Browsing Genomes with Ensembl
Ensembl Project

- Joint project
  - EMBL – European Bioinformatics Institute (EBI)
  - Wellcome Trust Sanger Institute
- Produce accurate, automatic genome annotation
- Focused on selected eukaryotic genomes
- Integrate external (distributed) biological data
- Presentation of the analysis to all via the Web at
  [http://www.ensembl.org](http://www.ensembl.org)
- Open distribution of the analysis to the community
- Development of open, collaborative software (databases and APIs)
Beyond classical *ab initio* gene prediction

- Ensembl automatic gene prediction relies on homology ‘supporting evidence’ to avoid overprediction.
- Classical *ab initio* gene prediction (e.g. GENSCAN) relies partly on global statistics of protein coding potentials, not used in the cell.
- Genes are just a series of short signals
  - Transcription start site
  - Translation start site
  - 5’ & 3’ Intron splicing signals
  - Termination signals
- Short signal sequences difficult to recognise over background noise in large genomes
Pre! and Archive! sites

http://pre.ensembl.org
http://www.ensembl.org
http://archive.ensembl.org
Open source open standards

• Object model
  – standard interface makes it easy for others to build custom applications on top of Ensembl data
• Open discussion of design (ensembl-dev@ebi.ac.uk)
• Most major pharma and many academics represented on mailing list and code is being actively developed externally
• Ensembl locally
  – Both industry & academia
Making genomes useful

- Interpretation
  - Where are the interesting parts of the genome?
  - What do they do?
  - How are they related to elements in other genomes?

- Access
  - for bench biologists
  - for non-programming mid-scale groups
  - for good programming groups
Ensembl

Analysis DB

Final DB

Supporting Databases

SNP

Manual Annotation

MartView
**Genome browsing**

**why present the whole genome?**

- Explore what is in a chromosome region
- See features in and around a specific gene
- Search & retrieve across the whole genome
- Investigate genome organization
- Compare to other genomes
Basic Genome Annotation

- Genes
  - Genomic location
  - Gene model structures
    - Exons
    - Introns
    - UTRs
  - Transcript(s)
    - Pseudogenes
    - Non-coding RNA
  - Protein(s)
  - Links to other sources of information
Advanced Genome Annotation

- Cytogenetic bands
- Polymorphic markers
  - Sequence Tagged Sites (STS)
- Genetic variation
  - Single Nucleotide Polymorphisms (SNPs)
  - Deletion-Insertion Polymorphisms (DIPs)
  - Short Tandem Repeats (STRs)
- Repetitive sequences
- Expressed Sequence Tags (ESTs)
- cDNAs or mRNAs from related species
- Regions of sequence homology
How to get started … …

- Species homepage
- Map View
- Text search
- BLAST
- SSAHA
BLAST and SSAHA

See blast hit on genome
Regions, maps and markers

ContigView
CytoView
SyntenyView
MultiContigView

MarkerView
SNPView
GeneSNPView
ContigView close-up

- Transcripts: red & black (Ensembl predictions)
- Blue (Vega) & gold (HAVANA, only in human)

Pop-up menu

Detailed view

Basepair view
ContigView - Navigation

- Click and drag mouse to select region
GeneSNP View

Variations in region of gene ENSG00000135744

[Diagram showing genetic variations and annotations for the gene ENSG00000135744]
Marker View

Chromosome Map Marker DXS9752

<table>
<thead>
<tr>
<th>Marker Source</th>
<th>82943 (database: unists)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker Location</td>
<td>Basepairs 138399786 - 138390044 on chromosome X</td>
</tr>
<tr>
<td>Marker Synonyms</td>
<td>Gdb: 737726, 738703, Genbank: 013338, Other: SHOC-11927, DXS9752, RH-0108</td>
</tr>
<tr>
<td>Marker Primers</td>
<td>Expected Product Size</td>
</tr>
<tr>
<td></td>
<td>259</td>
</tr>
</tbody>
</table>

Marker DXS9752 map locations

<table>
<thead>
<tr>
<th>Map Name</th>
<th>Synonym</th>
<th>Chromosome</th>
<th>Position</th>
<th>LOD Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>gm199g3</td>
<td>RH0100</td>
<td>X</td>
<td>4259</td>
<td>3.5</td>
</tr>
</tbody>
</table>
Genes & gene products

GeneView

TransView
ExonView
ProteinView

FamilyView

GOView
Ensembl GeneView

Chromosome X
136,336,415 - 138,371,137

Use Ensembl to...
- Run a BLAST search
- Search Ensembl
- Data mining (RefSeq)
- Upload your own data
- Download data
- Export data

Doses and downloads
- Information
- What's New
- About Ensembl
- Using Ensembl data
- Software

Other links
- Home
- Sitemap
- Archival sites
- Vega
- Trace server
- Stable (archived) link for this page
**Protein View**

**Ensembl Protein Report**

**Peptide**

AGT

This peptide is a member of the human CCDS set CCDS1566.

**Ensembl Peptide ID**

ENSP00000258224

**Translation Information**

This peptide is a translation of transcript ENST00000258224, which is a product of gene ENSG0000018774.

**Genomic Location**

This peptide can be found on Chromosome 1 at location: 227,145,822-227,153,331

This start of this peptide is located in **CDS**.

**Description**

Angiotensinogen precursor (Contains: Angiotensin I [Ang I]; Angiotensin II [Ang II]; Angiotensin III [Ang III] [De-Asp[1] Angiotensin III]).

**Prediction Method**

Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise model from a human/vertebrate protein, a set of aligned human cDNAs followed by GenomerWise for ORF prediction or from GenScan exons supported by protein, cDNA and EST evidence. GeneWise models are further combined with available aligned cDNAs to annotate UTRs.

