Dr. Lars Eijssen

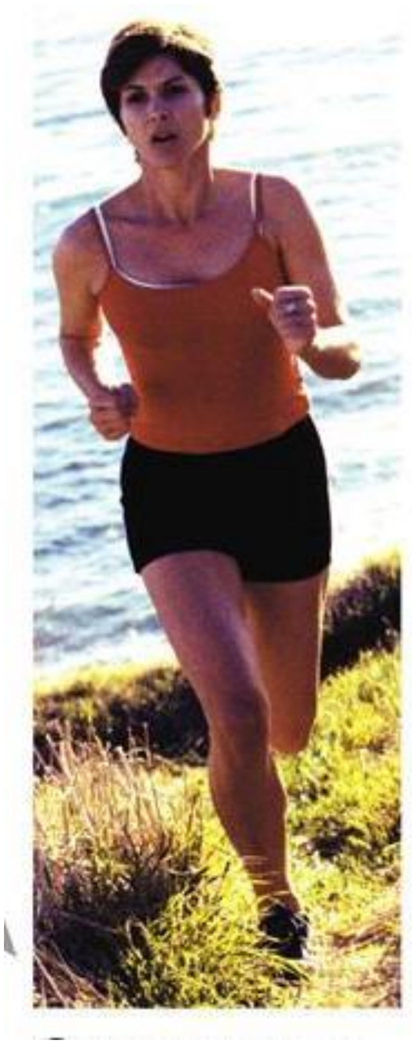
l.eijssen@maastrichtuniversity.nl

May 15th, 2017

Content

- Introduction
- Background
- Learning goals
- What is a database?
- What are biological sequence databases?
    - NCBI, Ensembl, UCSC
- Identifiers
- Human genome project
- ENCODE project
- Gene Ontology

What happens with the human body when you are running?



Organ systems work together

- Muscular system - pulls on the bones to enable you to move
- Respiratory system - makes sure your muscles have enough oxygen for respiration
- Cardiovascular system- provides oxygen and glucose to the skeletal muscle cells
- Nervous system – controls your movements and heart rate

Human body structure

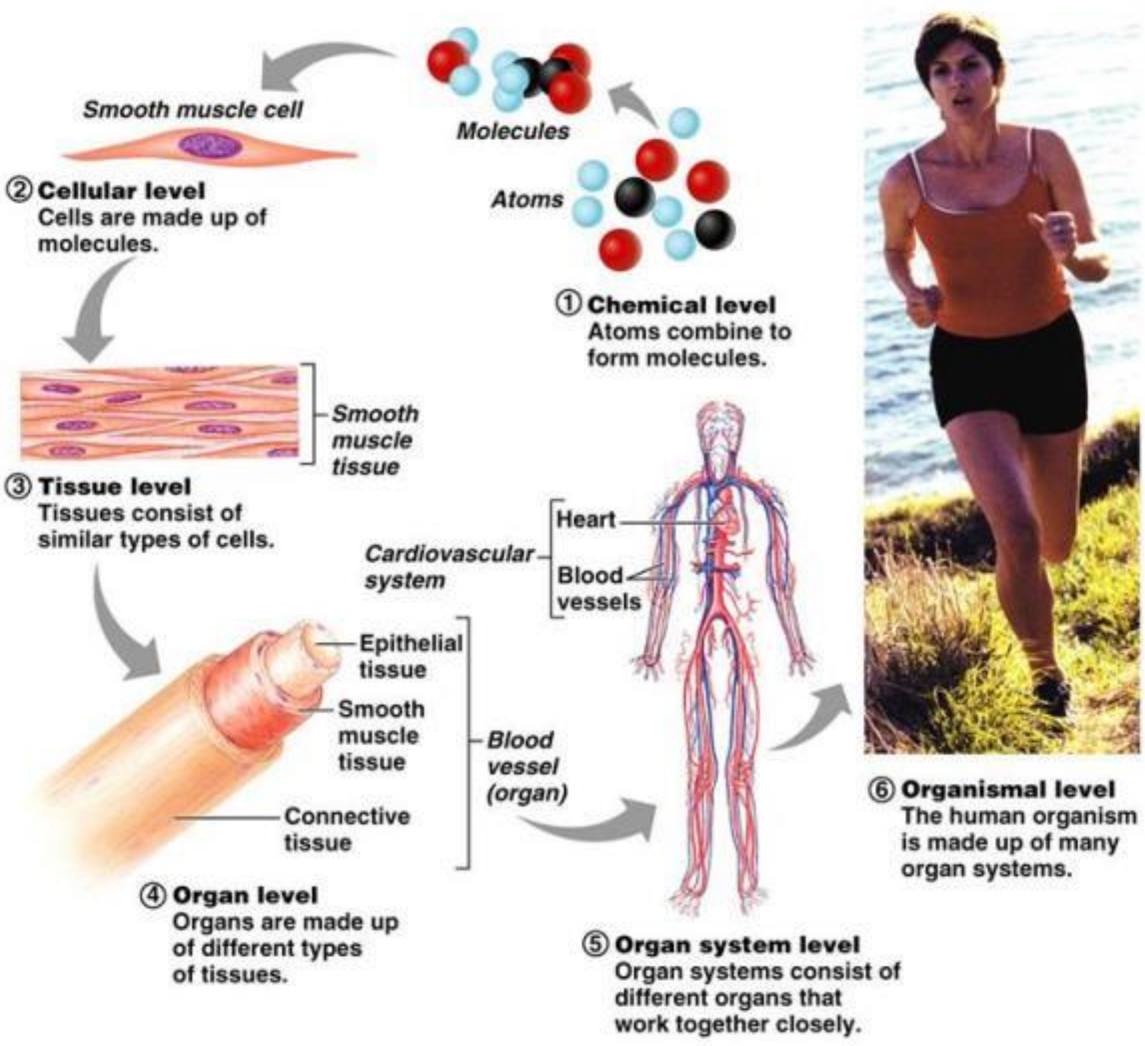
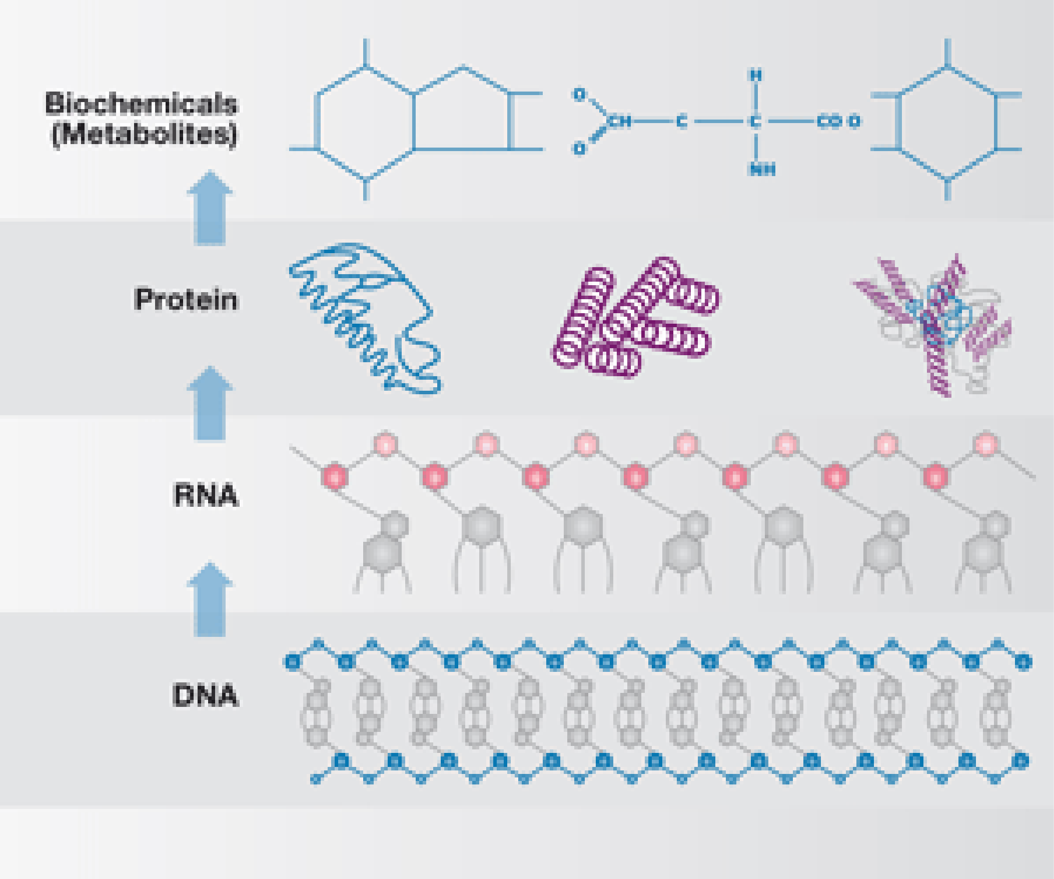
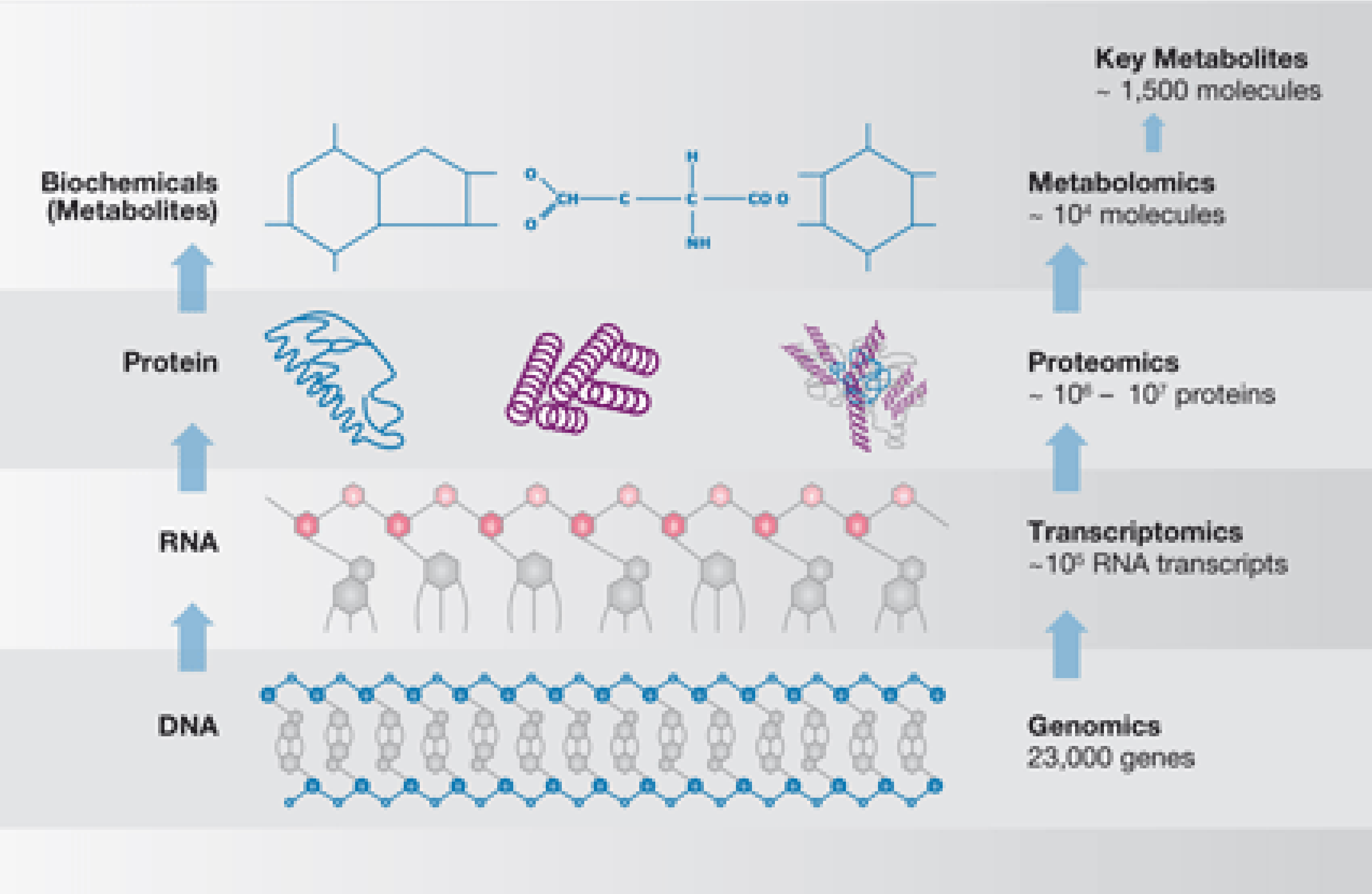

Smooth muscle cell
Molecules
Atoms
② Cellular level
Cells are made up of molecules.
① Chemical level
Atoms combine to form molecules.
Smooth muscle tissue
③ Tissue level
Tissues consist of similar types of cells.
Heart
Cardiovascular system
Blood vessels
Epithelial tissue
Smooth muscle tissue
Blood vessel (organ)
Connective tissue
④ Organ level
Organs are made up of different types of tissues.
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

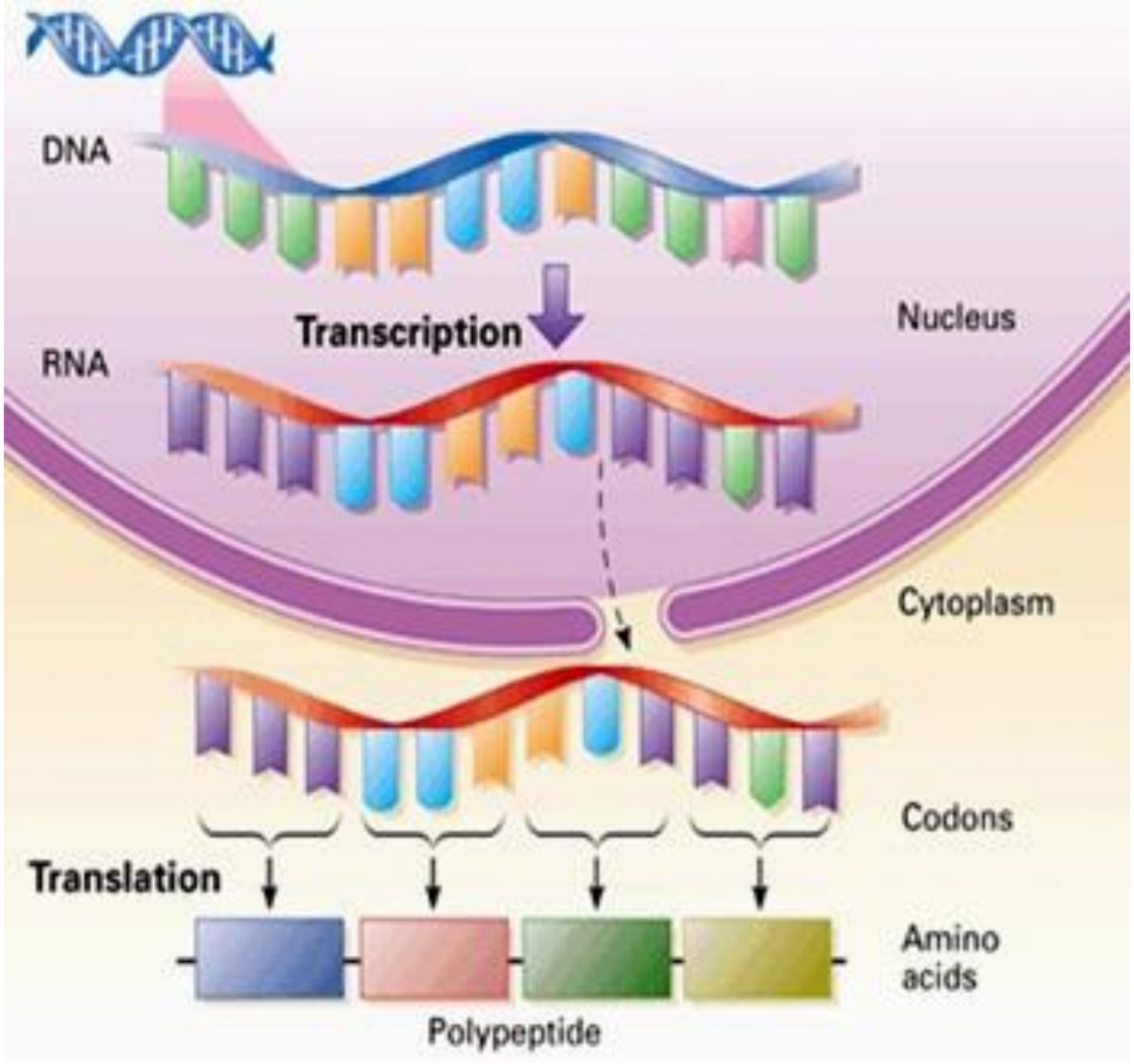
(Bio)Molecules
Individual players are important



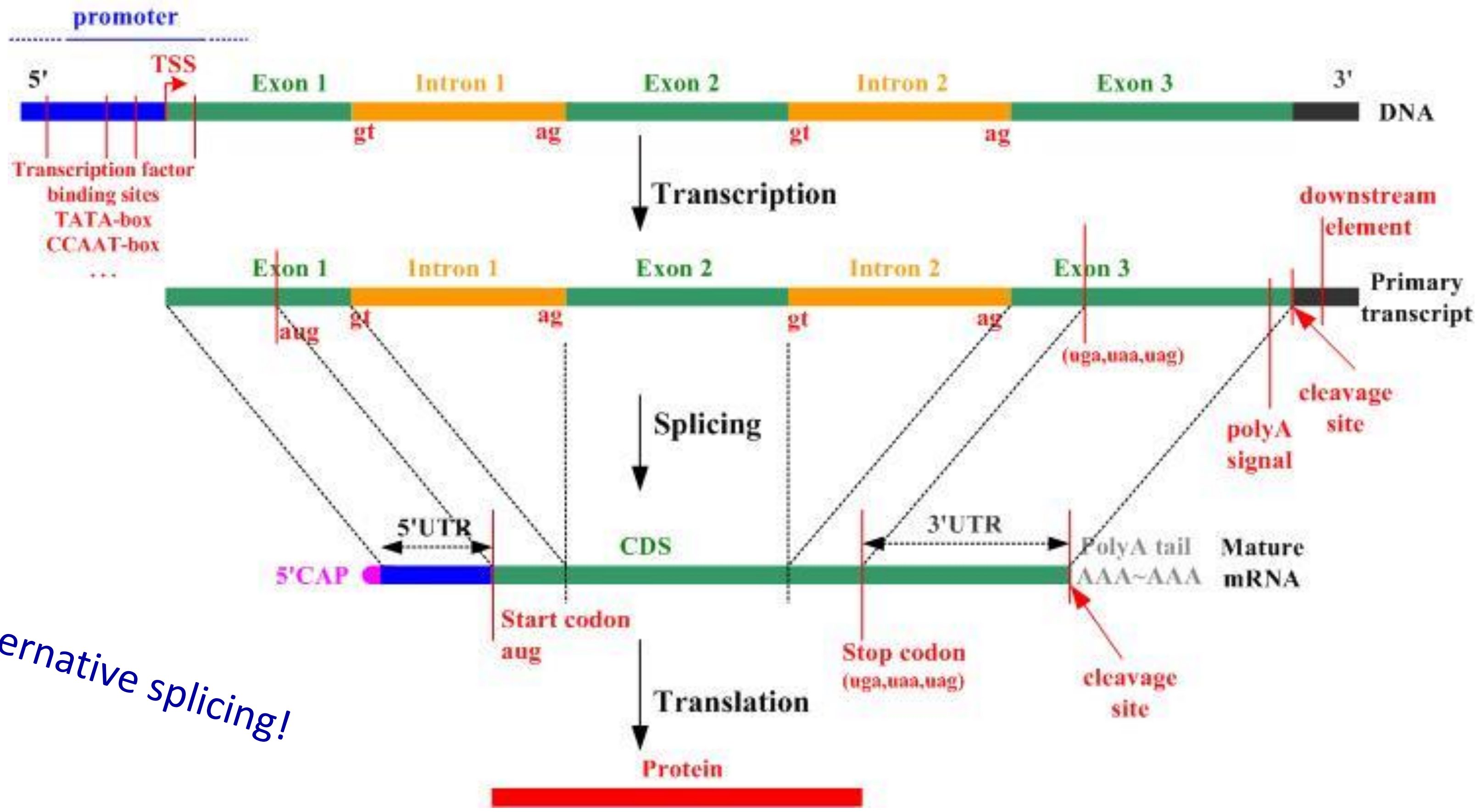
Heaps of knowledge on biomolecules online available.



