Dr. Susan Steinbusch

susan.coort@maastrichtuniversity.nl

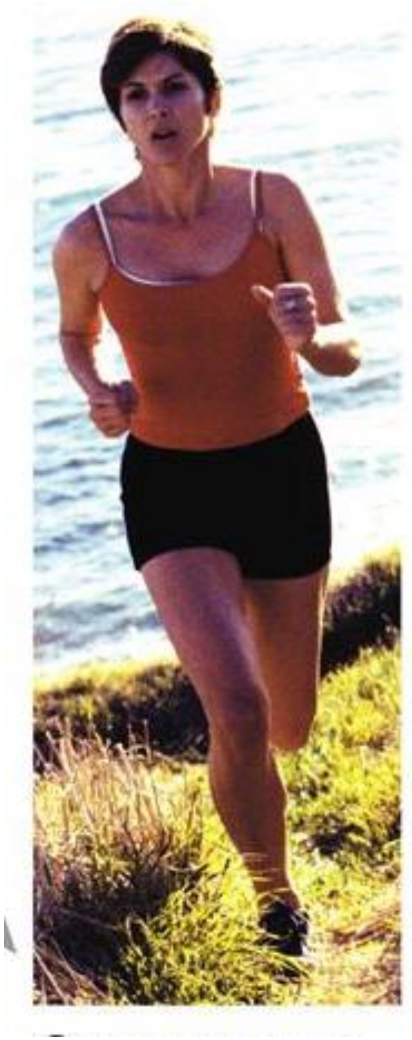
May 13th 2019

# Content

- Introduction
- Learning goals
- Biological sequence databases
  - Ensembl
  - NCBI
- Human genome project
- ENCODE project
- Genetic variation
- Gene Ontology
- WikiPathways

# Introduction

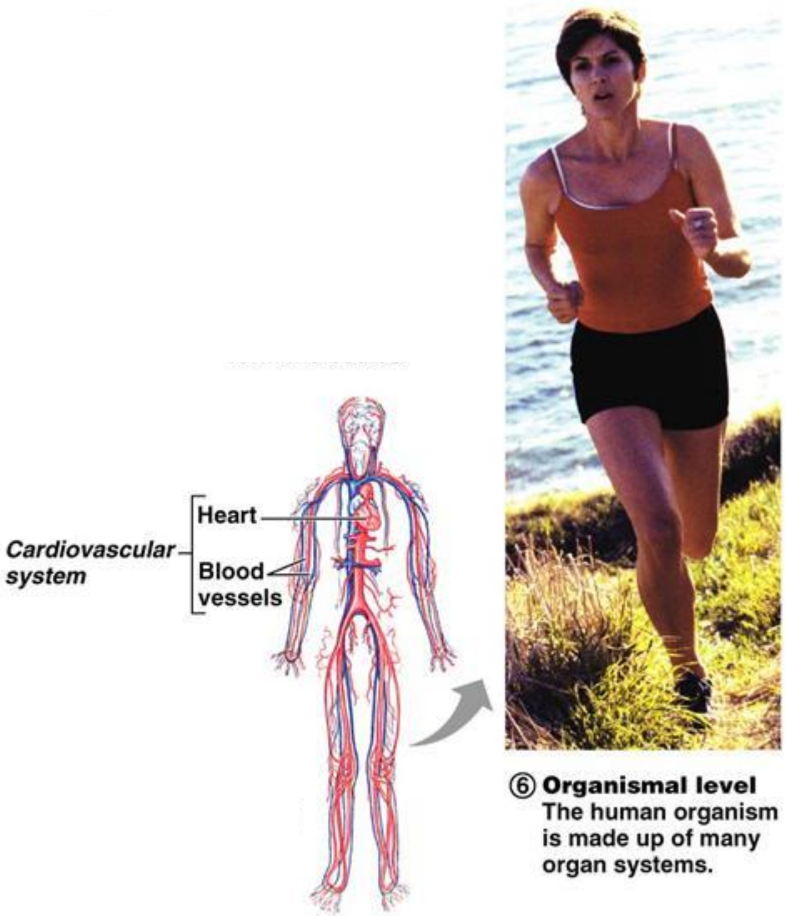
# What happens with the human body when you are running?



# 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

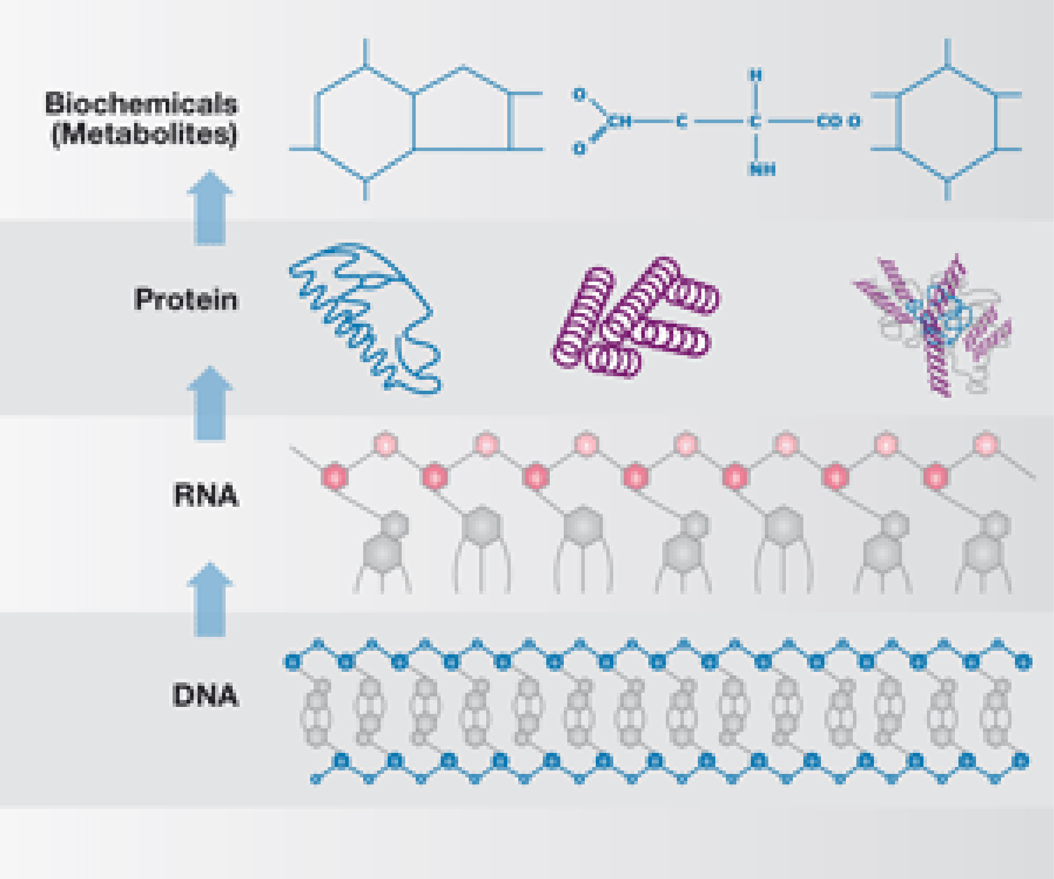


⑤ **Organ system level**
Organ systems consist of different organs that work together closely.

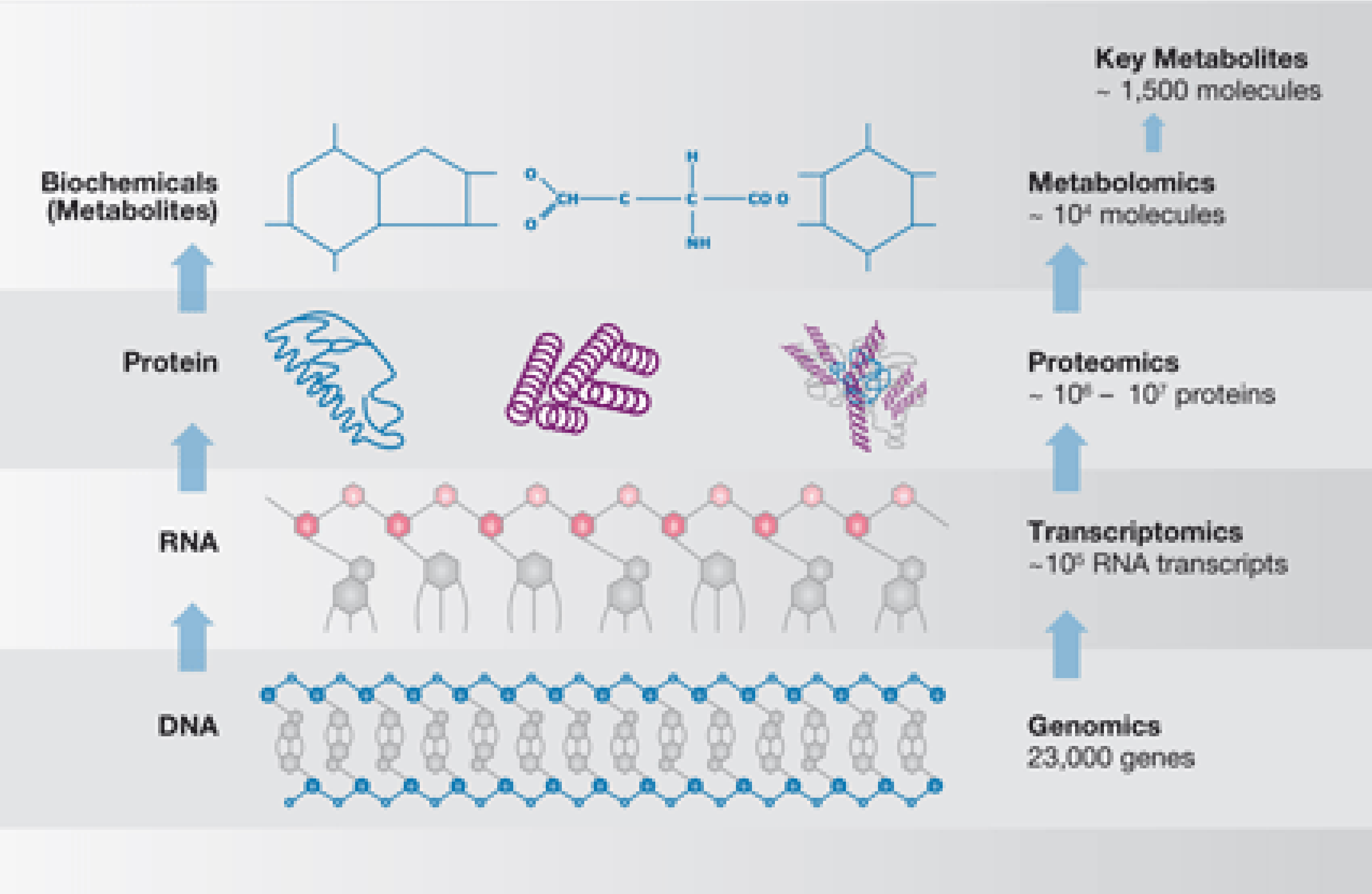
⑥ **Organismal level**
The human organism is made up of many organ systems.

Figure 1.1

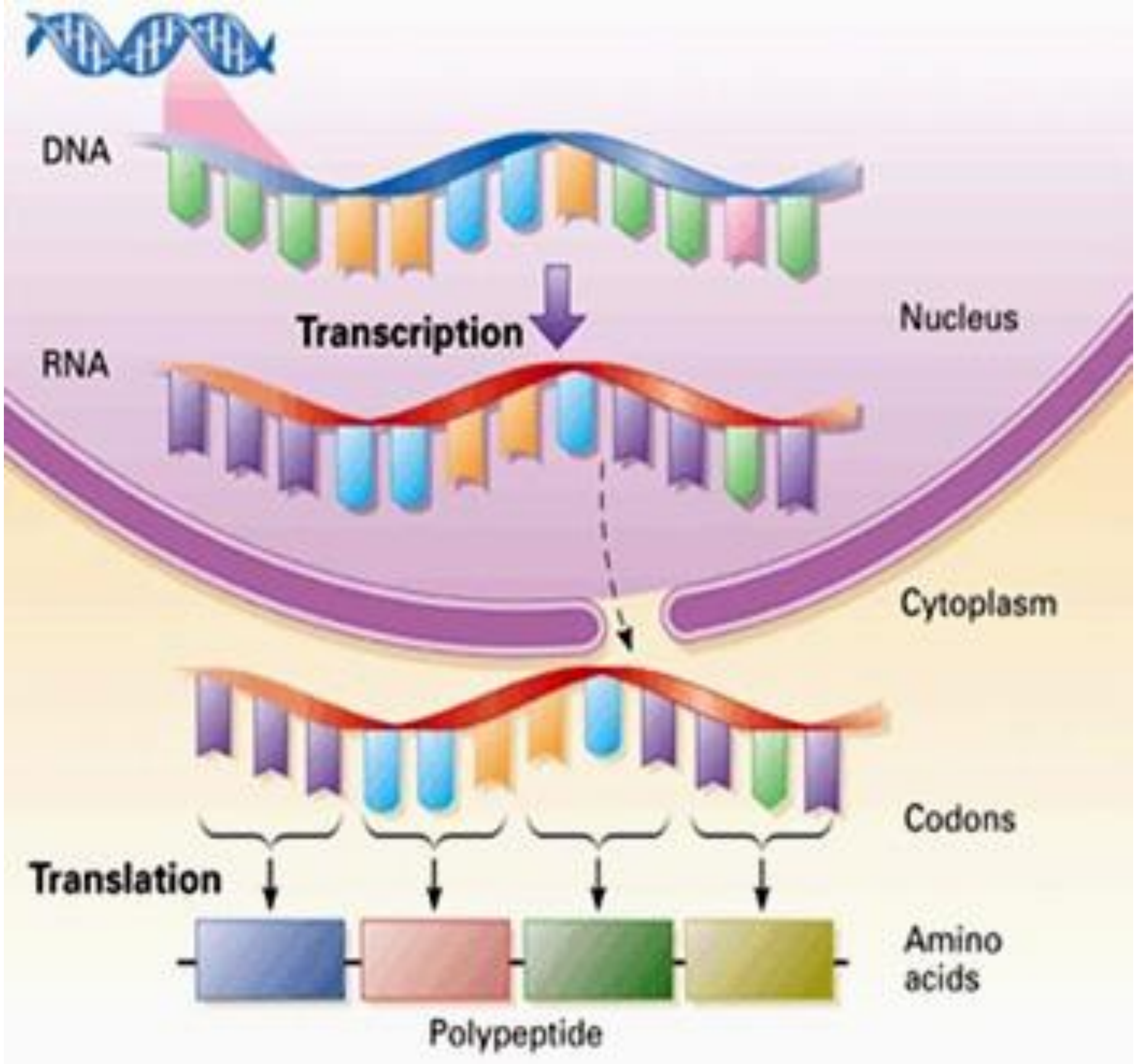
# (Bio)Molecules
# Individual players are important



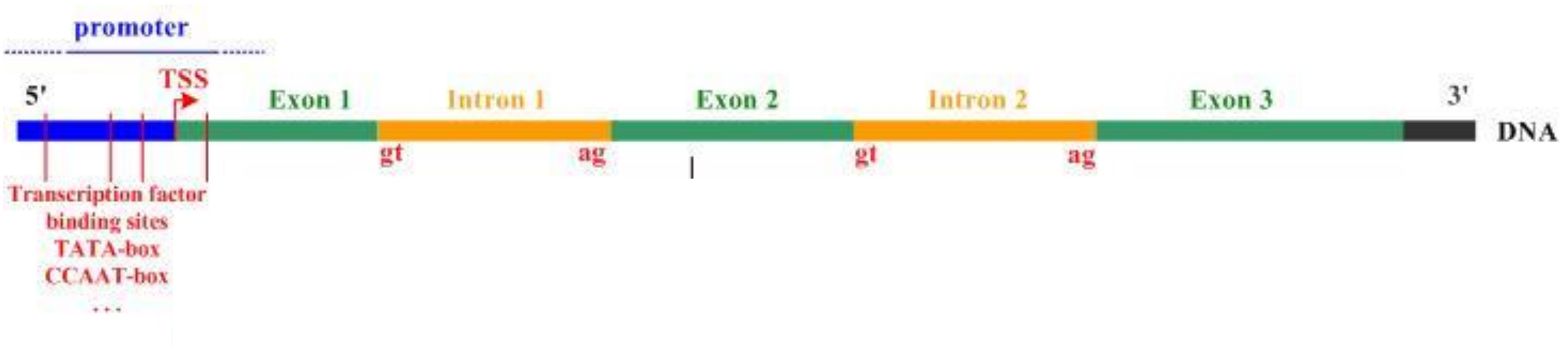
# Heaps of knowledge on biomolecules online available.