**InterPro**

PR000027: Angiotensinogen - [View other genes with this domain]
PR000021: Proteinase inhibitor 14, serpin - [View other genes with this domain]

**Protein Family**


This cluster contains 1 Ensembl gene member(s).

**Protein Features**

- **Prints**
- **Probe**
- **PM**
- **Signal peptide**
- **Low complexity**
- **Peptide**
- **SNP**
- **Scale (aa)**
- **SNP Legend**

**Peptide Sequence**

MKVRAQGSHMADAQWSLLELCLALAVAGLACVVYQGHFPGGTHEDIVTYCLAKMWB KMDPFTFLAPQAMTSPEVGLQIQULASDEETDLQLECLQHWNLPLFVISYDV LSAQALQVLNLEQGRKGAELLQEEQGFGPQGLNAITPPVGSLQED FELQPAKKEPROMATTGGCMLGSSYOSDAIPVWYFYQQWEMQGSLIAIQRF VPDSTMSTFVPVMGGCPQMDQXPDFITYQVGPSFETALGQSQAHKALWEGGTPQGLHMQNPKQTYQELQCMALKPCQPLIESPPQLEIWPPISTPTSTYQQNFKSVLVETILNPFLFYATTQQATANLFLGUVNPL7A
**Ensembl GO Search**

<table>
<thead>
<tr>
<th>GO Accession</th>
<th>GO:0004867 [serine-type endopeptidase inhibitor activity]</th>
</tr>
</thead>
</table>

**GO Database**
GO data is provided by the Gene Ontology Consortium.

**Search GO**

Search GO database for: GO:0004967 | e.g. GO:0004867, "vesicle, "calcium binding"

[Search]

**Go Graph**

<table>
<thead>
<tr>
<th>molecular_function</th>
<th>Ensembl Gene Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>enzyme regulator activity [GO:0030234]</td>
<td>5 gene(s)</td>
</tr>
<tr>
<td>enzyme inhibitor activity [GO:0004867]</td>
<td>22 gene(s)</td>
</tr>
<tr>
<td>protease inhibitor activity [GO:0004144]</td>
<td>1 gene(s)</td>
</tr>
<tr>
<td>endopeptidase inhibitor activity [GO:0004866]</td>
<td>17 gene(s)</td>
</tr>
<tr>
<td>serine-type endopeptidase inhibitor activity [GO:0004967]</td>
<td>80 gene(s)</td>
</tr>
<tr>
<td>chymotrypsin inhibitor activity [GO:0030989]</td>
<td>1 gene(s)</td>
</tr>
<tr>
<td>plasmin inhibitor activity [GO:0030968]</td>
<td>1 gene(s)</td>
</tr>
<tr>
<td>trypsin inhibitor activity [GO:0030304]</td>
<td>3 gene(s)</td>
</tr>
</tbody>
</table>
Data retrieval

BioMart

Export View

Data sets on ftp site
MySQL queries of databases
Perl API access to databases
Select region/feature to Export

Choose one or two features from the same chromosome as anchor points and display. Select "None" for the second feature, the display will be based around the first feature.

Please note that there is an upper limit of 2000 that we will export.

Region

- Chromosome name/fragment
- From (type): Base pair
- To (type): Base pair

Context

- Bp downstream
- Bp upstream

Output format

- Output format: FASTA, FEML, EMEL, CSW, GFF, TSV, Flat File

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Download data
- Export data

Does and downloads

- Information
- About Ensembl
- Using Ensembl data
- Software

Other links

- Home
- Sitemap
- What's New
- Stable (archive) link for this page

© 2005 Wellcome Trust Sanger Institute

ExportView
Help!

- context sensitive help pages - click
- access other documentation via generic home page
- email the helpdesk
### Ensembl Team

**Leaders**

- **Ewan Birney** (EBI), **Tim Hubbard** (Sanger Institute)

**Database Schema and Core API**

- **Glenn Proctor**, Andreas Kähäri, Ian Longden, Patrick Meidl

**BioMart**

- **Arek Kasprzyk**, Syed Haider, Damian Smedley, Richard Hollandr

**Distributed Annotation System (DAS)**

- Eugene Kulesha

**Outreach**

- **Xosé M Fernández**, Bert Overduin, Michael Schuster, Giulietta Spudich

**Web Team**

- **James Smith**, Fiona Cunningham, Anne Parker, Steve Trevanion (VEGA), Matt Wood

**Comparative Genomics**

- **Abel Ureta-Vidal**, Kathryn Beal, Benoît Ballester, Stephen Fitzgerald, Javier Herrero Sánchez, Albert Vilella

**Analysis and Annotation Pipeline**

- **Val Curwen**, **Steve Searle**, Browen Aken, Juilo Banet, Laura Clarke, Sarah Dyer, Jan-Hinnerck Vogel, Kevin Howe, Felix Kokocinski, Stephen Rice, Simon White

**Functional Genomics**

- **Paul Flicek**, Yuan Chen, Stefan Gräf, Nathan Johnson, Daniel Rios

**Zebrafish Annotation**

- **Kerstin Howe**, Mario Caccamo, Ian Sealy

**VectorBase Annotation**

- Martin Hammond, Dan Lawson, Karyn Megy

**Systems & Support**

- **Guy Coates**, Tim Cutts, Shelley Goddard

**Research**

- Damian Keefe, Guy Slater, Michael Hoffman, Alison Meynert, Benedict Paten, Daniel Zerbino, Dace Ruklisa

Feb 2007