Ensembl Genome Browser

Jana Vandrovcova
Michael K. Schuster

EMBL-European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, UK

EMBL-EBI

wellcome trust sanger institute

Workshop Outline

- Introduction to the Ensembl project
- Hands-on worked examples
- Data mining with BioMart
- Genome annotation, genes and transcripts
- Comparative genomics and proteomics
- Genetic Variation
- Time for your research

Ensembl Genome Browser

Background to Genomes and
Introduction to the Ensembl Project

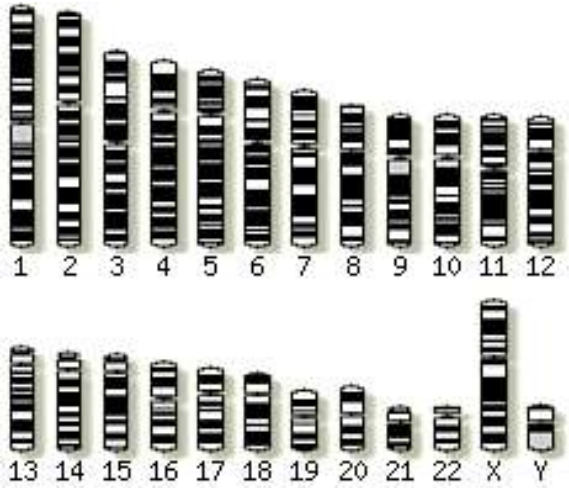
EMBL-EBI

wellcome trust sanger institute

Exploring Genomes

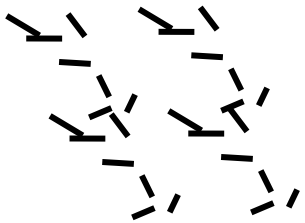
- Browse genes in genomic context
- Display features in and around a particular gene
- Explore larger chromosome regions
- Search and retrieve information on a gene- and genome-scale
- Investigate genome organization
- Compare genomes
- Understand vertebrate evolution

Mapping and Sequencing of the Human Genome



fragment

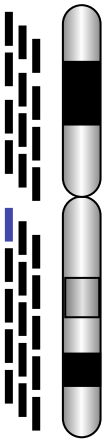
Shizuya *et al* 1992
Dib *et al* 1996
Deloukas *et al* 1998



BACs
bacterial artificial chromosomes
size ~150 kb

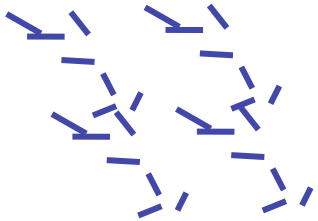
Osoegawa *et al* 2001

map



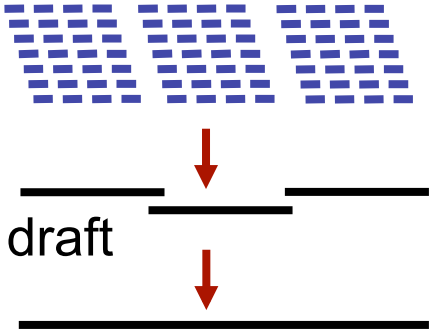
fragment

Bentley *et al* 2001
Bruls *et al* 2001
McPherson *et al* 2001
Montgomery *et al* 2001
Tilford *et al* 2001