Protein synthesis



Gene structure



Alternative splicing!

CDS = Coding DNA Sequence
UTR = UnTranslated region

www.carolguze.com

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?
- Two projects aimed at deciphering the content of the human genome, the human genome project & ENCODE.
- What is Gene Ontology?

What is a database

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

(till 2:58)

Genes in stead of persons

Name	Identifier	Sequence	Synonyms	Chromosomal location	Disease	Many more
Gene 1	2456	AGTCCCGT	DAH, HSD	4q12	Cancer	.....
Gene 2	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

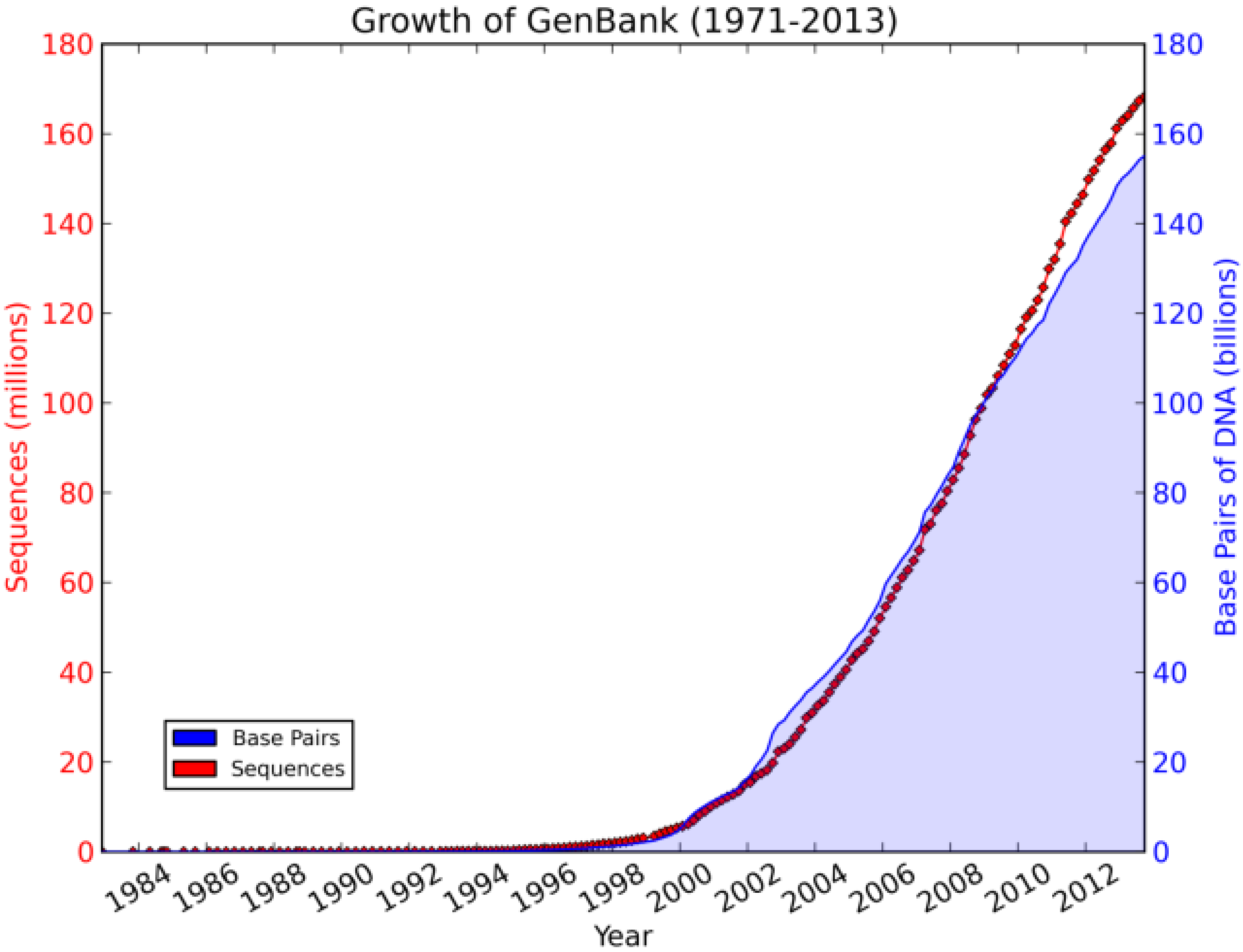
3. Protein Databases
   - UniProt

Lecture protein structures

NCBI nucleotide databases

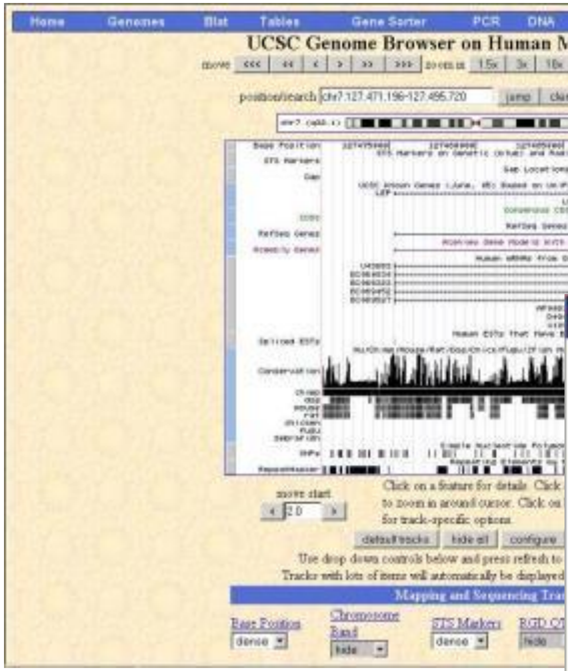
- GenBank
  - Individual submissions (DNA, mRNA, eiwit)
  - Bulk submissions (Genome centers)
    - High throughput sequencing (DNA)
    - Expressed Sequence Tags (mRNA)
- RefSeq
  - Curated subset of GenBank
  - “Reference” sequence
  - Single sequence per locus / molecule

Growth of GenBank



Genome-centered databases

UCSC



Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Help

UCSC Genome Browser on Human May 2004 Assembly

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

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

move start

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

default tracks hide all configure

Use drop down controls below and press refresh to Tracks with lots of items will automatically be displayed

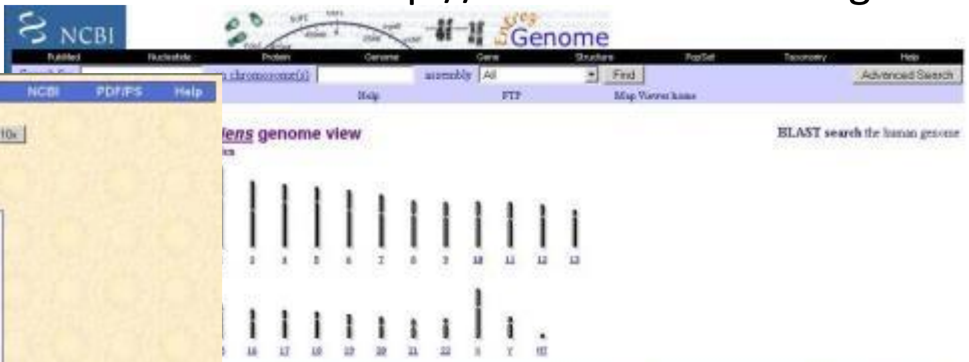
Mapping and Sequencing Tra

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

http://genome.ucsc.edu/

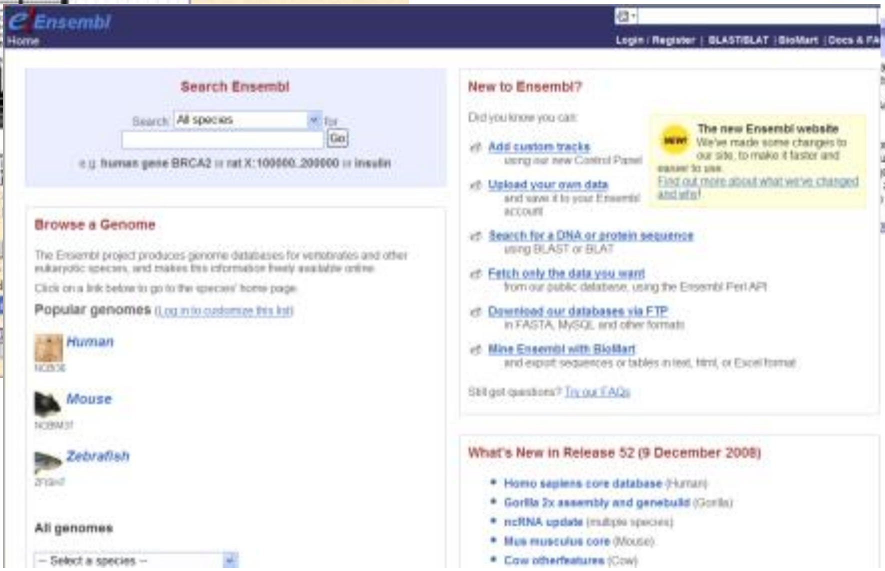
NCBI

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



...iens genome view

BLAST search the human genome



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 Zv7

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
- 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

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!

Still got questions? Try our FAQs

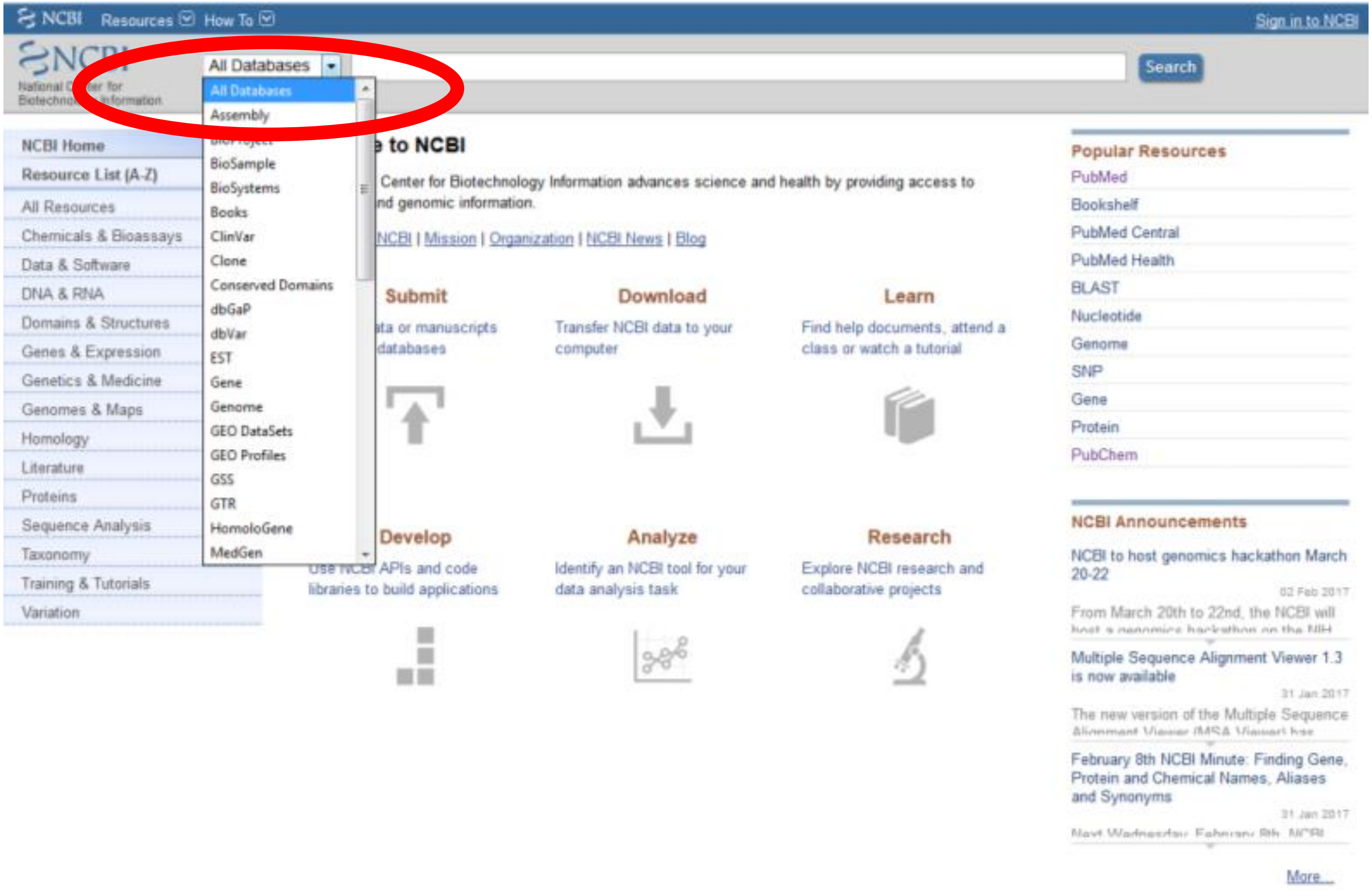
What's New in Release 52 (9 December 2008)

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

Ensembl

http://www.ensembl.org/

NCBI homepage



e to NCBI

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

NCBI | Mission | Organization | NCBI News | Blog

Submit
ta or manuscripts databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

NCBI Announcements

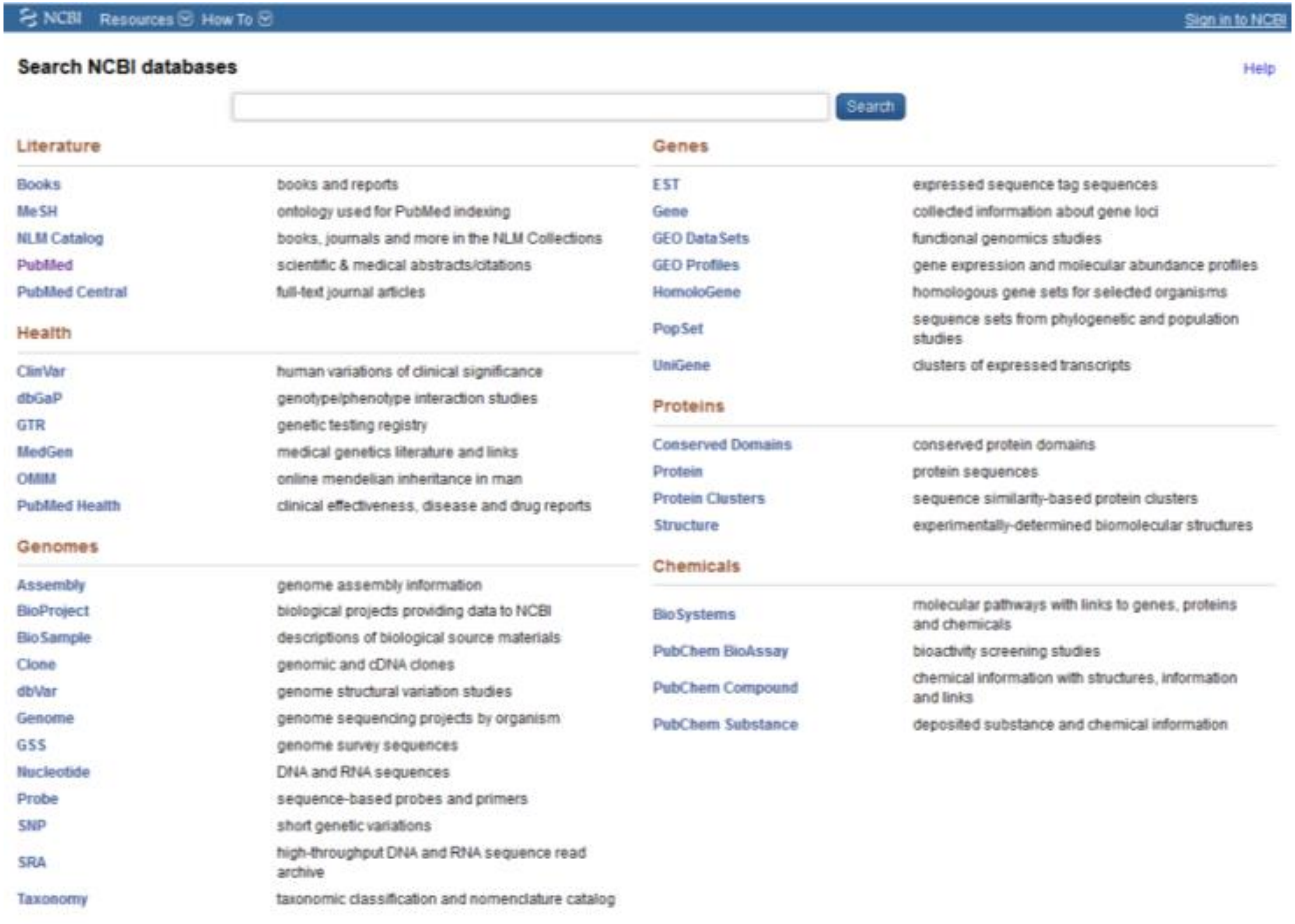
NCBI to host genomics hackathon March 20-22
02 Feb 2017
From March 20th to 22nd, the NCBI will host a genomics hackathon on the NIH

Multiple Sequence Alignment Viewer 1.3 is now available
31 Jan 2017
The new version of the Multiple Sequence Alignment Viewer (MSA Viewer) has

February 8th NCBI Minute: Finding Gene, Protein and Chemical Names, Aliases and Synonyms
31 Jan 2017
Next Wednesday, February 8th, NCBI

NCBI Global Cross-database search

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



Search NCBI databases

Literature

Books	books and reports
MeSH	ontology used for PubMed indexing
NLM Catalog	books, journals and more in the NLM Collections
PubMed	scientific & medical abstracts/citations
PubMed Central	full-text journal articles

Health

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

Genomes

Assembly	genome assembly information
BioProject	biological projects providing data to NCBI
BioSample	descriptions of biological source materials
Clone	genomic and cDNA clones
dbVar	genome structural variation studies
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