# Protein synthesis



# Gene structure



*Alternative splicing!*

CDS = Coding DNA Sequence

UTR = UnTranslated region

www.carolguze.com

# Learning goals

## 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?
- Two projects aimed at deciphering the content of the human genome, the human genome project & ENCODE.
- How to find information on genetic diseases
- What is gene ontology and WikiPathways?

# Biological sequence databases

# 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

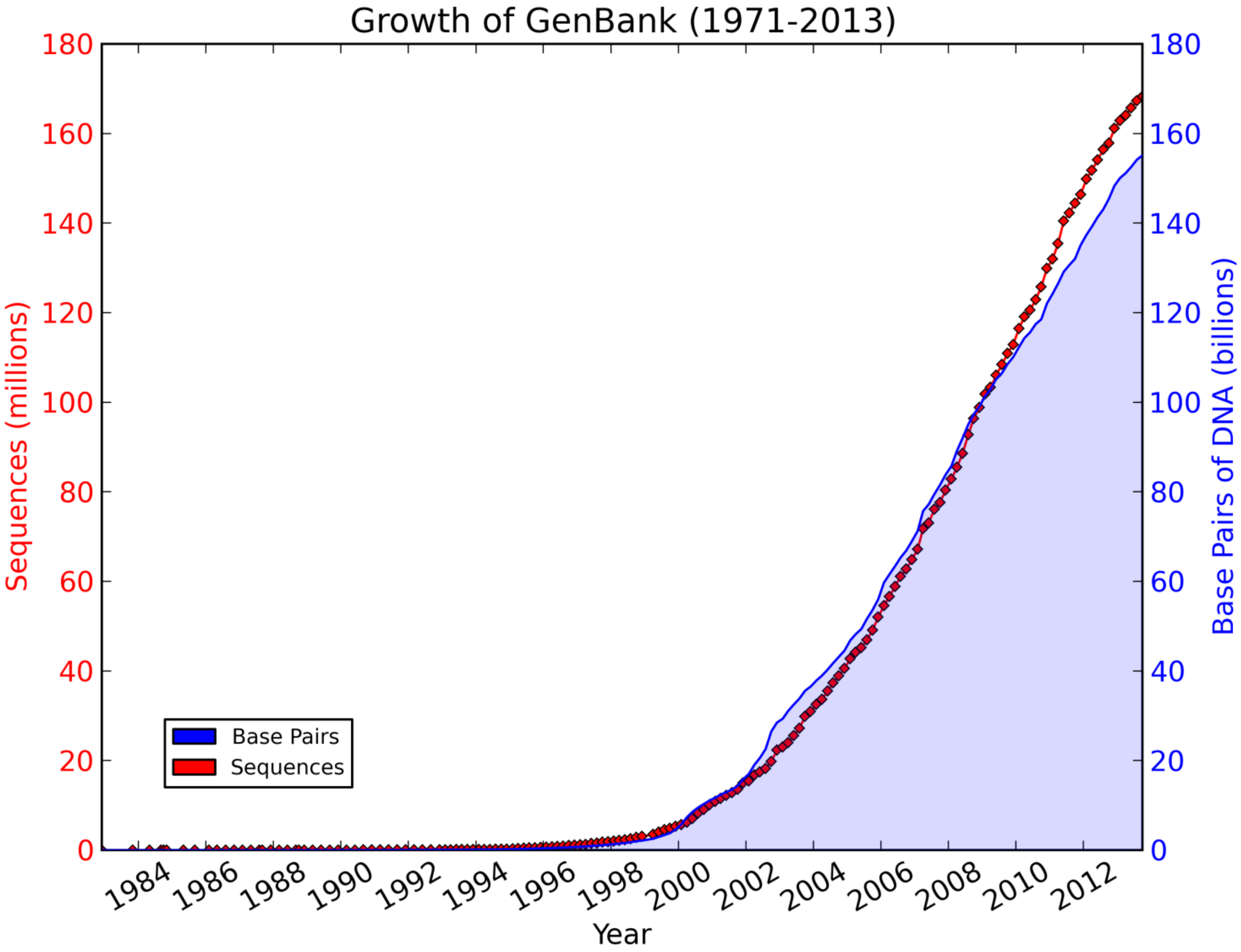
Lecture protein structures

# 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

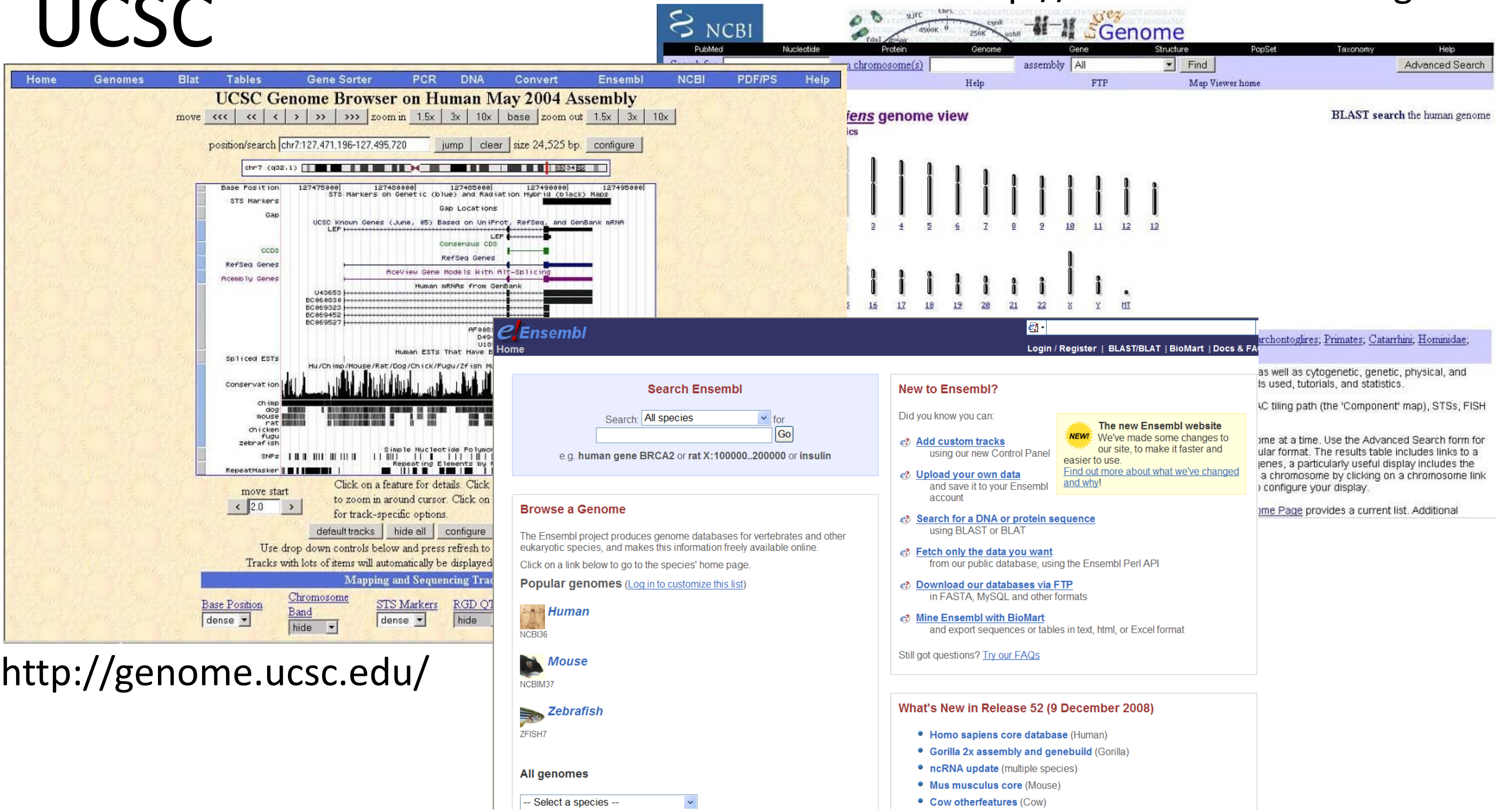
# Growth of GenBank



# Genome-centered databases

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

UCSC



*iens* genome view

BLAST search the human genome

rchontoglires; Primates; Catarrhini; Hominidae;

as well as cytogenetic, genetic, physical, and
is used, tutorials, and statistics.

AC tiling path (the 'Component' map), STSs, FISH

ome at a time. Use the Advanced Search form for
ular format. The results table includes links to a
genes, a particularly useful display includes the
a chromosome by clicking on a chromosome link
configure your display.

ome Page provides a current list. Additional

UCSC Genome Browser on Human May 2004 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
jump
clear
configure

position/search chr7:127,471,196-127,495,720 size 24,525 bp.

move start
2.0

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
Tracks with lots of items will automatically be displayed

Mapping and Sequencing Trac

| Base Position | Chromosome Band | STS Markers | RGD QT |
|---|---|---|---|
| dense | hide | dense | hide |

**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*
NCBIM37

*Zebrafish*
ZFISH7

**All genomes**

-- Select a species --

**New to Ensembl?**

Did you know you can:

**Add custom tracks**
using our new Control Panel

**Upload your own data**
and save it to your Ensembl account

NEW! **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!

**Search for a DNA or protein sequence**
using BLAST or BLAT

**Fetch only the data you want**
from our public database, using the Ensembl Perl API

**Download our databases via FTP**
in FASTA, MySQL and other formats

**Mine Ensembl with BioMart**
and export sequences or tables in text, html, or Excel format

Still got questions? Try our FAQs

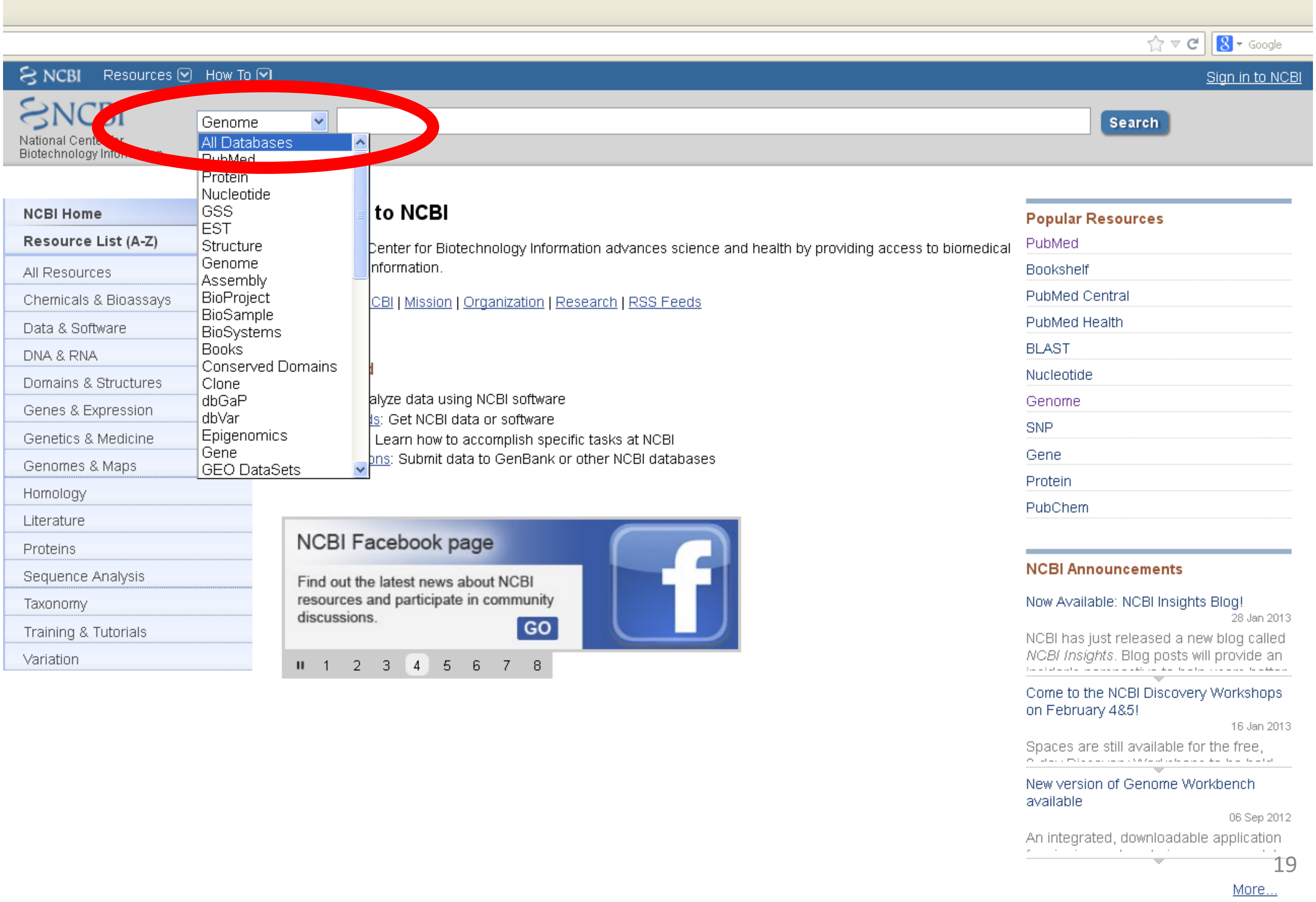
**What's New in Release 52 (9 December 2008)**

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

http://genome.ucsc.edu/

Ensembl http://www.ensembl.org/

# NCBI homepage



to NCBI

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

CBI | Mission | Organization | Research | RSS Feeds

alyze data using NCBI software

ls: Get NCBI data or software

Learn how to accomplish specific tasks at NCBI

ons: Submit data to GenBank or other NCBI databases

NCBI Facebook page

Find out the latest news about NCBI resources and participate in community discussions.

