Browsing Genes and Genomes with Ensembl

Katarina Truvé
Department of Animal Breeding and Genetics,
Swedish University of Agricultural Science.

References:

Bert Overduin
Ensembl User Support
EMBL Outstation
European Bioinformatics Institute
WellcomeTrust Genome Campus
Hinxton, Cambridge, UK

Outline

1. Introductory lecture about Ensembl.

2. A web-site walk through the key Ensembl web pages. Interactive session.

3. Ensembl assignment: Answering questions by using Ensembl (course homepage)

Ensembl - Goals

- Provide automatic annotation of genomic sequence
- Integrate other biological data
- Make data available to all via the web

Ensembl - Organisation

- Joint project between the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI)
- Started in 1999 for the Human Genome Project
- Funded primarily by the Wellcome Trust, with additional funding by EMBL, EU, NIH-NIAID, BBSRC and MRC
- Team of ca. 50 people, led by Ewan Birney (EBI) and Tim Hubbard (WTSI)

The big Genome Browsers

- Ensembl Genome browser
  http://www.ensembl.org

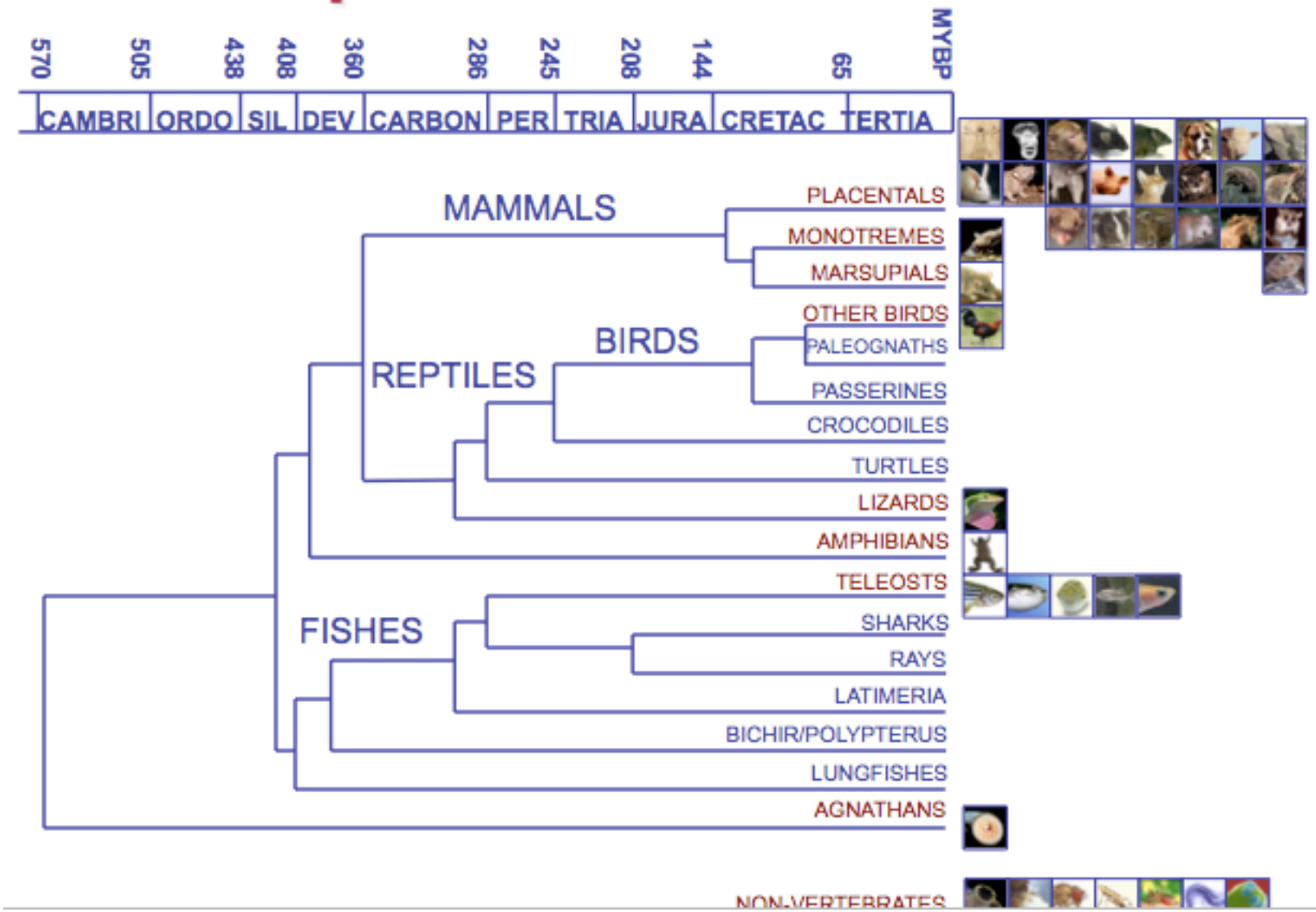
- NCBI Map Viewer
  http://www.ncbi.nlm.nih.gov/mapview/