Genes

EST	expressed sequence tag sequences
Gene	collected information about gene loci
GEO DataSets	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

Proteins

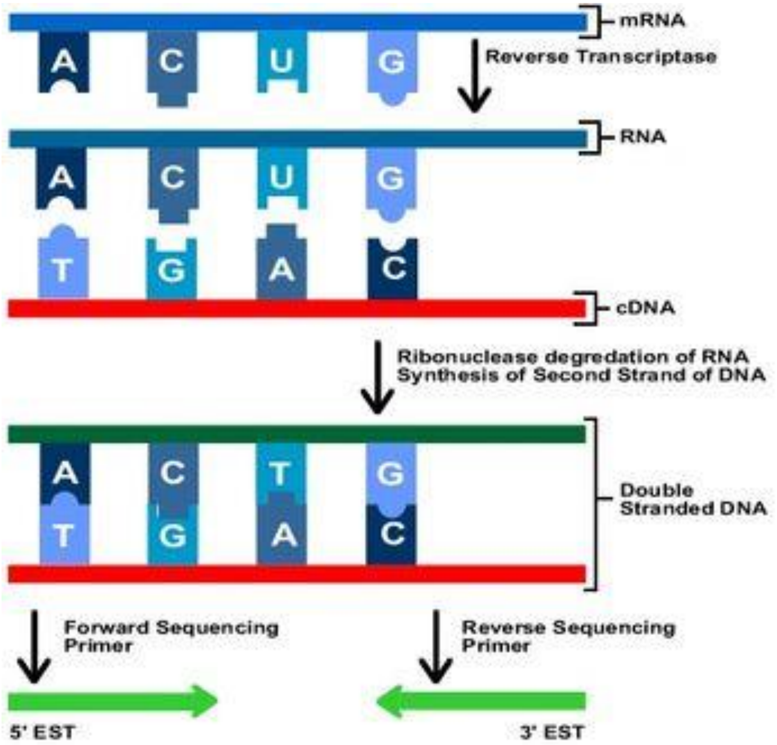
Conserved Domains	conserved protein domains
Protein	protein sequences
Protein Clusters	sequence similarity-based protein clusters
Structure	experimentally-determined biomolecular structures

Chemicals

BioSystems	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	bioactivity screening studies
PubChem Compound	chemical information with structures, information and links
PubChem Substance	deposited substance and chemical information

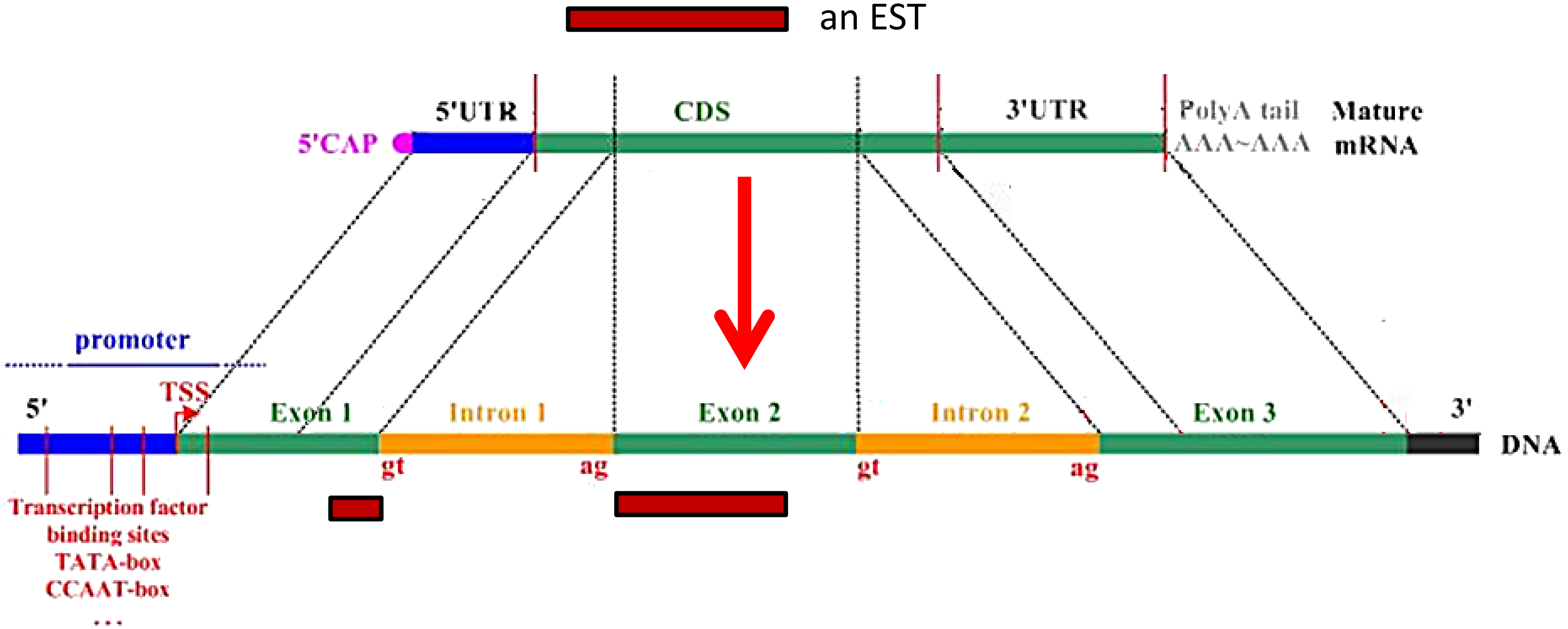
UniGene

- EST (=expressed sequence tag):
  - DNA sequence corresponding to mRNA from expressed gene
  - ~500 base pairs long
  - Sequenced from a cDNA library

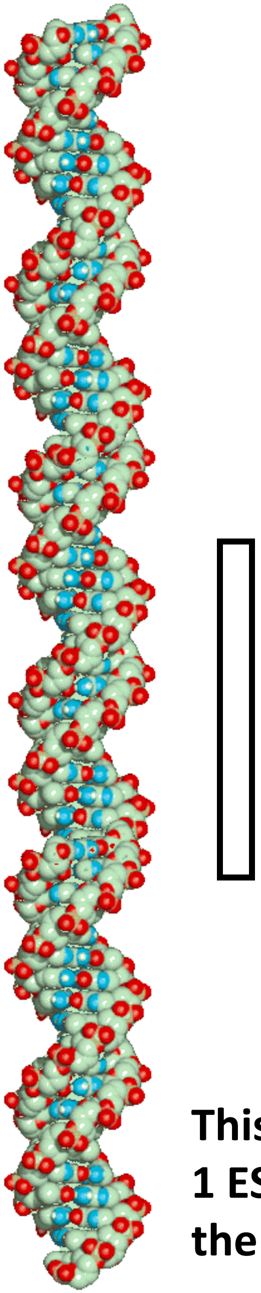


- Predict genes based on ESTs.
- Cluster ESTs from many cDNA libraries to predict distinct genes

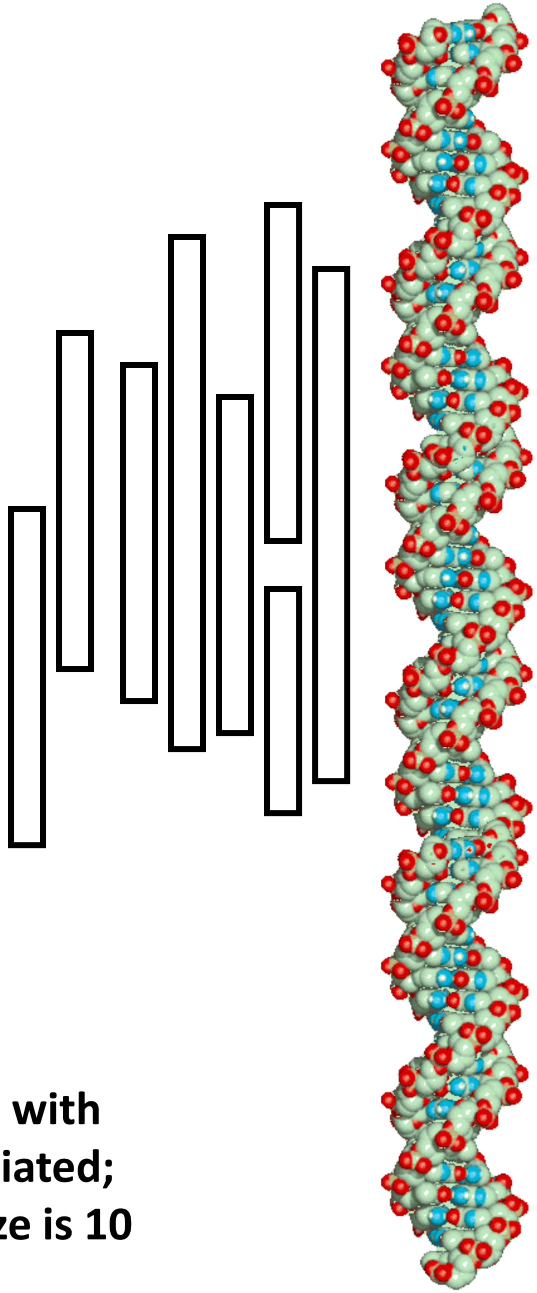
Map mRNA (EST) back to DNA



EST clusters

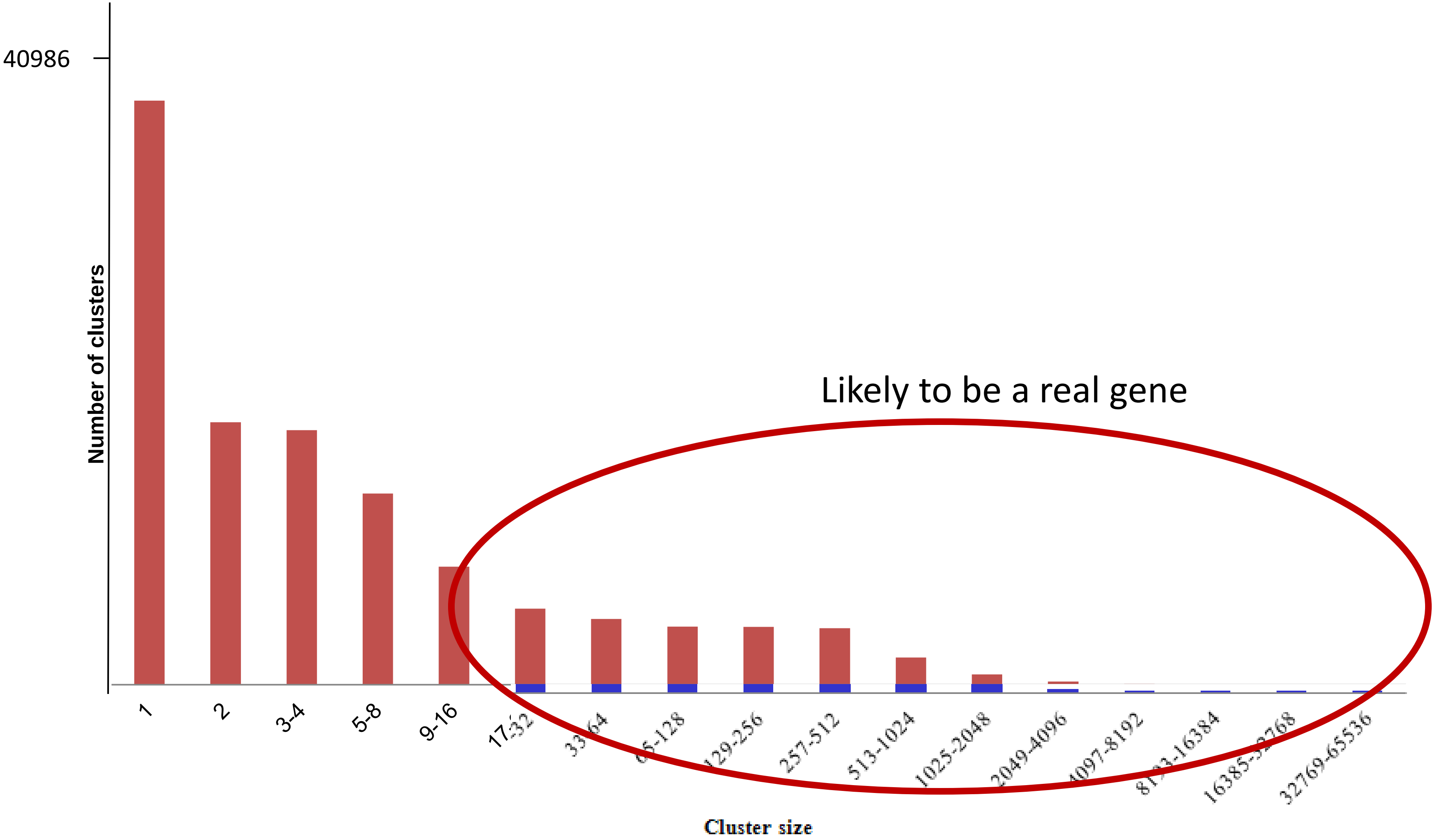


This is a gene with
1 EST associated;
the cluster size is 1



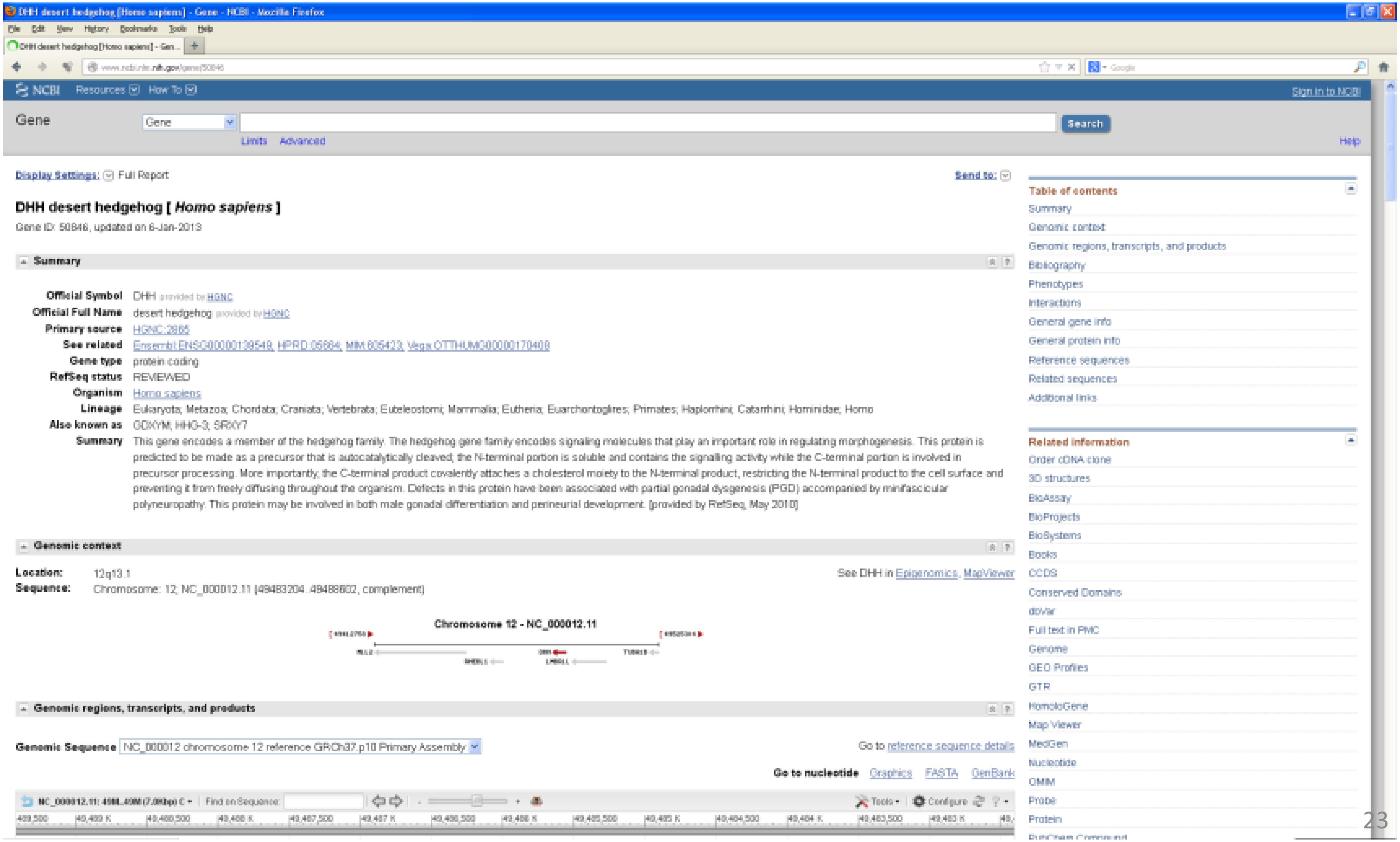
This is a gene with
10 ESTs associated;
the cluster size is 10

UniGene clusters



Likely to be a real gene

Gene (NCBI)
DHH as example



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

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

Display Settings: Full Report

Send to:

DHH desert hedgehog [*Homo sapiens*]

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

Summary

Official Symbol	DHH provided by HGNC
Official Full Name	desert hedgehog provided by HGNC
Primary source	HGNC:2865
See related	Ensembl:ENSG00000139549; HPRD:05684; MIM:605423; Vega:OTTHUMG00000170408
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	GDXYM; HHG-3; SRXY7
Summary	This gene encodes a member of the hedgehog family. The hedgehog gene family encodes signaling molecules that play an important role in regulating morphogenesis. This protein is predicted to be made as a precursor that is autocatalytically cleaved; the N-terminal portion is soluble and contains the signalling activity while the C-terminal portion is involved in precursor processing. More importantly, the C-terminal product covalently attaches a cholesterol moiety to the N-terminal product, restricting the N-terminal product to the cell surface and preventing it from freely diffusing throughout the organism. Defects in this protein have been associated with partial gonadal dysgenesis (PGD) accompanied by minifascicular polyneuropathy. This protein may be involved in both male gonadal differentiation and perineurial development. [provided by RefSeq, May 2010]