Now Available: NCBI Insights Blog!
28 Jan 2013

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

Come to the NCBI Discovery Workshops on February 4&5!
16 Jan 2013

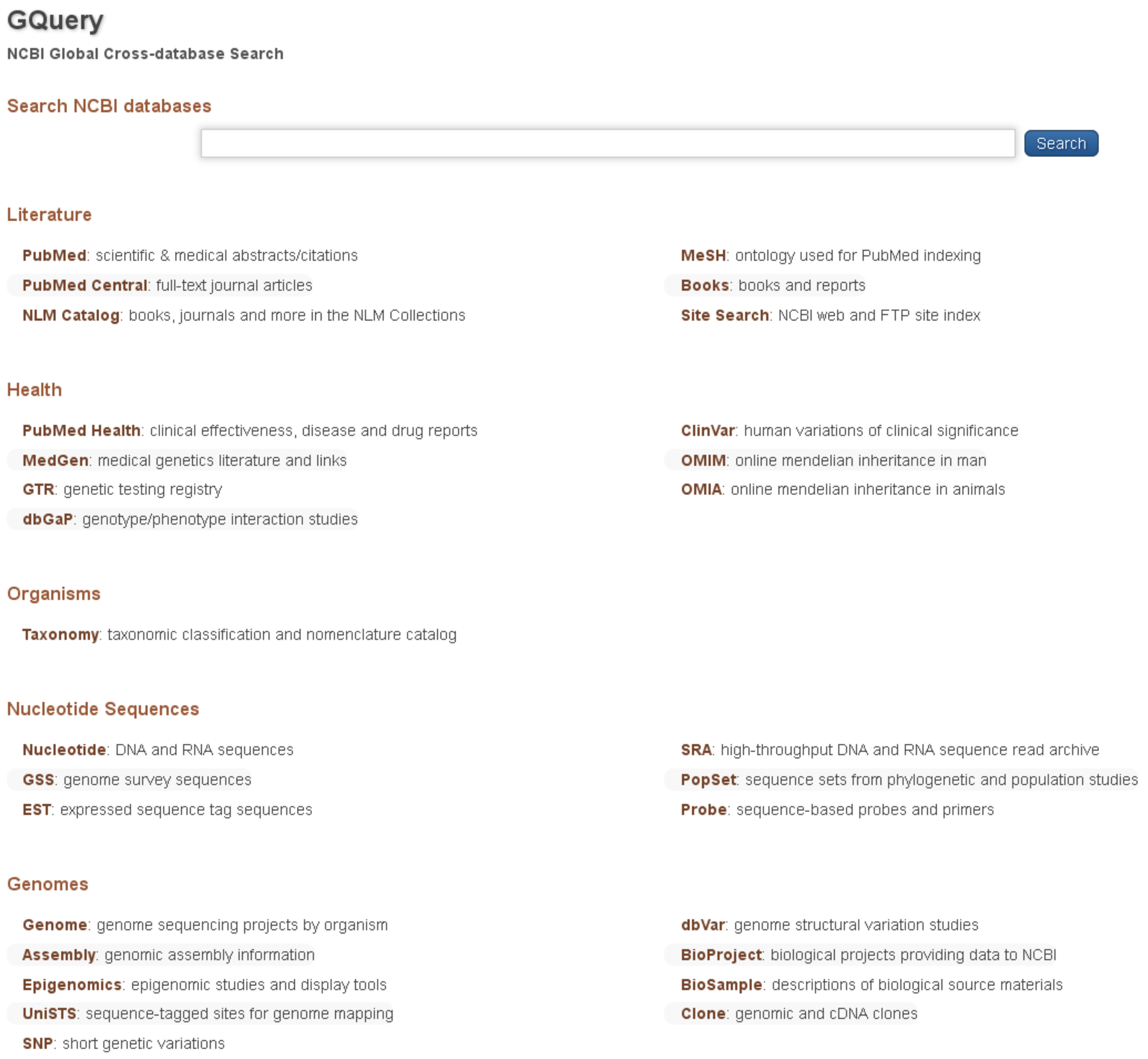
Spaces are still available for the free,

New version of Genome Workbench available
06 Sep 2012

An integrated, downloadable application

# NCBI Global Cross-database search

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



### GQuery

**NCBI Global Cross-database Search**

#### Search NCBI databases

#### 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

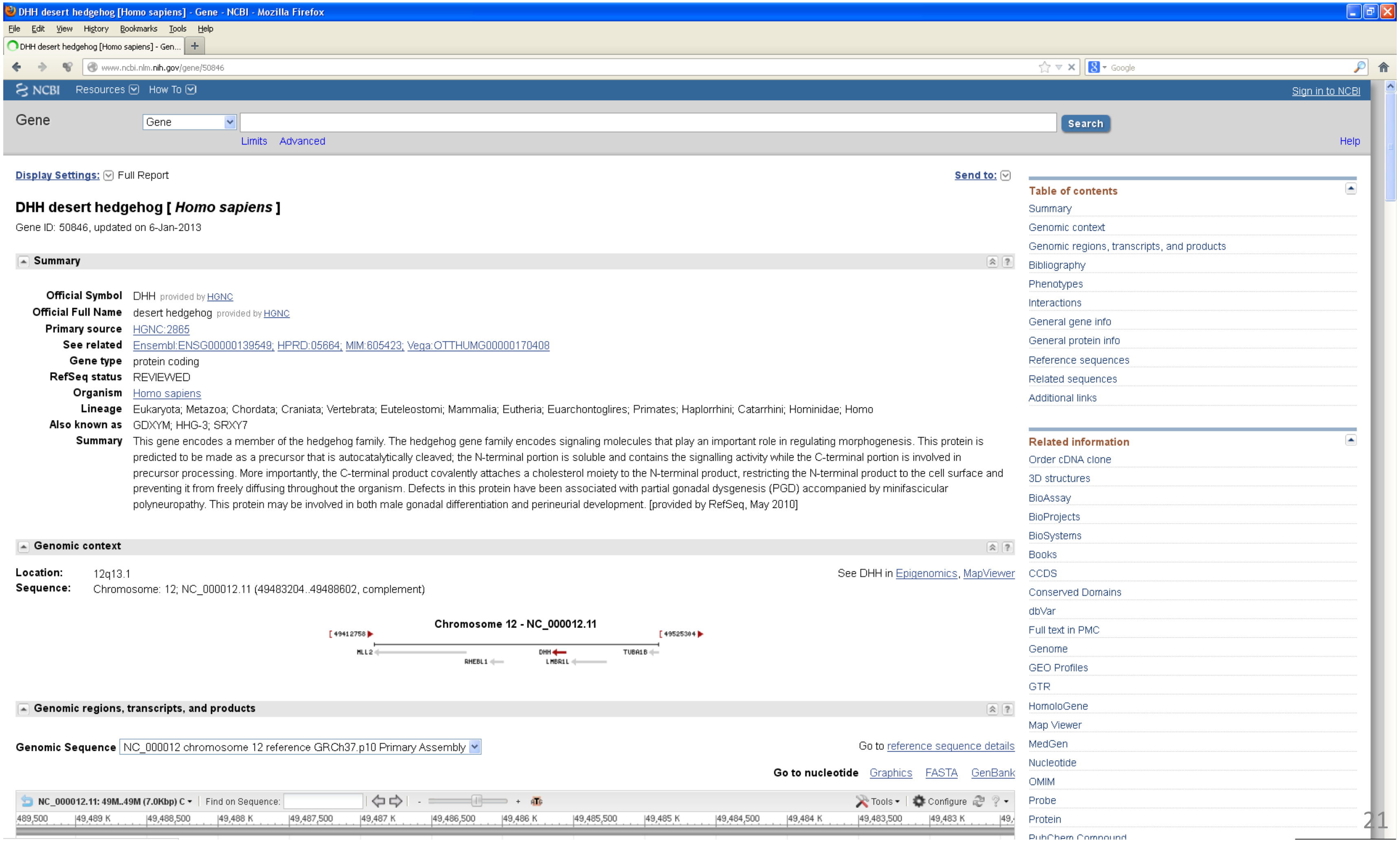
**dbVar**: genome structural variation studies

**BioProject**: biological projects providing data to NCBI

**BioSample**: descriptions of biological source materials

**Clone**: genomic and cDNA clones

# Gene (NCBI)
# DHH as example



## 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

See DHH in Epigenomics, MapViewer

**Sequence:** Chromosome: 12; NC_000012.11 (49483204..49488602, complement)

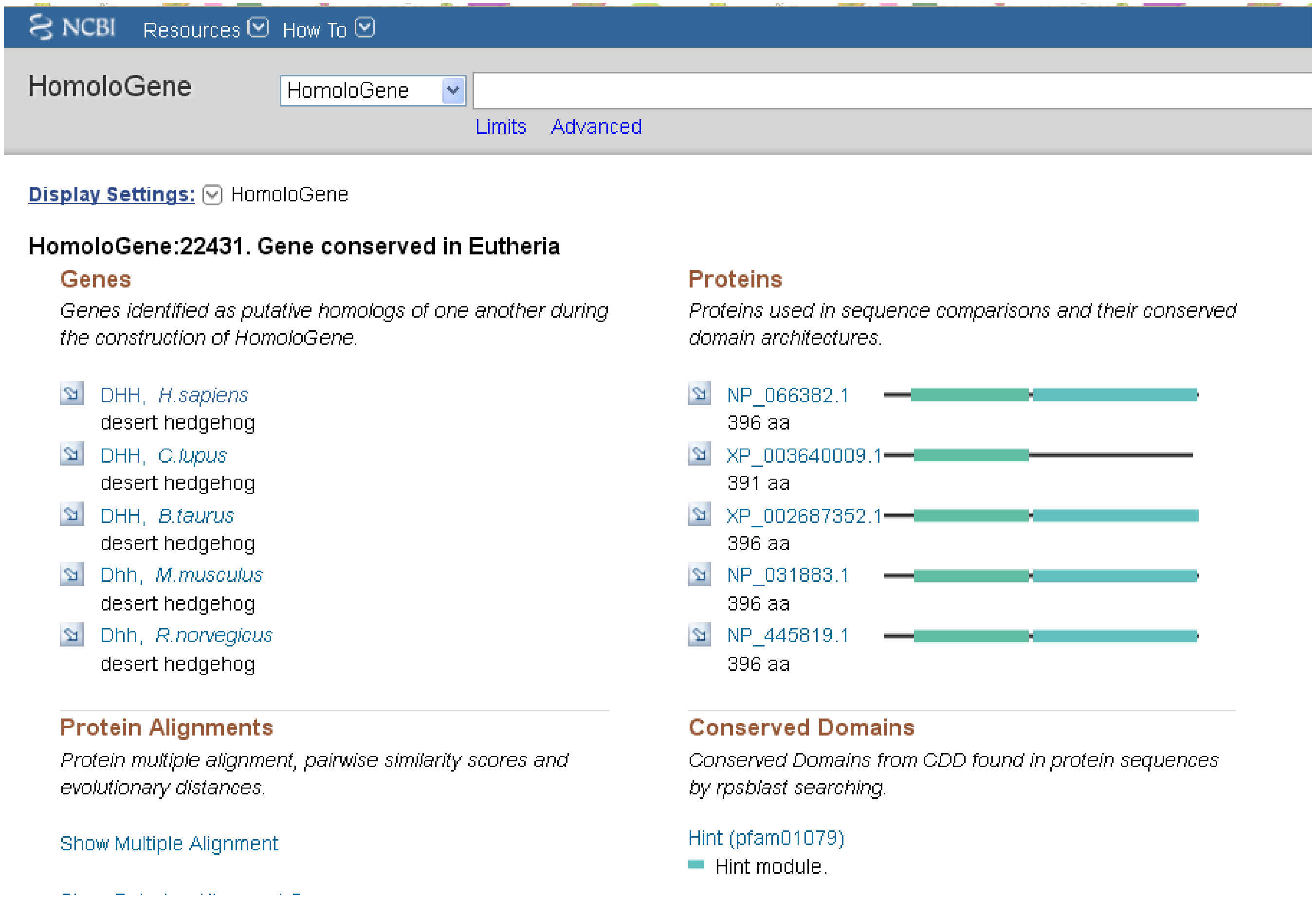
### Genomic regions, transcripts, and products

**Genomic Sequence** NC_000012 chromosome 12 reference GRCh37.p10 Primary Assembly

Go to reference sequence details

**Go to nucleotide** Graphics FASTA GenBank

# Homologene



**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**

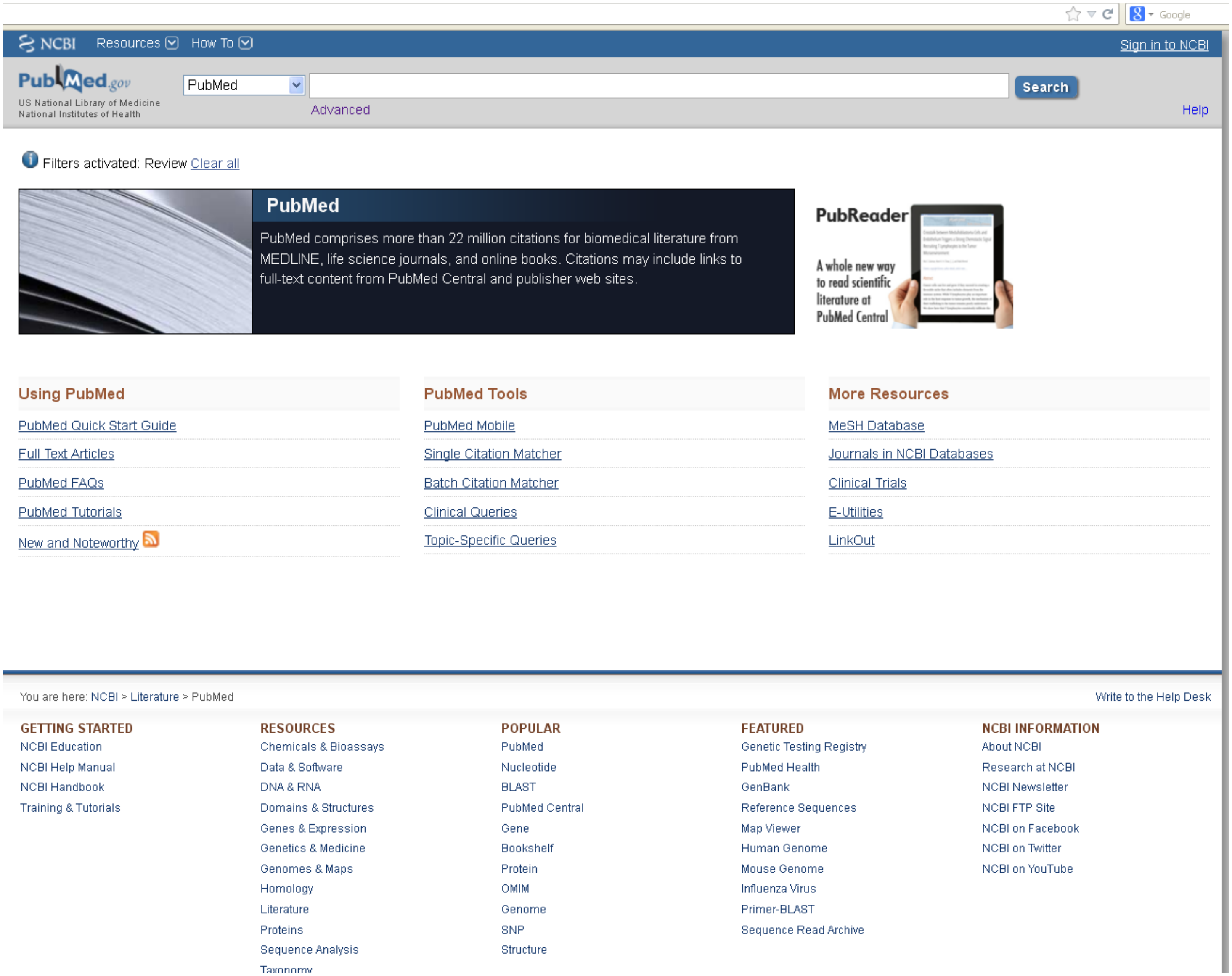
*Conserved Domains from CDD found in protein sequences by rpsblast searching.*

Hint (pfam01079)

🟦 Hint module.

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

# PubMed (NCBI)



Google
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

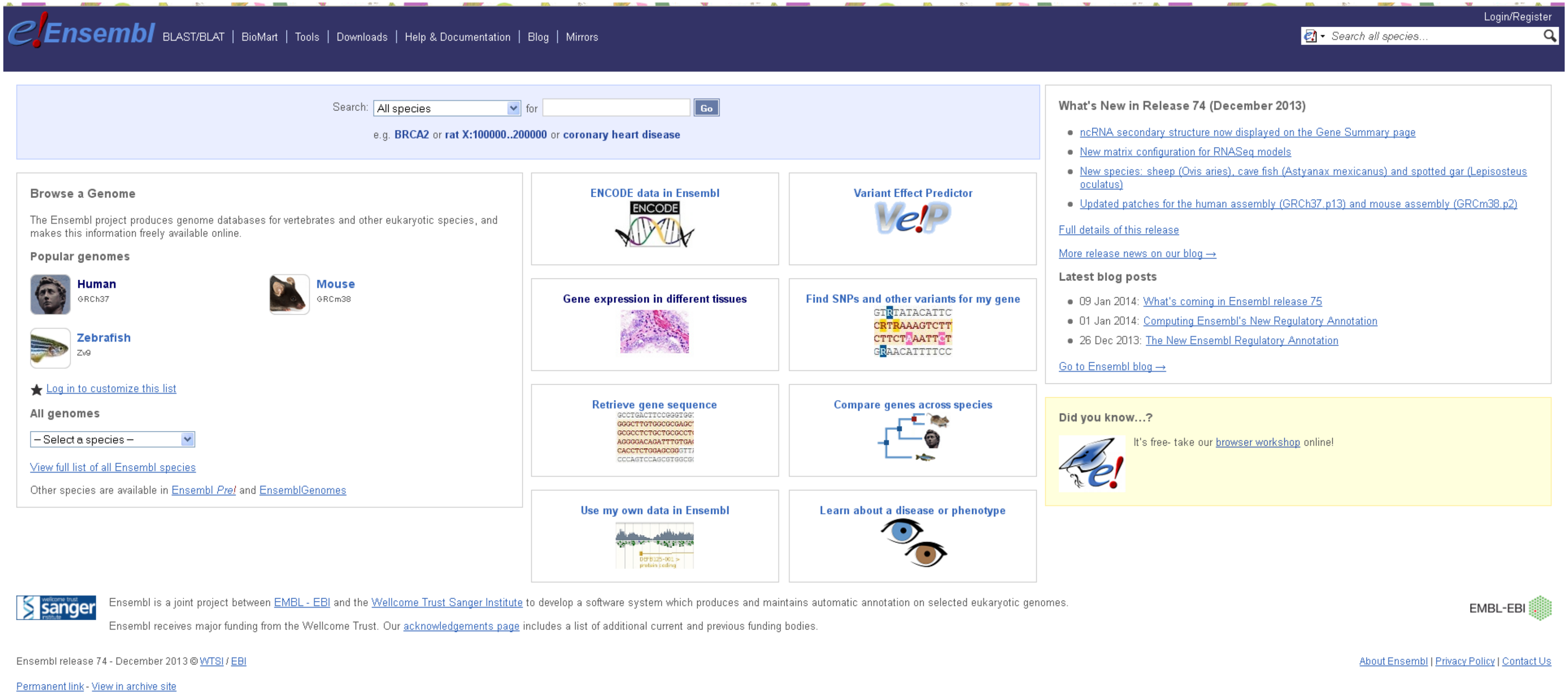
**FEATURED**

- 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



e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login/Register
Search all species...