- UCSC Genome Browser
  http://genome.ucsc.edu

Ensembl / NCBI Map Viewer / UCSC

- All allow access of multiple organisms
- All are based on same data
- Annotations are different
- Assembly versions may differ
- Some organisms specific to only a certain browser

Species in Ensembl



Species in Ensembl

- 46 chordates, ranging from human to two *Ciona* species

- 3 key eukaryote model organisms:

  Drosophila melanogaster
  Caenorhabditis elegans
  Saccharomyces cerevisiae

- 2 insect pathogen vectors:

  Anopheles gambiae
  Aedes aegypti

The Ensembl Genebuild

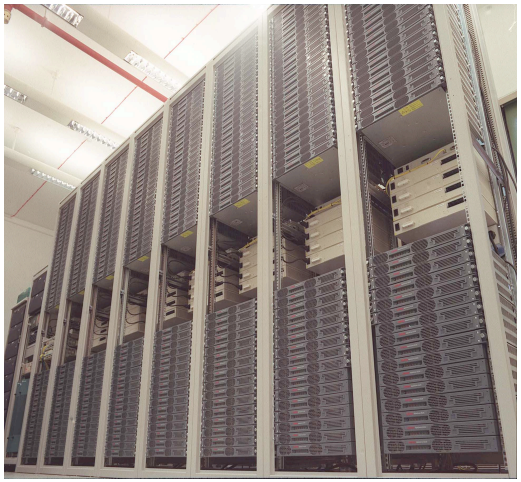
Genome assembly

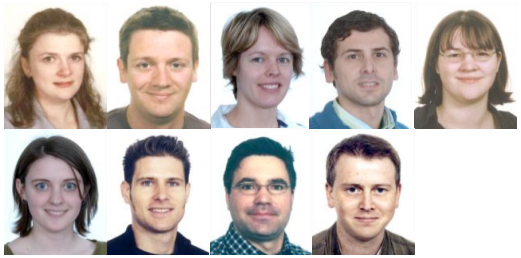
+

Experimental evidence

+

Computer programs





Ensembl Genes

Genome Assemblies

Genome assemblies are not created by Ensembl, but provided by other institutes / consortia, e.g.

- NCBI: human, mouse
- Rat Genome Sequencing Consortium: rat
- Sanger: zebrafish
- Broad Institute: mammals
- Baylor College: cow
- Washington University: chicken

  etc. etc.

Biological Evidence

All Ensembl gene predictions are based on experimental evidence:

- UniProt/Swiss-Prot

  A manually curated database and therefore of highest accuracy
- NCBI RefSeq

  A partially manually curated database
- UniProt/TrEMBL

  Automatically annotated translations of EMBL coding sequence (CDS) features
- EMBL / GenBank / DDBJ

  Primary nucleotide sequence repository

What annotation is available?

- Gene/transcript/peptide models (coding and noncoding (ncRNAs))
- IDs in other databases
- Mapped cDNAs, peptides, micro array probes, BAC clones etc.
- Other features of the genome:
  cytogenetic bands, markers, repeats etc.
- → Comparative data:
  orthologues and paralogues, protein families, whole genome alignments, syntenic regions
- → Variation data:
  SNPs
- Regulatory data:
  "best guess" set of regulatory elements from ENCODE

Comparative Genomics

- Allows us to achieve a greater understanding of vertebrate evolution
- Tells us what is common and what is unique between different species at the genome level
- The function of human genes and other regions may be revealed by studying their counterparts in lower organisms
- Helps identify both coding and non-coding genes and regulatory elements

Comparative genomics

- Functional sequences evolve more slowly than non-functional sequences, therefore sequences that remain conserved *may* perform a biological function.
- Comparing genomic sequences from species at different evolutionary distances allows us to identify:
  - Coding genes
  - Non-coding genes
  - Non-coding regulatory sequences