Genomic context

Location: 12q13.1

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

See DHH in Epigenomics, MapViewer

Chromosome 12 - NC_000012.11

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

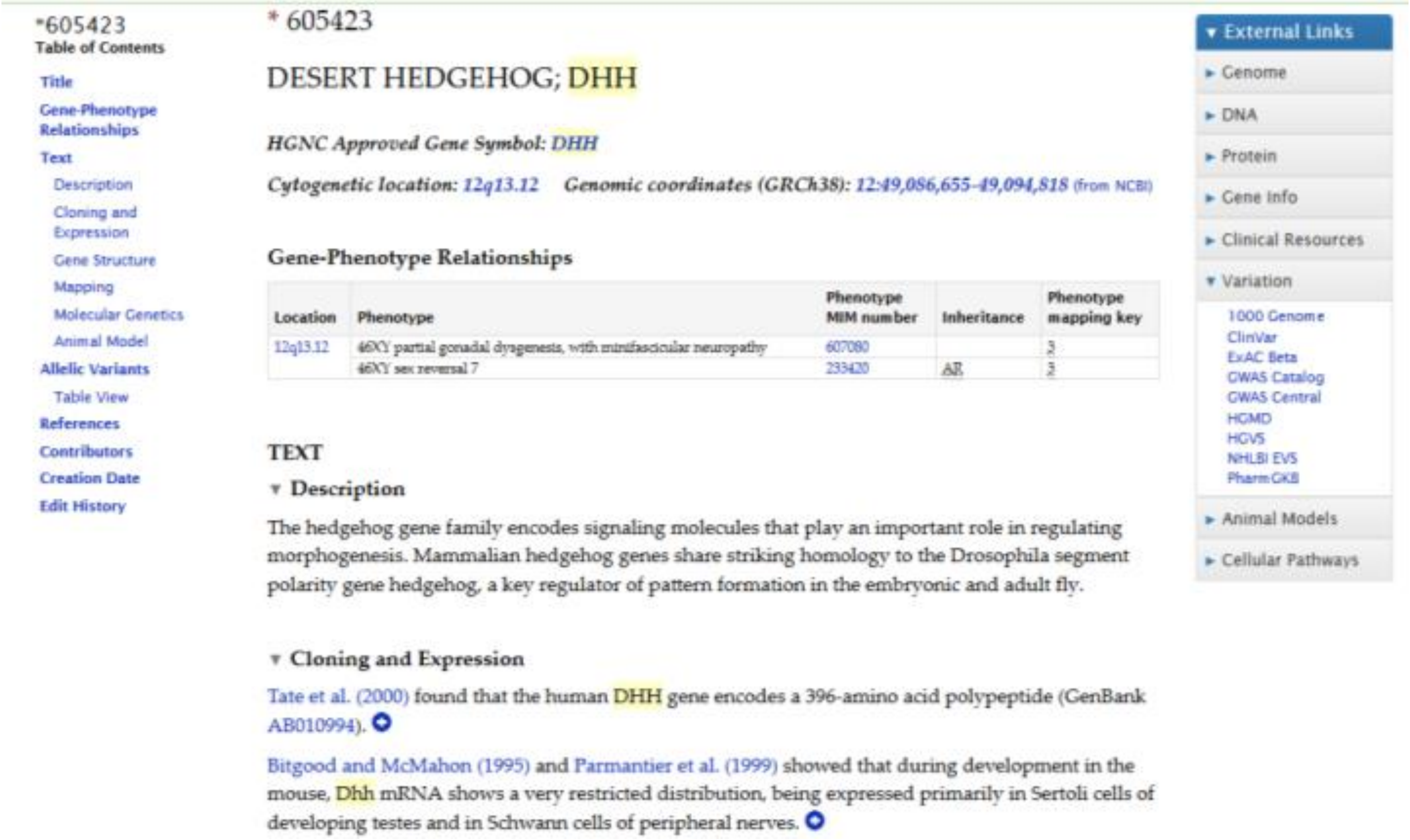
Table of contents

- 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)
Online Mendelian Inheritance in Man



*605423
Table of Contents

* 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

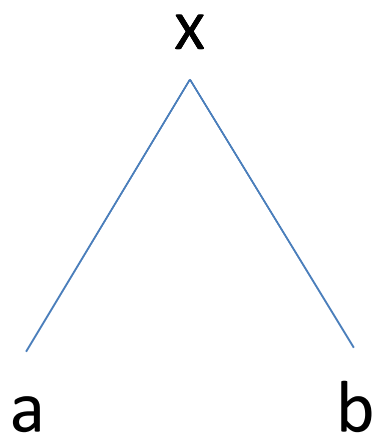
Tate et al. (2000) found that the human DHH gene encodes a 396-amino acid polypeptide (GenBank AB010994).

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.

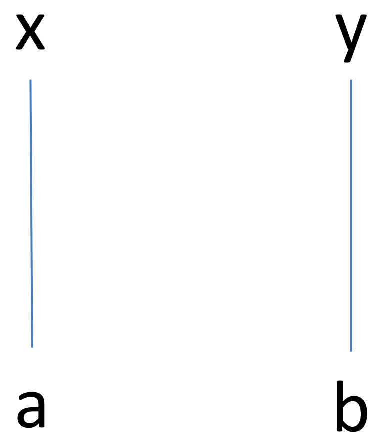
An Online Catalog of Human Genes and Genetic Disorders

Homology

- Homologous protein or DNA sequences share common ancestry.
- Homology need not imply similar function.
- A pair of sequences is either homologous or not homologous.

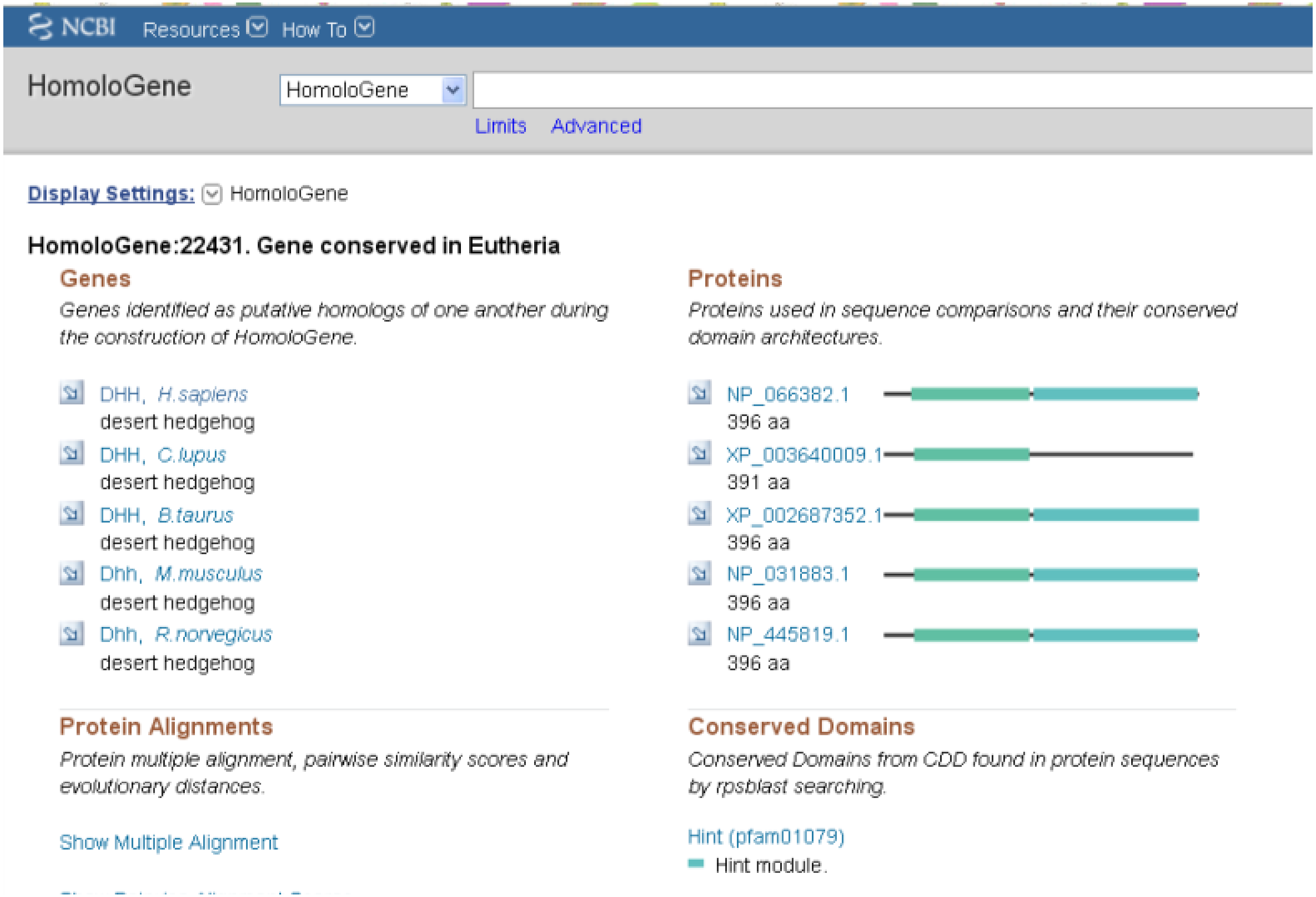


a, b homologous



a, b NOT homologous

HomoloGene (NCBI)



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

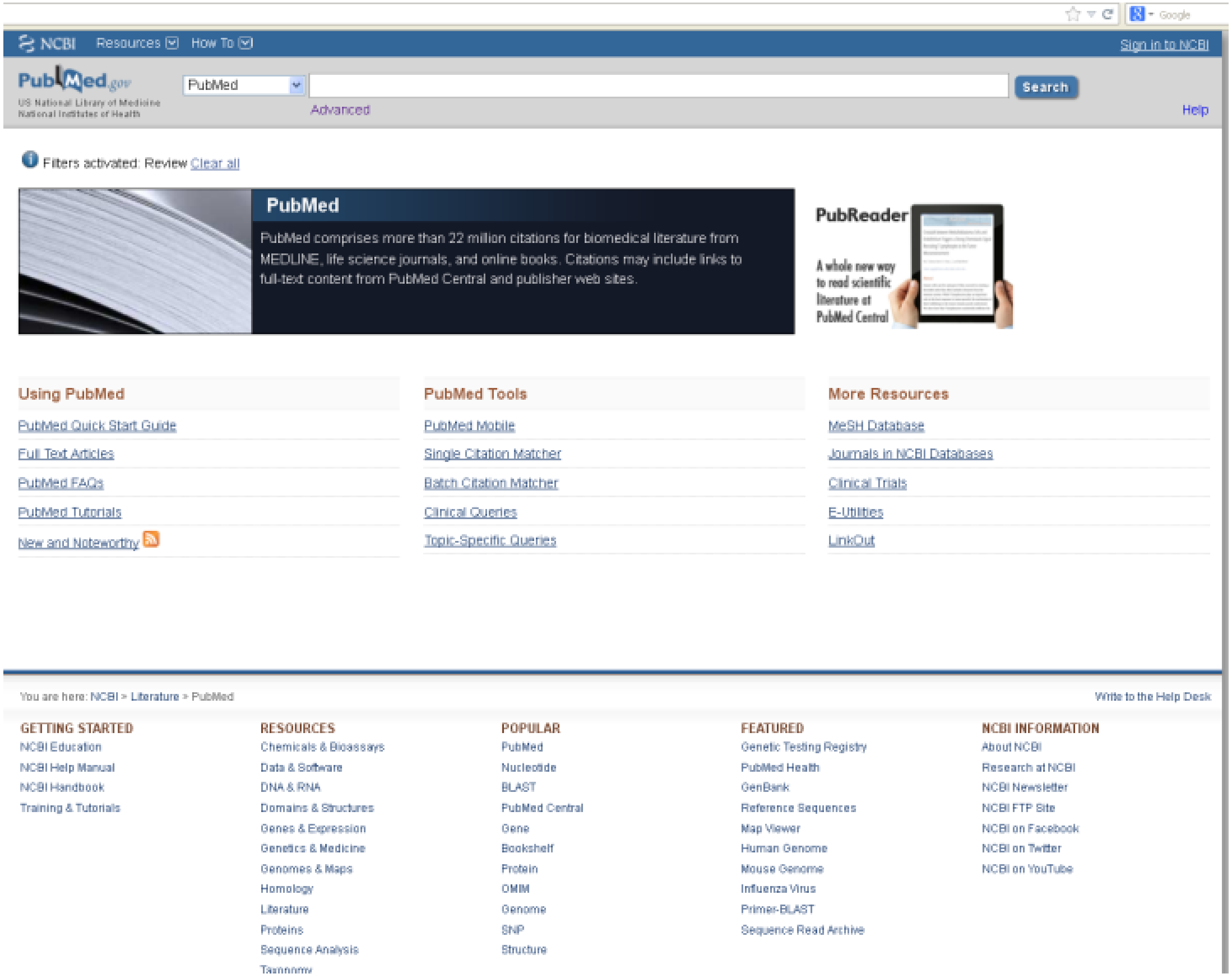
Conserved Domains

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

Hint (pfam01079)

▬ Hint module.

PubMed (NCBI)



NCBI Resources How To Sign in to NCBI

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed Search
Advanced Help

Filters activated: Review Clear all

PubMed

PubMed comprises more than 22 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

PubReader

A whole new way to read scientific literature at PubMed Central

Using PubMed

- PubMed Quick Start Guide
- Full Text Articles
- PubMed FAQs
- PubMed Tutorials
- New and Noteworthy

PubMed Tools

- PubMed Mobile
- Single Citation Matcher
- Batch Citation Matcher
- Clinical Queries
- Topic-Specific Queries

More Resources

- MeSH Database
- Journals in NCBI Databases
- Clinical Trials
- E-Utilities
- LinkOut

You are here: NCBI > Literature > PubMed

Write to the Help Desk

GETTING STARTED

NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials

RESOURCES

Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy

POPULAR

PubMed
Nucleotide
BLAST
PubMed Central
Gene
Bookshelf
Protein
OMIM
Genome
SNP
Structure

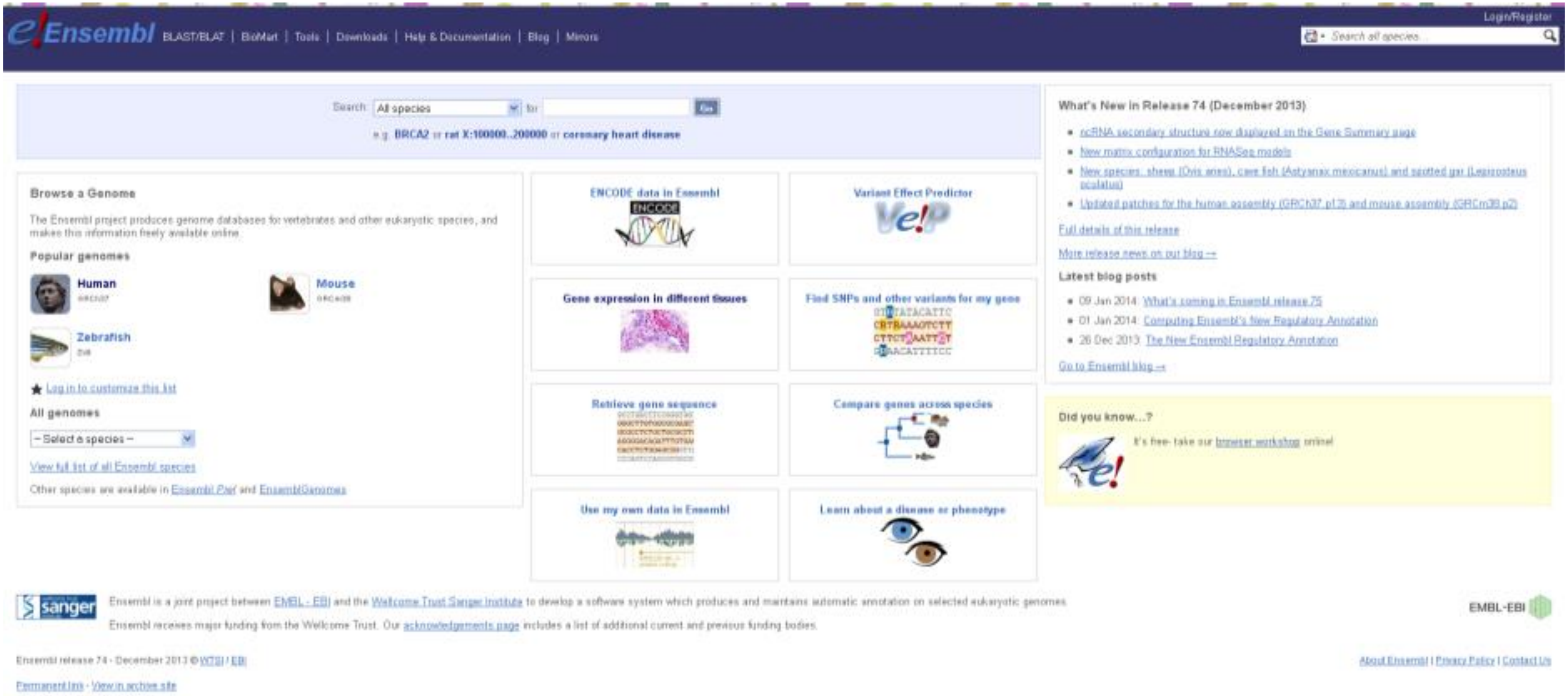
FEATURED

Genetic Testing Registry
PubMed Health
GenBank
Reference Sequences
Map Viewer
Human Genome
Mouse Genome
Influenza Virus
Primer-BLAST
Sequence Read Archive

NCBI INFORMATION

About NCBI
Research at NCBI
NCBI Newsletter
NCBI FTP Site
NCBI on Facebook
NCBI on Twitter
NCBI on YouTube

Ensembl homepage



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.

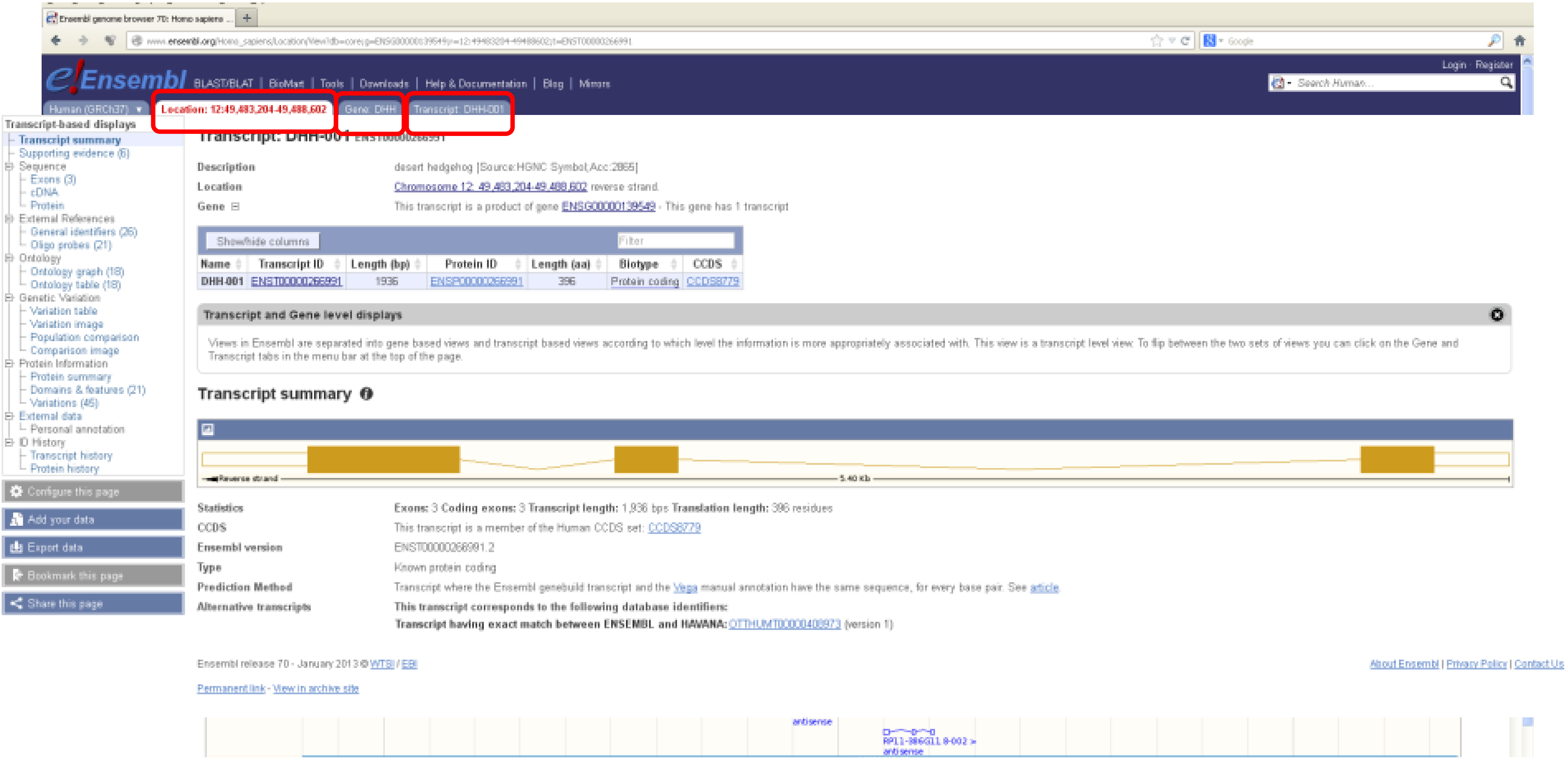
Ensembl receives major funding from the Wellcome Trust. Our acknowledgements page includes a list of additional current and previous funding bodies.