Search: All species for Go

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
ENCODE
Variant Effect Predictor
VeP
Gene expression in different tissues
Find SNPs and other variants for my gene
Retrieve gene sequence
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.

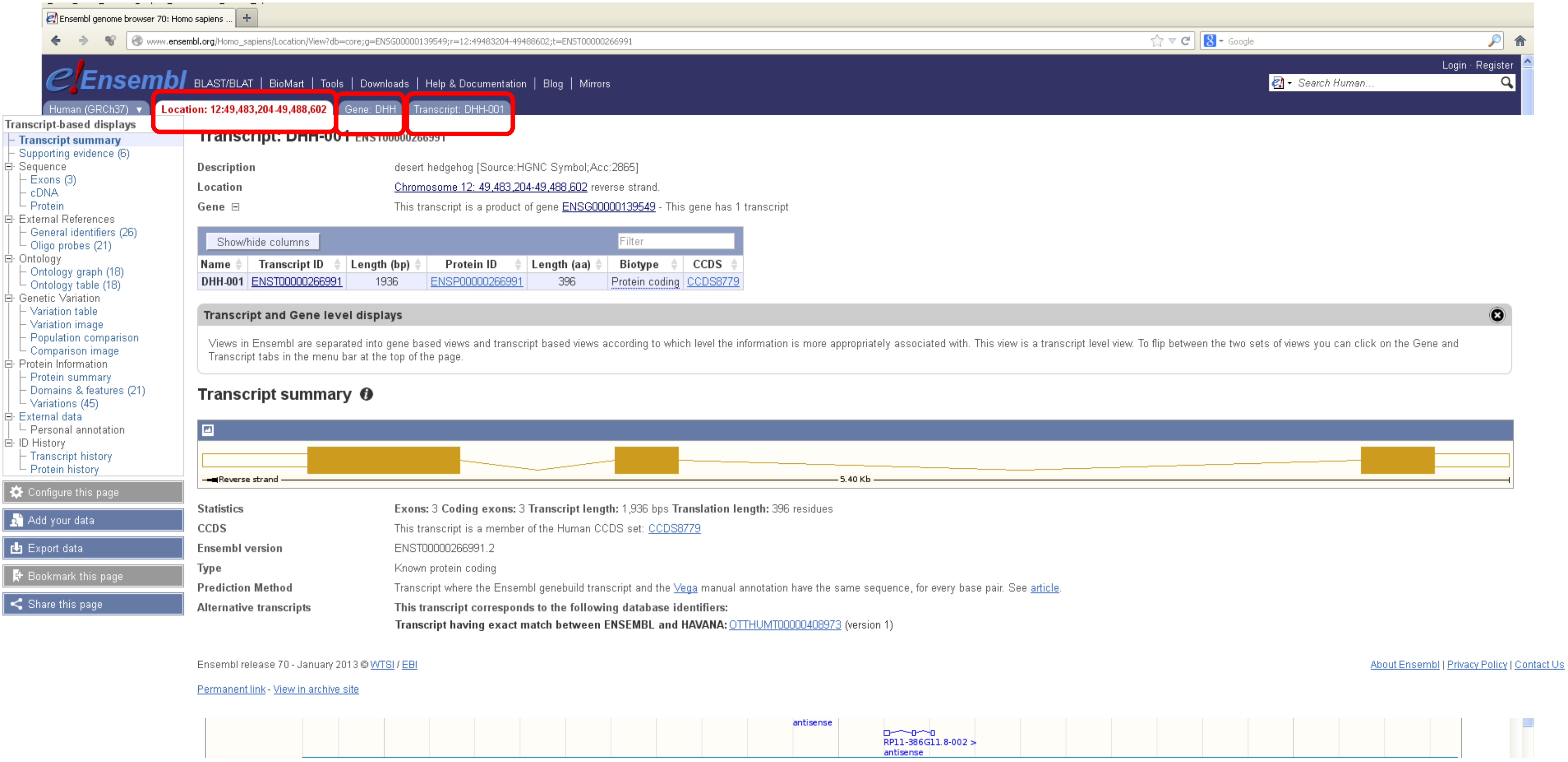
Ensembl receives major funding from the Wellcome Trust. Our acknowledgements page includes a list of additional current and previous funding bodies.

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
Google
e!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

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 |

| 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**

| | |
|---|---|
| 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:**<br>**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

# 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.

# Human Genome & ENCODE project

# 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 |

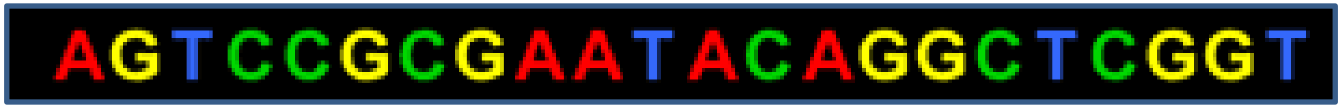
# HGP and ENCODE

- 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

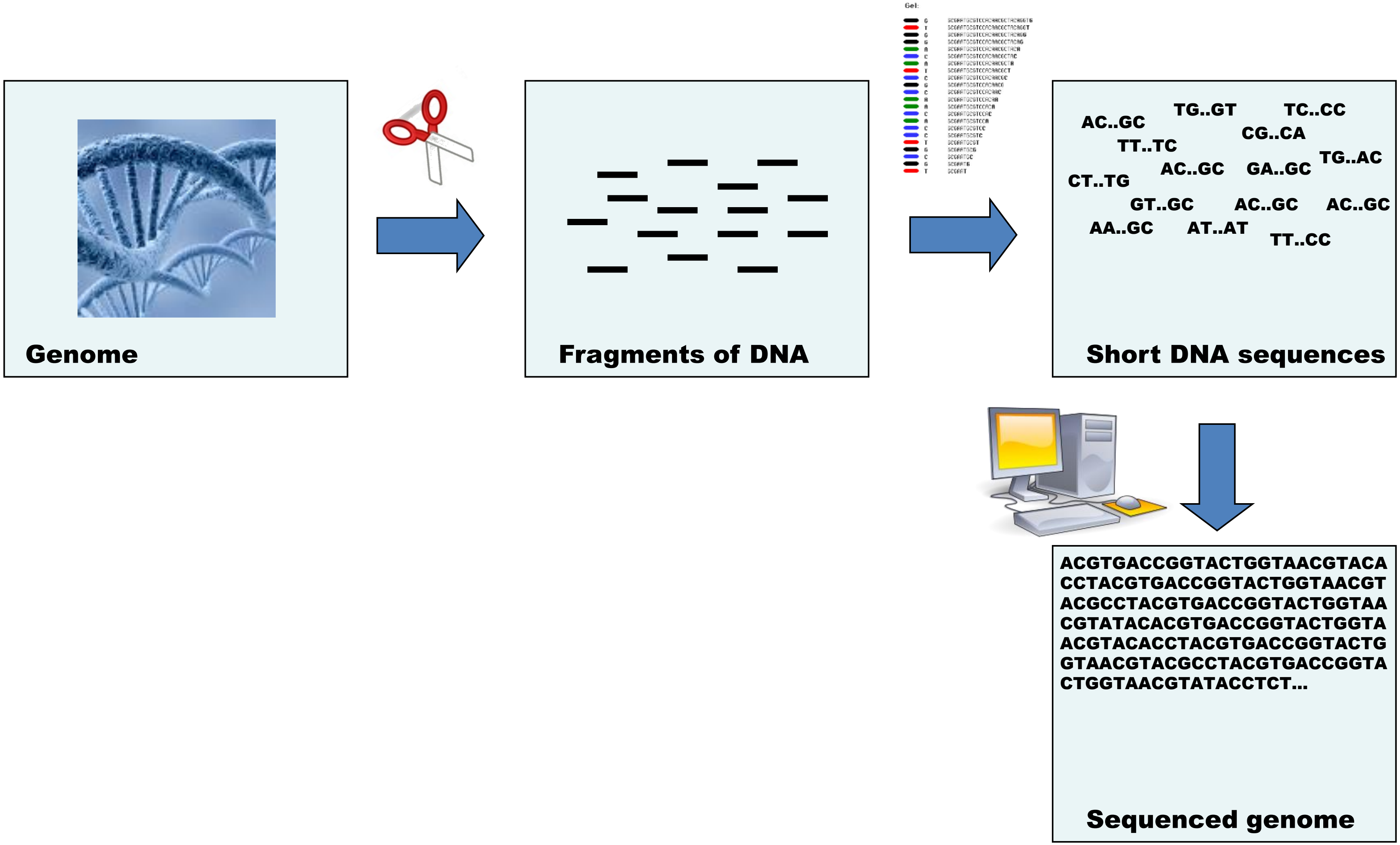
# the Human Genome Project



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



Genome

Fragments of DNA

AC..GC TG..GT TC..CC
TT..TC CG..CA
CT..TG AC..GC GA..GC TG..AC
GT..GC AC..GC AC..GC
AA..GC AT..AT TT..CC

Short DNA sequences

ACGTGACCGGTACTGGTAACGTACA
CCTACGTGACCGGTACTGGTAACGT
ACGCCTACGTGACCGGTACTGGTAA
CGTATACACGTGACCGGTACTGGTA
ACGTACACCTACGTGACCGGTACTG
GTAACGTACGCCTACGTGACCGGTA
CTGGTAACGTATACCTCT...

Sequenced genome

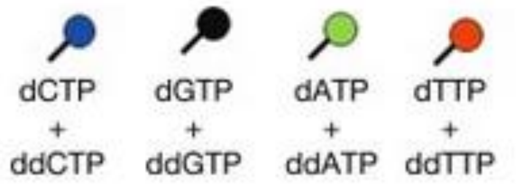
# Sanger sequencing (chain termination)

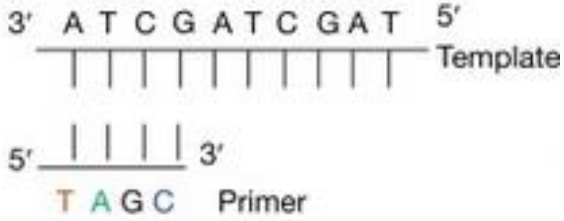
The Sanger sequencing method is the most commonly used analysis technique in genetic diagnostics.

It was also used to sequence the whole human genome.

The following are mixed in a test tube:

1. DNA template
2. One primer
3. DNA polymerase
4. dNTPs: the DNA building blocks A, C, G, T,
   -> a mixture of normal nucleotides
5. ddNTPs: Modified nucleotides with
   fluorescent markers.
   -> do not allow the chain to lengthen, so they stop the reaction.





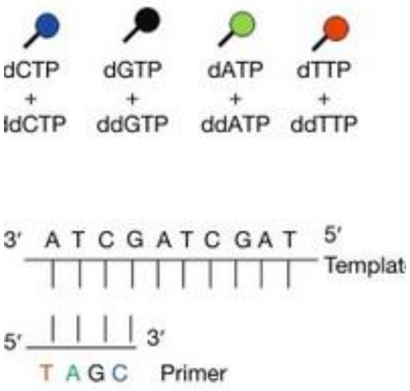
http://www.sciencedirect.com/topics/page/Sanger_sequencing

# Sanger sequencing

Step 1: The primer recognizes and binds to a complementary piece of DNA.

Step 2: DNA polymerase transcribes the code using letters that are freely 'swimming around'.

Step 3: A fluorescent letter is inserted at random in a specific place -> transcription stops.

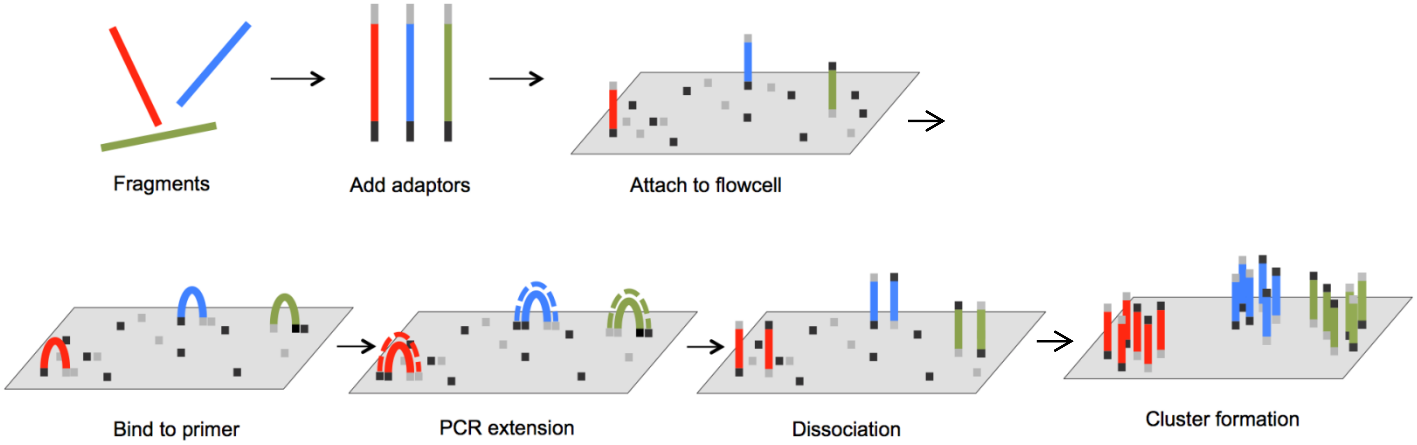


Step 4:
These fragments are arranged in order of length and separated . The fluorescent signals (which are different for each of the 4 nucleotides) are successively received by the sequencer. In this way the original code is 'assembled' by the computer.

http://www.brusselsgenetics.be/

# Next generation sequencing (=Massive parallel sequencing)

- Subsequently the different samples - each with their unique bar code - are pooled
- Then every individual DNA piece from the DNA-library is replicated using its adaptor on a glass slide (flowcell). This process is called **‘clonal’ amplification**.