pUCs
size ~2-4 kb

sequence assembly

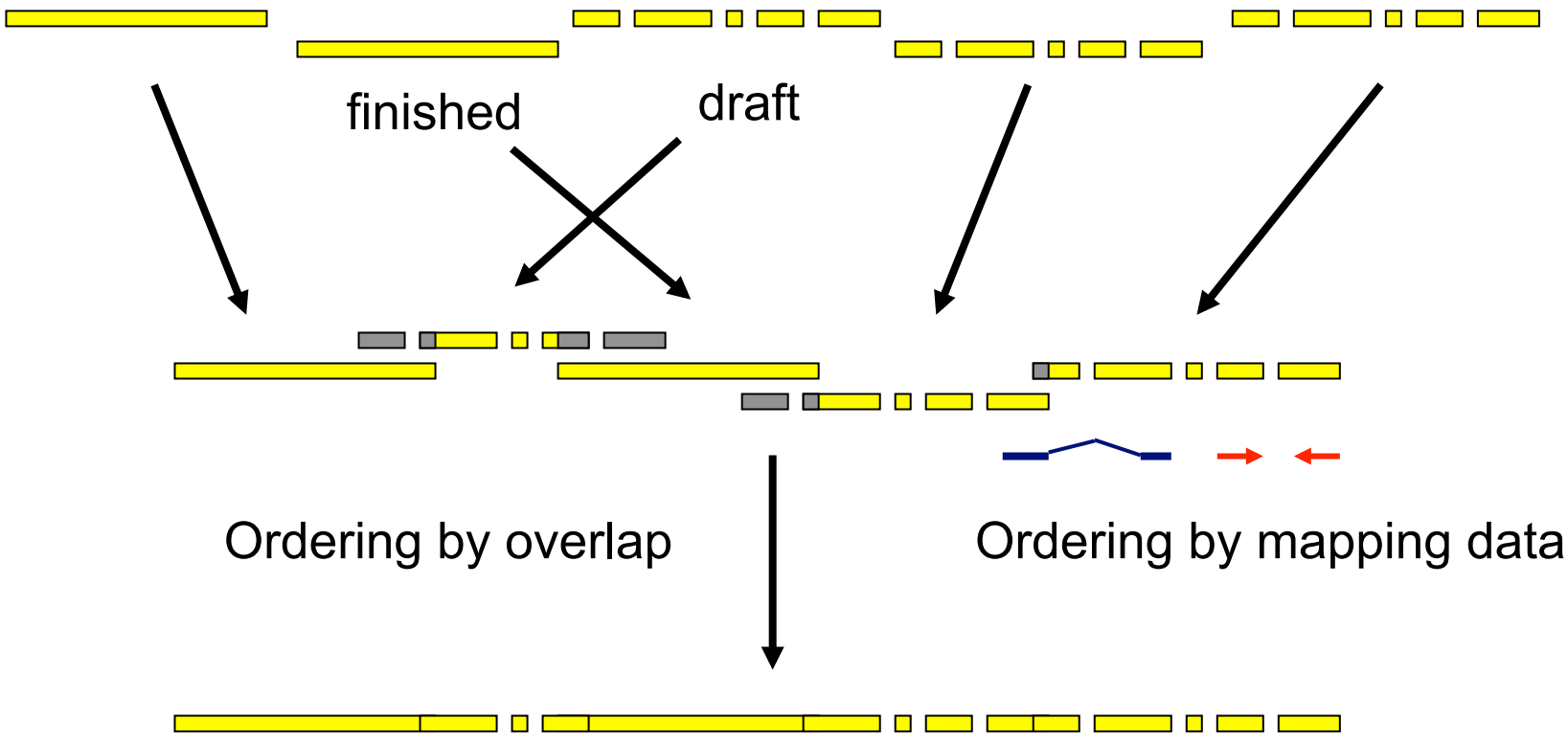


WGS

draft

finished BAC

Genome Sequence Assembly



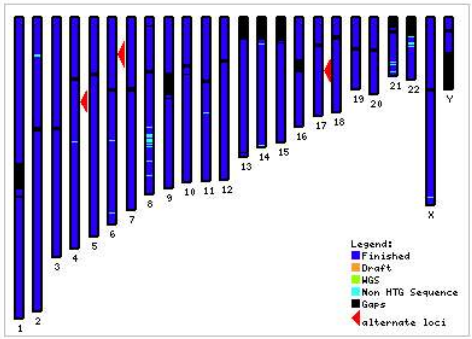
Genome Reference Consortium

GRC Home | Human | Mouse | Help | Report an Issue | Contact Us | Curators Only

Overview | Issues under Review | Assembly Data | Report a problem

Human Genome Overview

Information concerning continuing improvement of the human genome.



GRCh37: A graphical representation of the latest human assembly. The genome is colored with respect to the genomic component used to build the genome assembly at that location. The red triangles mark regions where alternate loci have been provided.

The most recent assembly for human is GRCh37 (download the assembly). This is the first assembly produced by the GRC and is considered the next version of NCBI Build 36 (also known as hg18). Improvements in this assembly include:

- **Closure of 25 unspanned gaps found in Build 36**
- **Resolution of over 150 issues reported as problems in Build 36**
- **Addition of alternate loci for three complex regions, including the MHC region.**
- **Standardization of AGPs, including the addition of biological gap information.**

GRCh37 is a haploid assembly, constructed from multiple individuals and can be divided into a ' primary assembly ' and a set of ' alternate loci '. The primary assembly represents the assembled chromosomes, plus any unlocalized or unplaced sequence that represent the non-redundant, haploid assembly .The alternate loci represent regions for which there is large scale variation and an alternate tilng path is available for this region. An example of such a region can be found at chromosome 17q21.31, often known as the MAPT locus. This region was described as carrying an inversion polymorphism (PMID: 15654335) and has been associated with various phenotypes (PMID: 16718704 ; PMID: 18628315). The version of this region in Build 36 was actually a mosaic of both haplotypes (as tracked in HG-77) and has been resolved in GRCh37 thanks to data described in Zody et al., 2008 (PMID: 19165922).

Information on alternate loci

Chromosome region with alternate loci	Length of region	Number of alternate contigs in region	View Region
UGT2B17 region (chr4:69,170,077-69,877,175)	707,099 bp	1 contig	view
MHC region (chr6: 28,477,797-33,448,354)	4,970,558 bp	7 contigs	view
MAPT region (chr17: 43,384,864-44,913,631)	1,528,768 bp	1 contig	view

References relevant to the genome assembly

Adaptive evolution of UGT2B17 copy-number variation (Xue et al., 2008)
Evolutionary toggling of MAPT 17q21.31 inversion region (Zody et al., 2008)
Closing gaps in the human genome using sequencing by synthesis (Garber et al., 2009)

GRC News and Updates

GRCh37 is now available in Map Viewer
Fri, 14 Aug 2009

NCBI has annotated and released the latest version of the public human genome assembly (GRCh37).