Ensembl release 74 - December 2013 © WTSI / EBI

About Ensembl | Privacy Policy | Contact Us

Permanent link - View in archive site

Ensembl
example DHH (human)



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

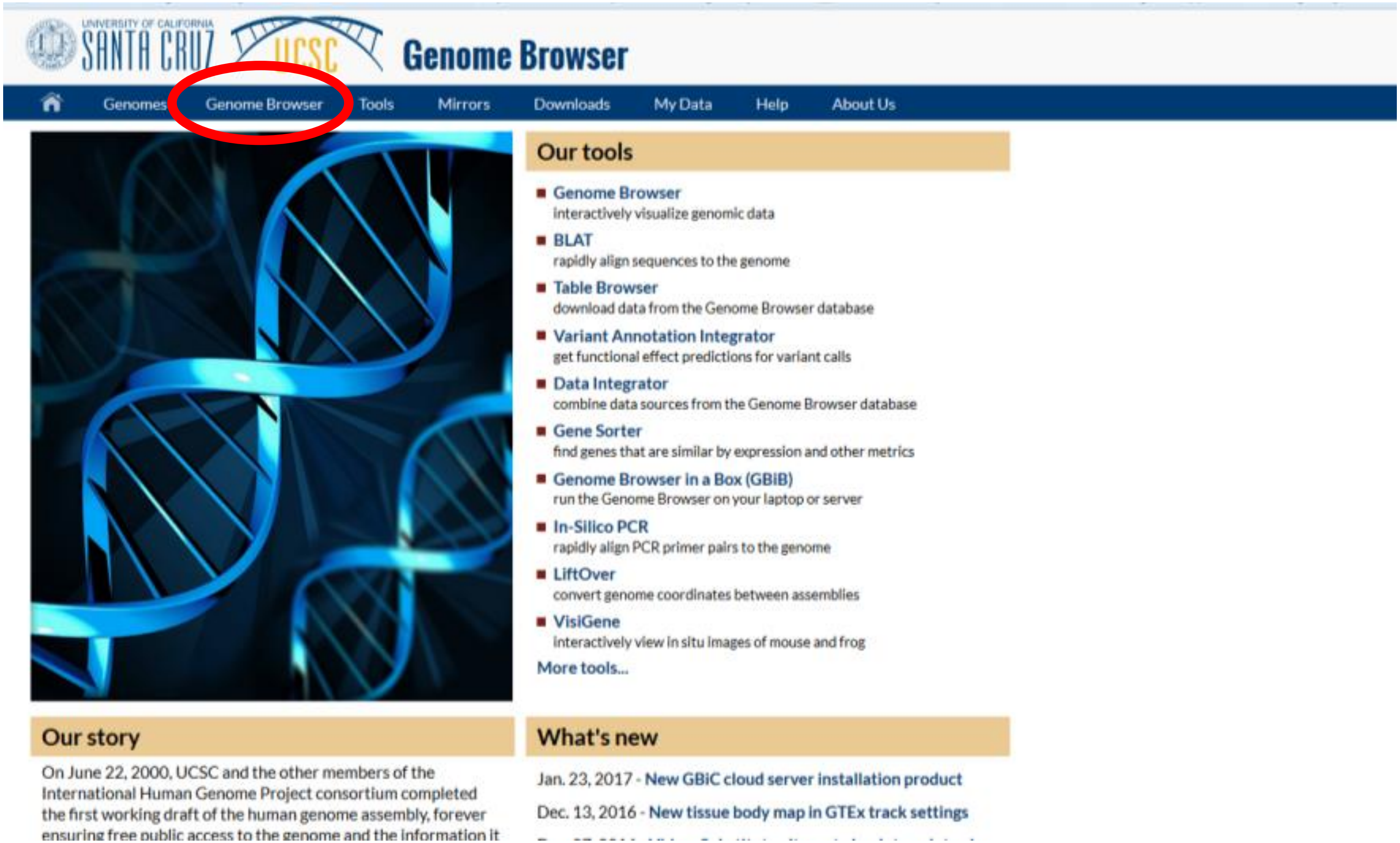
Statistics	**Exons:** 3 **Coding exons:** 3 **Transcript length:** 1,936 bps **Translation length:** 396 residues
CCDS	This transcript is a member of the Human CCDS set: CCDS8779
Ensembl version	ENST00000266991.2
Type	Known protein coding
Prediction Method	Transcript where the Ensembl genebuild transcript 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

UCSC homepage



Our tools

- Genome Browser
  interactively visualize genomic data
- BLAT
  rapidly align sequences to the genome
- Table Browser
  download data from the Genome Browser database
- Variant Annotation Integrator
  get functional effect predictions for variant calls
- Data Integrator
  combine data sources from the Genome Browser database
- Gene Sorter
  find genes that are similar by expression and other metrics
- Genome Browser in a Box (GBiB)
  run the Genome Browser on your laptop or server
- In-Silico PCR
  rapidly align PCR primer pairs to the genome
- LiftOver
  convert genome coordinates between assemblies
- VisiGene
  interactively view in situ images of mouse and frog

More tools...

Our story

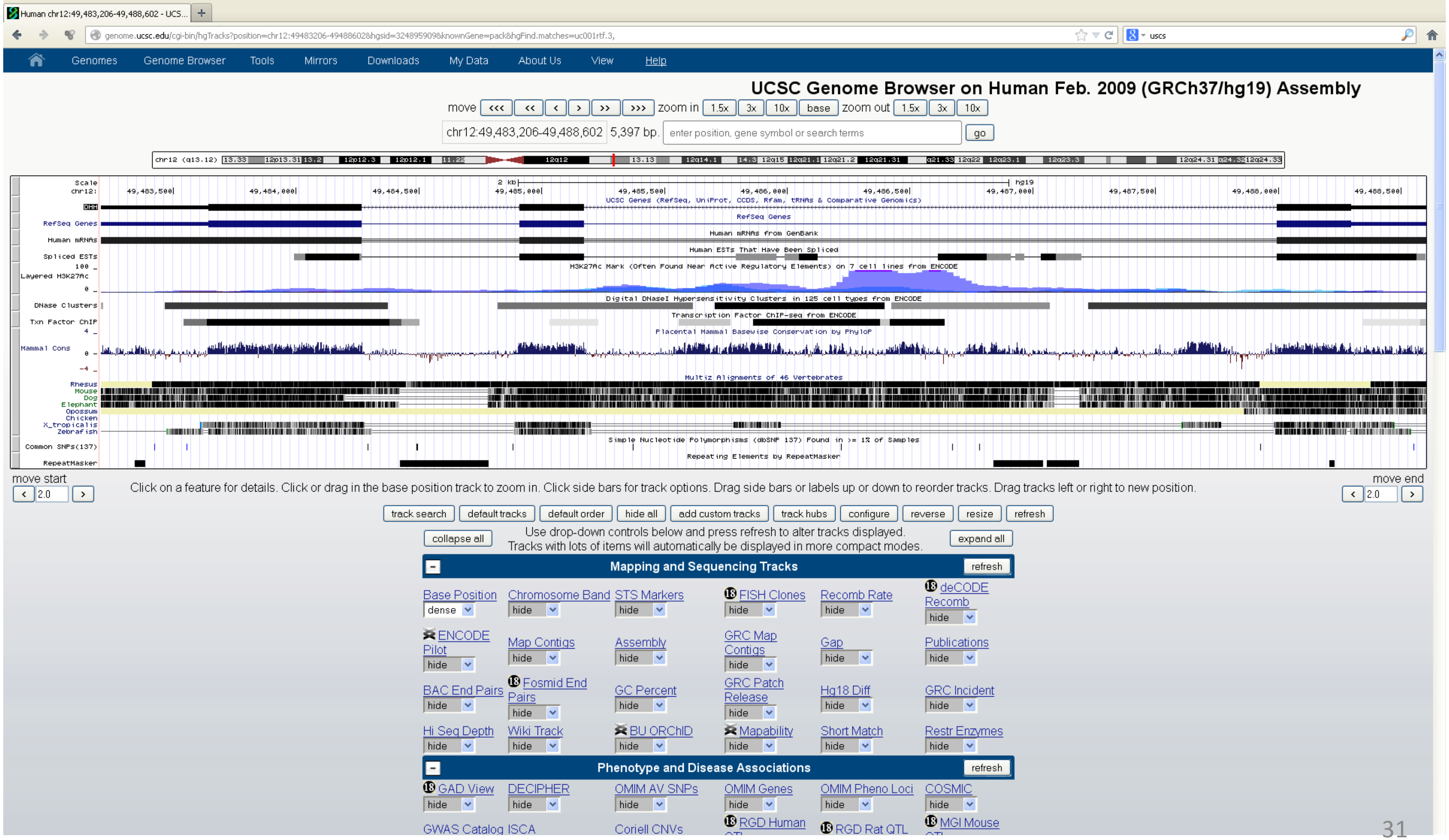
On June 22, 2000, UCSC and the other members of the International Human Genome Project consortium completed the first working draft of the human genome assembly, forever ensuring free public access to the genome and the information it

What's new

Jan. 23, 2017 - New GBiC cloud server installation product

Dec. 13, 2016 - New tissue body map in GTEx track settings

UCSC: Entry page (DHH)


Human chr12:49,483,206-49,488,602 - UCS...
genome.ucsc.edu/cgi-bin/hgTracks?position=chr12:49483206-49488602&hgsid=324895909&knownGene=pack&hgFind.matches=uc001rtf.3,
uscs
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
About Us
View
Help
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
enter position, gene symbol or search terms
go
chr12 (q13.12) 13.33 12p13.31 13.2 12p12.3 12p12.1 11.22 12q12 13.13 12q14.1 14.3 12q15 12q21.1 12q21.2 12q21.31 q21.33 12q22 12q23.1 12q23.3 12q24.31 q24.32 12q24.33
Scale
chr12:
2 kb
hg19
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
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
DHH
RefSeq Genes
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
100
Layered H3K27Ac
0
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
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
4
Mammal Cons
0
-4
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Common SNPs(137)
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker
move start
< 2.0 >
move end
< 2.0 >
track search
default tracks
default order
hide all
add custom tracks
track hubs
configure
reverse
resize
refresh
collapse all
expand all
refresh
refresh

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

chr12:49,483,206-49,488,602 5,397 bp.

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.

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

Mapping and Sequencing Tracks

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

Phenotype and Disease Associations

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

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 (Entrez) gene ID:
  - Number
- OMIM ID:
  - Number
- Pubmed ID:
  - Number

Ensembl identifiers

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

- For other species than human a suffix is added:

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

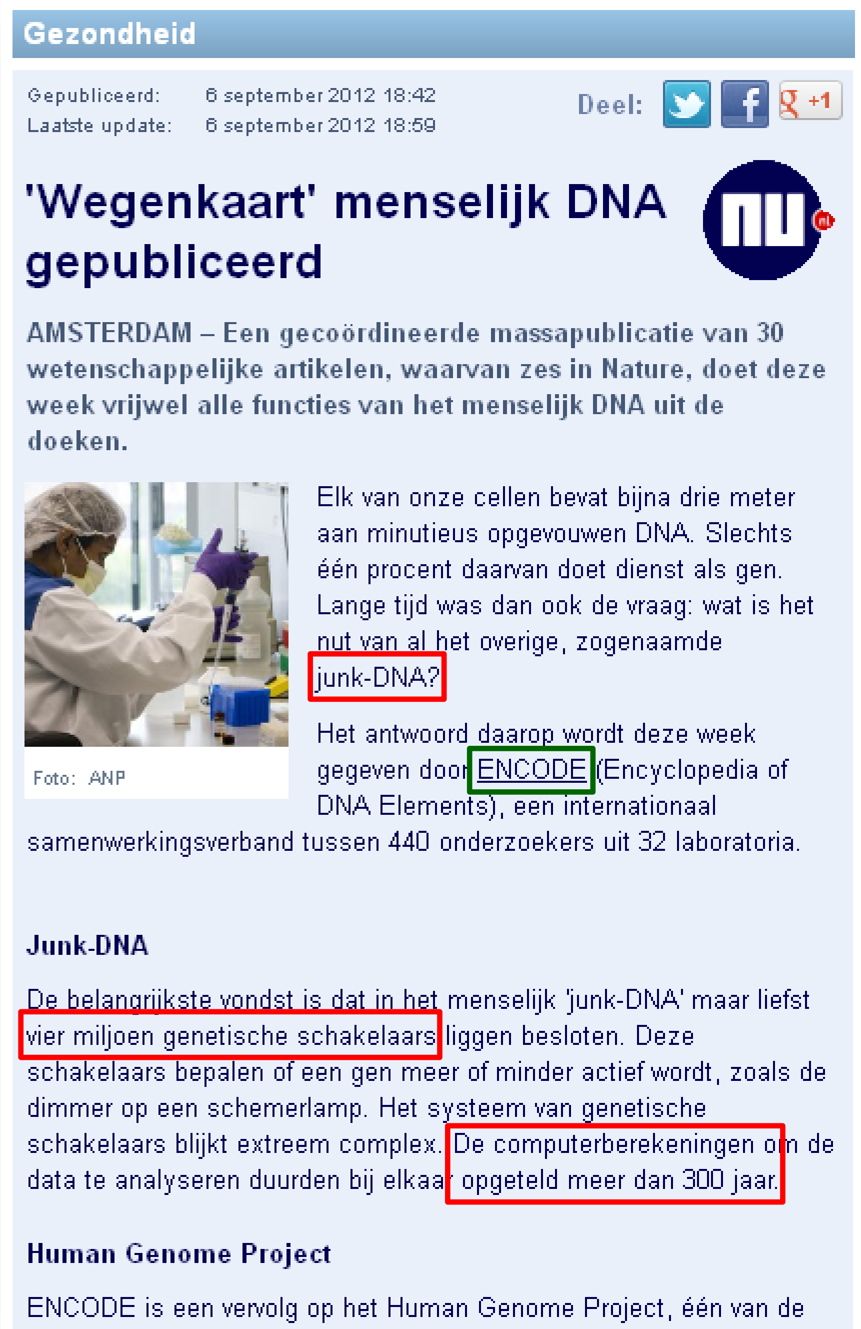
Where does all this information come from?

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

Functional genomics

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

nu.nl – Sept. 6th 2012



Gezondheid

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

'Wegenkaart' menselijk DNA gepubliceerd

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

Foto: ANP

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

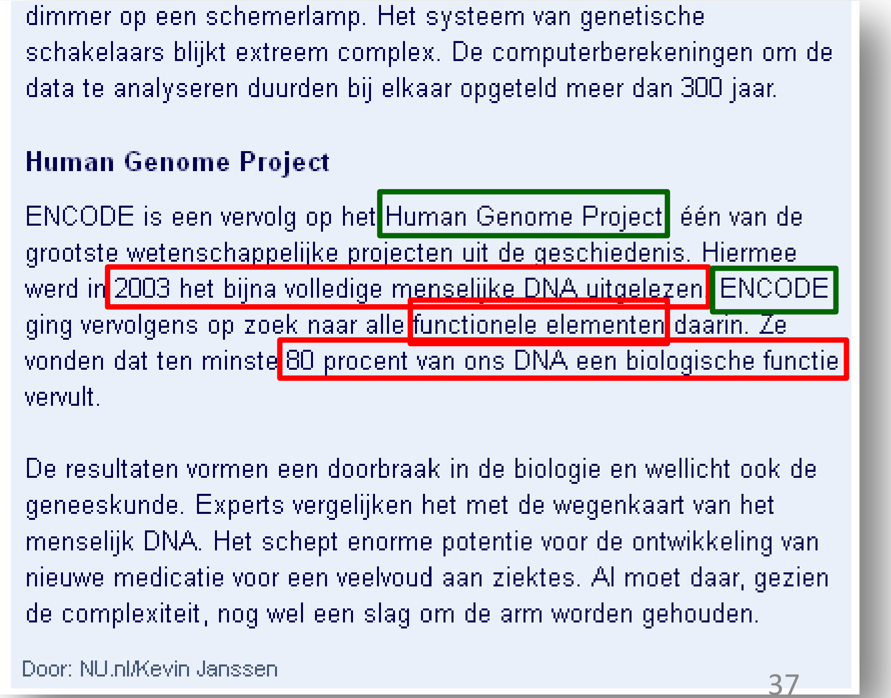
Het antwoord daarop wordt deze week gegeven door ENCODE (Encyclopedia of DNA Elements), een internationaal samenwerkingsverband tussen 440 onderzoekers uit 32 laboratoria.

Junk-DNA

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

Human Genome Project

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



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

Human Genome Project

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

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

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

Door: NU.nl/Kevin Janssen

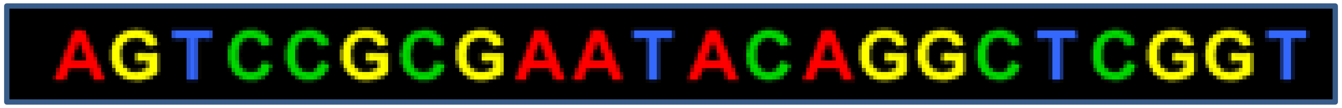
HGP and ENCODE

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

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

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

the Human Genome Project



AGTCCGCGAATACAGGCTCGGT

movie

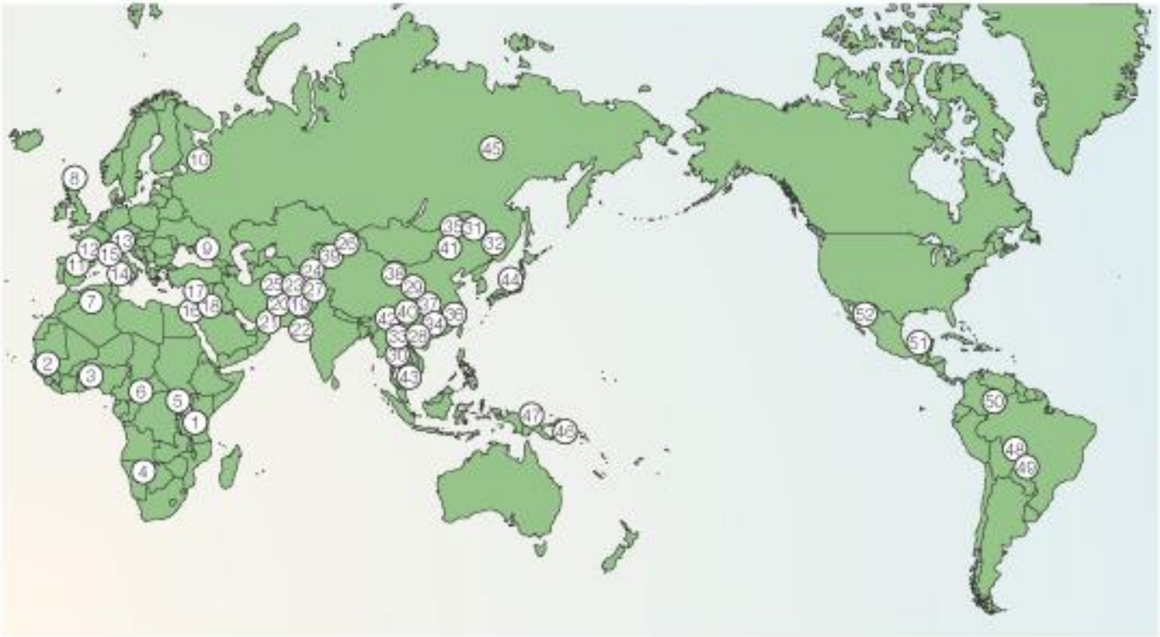
International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. Nature 431, 931-945 (21 October 2004).