Homology

- Orthologues :

  any gene pairwise relation where the ancestor node is a speciation event

- Paralogues :

  any gene pairwise relation where the ancestor node is a duplication event

Variation between individuals
Single nucleotide polymorphism (SNPs)

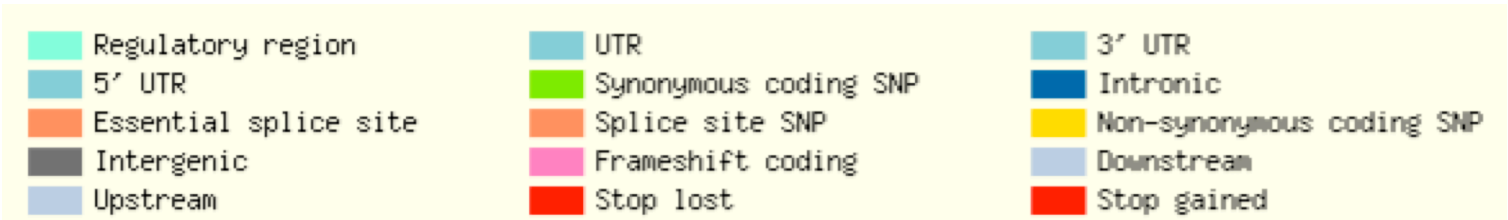
- Two human genomes differ by ~0.1%
- Polymorphism: a DNA variation in which each possible sequence is present in at least 1% of people
- Most polymorphisms (~90%) take the forms of SNPs: variations that involve just one nucleotide
  - ~1 out of every 300 bases in the human genome
  - ~10 million in the human genome

Functional Consequences

Type	Consequence
SNPs in coding area that alter aa sequence	Cause of most monogenic disorders, e.g: Hemochromatosis (HFE) Cystic fibrosis (CFTR) Hemophilia (F8)
SNPs in coding areas that don't alter aa sequence	May affect splicing
SNPs in promoter or regulatory regions	May affect the level, location or timing of gene expression
SNPs in other regions	No direct known impact on phenotype Useful as markers

SNPs in Ensembl - Types

Non-synonymous	In coding sequence, resulting in an aa change
Synonymous	In coding sequence, not resulting in an aa change
Frameshift	In coding sequence, resulting in a frameshift
Stop lost	In coding sequence, resulting in the loss of a stop codon
Stop gained	In coding sequence, resulting in the gain of a stop codon
Essential splice site	In the first 2 or the last 2 basepairs of an intron
Splice site	1-3 bps into an exon or 3-8 bps into an intron
Upstream	Within 5 kb upstream of the 5'-end of a transcript
Regulatory region	In regulatory region annotated by Ensembl
5' UTR	In 5' UTR
Intronic	In intron
3' UTR	In 3' UTR
Downstream	Within 5 kb downstream of the 3'-end of a transcript
Intergenic	More than 5 kb away from a transcript



Practical Applications

- Disease diagnosis
- Association studies
- Pharmacogenomics
- Forensic testing
- Population genetics and evolutionary studies
- Marker-assisted selection

UCSC Genome Bioinformatics

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

Genome Browser
ENCODE
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
VisiGene
Proteome Browser
Utilities
Downloads
Release Log
Custom

About the UCSC Genome Bioinformatics Site

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

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

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

News

News Archives ▶

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

27 April 2009 - New Human Browser Released

We are pleased to announce the release of the February 2009 human genome browser, UCSC version hg19.

Starting with this assembly, the human genome sequence is now provided by the Genome Reference Consortium, whose goal is to correct the small number of regions in the reference that are currently misrepresented, to close as many remaining gaps as possible and to produce alternative assemblies of structurally variant loci when necessary. The hg19 browser corresponds to GRCh37.

Statistics for the GRCh37 build assembly can be found on the NCBI Build 37.1 Statistics web page.

The hg19 browser contains 9 haplotypes. See the Wellcome Trust Sanger Institute MHC Haplotype Project web site for additional information on the chr6

Home Genomes Blat Tables Gene Sorter PCR Session FAQ Help

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.

clade	genome	assembly	position or search term	image width	
Mammal	Human	Mar. 2006	chr8:122,882,337-122,882,339	1200	submit

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

add custom tracks | configure tracks and display | clear position

About the Human Mar. 2006 (hg18) assembly (sequences)

The March 2006 human reference sequence (NCBI Build 36.1) was produced by the International Human Genome Sequencing Consortium.