GRCh37 now available at Ensembl
Fri, 14 Aug 2009

Ensembl has annotated and released the latest version of the public human genome assembly (GRCh37). This is available as part of Ensembl release 55.

see all

Recently Resolved Human Issues

Human (HG-333) *Aug14, 2009*

This ticket is the same as ticket number HG-426 in which the reference contains and extra C. The accession AC027220.9 was updated to accession AC027220.10 in NCBI and the extra C at bp position 63,704,670 was removed. The reference will now match the mRNA.

Human (HG-426) *Aug14, 2009*

The sequence was fixed by MIT, but was not uploaded to Genbank for Build 36. Build 37 should reflect the correct sequence.

see all

References

Whole Genome Papers
The HGP Reference Assembly
The Venter Genome Assembly

Human Chromosome Papers
Chr1 Chr2 Chr3 Chr4 Chr5 Chr6
Chr7 Chr8 Chr9 Chr10 Chr11
Chr12 Chr13 Chr14 Chr15 Chr16
Chr17 Chr18 Chr19 Chr20 Chr21
Chr22 ChrX ChrY

FTP | NHGRI | The Wellcome Trust | HHS | NIH | Accessibility |

Page last updated: June 4, 2009

wellcome trust sanger institute

EMBL-EBI

1200 base pairs out of 3,000,000,000…

GCCCCTACAGGCCCCACCCACGACGCTGGCGAGGGATCGGGCGGTCACCGGGAATCGTCT
TAATGCGCGGCAAGGCGCGGGCCTCTCCCTCTCCGCCCGTGAGCCCCGGTGGGAGCGCGC
CGGCGCCCAACTCAAGCGAAACCGCGGGCGTCCCGCCCCGCCCGGCCGCGCCCCGCCTCG
CCTCACGCTAGACTGGGGAGGCGGGACCAATCAGCGAGCGACGTCTCCCTTCCGATTCGA
GGCCCCCGATGCGCGGCTCACACCCCGAGCTTCCCTCGTGCTGATTGGCTGCGGGCGCCG
CCGGTCCGGCCGGGAGGCGGGGCGGGCCGTAGGCAAAGGGAGGTGGGGAGGCGGTGGCCG
GCGACTCCCCGCGCCCCGCTCGCCCCCCGGCCCTTCCCGCGGTGCTCGGCCTCGTTCCTT
TCCTCCTCCGCTCCCTCCGTCTTCCATACCCGCCCCGCGCGGCTTTCGGCCGGCGTGCCT
CGCGCCCTAACGGGCGGCTGGAGGCGCCAATCAGCGGGCGGCAGGGTGCCAGCCCCGGGG
CTGCGCCGGCGAATCGGCGGGGCCCGCGGCCCAGGGTGGCAGGCGGGTCTACCCGCGCGG
CCGCGGCGGCGGAGAAGCAGCTCGCCAGCCAGCAGCCCGCCAGCCGCCGGGAGGTGGGTG
CGTGGCGCCGCGGCGGCCGGCGGCCGAGGGCGGAGGGCGGAAGCGGAGGTGGGCTGGCGG
GGGAGGGCGCGGCCGTGCGGGCGGCCGGTAGGGCTGCGGGCGCGCGCCTGAGGGGAGGAG
GGGCAGCGCGGGCGCGCGCGTCCTCACCCCCTCCTTCCCCGCGGGCGGCGGCCAGGCTCC
CTCCCCTCCCCTTCCCTCTCCTCCCCTCCCCTCCCCTCTCTTCCCCTACCCTCCCGCGCG
CCCGGGCCGCCGGCCGGGCCCGGGCCTGGGGGCGGGGCGGGAAGACGGCGGCCGGGAGTG
TTTTCAGTTCCGCCTCCAATCGCCCATTCCCCTCTTCCCCTCCCAGCCCCCTCCATCCCA
TCGGAAGAGGAAGGAACAAAAGGTCCCGGACCCCCCGGATCTGACGGGGCGGGACCTGGC
GCCACCTTGCAGGTAAAGCCTGGGCGCCCGCGGGCCTCCAGCTAGGGAAGTGTTTGCGTG
CGTCCGCGGCCGGGGCGATGGGCCGTGTCACATGGCCGCTGCGGGTGGGGGCTGGGGTGT
. . .

Basic Genome Annotation

- Genes
  - Genomic location
  - Gene model structures
    - Exons
    - Introns
    - UTRs
  - Transcripts
    - Protein-coding
    - Pseudo-transcripts
    - non-coding RNA
  - Proteins
  - Links to other sources of information