The human genome project

HGP aim: sequence the entire human genome and provide the data free to the world.

First major global collaboration of its kind and the largest biological research project ever undertaken, involving thousands of staff in institutes across the globe.

By assigning different portions of the genome to different research groups in a coordinated and efficient way, the HGP researchers were able to overcome this challenge.



Africans
- 1 Bantu
- 2 Mandenka
- 3 Yoruba
- 4 San
- 5 Mbuti pygmy
- 6 Biaka
- 7 Mozabite

Europeans
- 8 Orcadian
- 9 Adygei
- 10 Russian
- 11 Basque
- 12 French
- 13 North Italian
- 14 Sardinian
- 15 Tuscan

Western Asians
- 16 Bedouin
- 17 Druze
- 18 Palestinian

Central and Southern Asians
- 19 Balochi
- 20 Brahui
- 21 Makrani
- 22 Sindhi
- 23 Pathan
- 24 Burusho
- 25 Hazara
- 26 Uygur
- 27 Kalash

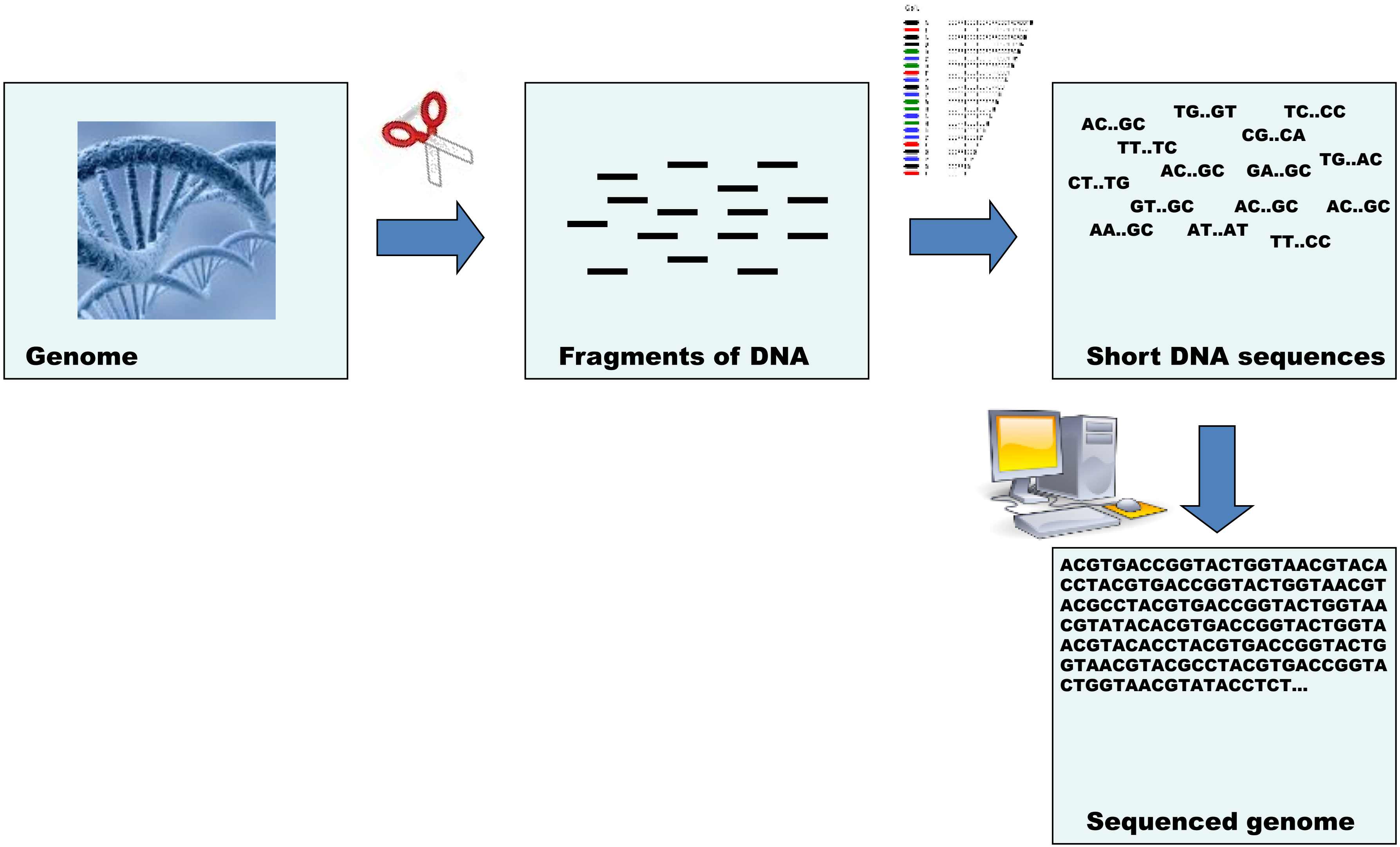
Eastern Asians
- 28 Han (S. China)
- 29 Han (N. China)
- 30 Dai
- 31 Daur
- 32 Hezhen
- 33 Lahu
- 34 Miao
- 35 Oroqen
- 36 She
- 37 Tujia
- 38 Tu
- 39 Xibo
- 40 Yi
- 41 Mongola
- 42 Naxi
- 43 Cambodian
- 44 Japanese
- 45 Yakut

Oceanians
- 46 Melanesian
- 47 Papuan

Native Americans
- 48 Karitiana
- 49 Surui
- 50 Colombian
- 51 Maya
- 52 Pima

Nature Reviews | Genetics

Genome sequencing: general principle



Genome

Fragments of DNA

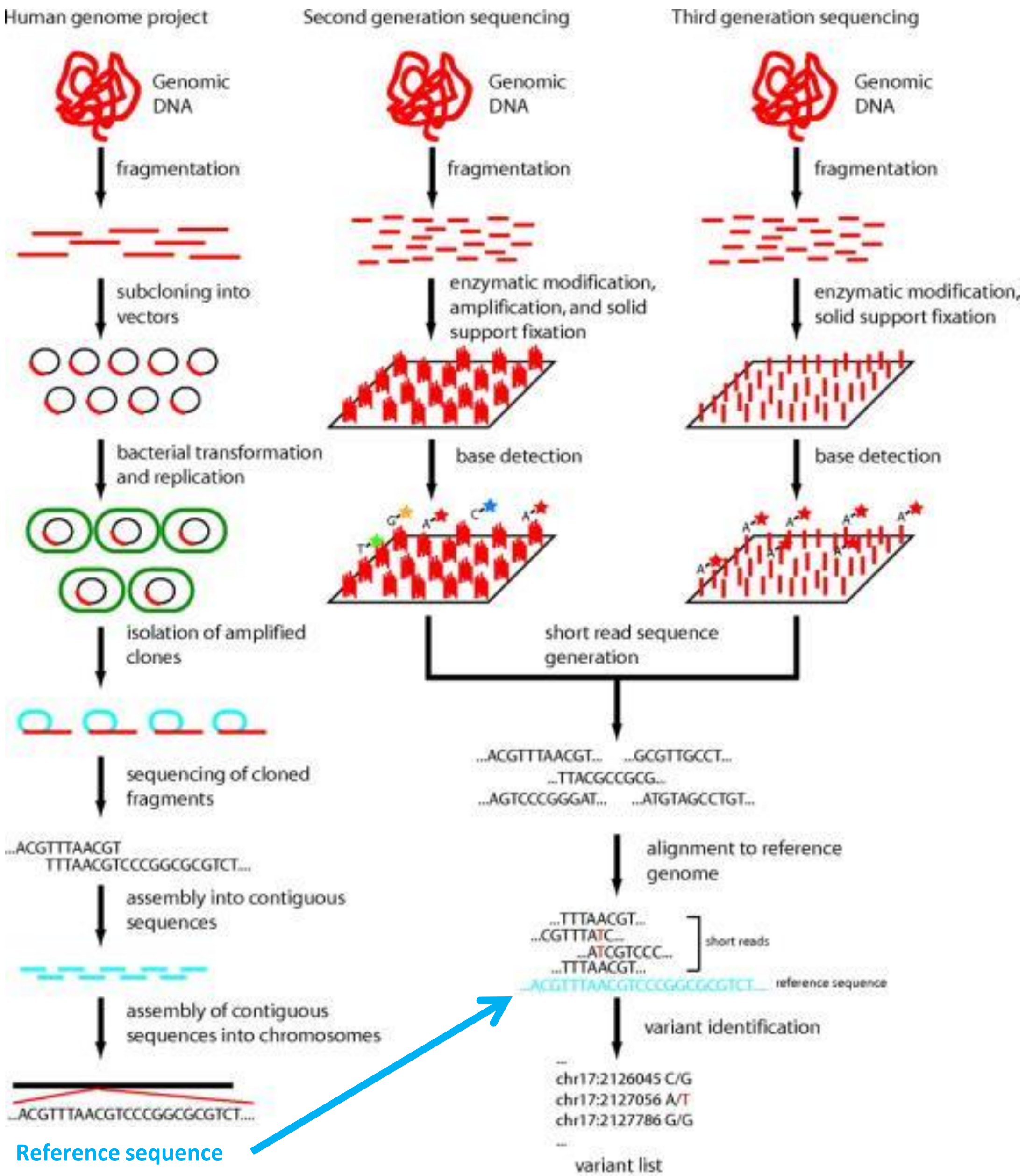
Short DNA sequences

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

ACGTGACCGGTACTGGTAACGTACA
CCTACGTGACCGGTACTGGTAACGT
ACGCCTACGTGACCGGTACTGGTAA
CGTATACACGTGACCGGTACTGGTA
ACGTACACCTACGTGACCGGTACTG
GTAACGTACGCCTACGTGACCGGTA
CTGGTAACGTATACCTCT...

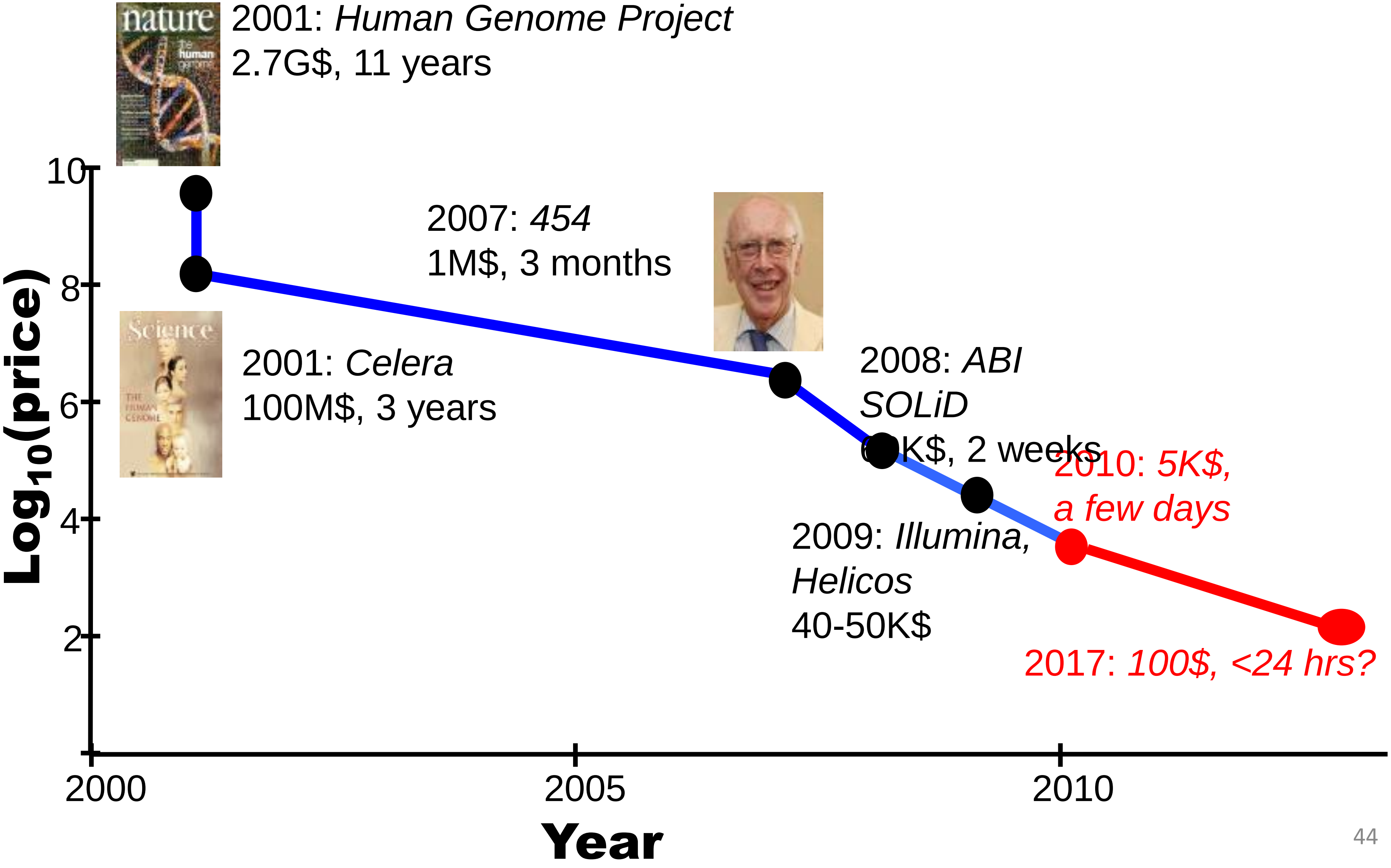
Sequenced genome

How was the human genome sequenced?


Human genome project
Second generation sequencing
Third generation sequencing
Genomic DNA
fragmentation
subcloning into vectors
bacterial transformation and replication
isolation of amplified clones
sequencing of cloned fragments
...ACGTTTAACGT
TTTAACGTCCCGGCGCGTCT....
assembly into contiguous sequences
assembly of contiguous sequences into chromosomes
...ACGTTTAACGTCCCGGCGCGTCT....
Genomic DNA
fragmentation
enzymatic modification, amplification, and solid support fixation
base detection
G
A
C
A
T
Genomic DNA
fragmentation
enzymatic modification, solid support fixation
base detection
A
A
A
A
A
A
short read sequence generation
...ACGTTTAACGT... ...GCGTTGCCT...
...TTACGCCGCG...
...AGTCCCGGGAT... ...ATGTAGCCTGT...
alignment to reference genome
...TTTAACGT...
...CGTTTATC...
...ATCGTCCC...
...TTTAACGT...
short reads
...ACGTTTAACGTCCCGGCGCGTCT....
reference sequence
variant identification
...
chr17:2126045 C/G
chr17:2127056 A/T
chr17:2127786 G/G
...
variant list

Reference sequence

Costs of sequencing the human genome



2001: *Human Genome Project*
2.7G$, 11 years

2001: *Celera*
100M$, 3 years

2007: *454*
1M$, 3 months

2008: *ABI SOLiD*
60K$, 2 weeks

2009: *Illumina, Helicos*
40-50K$

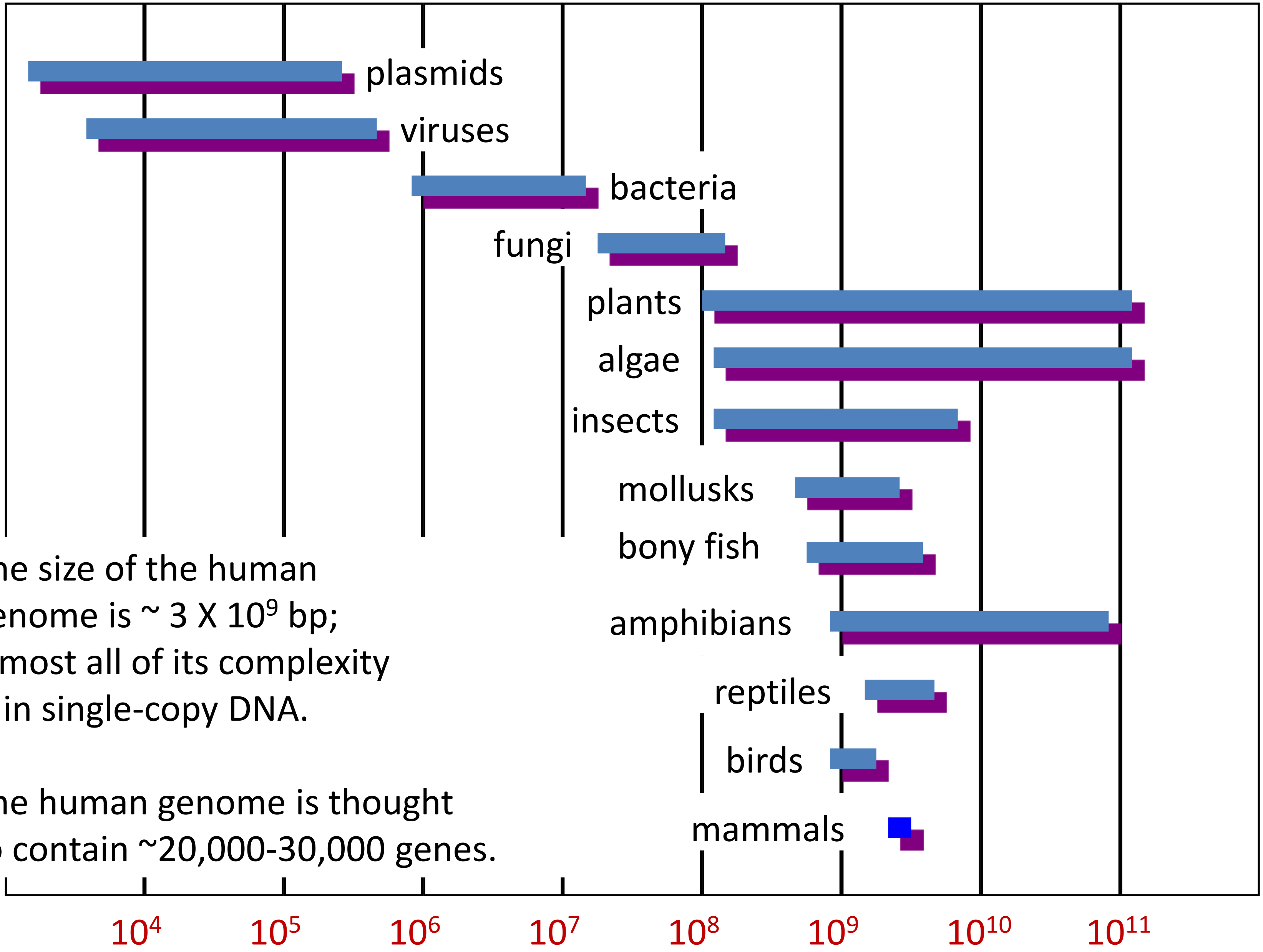
2010: *5K$, a few days*

2017: *100$, <24 hrs?*

When has a genome been fully sequenced?