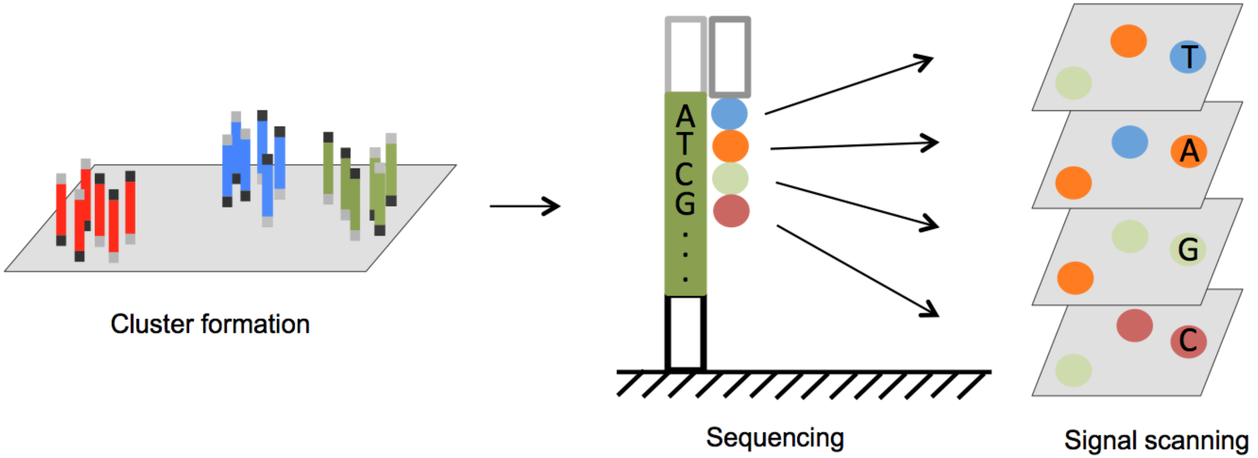


Next step

http://www.brusselsgenetics.be/

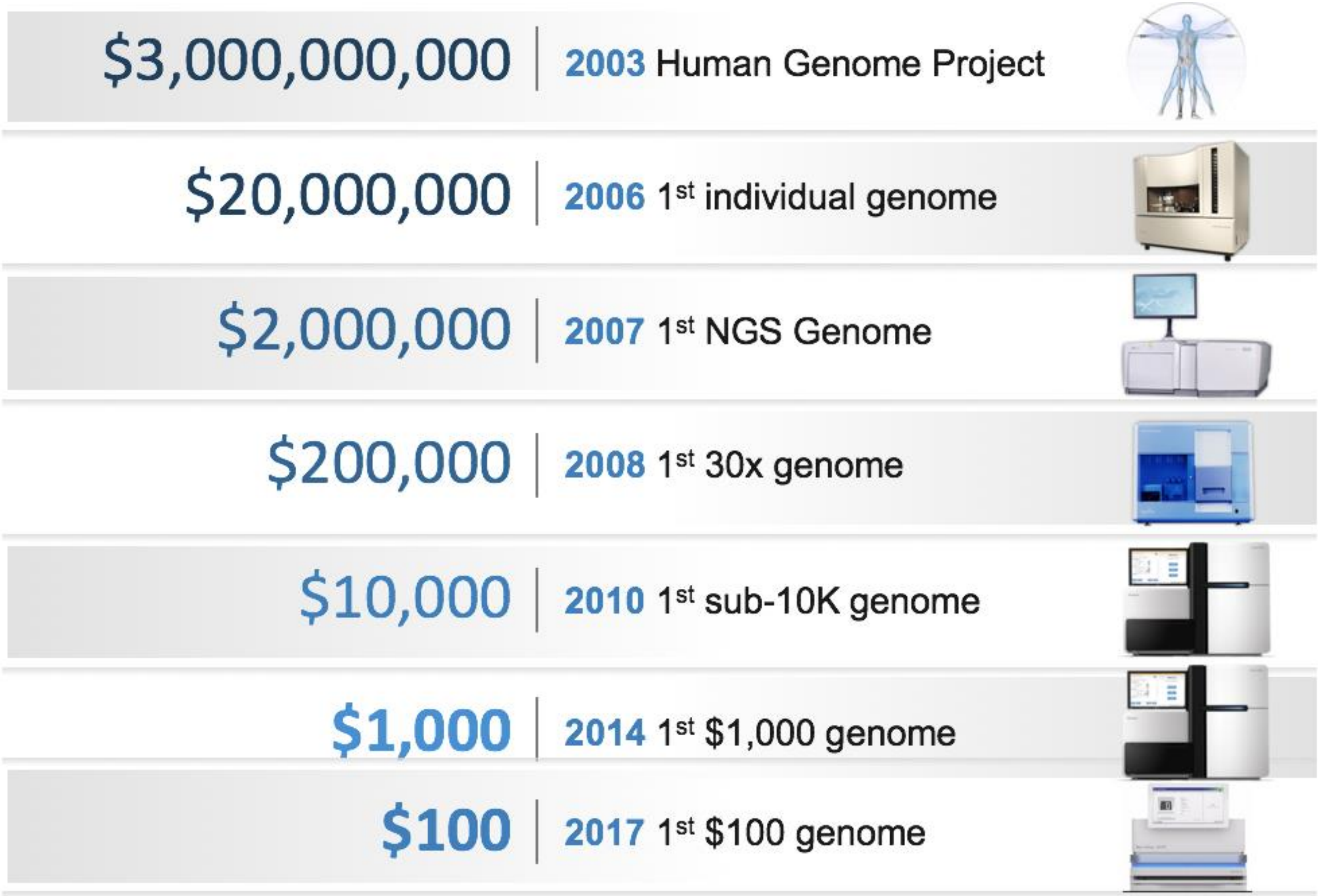
# Next generation sequencing (=Massive parallel sequencing)

Finally the sequence is determined whereby all replicated (amplified) DNA fragments from the DNA-library are sequenced using sequencing primers, a polymerase enzyme and the simultaneous addition of the 4 fluorescent labelled DNA building blocks (sequencing by synthesis).

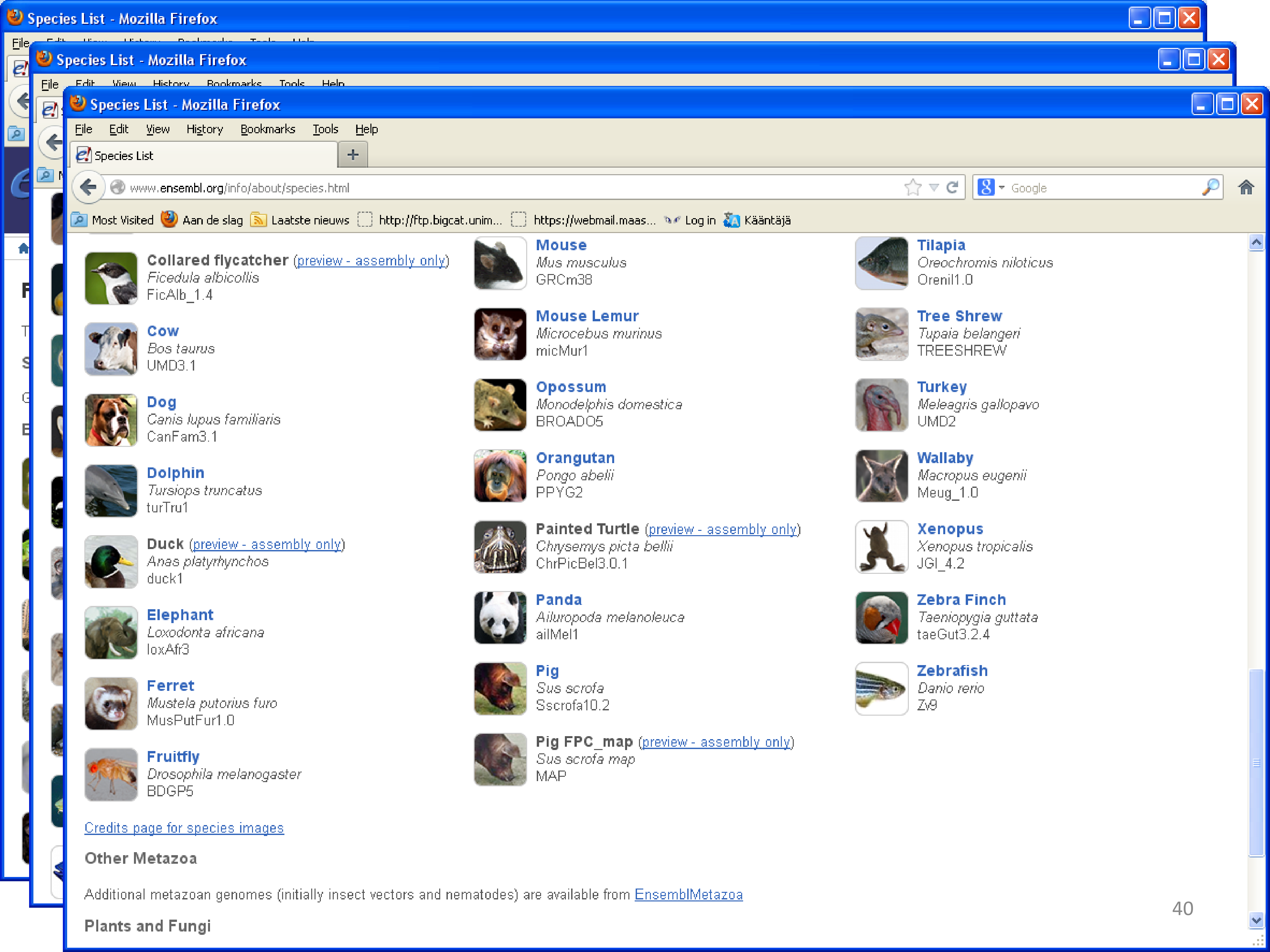


http://www.brusselsgenetics.be/

# Sequencing the Human Genome



| | | |
|---|---|---|
| $3,000,000,000 | 2003 | Human Genome Project |
| $20,000,000 | 2006 | 1st individual genome |
| $2,000,000 | 2007 | 1st NGS Genome |
| $200,000 | 2008 | 1st 30x genome |
| $10,000 | 2010 | 1st sub-10K genome |
| $1,000 | 2014 | 1st $1,000 genome |
| $100 | 2017 | 1st $100 genome |


Species List - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Species List
www.ensembl.org/info/about/species.html
Google
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*
BROADO5

**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*
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**

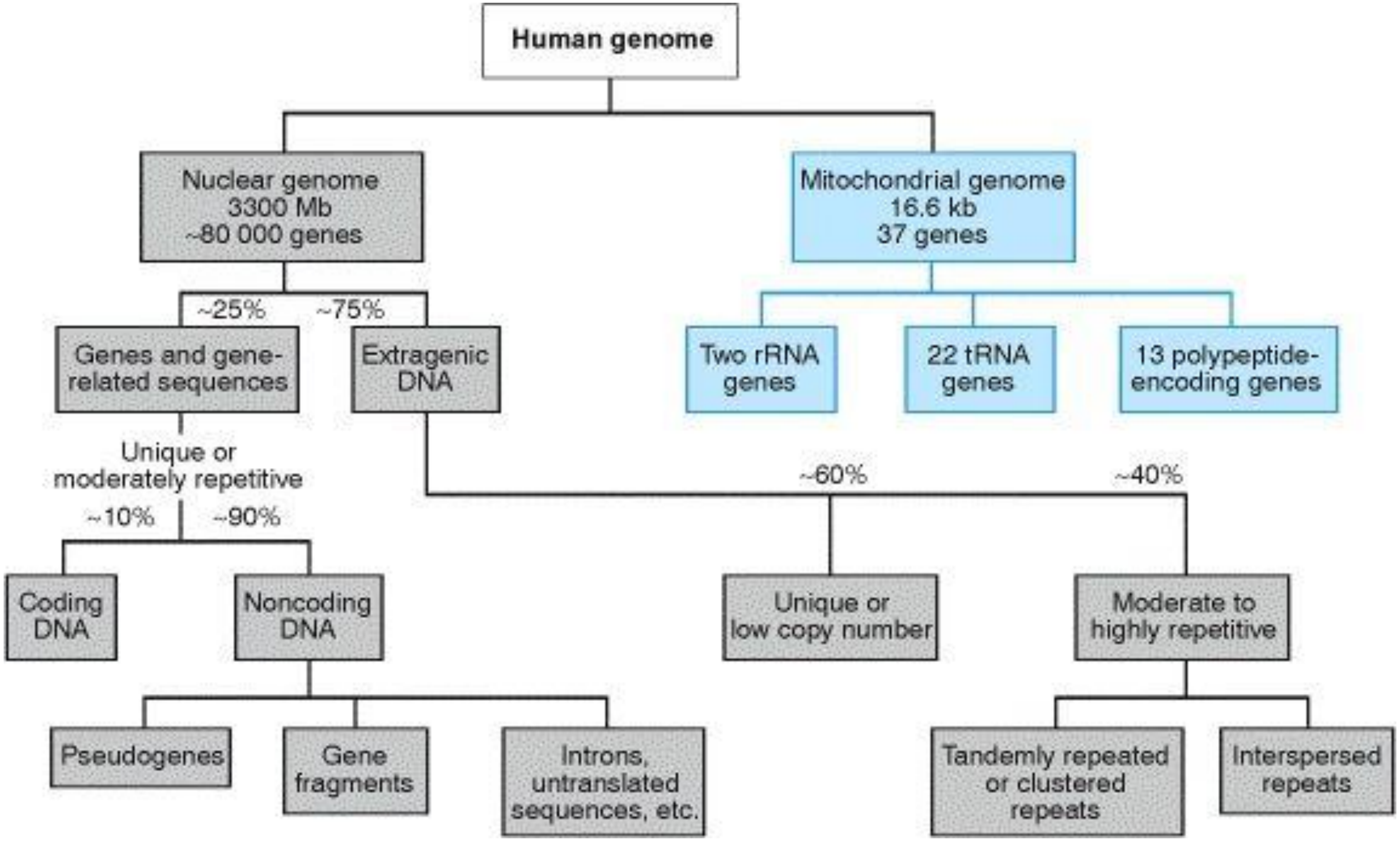
# Number of genes

| **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

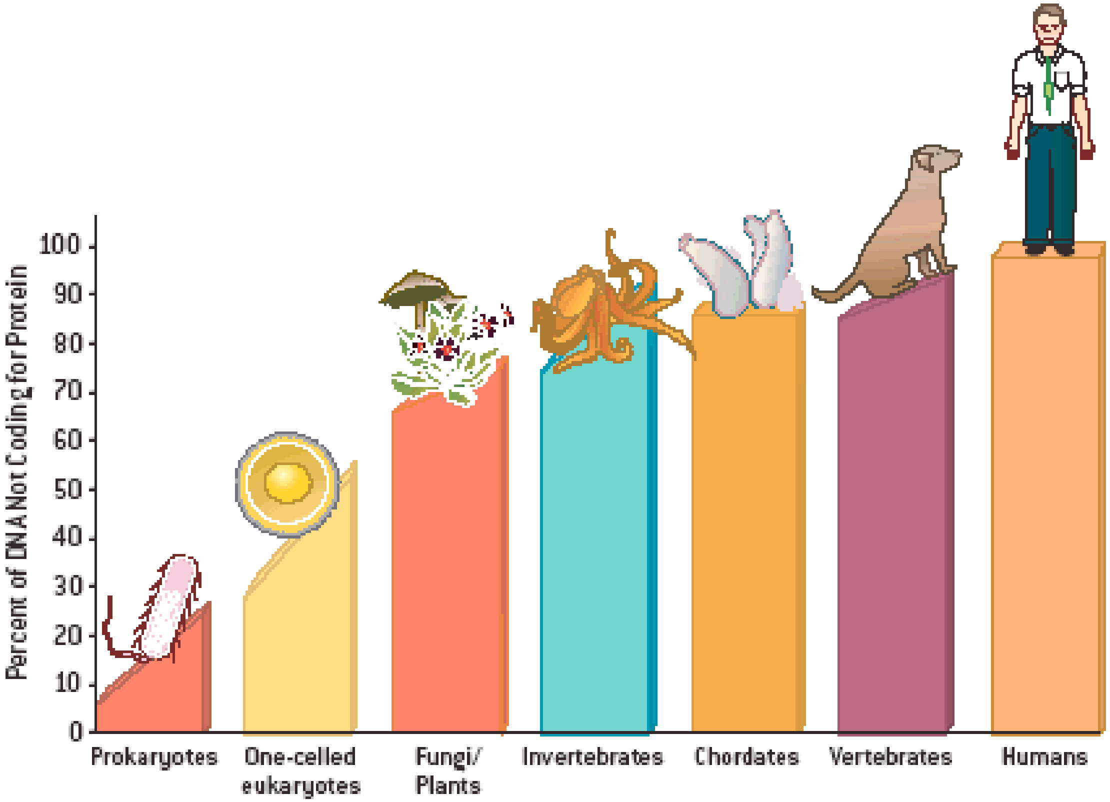
Pray, L. (2008) Eukaryotic genome complexity. Nature Education 1(1)

# Organization of the human genome


Human genome
Nuclear genome
3300 Mb
~80 000 genes
Mitochondrial genome
16.6 kb
37 genes
~25%
~75%
Genes and gene-related sequences
Extragenic DNA
Two rRNA genes
22 tRNA genes
13 polypeptide-encoding genes
Unique or moderately repetitive
~10%
~90%
~60%
~40%
Coding DNA
Noncoding DNA
Unique or low copy number
Moderate to highly repetitive
Pseudogenes
Gene fragments
Introns, untranslated sequences, etc.
Tandemly repeated or clustered repeats
Interspersed repeats

Human Genome Structure
from Strachan & Read, Human Molecular Genetics 2E, Wiley-Liss, 1999

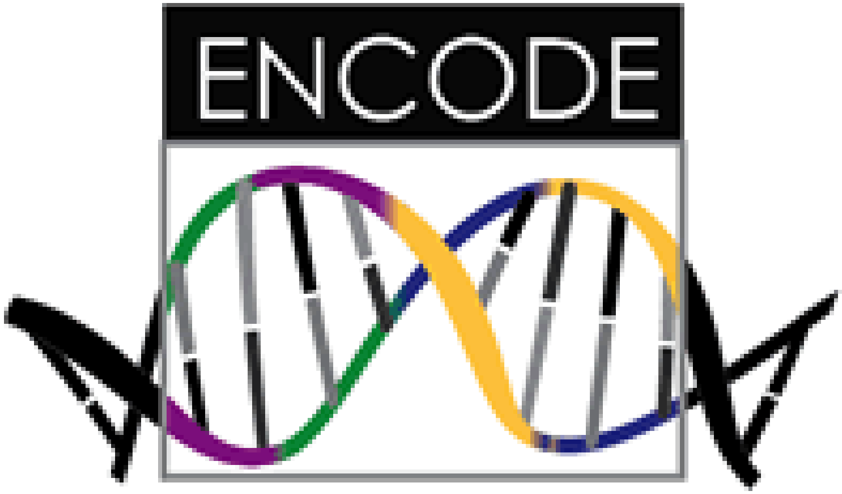
# Non-Protein coding DNA



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.