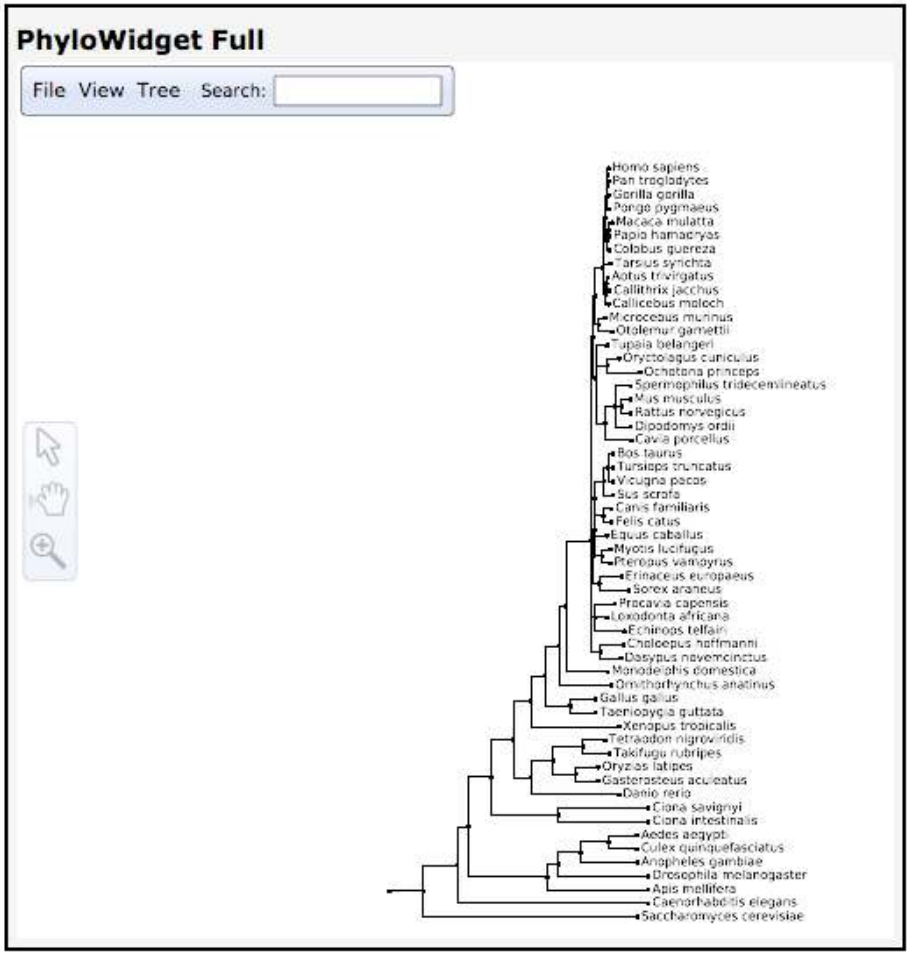
Advanced Genome Annotation

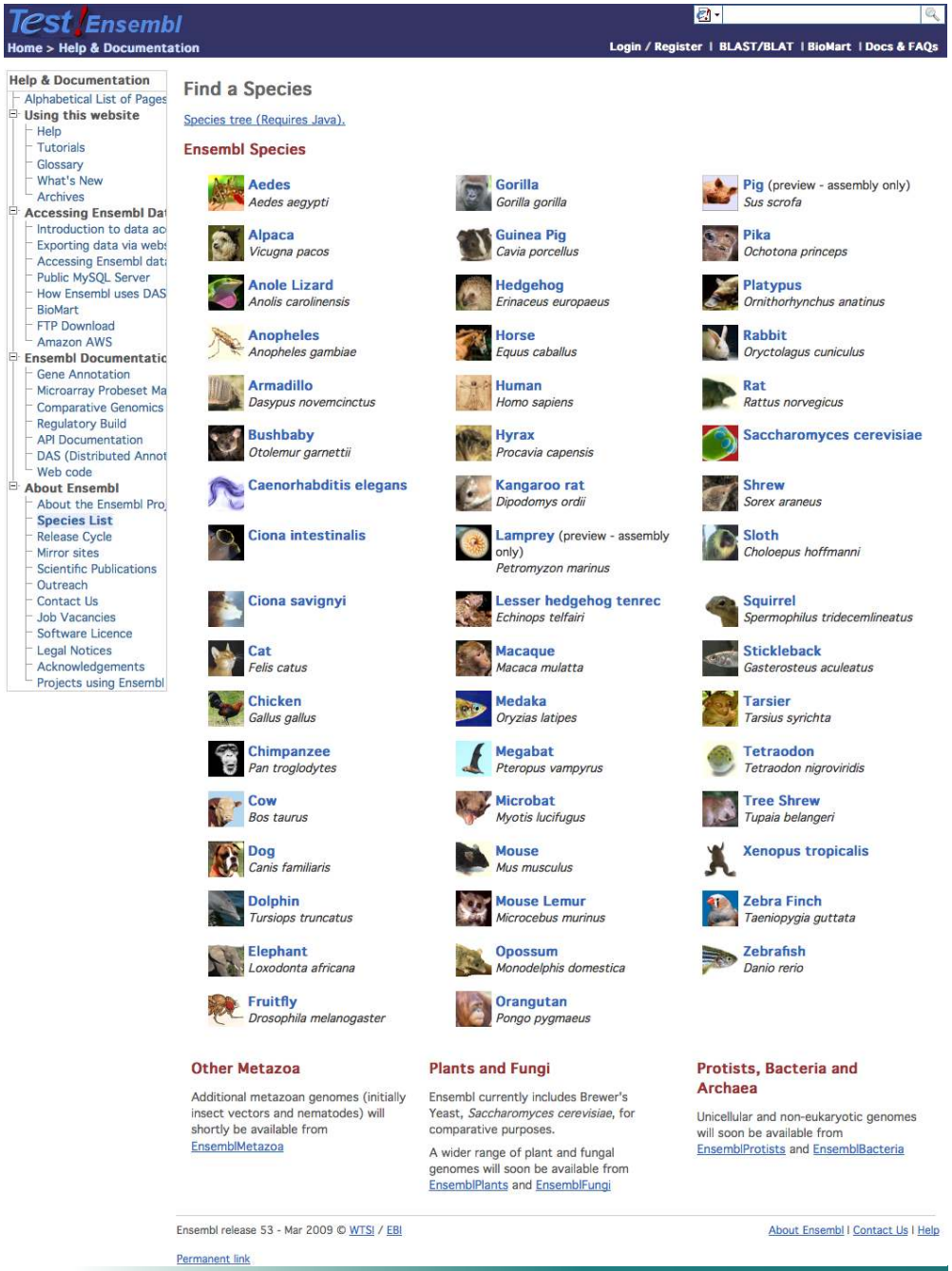
- Cytogenetic bands
- Polymorphic markers
  - Sequence Tagged Sites (STS)
- Genetic variation
  - Single Nucleotide Polymorphisms (SNPs)
  - Deletion-Insertion Polymorphisms (DIPs)
  - Short Tandem Repeats (STRs)
- Repetitive sequences
- Expressed Sequence Tags (ESTs)
- cDNAs or mRNAs from related species
- Expression array probe set mapping