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

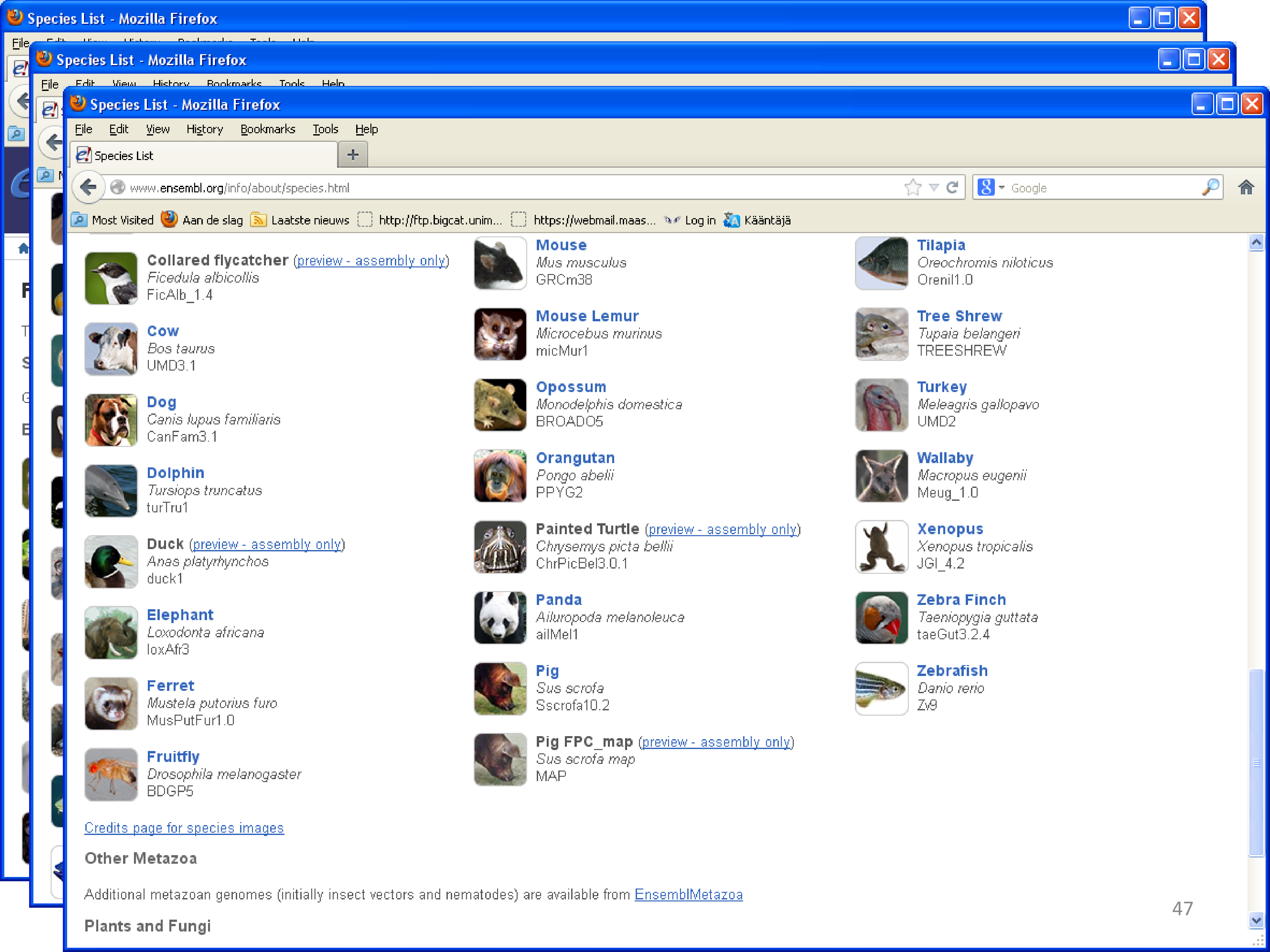
Genome sizes in nucleotide base pairs (log scale)



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

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

http://www3.kumc.edu/jcalvet/PowerPoint/bioc801b.ppt


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

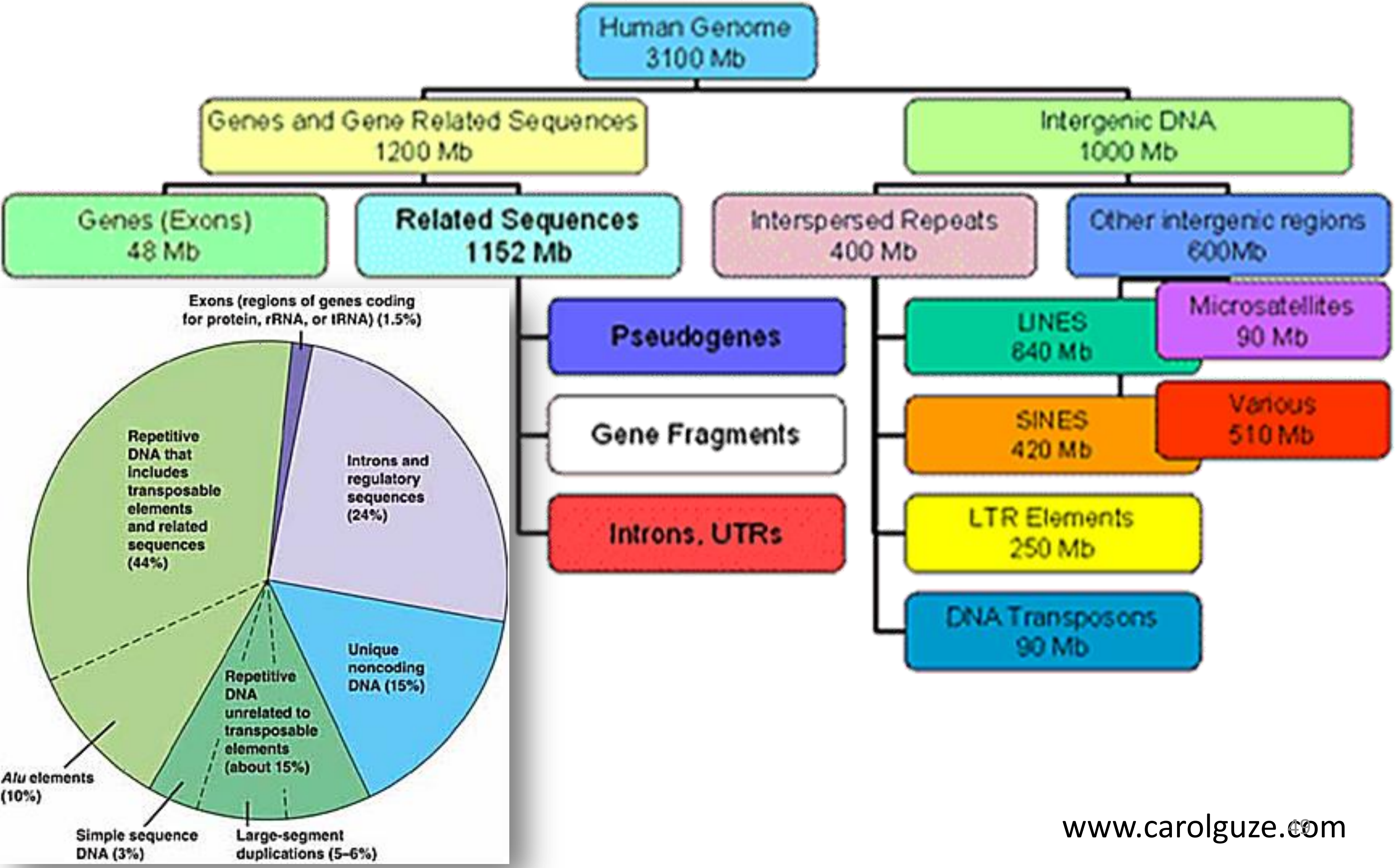
Number of genes

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



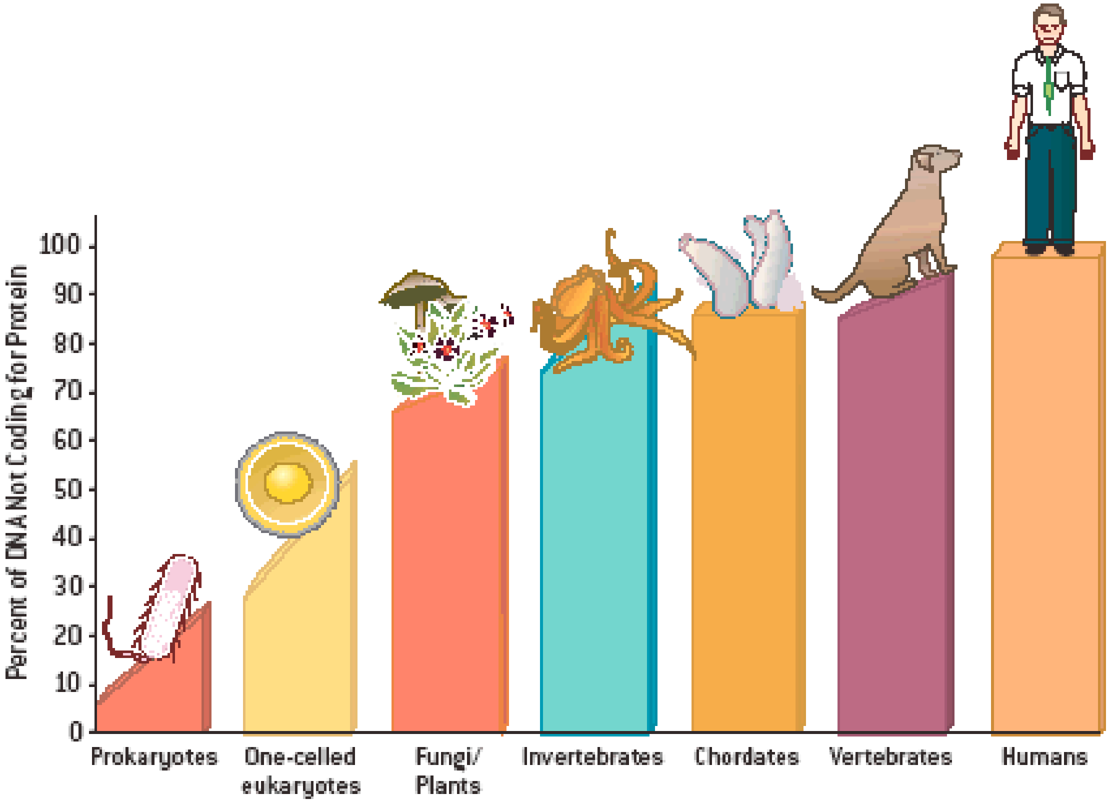
Human Genome 3100 Mb

- Genes and Gene Related Sequences 1200 Mb
  - Genes (Exons) 48 Mb
  - Related Sequences 1152 Mb
    - Pseudogenes
    - Gene Fragments
    - Introns, UTRs
- Intergenic DNA 1000 Mb
  - Interspersed Repeats 400 Mb
    - LINES 640 Mb
    - SINES 420 Mb
    - LTR Elements 250 Mb
    - DNA Transposons 90 Mb
  - Other intergenic regions 600Mb
    - Microsatellites 90 Mb
    - Various 510 Mb

Exons (regions of genes coding for protein, rRNA, or tRNA) (1.5%)
Introns and regulatory sequences (24%)
Unique noncoding DNA (15%)
Repetitive DNA unrelated to transposable elements (about 15%)
Repetitive DNA that includes transposable elements and related sequences (44%)
Alu elements (10%)
Simple sequence DNA (3%)
Large-segment duplications (5–6%)

www.carolguze.com

Non-Protein coding DNA

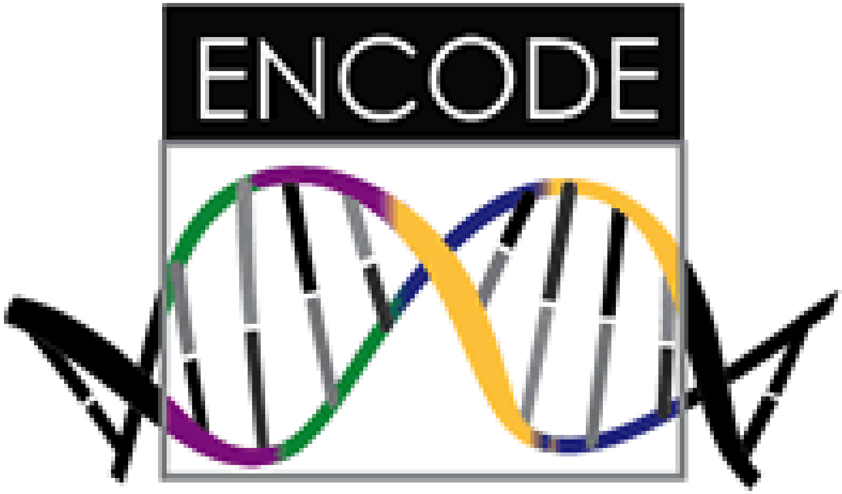


NONPROTEIN-CODING SEQUENCES make up only a small fraction of the DNA of prokaryotes. Among eukaryotes, as their complexity increases, generally so, too, does the proportion of their DNA that does not code for protein. The noncoding sequences have been considered junk, but perhaps it actually helps to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

www.carolguze.com

The ENCODE Project: ENCyclopedia Of DNA Elements

A public research consortium

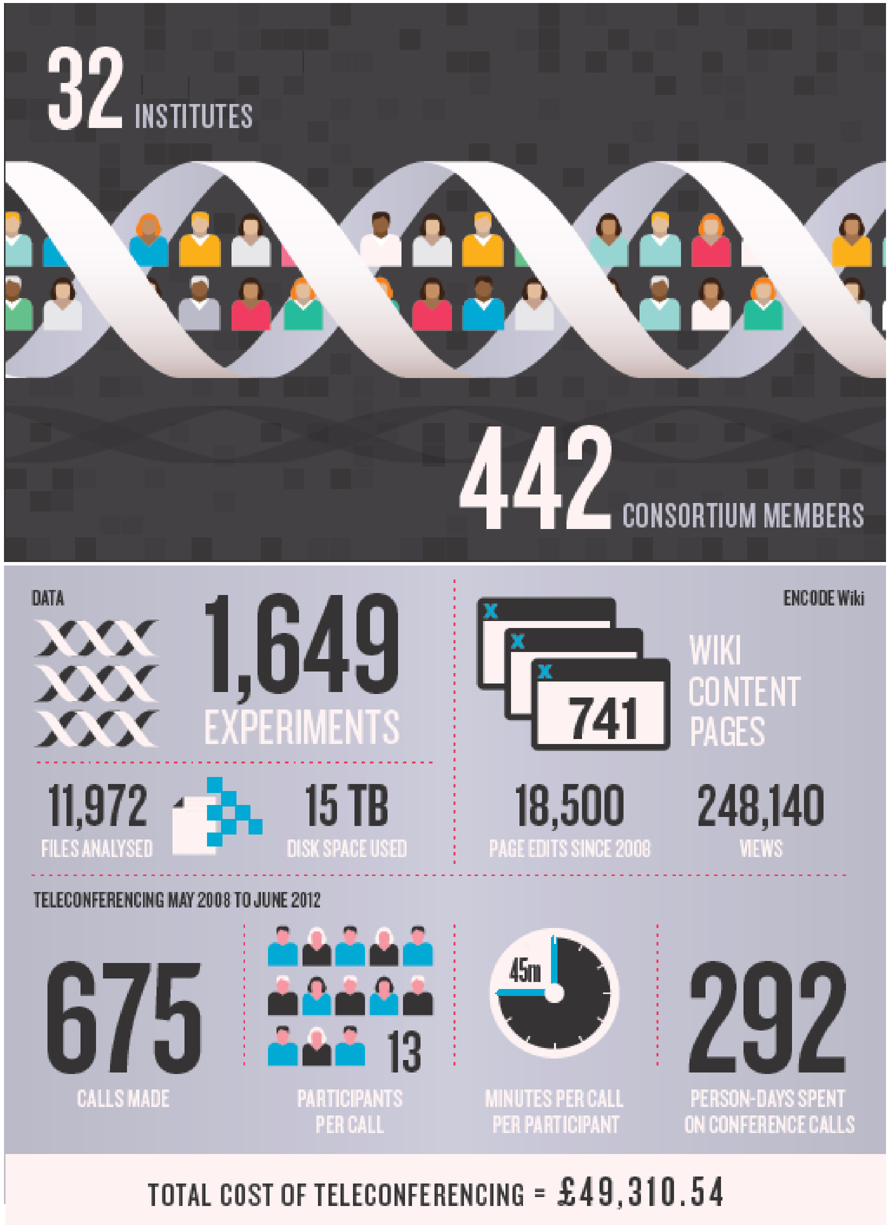


Launched: September 2003, upgraded to the entire genome September 2007.

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.

