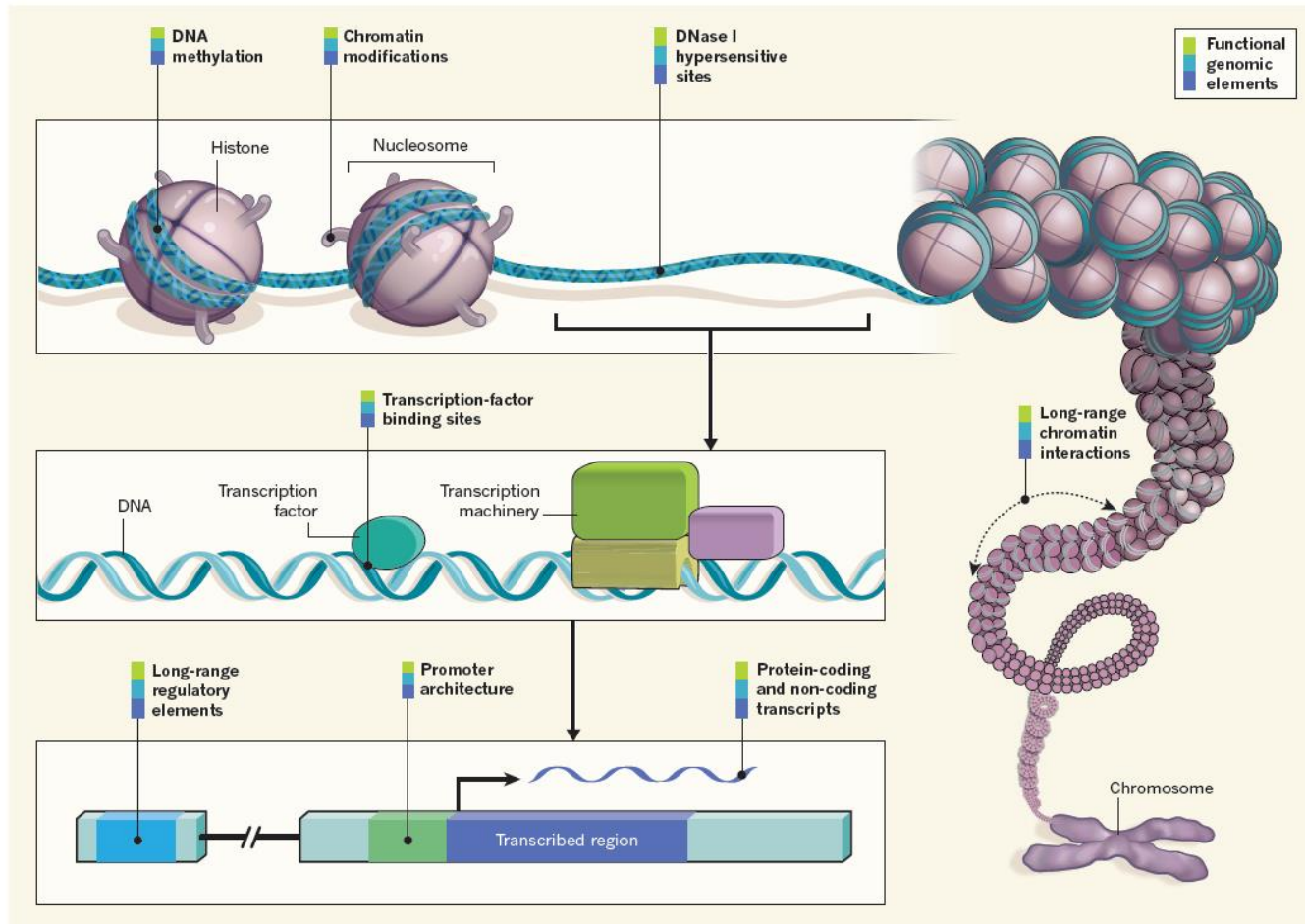
Understanding of the human genome is far from complete. We are missing knowledge on:

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



ENCODE data in Ensembl



BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search: for

e.g. BRCA2 or rat X:100000-200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes



Human
GRCh37



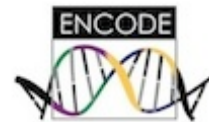
Mouse
GRCm38



Zebrafish
Zv9

★ [Log in to customize this list](#)

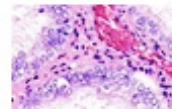
ENCODE data in Ensembl



Variant Effect Predictor



Gene expression in different tissues



Find SNPs and other variants for my gene

```
GTATACATTC  
CRTRAAAGTCTT  
CTTCTAAATTCT  
GRAACATTTCC
```

[Retrieve gene sequence](#)

[Compare genes across](#)

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

The screenshot shows the Gene Ontology (GO) tree view for muscle contraction. The interface includes a navigation bar with tabs for "Ancestors and Children", "Inferred Tree View", "Graph View", "Other Views", "Downloads", and "Mappings". The "Inferred Tree View" tab is selected. The tree structure is as follows:

- GO:0008150 biological_process [500154 gene products]
 - GO:0032501 multicellular organismal process [60501 gene products]
 - GO:0044699 single-organism process [237780 gene products]
 - GO:0044707 single-multicellular organism process [57987 gene products]
 - GO:0003008 system process [14138 gene products]
 - GO:0003012 muscle system process [1789 gene products]
 - GO:0006936 muscle contraction [1553 gene products]**
 - GO:0030049 muscle filament sliding [81 gene products]
 - GO:0045932 negative regulation of muscle contraction [130 gene products]
 - GO:0045933 positive regulation of muscle contraction [259 gene products]
 - GO:0006937 regulation of muscle contraction [878 gene products]
 - GO:0006939 smooth muscle contraction [572 gene products]
 - GO:0006941 striated muscle contraction [673 gene products]

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	Data source	Species
All	All	All
complex	ASAP	Arabidopsis thaliana
gene	AspGD	Aspergillus fumig...
gene product	CGD	Aspergillus fumig...

Filter by Association Evidence Code:

Evidence Code
All
IBA
KR
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
<input type="checkbox"/>	Aldoa aldolase A, fructose-bisphosphate	15 associations protein from <i>Mus musculus</i>		ISO With UniProtKB:P04075	MGI:MGI:4834177	MGI
<input type="checkbox"/>	Aldoa aldolase A, fructose-bisphosphate	27 associations gene from <i>Rattus norvegicus</i> BLAST		ISO With RGD:735815	RGD:1624291	RGD
<input type="checkbox"/>	ALDOA Fructose-bisphosphate aldolase	12 associations protein from <i>Bos taurus</i> BLAST		IEA With Ensembl:ENSP00000378669	GO REF:0000019	Ensembl (via UniProtKB)
<input type="checkbox"/>	ALDOA Fructose-bisphosphate aldolase A	29 associations protein from <i>Homo sapiens</i> BLAST		IMP	PMID:14615364	BHF-UCL (via UniProtKB)
<input type="checkbox"/>	Arg2 arginase 2	35 associations gene from <i>Rattus norvegicus</i> BLAST		IEA With Ensembl:ENSMUSP00000021550	RGD:1600115	Ensembl (via RGD)
				ISO With RGD:736823	RGD:1624291	RGD
<input type="checkbox"/>	Arg2 arginase type II	13 associations protein from <i>Mus musculus</i> BLAST		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	parentage
is_a: GO:0006092	

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