Modern Genome Annotation

- Comparative Genomics
  - Sequence homology
  - Conservation scores
  - Multiple sequence alignments
- Functional Genomics
  - Paired-end tags (PET)
  - Cap analysis gene expression (CAGE) tags
  - ChIP on CHIP or ChIP Seq data
    - Histone modifications
    - Transcription factor binding sites
    - DNAse I hypersensitivity sites
    - DNA methylation
- Structural genome variation
  - Copy number variations (CNVs)

Species in Ensembl





wellcome trust sanger institute

EMBL-EBI

Ensembl - Project Aims

- Funded to provide metazoan genomes to the world
- Aims to provide automated genome annotation system
- Graphical representation of complex data sets
- Large-scale data export tools
- All software, data and results freely available

Ensembl - Project Background

- Joint project between
  - European Bioinformatics Institute
  - Wellcome Trust Sanger Institute
- Group of ca 40 people led by **Paul Flicek** (EBI) and **Steve Searle** (WTSI)
- Ensembl strategy committee: **Ewan Birney** (EBI), **Tim Hubbard** (WTSI) and **Richard Durbin** (WTSI)
- Mainly Wellcome Trust funded
- Additional EMBL, EU, BBSRC and NIH-NIAID funding

Ensembl Team

D800–D806 Nucleic Acids Research, 2011, Vol. 39, Database issue *Published online 2 November 2010*
doi:10.1093/nar/gkq1064

Ensembl 2011

Paul Flicek[1,2,*], M. Ridwan Amode[2], Daniel Barrell[2], Kathryn Beal[1], Simon Brent[2], Yuan Chen[1], Peter Clapham[2], Guy Coates[2], Susan Fairley[2], Stephen Fitzgerald[1], Leo Gordon[1], Maurice Hendrix[2], Thibaut Hourlier[2], Nathan Johnson[1], Andreas Kähäri[1], Damian Keefe[1], Stephen Keenan[1], Rhoda Kinsella[1], Felix Kokocinski[2], Eugene Kulesha[1], Pontus Larsson[1], Ian Longden[1], William McLaren[1], Bert Overduin[1], Bethan Pritchard[2], Harpreet Singh Riat[2], Daniel Rios[1], Graham R. S. Ritchie[1], Magali Ruffier[2], Michael Schuster[1], Daniel Sobral[1], Giulietta Spudich[1], Y. Amy Tang[2], Stephen Trevanion[2], Jana Vandrovcova[1], Albert J. Vilella[1], Simon White[2], Steven P. Wilder[1], Amonida Zadissa[2], Jorge Zamora[1], Bronwen L. Aken[2], Ewan Birney[1], Fiona Cunningham[1], Ian Dunham[1], Richard Durbin[2], Xosé M. Fernández-Suarez[1], Javier Herrero[1], Tim J. P. Hubbard[2], Anne Parker[2], Glenn Proctor[1], Jan Vogel[2] and Stephen M. J. Searle[2]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton Cambridge CB10 1SD and [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Received October 7, 2010; Accepted October 13, 2010