Sample position queries

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

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000

NCBI HomePage

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

NCBI

Introductor...th May 2009 | Burgoyne Sh...o Tradition | Perl Hash Howto | Mosaik | maqview.1 | IUPAC Co

NCBI

National Center for Biotechnology Information

National Library of Medicine | National Institutes of Health

PubMed | All Databases | BLAST | OMIM | Books | TaxBrowser | Structure

Search All Databases for Go

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More about NCBI...

NLM/NCBI H1N1 Flu Resources

- Newest H1N1 influenza sequences
- Citations recently added to PubMed
- MedlinePlus (consumer health information)
- Enviro-Health Links

Hot Spots

- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources

NCBI

Entrez, The Life Sciences Search Engine

HOME | SEARCH | SITE MAP | PubMed | All Databases | Human Genome | GenBank | Map Viewer | BLAST

Search across databases IL2 GO Clear Help

- Result counts displayed in gray indicate one or more terms not found

2267	**PubMed:** biomedical literature citations and abstracts	327	**Books:** online books
2027	**PubMed Central:** free, full text journal articles	206	**OMIM:** online Mendelian Inheritance in Man
2	**Site Search:** NCBI web and FTP sites	none	**OMIA:** online Mendelian Inheritance in Animals

Wait	**Nucleotide:** Core subset of nucleotide sequence records	none	**dbGaP:** genotype and phenotype
20068	**EST:** Expressed Sequence Tag records	40	**UniGene:** gene-oriented clusters of transcript sequences
4	**GSS:** Genome Survey Sequence records	4	**CDD:** conserved protein domain database
1431	**Protein:** sequence database	47	**3D Domains:** domains from Entrez Structure
27	**Genome:** whole genome sequences	27	**UniSTS:** markers and mapping data
9	**Structure:** three-dimensional macromolecular structures	22	**PopSet:** population study data sets
none	**Taxonomy:** organisms in GenBank	Wait	**GEO Profiles:** expression and molecular abundance profiles
Wait	**SNP:** single nucleotide polymorphism	26	**GEO DataSets:** experimental sets of GEO data
106	**Gene:** gene-centered information	Wait	**Cancer Chromosomes:** cytogenetic databases

NCBI Entrez Gene

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals | Books

Search Gene for IL2 Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Sort by Relevance Send to

All: 106 | Current Only: 103 | Genes Genomes: 93 | SNP GeneView: 71

Items 1 - 20 of 106 Page 1 of 6 Next

1: IL2 Order cDNA clone, Links

Official Symbol IL2 and **Name:** interleukin 2 [*Homo sapiens*]
Other Aliases: IL-2, TCGF, lymphokine
Other Designations: T cell growth factor; aldesleukin; interleukin-2; involved in regulation of T-cell clonal expansion
Chromosome: 4; **Location:** 4q26-q27
Annotation: Chromosome 4, NC_000004.10 (12
MIM: 147680
GeneID: 3558

2: Il2

Official Symbol Il2 and **Name:** interleukin 2 [*Mu*
Other Aliases: DN-144H19.3, Il-2
Other Designations: OTTMUSP00000007837;
Chromosome: 3; **Location:** 3 19.2 cM
Annotation: Chromosome 3, NC_000069.5 (37(
GeneID: 16183

3: IL2

interleukin 2 [*Bos taurus*]
Other Aliases: IL-2
Other Designations: interleukin-2
Chromosome: 17
Annotation: Chromosome 17, NC_007315.3 (36
GeneID: 280822

1: IL2 interleukin 2 [*Homo sapiens*]

GeneID: 3558 updated 03-May-2009

Summary

Official Symbol	IL2 provided by HGNC
Official Full Name	interleukin 2 provided by HGNC
Primary source	HGNC:6001
See related	Ensembl:ENSG00000109471; HPRD:00979; MIM:147680
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	IL-2; TCGF; lymphokine; IL2
Summary	The protein encoded by this gene is a secreted cytokine that is important for the proliferation of T and B lymphocytes. The receptor of this cytokine is a heterotrimeric protein complex whose gamma chain is also shared by interleukin 4 (IL4) and interleukin 7 (IL7). The expression of this gene in mature thymocytes is monoallelic, which represents an unusual regulatory mode for controlling the precise expression of a single gene. The targeted disruption of a similar gene in mice leads to ulcerative colitis-like disease, which suggests an essential role of this gene in the immune response to antigenic stimuli. [provided by RefSeq]