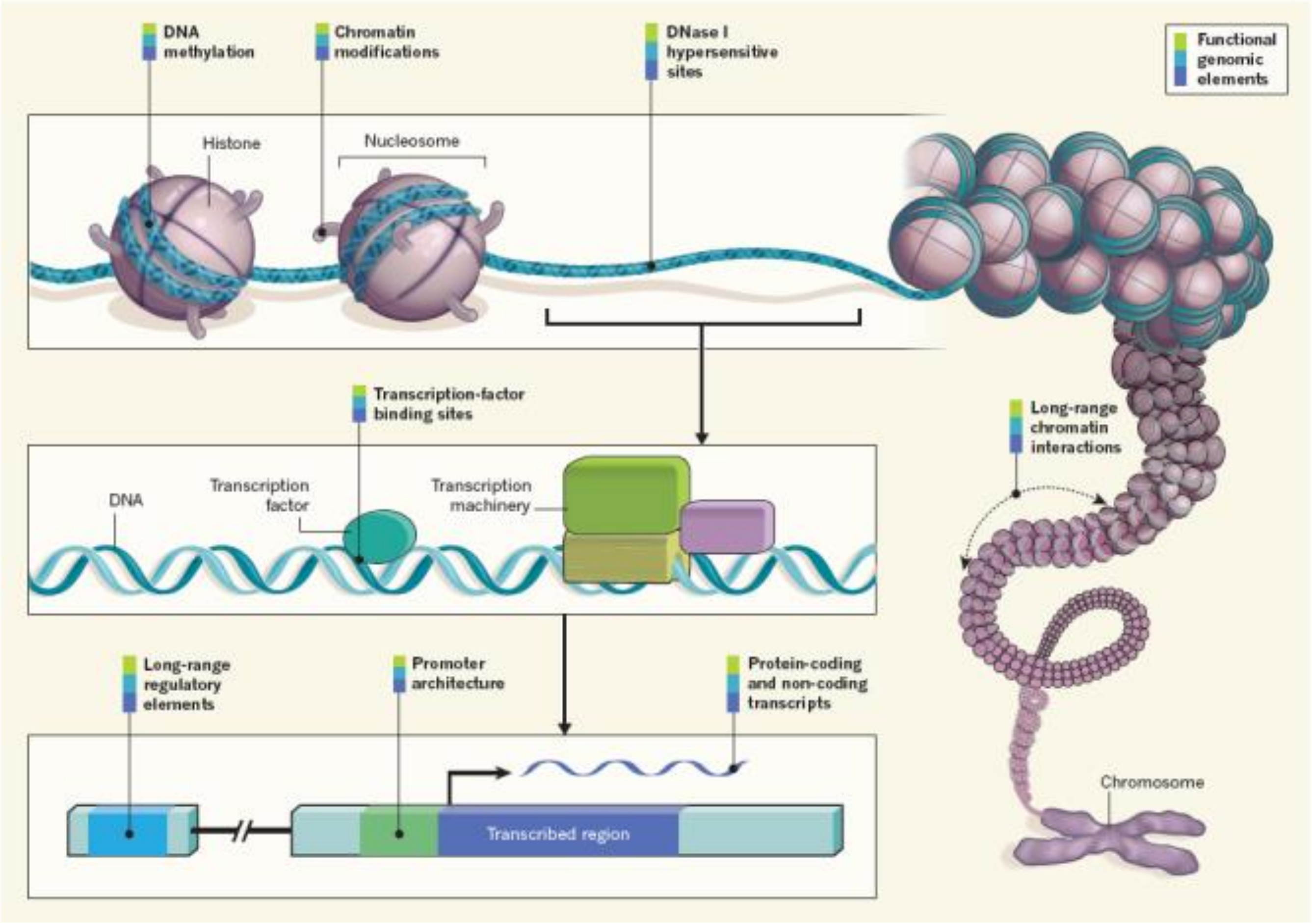

32 INSTITUTES
442 CONSORTIUM MEMBERS
DATA
1,649 EXPERIMENTS
ENCODE Wiki
741
WIKI CONTENT PAGES
11,972 FILES ANALYSED
15 TB DISK SPACE USED
18,500 PAGE EDITS SINCE 2008
248,140 VIEWS
TELECONFERENCING MAY 2008 TO JUNE 2012
675 CALLS MADE
13 PARTICIPANTS PER CALL
45m MINUTES PER CALL PER PARTICIPANT
292 PERSON-DAYS SPENT ON CONFERENCE CALLS
TOTAL COST OF TELECONFERENCING = £49,310.54

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

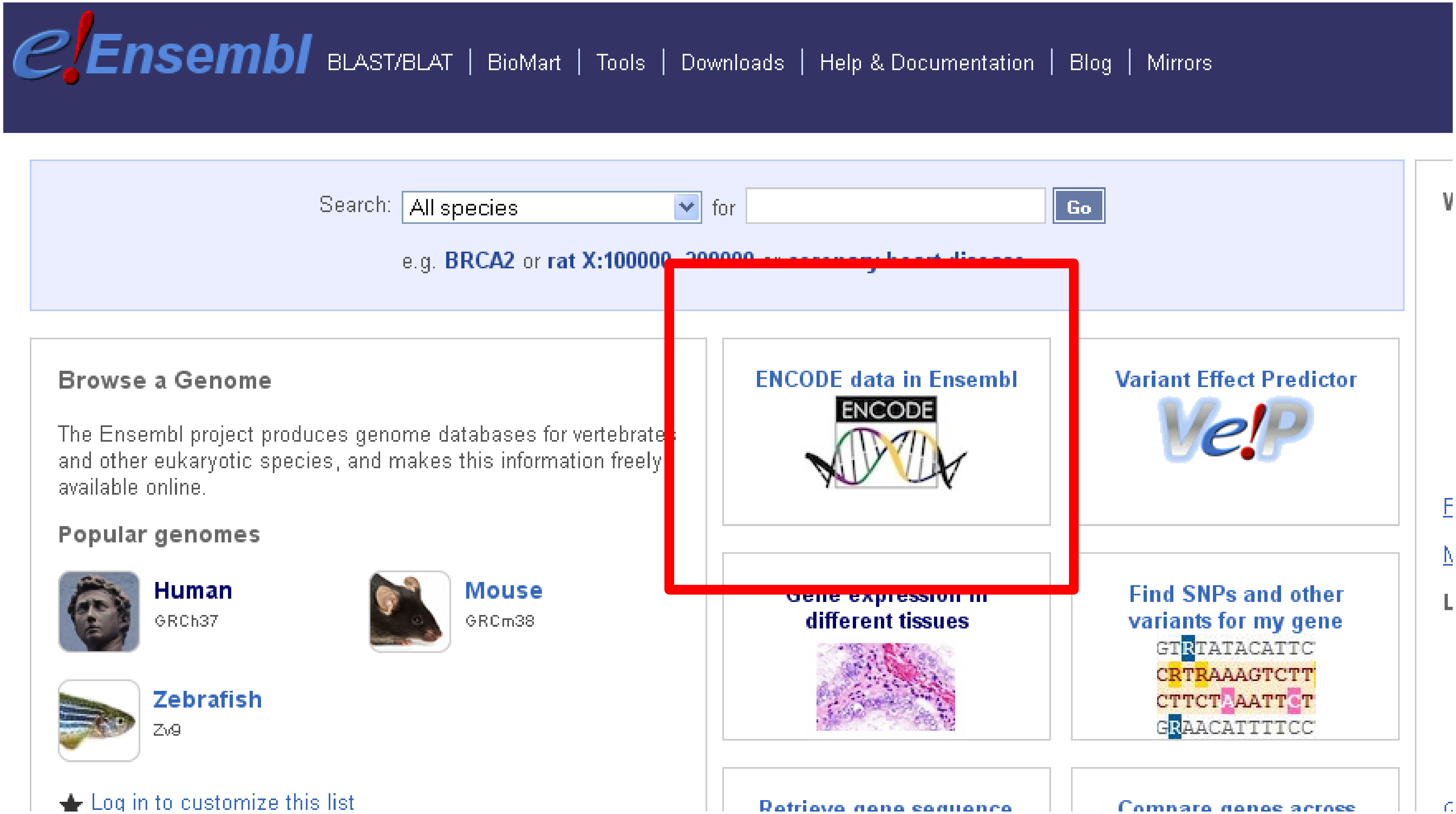
The making of ENCODE: Lessons for big-data projects. Birney E.
Nature. 2012 Sep 6;489(7414):49-51

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

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
ENCODE
Variant Effect Predictor
Ve!P
Gene expression in different tissues
Find SNPs and other variants for my gene
GTRTATACATTC
CRTRAAAGTCTT
CTTCTAAATTCT
GRAACATTTTCC
Retrieve gene sequence
Compare genes across

Gene Ontology

- Built for a very specific purpose:

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

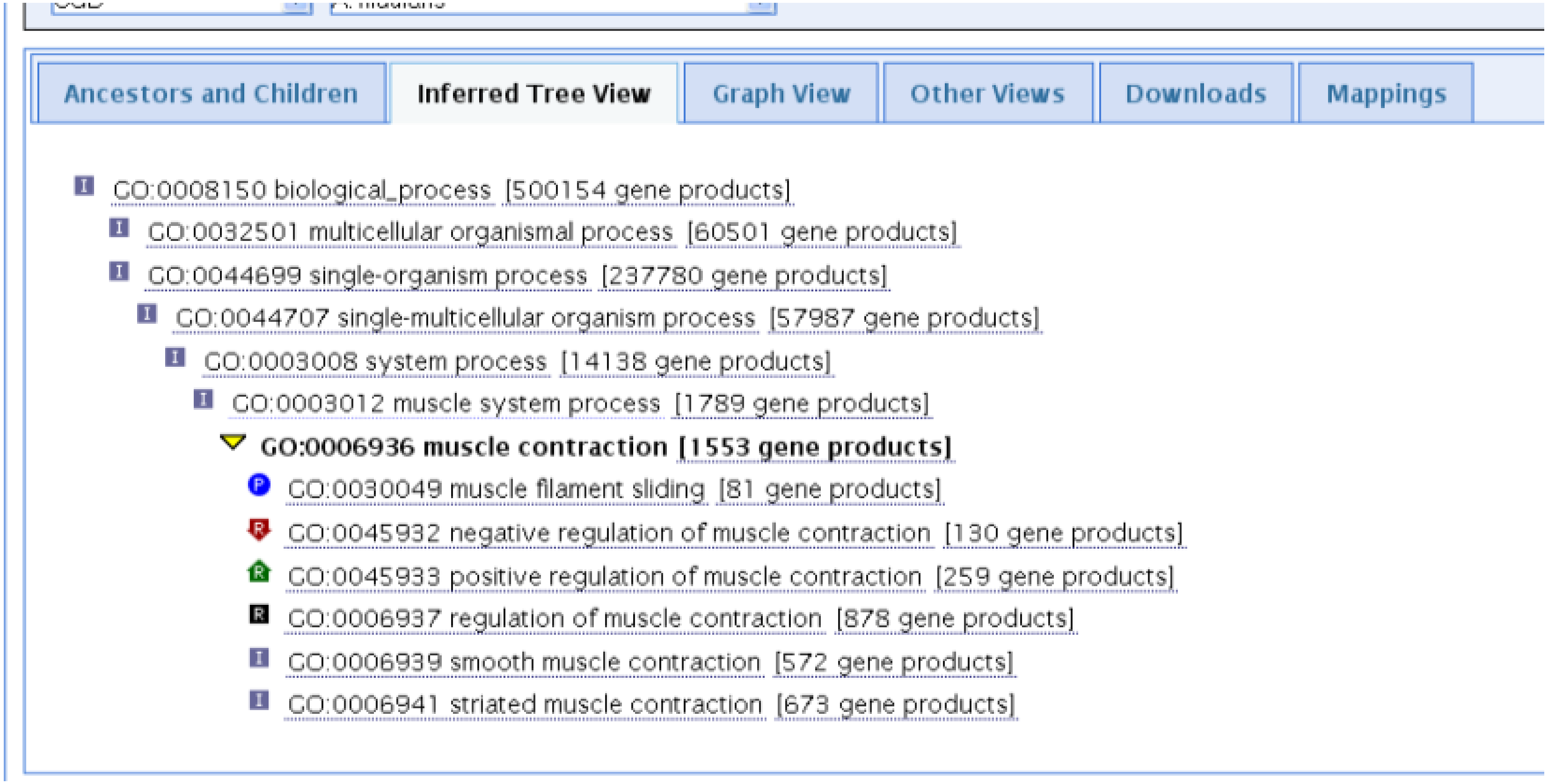
- Applicable to all species

- GO covers ‘normal’ functions and processes
  - No pathological processes
  - No experimental conditions

The 3 Gene Ontologies

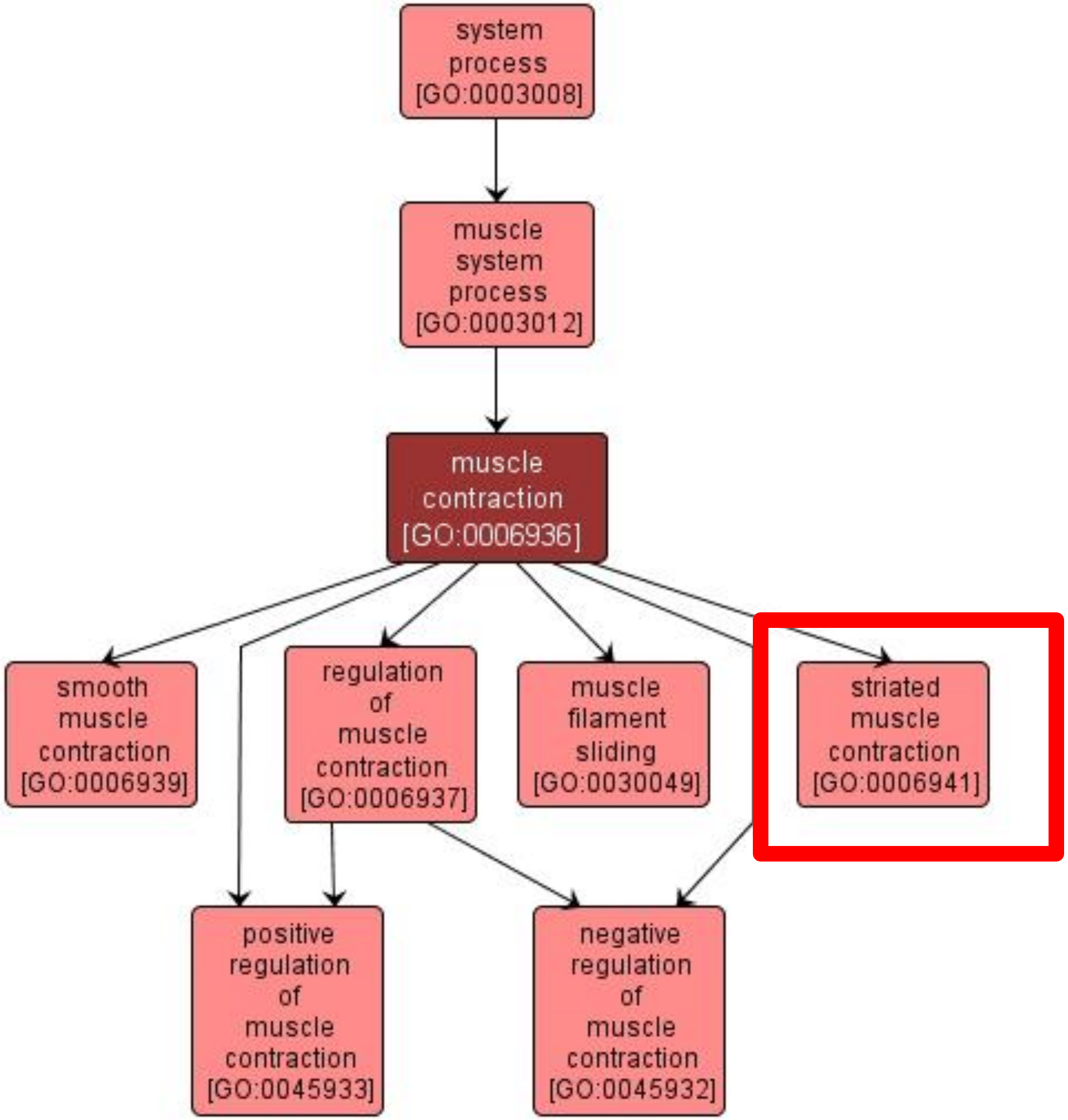
- **Molecular Function** = elemental activity/task
  - the tasks performed by individual gene products; examples are ***carbohydrate binding*** and ***ATPase activity***
- **Biological Process** = biological goal or objective
  - broad biological goals, such as ***mitosis*** or ***purine metabolism***,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

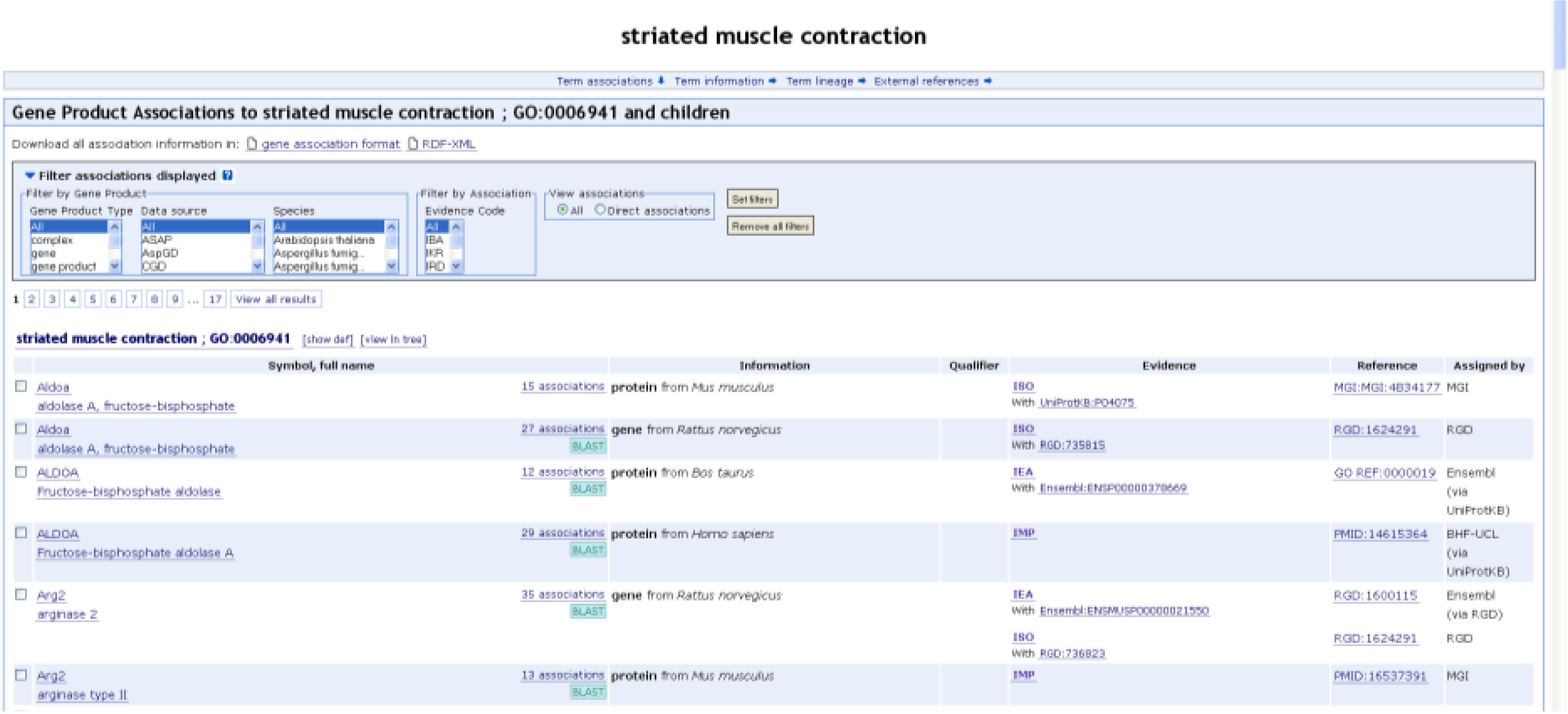


- I GO:0008150 biological_process [500154 gene products]
  - I GO:0032501 multicellular organismal process [60501 gene products]
  - I GO:0044699 single-organism process [237780 gene products]
    - I GO:0044707 single-multicellular organism process [57987 gene products]
      - I GO:0003008 system process [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)



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

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

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

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

Searching and Browsing GO

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

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

Practical session

- Ensembl tutorials
- Ensembl genome browser

- Several NCBI databases
  - Gene
  - OMIM

- Gene Ontology

Questions