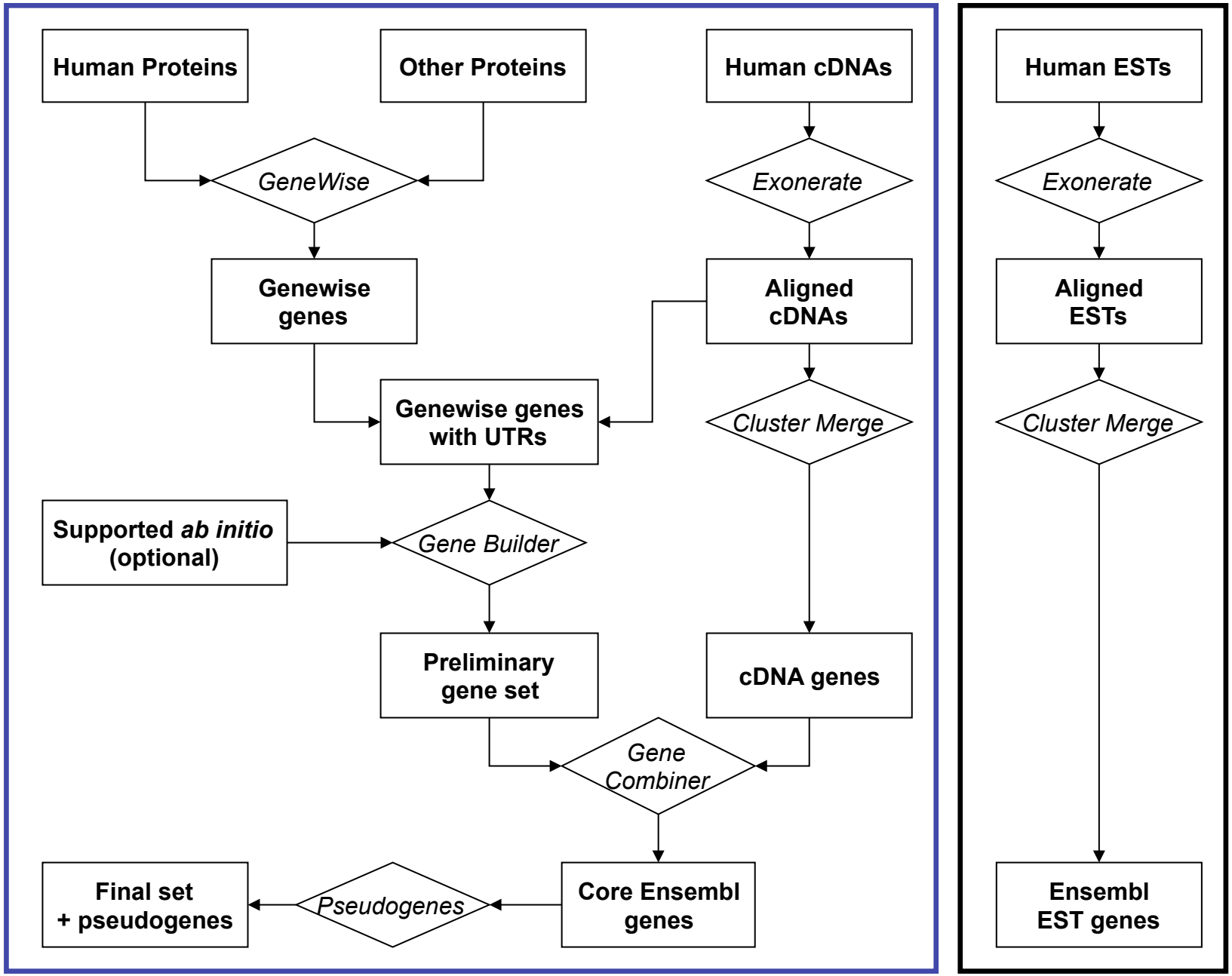
The Ensembl System

- Scalable database system for storing genomic data and annotation
- Automatic annotation method based on a rule set of heuristics
- Public web interface for genome annotation display **Ensembl Genome Browser**