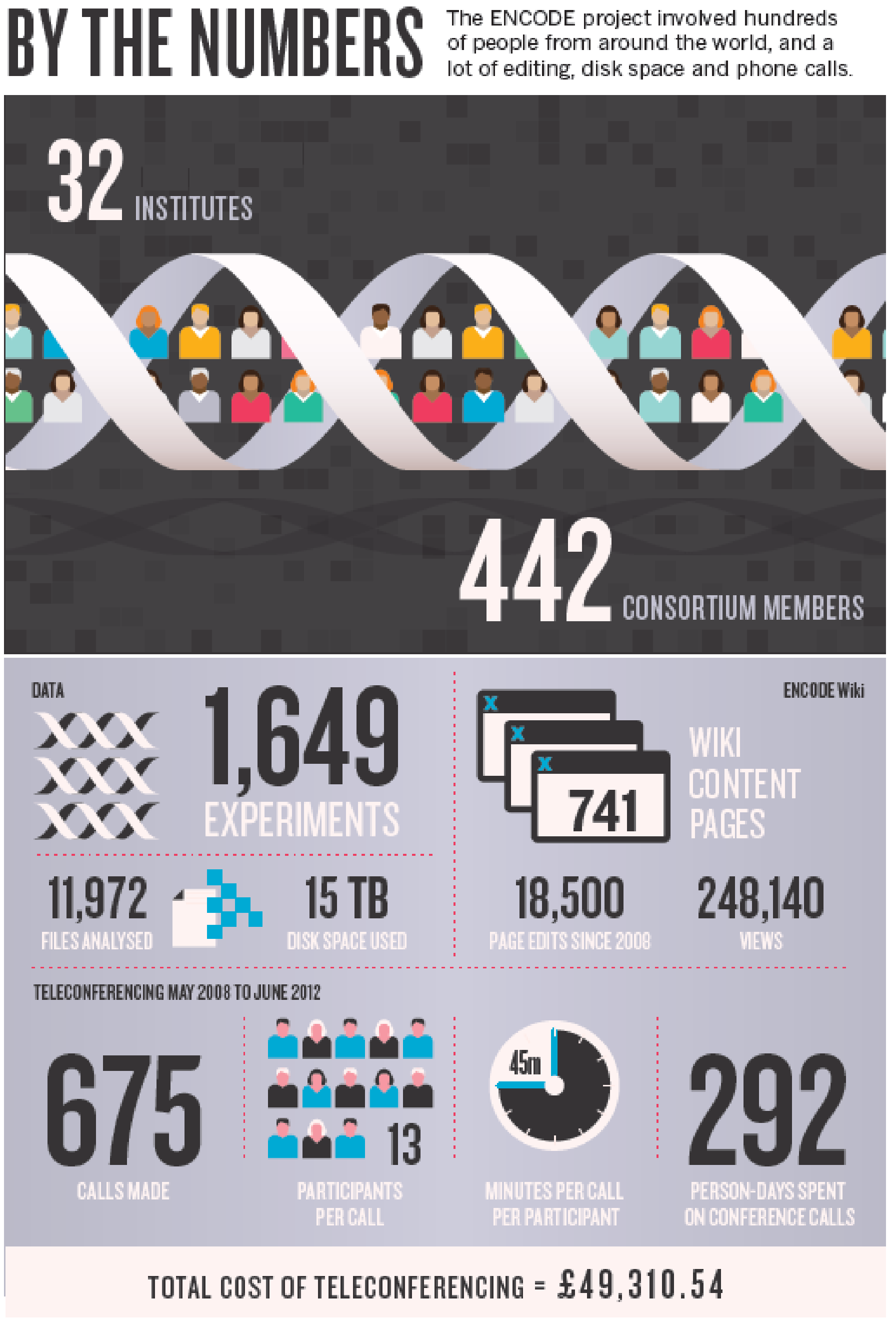
# The ENCODE Project: ENCyclopedia Of DNA Elements

## 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.

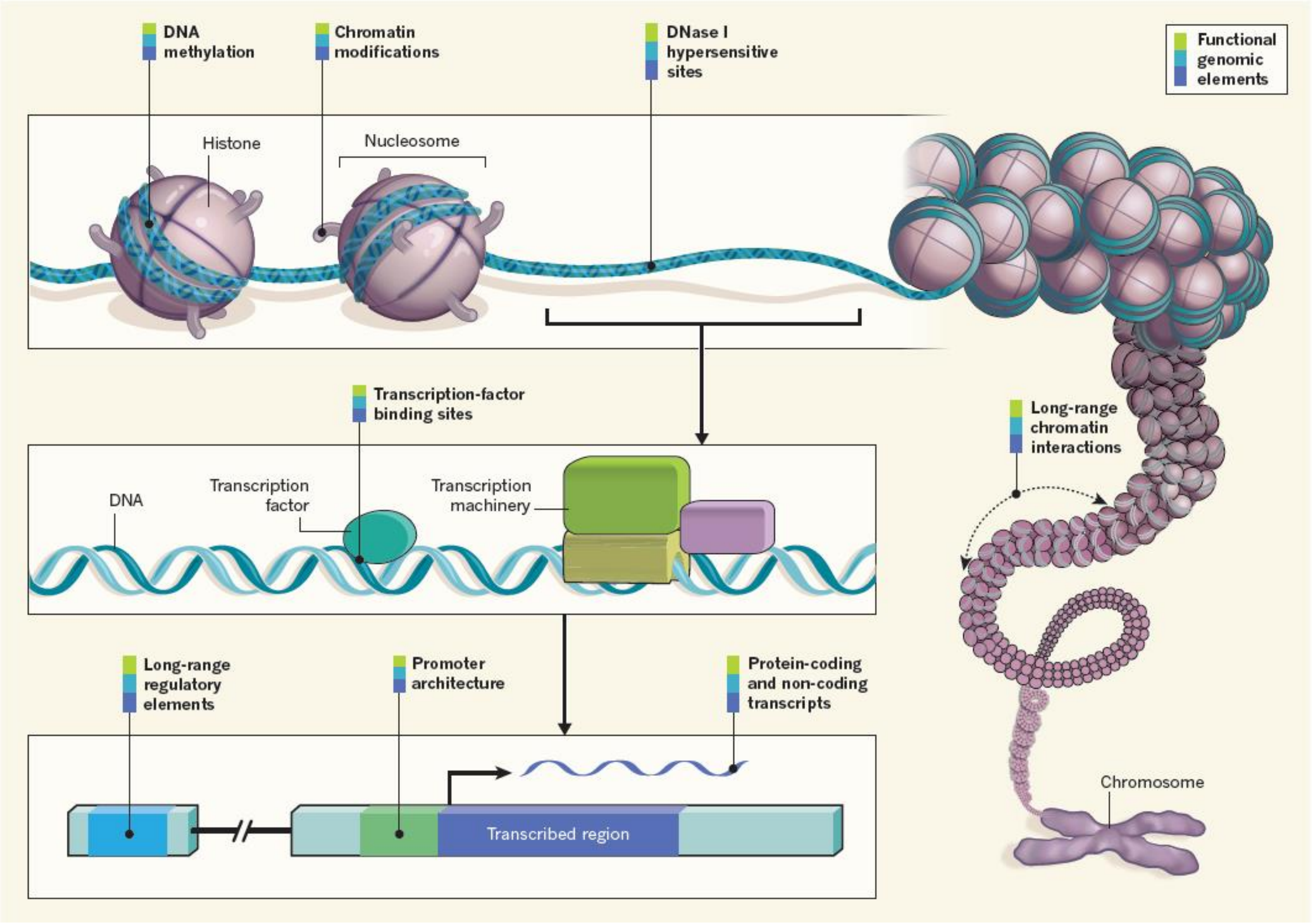


Understanding of the human genome is far from complete. We are missing knowledge on:

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

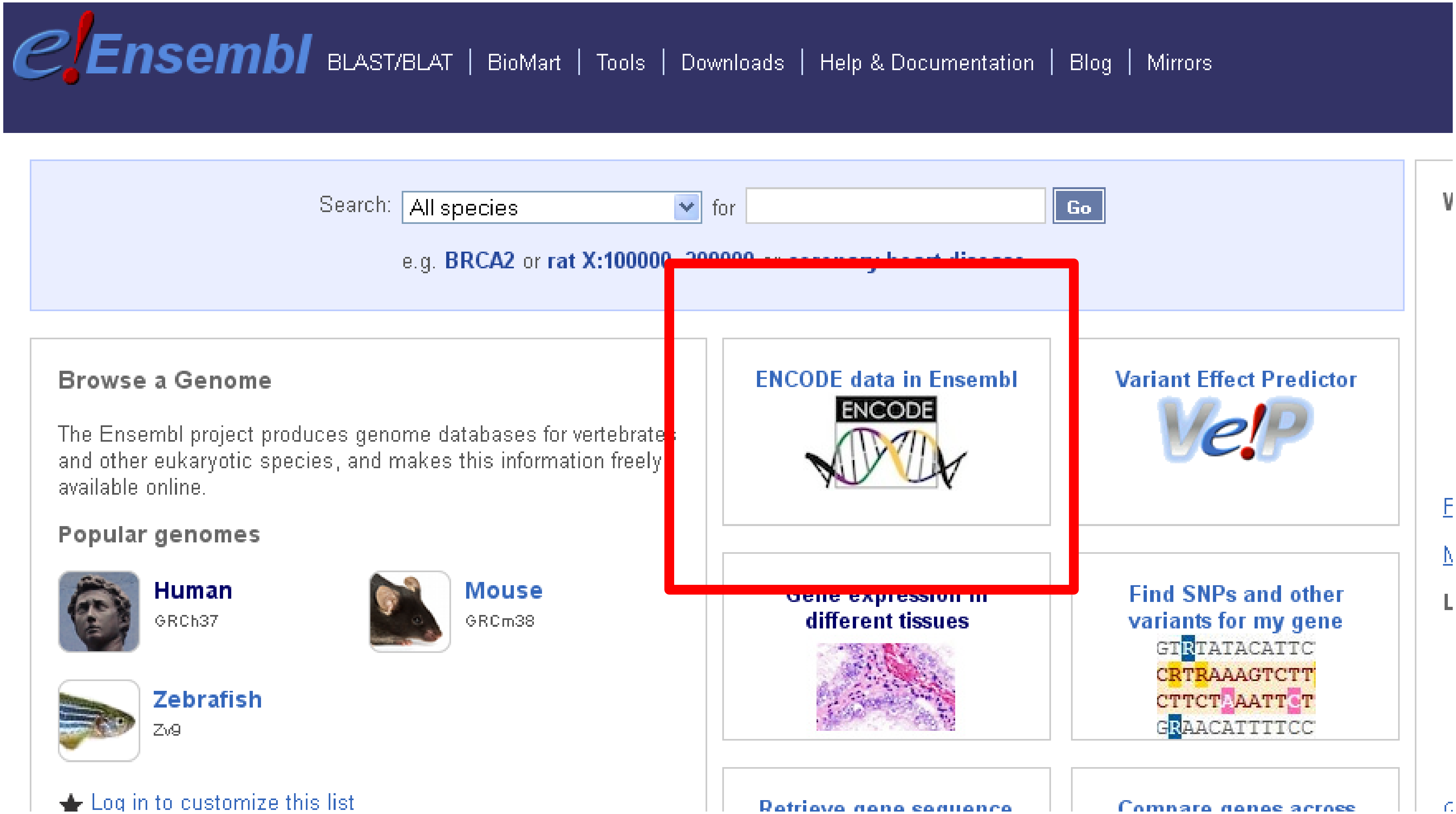
The making of ENCODE: Lessons for big-data projects. Birney E.
Nature. 2012 Sep 6;489(7414):49-51

# Data retrieved from ENCODE project


DNA methylation
Chromatin modifications
DNase I hypersensitive sites
Functional genomic elements
Histone
Nucleosome
Transcription-factor binding sites
DNA
Transcription factor
Transcription machinery
Long-range chromatin interactions
Long-range regulatory elements
Promoter architecture
Protein-coding and non-coding transcripts
Transcribed region
Chromosome

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



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

Search: All species for Go

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

Retrieve gene sequence

Compare genes across

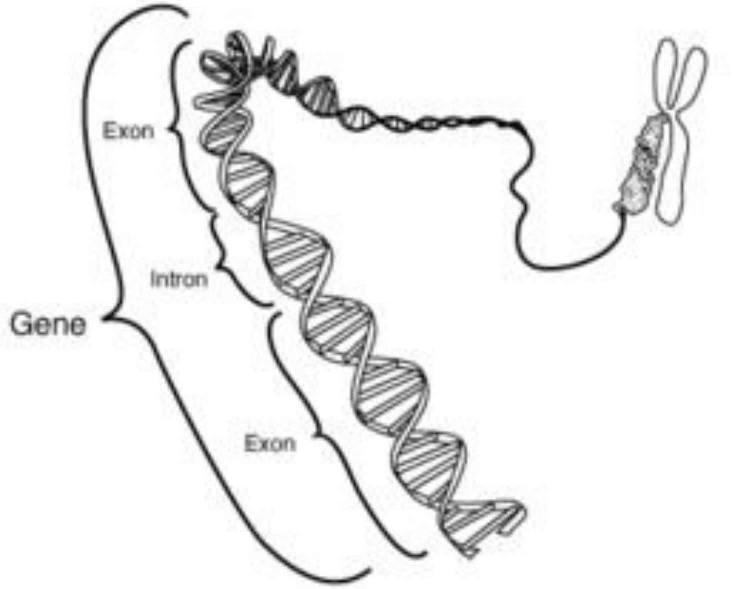
# Genetic Variation

# Genetic variations

- In human beings, 99.9 percent of the bases are the same.

- Remaining 0.1 percent makes a person unique.
  - Different attributes / characteristics / traits
    - how a person looks
    - diseases he or she develops

- Most of those variations are in non-coding regions
  - This does not mean they have no effect!