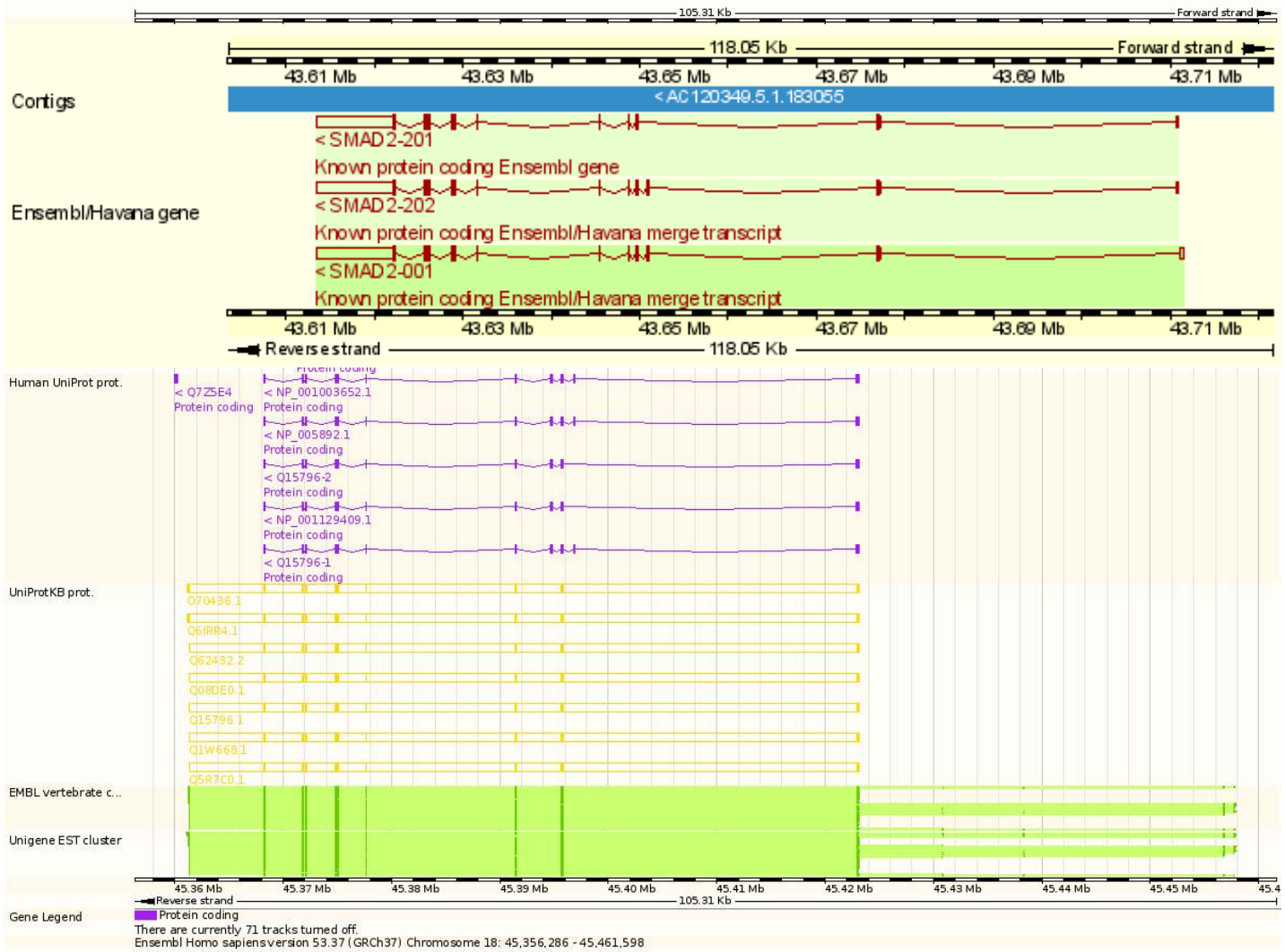
Biological Evidence

All Ensembl gene predictions are based on **experimental evidence**

- UniProtKB/Swiss-Prot
  - A manually curated database and therefore of highest accuracy
- NCBI RefSeq
  - A partially manually curated database
- UniProtKB/TrEMBL
  - Automatically annotated translations of EMBL coding sequence (CDS) features
- European Nucleotide Archive (ENA)
  - Comprehensive set of nucleotide sequences owned by original submitters
  - International Nucleotide Sequence Database Collaboration GenBank (NCBI) and DDBJ (Japan)


Human Proteins
Other Proteins
Human cDNAs
Human ESTs
GeneWise
Exonerate
Exonerate
Genewise genes
Aligned cDNAs
Aligned ESTs
Genewise genes with UTRs
Cluster Merge
Cluster Merge
Supported ab initio (optional)
Gene Builder
Preliminary gene set
cDNA genes
Gene Combiner
Final set + pseudogenes
Pseudogenes
Core Ensembl genes
Ensembl EST genes

Non-redundant Set of Alternatively Spliced Transcripts



Ensembl Transcripts

- Ensembl genes or transcripts
  - An automatically annotated gene set
- Havana genes or transcripts
  - A manually curated gene set
- Ensembl – Havana merged transcripts
  - Coding sequence and exon boundary overlap
- Known genes or transcripts
  - Based on species-specific evidence
- Novel genes or transcripts
  - Inferred from closely related species
- EST genes or transcripts
  - Predicted on the basis of EST evidence
- GENSCAN or SNAP transcripts
  - Based on *ab initio* transcript model predictors

Ensembl - Open Source

- Data and Software freely-available
- Developer community of about 300 people, including companies
- Over 50 Ensembl installs worldwide
  - Official Mirrors
    - http://uswest.ensembl.org/
    - http://ensembl.genomics.org.cn/
  - Other species and data sets
    - Arabidopsis (NASC, UK)
    - CADRE (Manchester, UK)
    - Gramene (CSHL, US)
    - Fugu (ICMB, SG)
    - Ciona (Temasek, SG)