# Consequences of genetic variations

- Variations can be:
  - Harmless (change in phenotype)
  - Harmful (diabetes, cancer, heart disease, Huntington's disease, and hemophilia )
  - Latent (variations found in coding and regulatory regions that are not harmful on their own, and the change in each gene only becomes apparent under certain conditions, *e.g.* susceptibility to lung cancer)

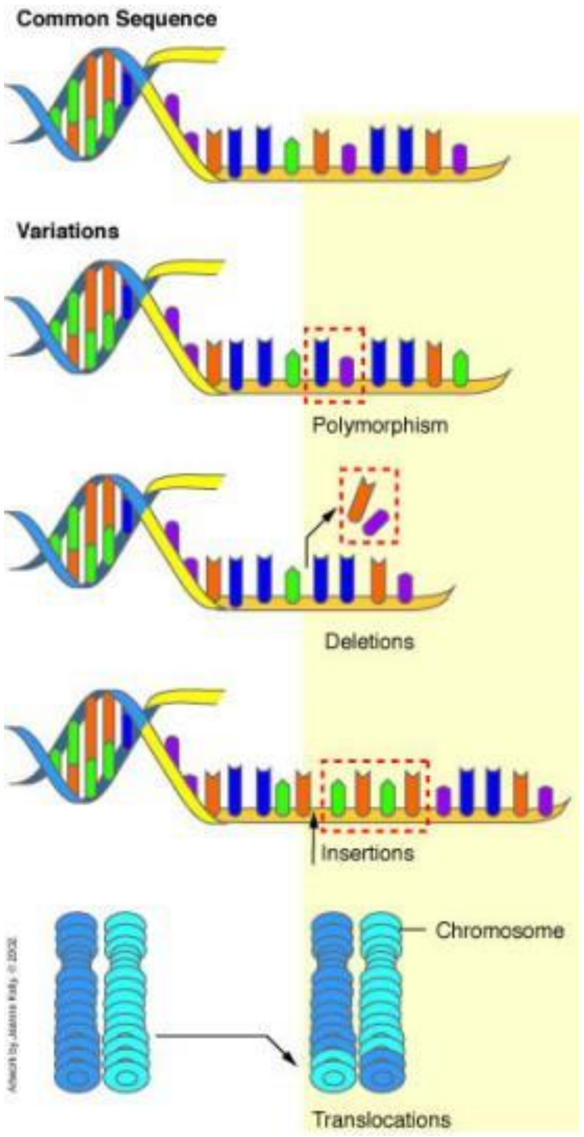
# Types of genetic variation

SNPs

Deletions

Insertions

Translocations



Artwork by Jeanne Kelly © 2002

# Single Nucleotide Polymorphisms (SNP)

- A SNP (single nucleotide polymorphism) is defined as a single base change in a DNA sequence *that occurs in a significant proportion* (more than 1 percent) of a large population

- Currently (2017), dbSNP at NCBI (build 151) has > 100 million validated human SNPs
  - The minimal frequency criterion is <u>not</u> used

# SNP facts

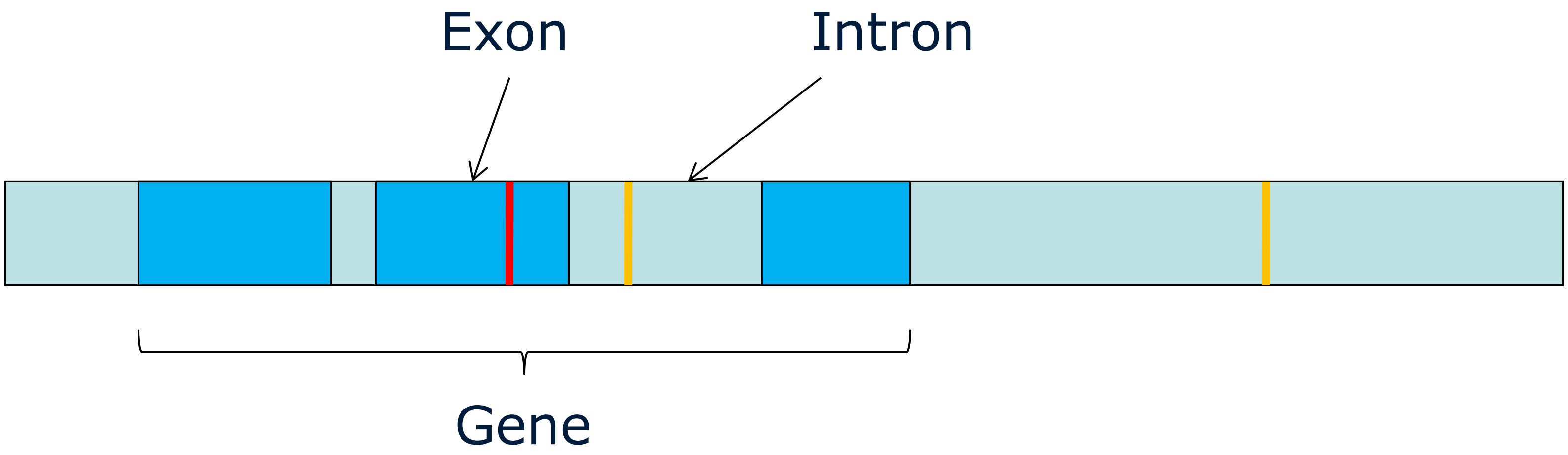
- SNPs are found in
  - coding and (mostly) non-coding regions.

- Occur with a very high frequency
  - about 1 in 1000 bases to 1 in 100 to 300 bases.

- The abundance of SNPs and the ease with which they can be measured make these genetic variations significant.

- SNPs in coding regions alter the protein sequence made by that coding region:
  - **Synonymous** SNP: no protein sequence alteration
  - **Non-synonymous** SNP: protein sequence alteration -> also known as **missense** mutation
    - Special case: a truncating SNP: premature end of protein -> also known as **nonsense** mutation

# Types of SNPs in a gene

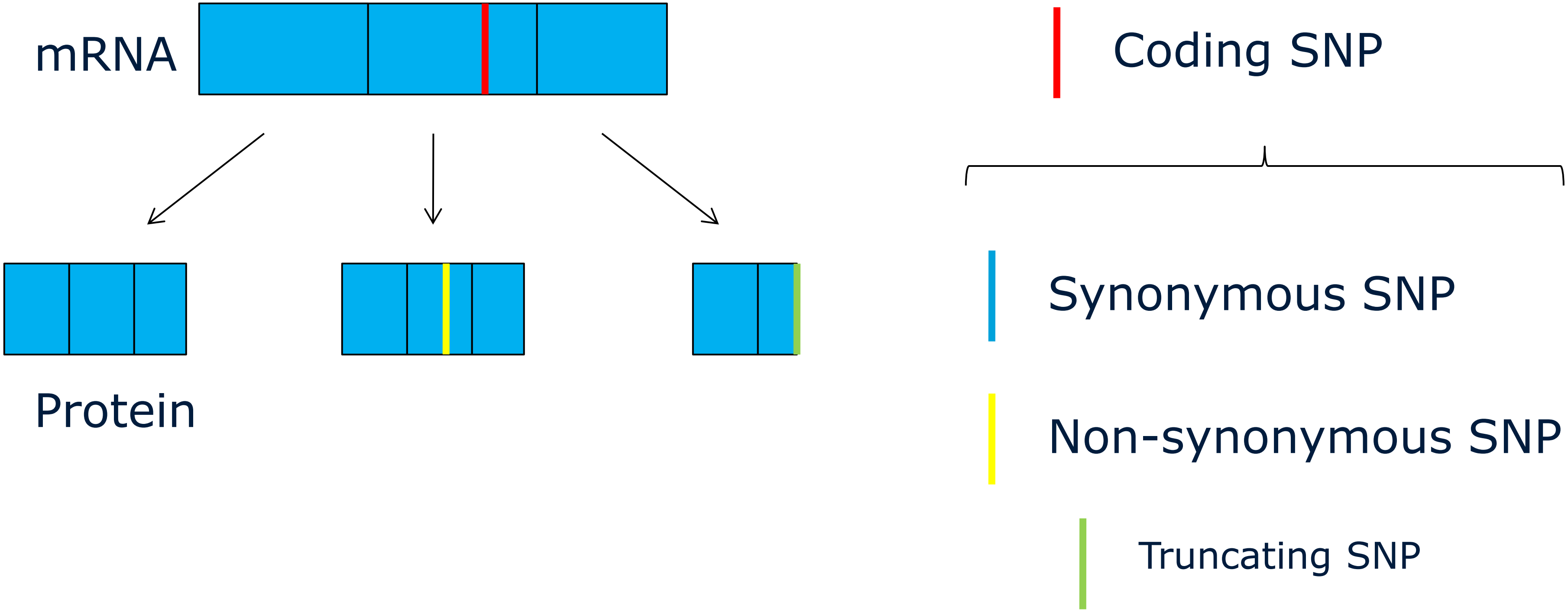


🟨 Non-coding SNP

>

🟥 Coding SNP

# (Coding) SNPs in a protein



mRNA

Protein

🟥 Coding SNP

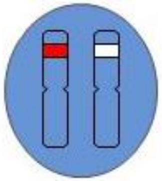
🟦 Synonymous SNP

🟨 Non-synonymous SNP

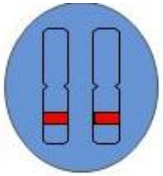
🟩 Truncating SNP

# Inheritance of single-gene disorders

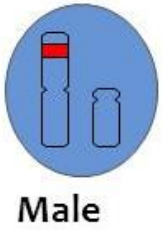
- Errors in DNA sequences
  - Autosomal dominant



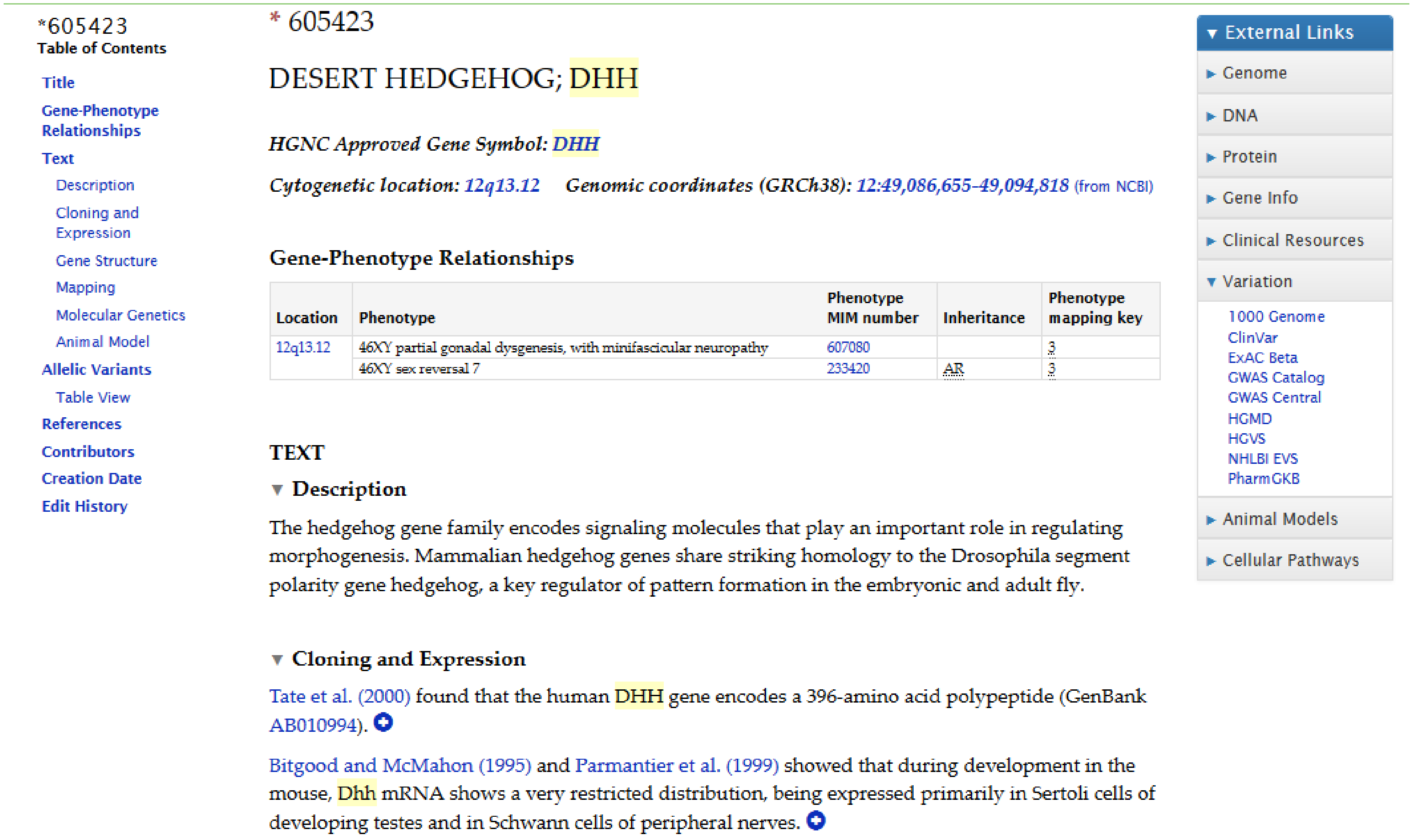
  - Autosomal recessive



  - X-linked recessive
  - X-linked dominant
  - Y-linked (holandric)



# NCBI - OMIM
# Online Mendelian Inheritance in Man



*605423
**Table of Contents**

- Title
- Gene-Phenotype Relationships
- Text
  - Description
  - Cloning and Expression
  - Gene Structure
  - Mapping
  - Molecular Genetics
  - Animal Model
- Allelic Variants
  - Table View
- References
- Contributors
- Creation Date
- Edit History

* 605423

## DESERT HEDGEHOG; DHH

*HGNC Approved Gene Symbol: DHH*

*Cytogenetic location: 12q13.12* *Genomic coordinates (GRCh38): 12:49,086,655-49,094,818* (from NCBI)

### Gene-Phenotype Relationships

| Location | Phenotype | Phenotype MIM number | Inheritance | Phenotype mapping key |
|---|---|---|---|---|
| 12q13.12 | 46XY partial gonadal dysgenesis, with minifascicular neuropathy | 607080 | | 3 |
| | 46XY sex reversal 7 | 233420 | AR | 3 |

### 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 and Expression

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.

# OMIM Content: Scope of Phenotypes

- Single-gene mendelian disease/disorders/phenotypes

(including: cystic fibrosis, sickle cell anemia, achondroplasia, phenotypic traits such as hair and eye color, susceptibility to drug reaction as in malignant hyperthermia and warfarin sensitivity, altered reaction to infection such as herpes simplex encephalitis and progression to AIDS in HIV infection, germline susceptibilities to cancer such as BRCA1 and breast/ovarian cancer, etc.)

- Complex diseases with significant single gene contribution (

such as: complement factor H and age related macular degeneration)

- Descriptions of recurrent deletion and duplication syndromes

(e.g., Potocki-Shaffer syndrome, and chromosome 10q26 deletion syndrome)

# How to name a SNP? – SNP identifiers

- A standard ID for SNPs is the dbSNP ID
  - also called "rs number"
  - example: rs4986852
  - Standardised, unique, stable

- An alternative for disease related SNPs is the OMIM variation ID
  - example: 113705.0011 (this is: gene_number.SNP_number)
  - Standardised, unique, stable

- A final possibility is the
  - For non-coding or coding SNPs: variation
    - Example: BRCA1, 2978G>A
  - For coding SNPs (also): mutation
    - Example: BRCA1, SER1040ASN
  - Easier to interpret, but not stable

# Gene Ontology

# Gene Ontology

- Built for a very specific purpose:

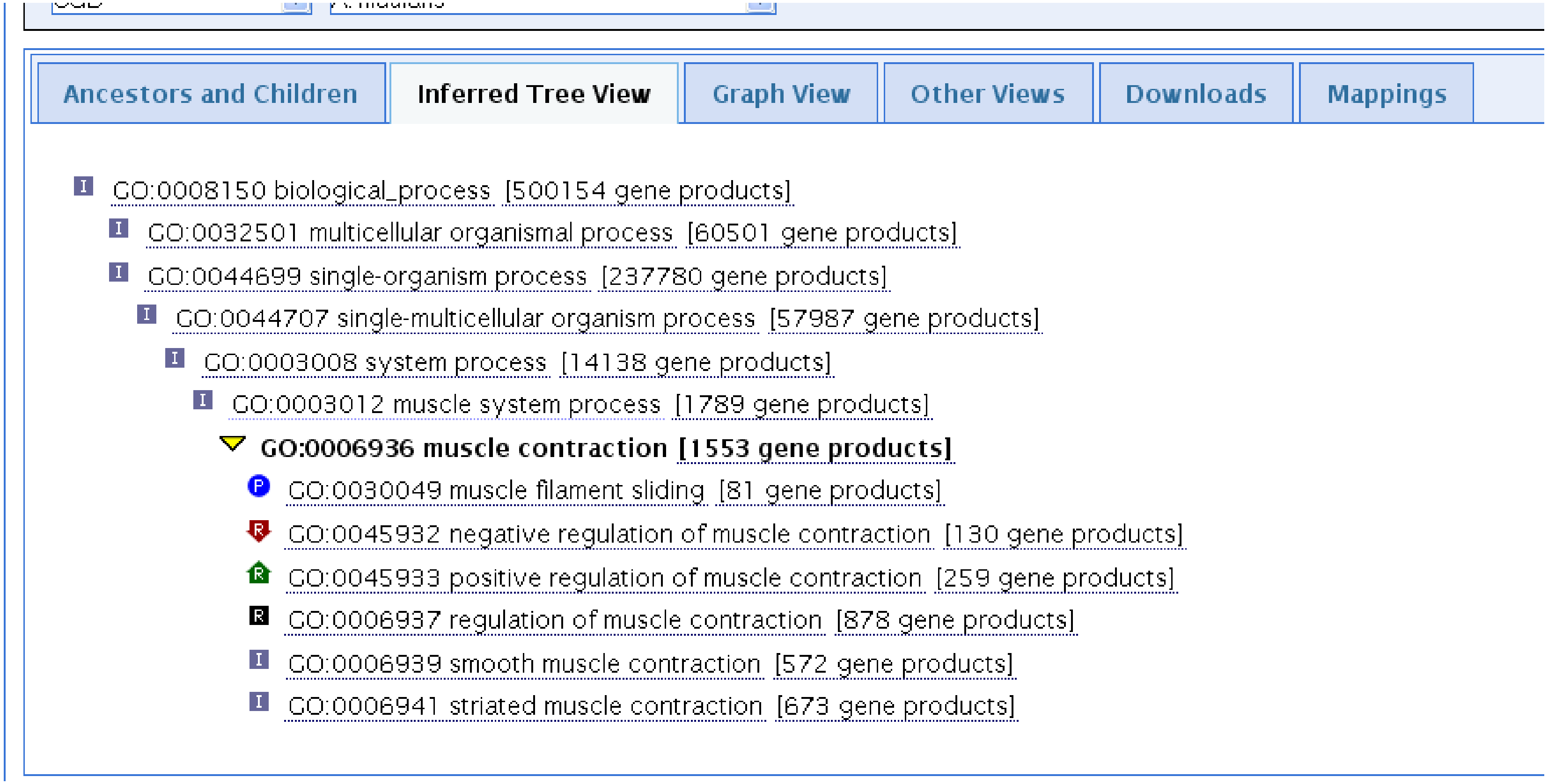
“annotation of genes and proteins in genomic and protein databases”

- Applicable to all species

# The 3 Gene Ontologies

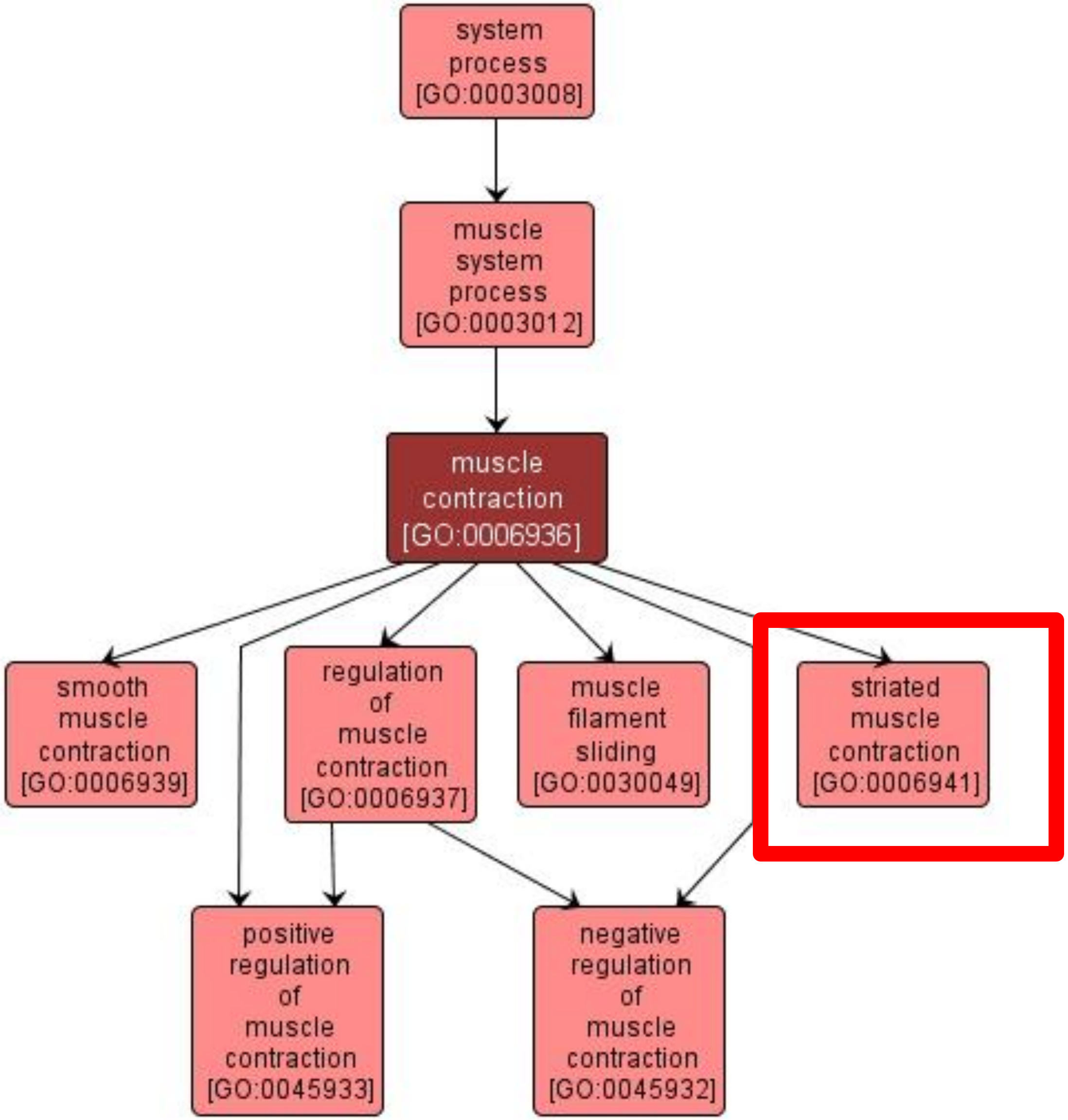
- **Molecular Function** = elemental activity/task
  - the tasks performed by individual gene products; examples are ***carbohydrate binding*** and ***ATPase activity***
- **Biological Process** = biological goal or objective
  - broad biological goals, such as ***mitosis*** or ***purine metabolism***, 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

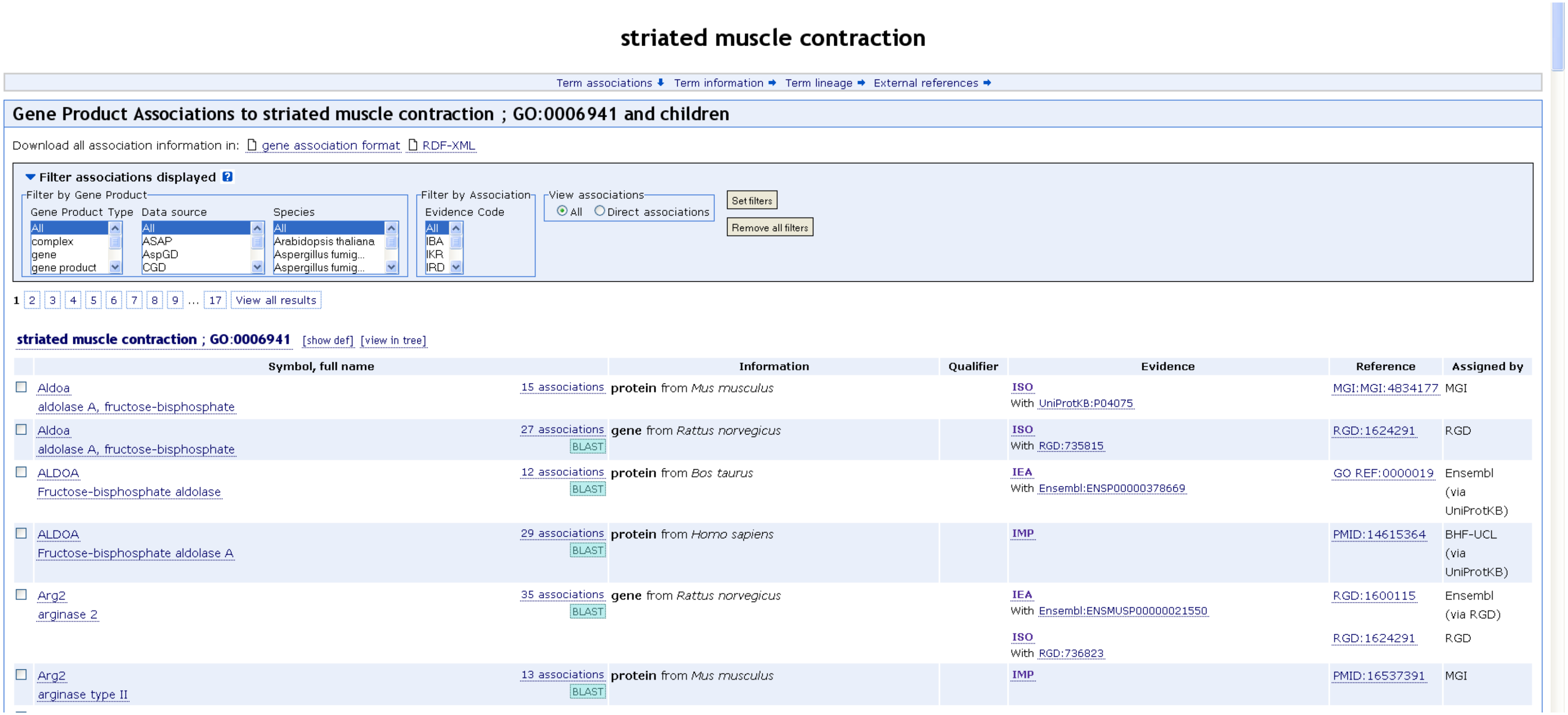


```
I GO:0008150 biological_process [500154 gene products]
  I GO:0032501 multicellular organismal process [60501 gene products]
  I GO:0044699 single-organism process [237780 gene products]
    I GO:0044707 single-multicellular organism process [57987 gene products]
      I GO:0003008 system process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

# GO muscle contraction – tree view


system process [GO:0003008]
muscle system process [GO:0003012]
muscle contraction [GO:0006936]
smooth muscle contraction [GO:0006939]
regulation of muscle contraction [GO:0006937]
muscle filament sliding [GO:0030049]
striated muscle contraction [GO:0006941]
positive regulation of muscle contraction [GO:0045933]
negative regulation of muscle contraction [GO:0045932]

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

# Searching and Browsing GO

- Gene Ontology consortium:
  http://geneontology.org/

- AmiGO 2
  http://amigo.geneontology.org/amigo

# WikiPathways

# WikiPathways

- Biological pathway database

  www.wikipathways.org
- Founded in 2008 by Gladstone Institutes and the Department of Bioinformatics in Maastricht

- **Wiki**Pathways - What is a wiki?

  "A wiki is an application, typically a web application, which allows collaborative modification, extension, or deletion of its content and structure."

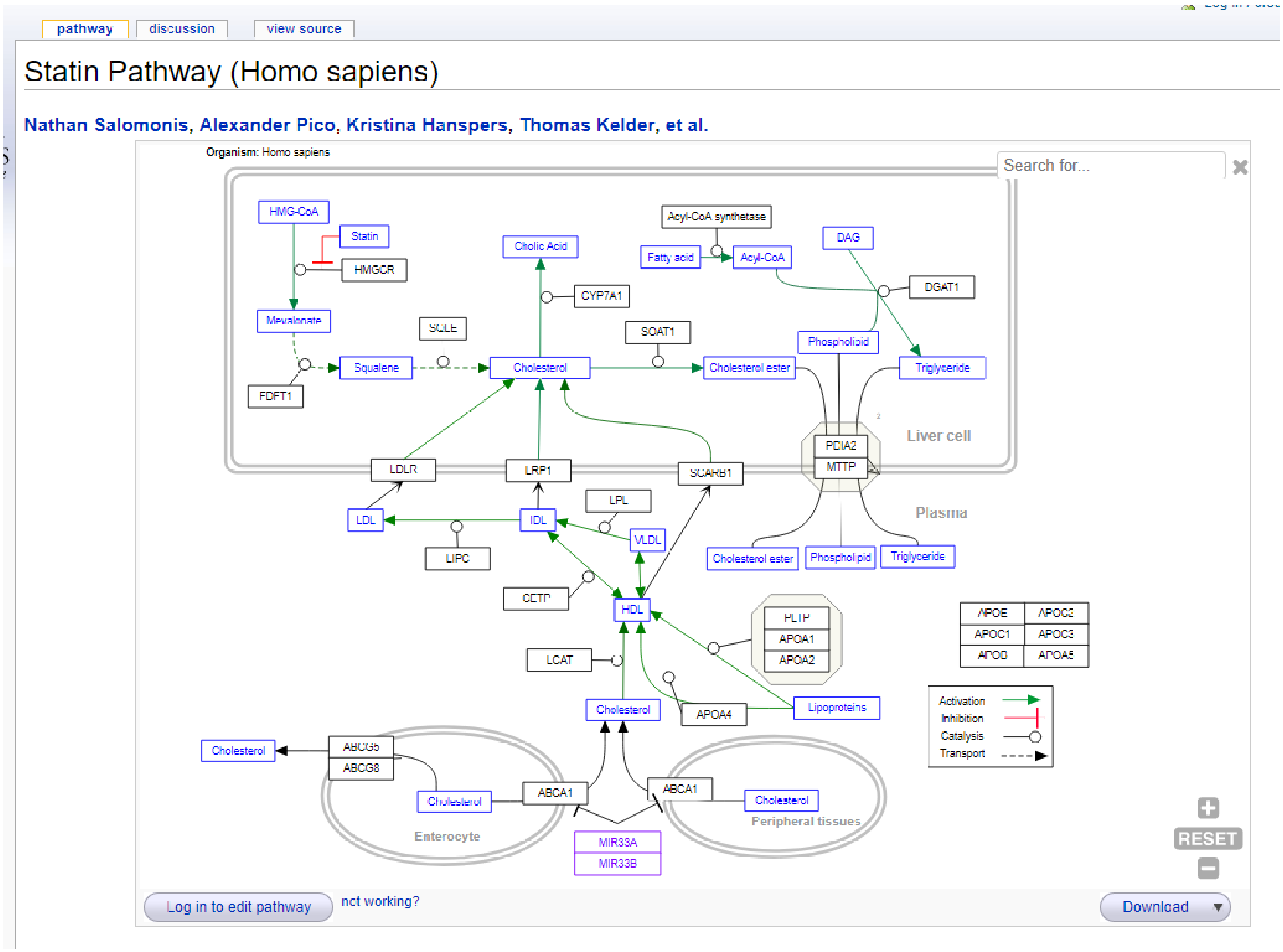
*Definition Wikipedia 2017*

Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, Melius J, Waagmeester A, Sinha SR, Miller R, Coort SL, Cirillo E, Smeets B, Evelo CT, Pico AR **WikiPathways: capturing the full diversity of pathway knowledge.** *Nucleic Acids Research.* 2015 Oct 19;44(D1):D488-94. doi:10.1093/nar/gkv1024

# WikiPathways

- A Wikipedia for pathways
  - Collection and curation of knowledge
  - Community curated
    - Everybody can contribute pathways
    - Everybody can edit and curate pathways
    - Everybody can use the pathway collections
  - Tools
    - Not just images but fully annotated models
    - Interactive pathway viewer
    - Full pathway editor and analysis software: PathVisio
  - New findings can be added immediately - fast!

# Pathway pages



Statin Pathway (Homo sapiens)

Nathan Salomonis, Alexander Pico, Kristina Hanspers, Thomas Kelder, et al.

https://www.wikipathways.org/index.php/Pathway:WP430

# Questions

# Practical session

- Ensembl tutorials
- Ensembl genome browser

- Several NCBI databases
  - Gene
  - OMIM

- WikiPathways

# QUIZ at GoSoapBox

- Go to **app.gosoapbox.com** on your own computer, tablet, or smartphone.
- Type in **233-291-104** in the Access Code field.
- Enter your name prior to joining.