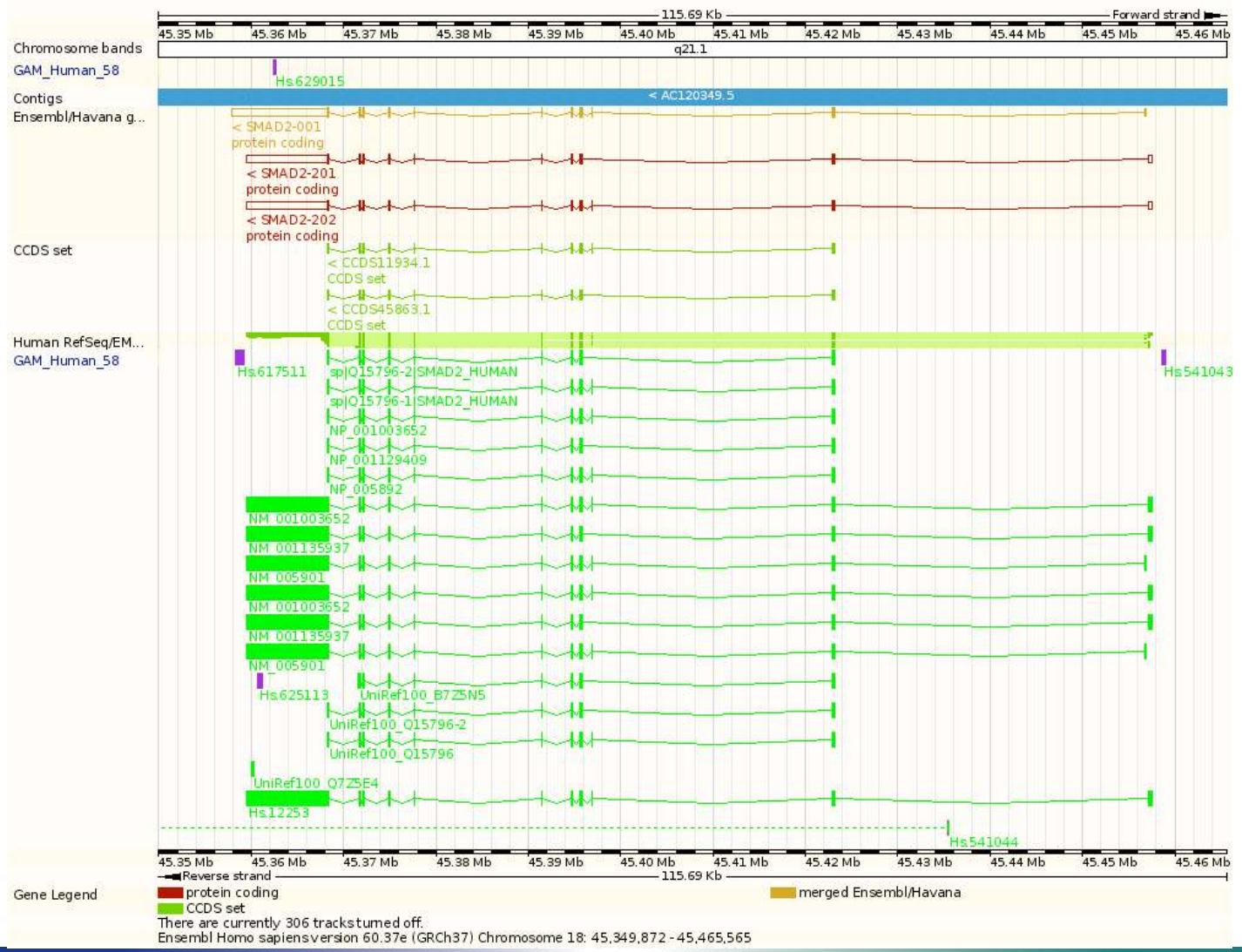
Access for Laboratory-Based Scientists

- Focussed on one or a few related genes
- Access mainly via the web site
- Web site designed for non-programming, not that genome aware biologists
  - Basic annotation is simple to find
  - Consistency of layout, colour schema and information
  - Integrated genome browser supporting several species
  - Simple feature and sequence retrieval

The Ensembl Web Site

- “Public face” of Ensembl
  - Contact point for the Ensembl project
- Visual display of Ensembl genome annotation data
  - Graphical, intuitive displays aimed at biologists
- Web-based tools
  - Sequence similarity searches (BLAST and BLAT)
  - Variant Effect Predictor
  - Gene name and transposon registry for *Anopheles gambiae*
- A framework to integrate user data
  - Simple data upload: GFF, GTF, BED, PSL, now BAM
  - Distributed Annotation System (DAS) and DAS Registry
  - Local data integration via Plug-Ins, Registry and Adaptors
- Local site installation
  - Free, open-source, supported

Distributed Annotation System (DAS)



Web Access to Genome Annotation

- Current Release
  - Fully annotated genomes
  - Complete integration (Comparative Genomics, BioMart)
  - http://www.ensembl.org/
- Pre-Release
  - Preliminary data sets for new assemblies
  - http://pre.ensembl.org/
- Archives
  - Previous releases
  - Serve as reference points for publications
  - Kept for at least two years
  - http://archive.ensembl.org/

Access for Mid-Scale Groups

- Work with 50 to 1,000 genes, genome regions, expression data
- Little in-house programming
  - Some web views particularly designed for this group
  - BioMart focuses on this group

Exporting Data Subsets

- Genome Browser Export Dialogue
  - Region-focused
  - EMBL, GenBank annotated flat files (transcripts, variation, repetitive DNA elements ...)
  - FASTA simple sequence export (sequence only format)
  - GFF annotation export (annotation only format)
  - Images from graphical displays in PNG, SVG or PDF format
- BioMart
  - Feature-focused
  - Mix and match queries
  - “Instant” refresh of selected set
  - Flexible output to HTML table, FASTA, CSV, TSV, Excel …
  - All Ensembl genes on chromosome 5 in GTF format, etc…

Access for Large-Scale Groups

- Full use of the genome, experienced bioinformaticians
- Complete openness of Ensembl
  - Open data
  - Open software
  - Open MySQL server on the internet
  - Expect everything to be portable
  - Participate in standards and adopt other standards (DAS, UCSC upload)

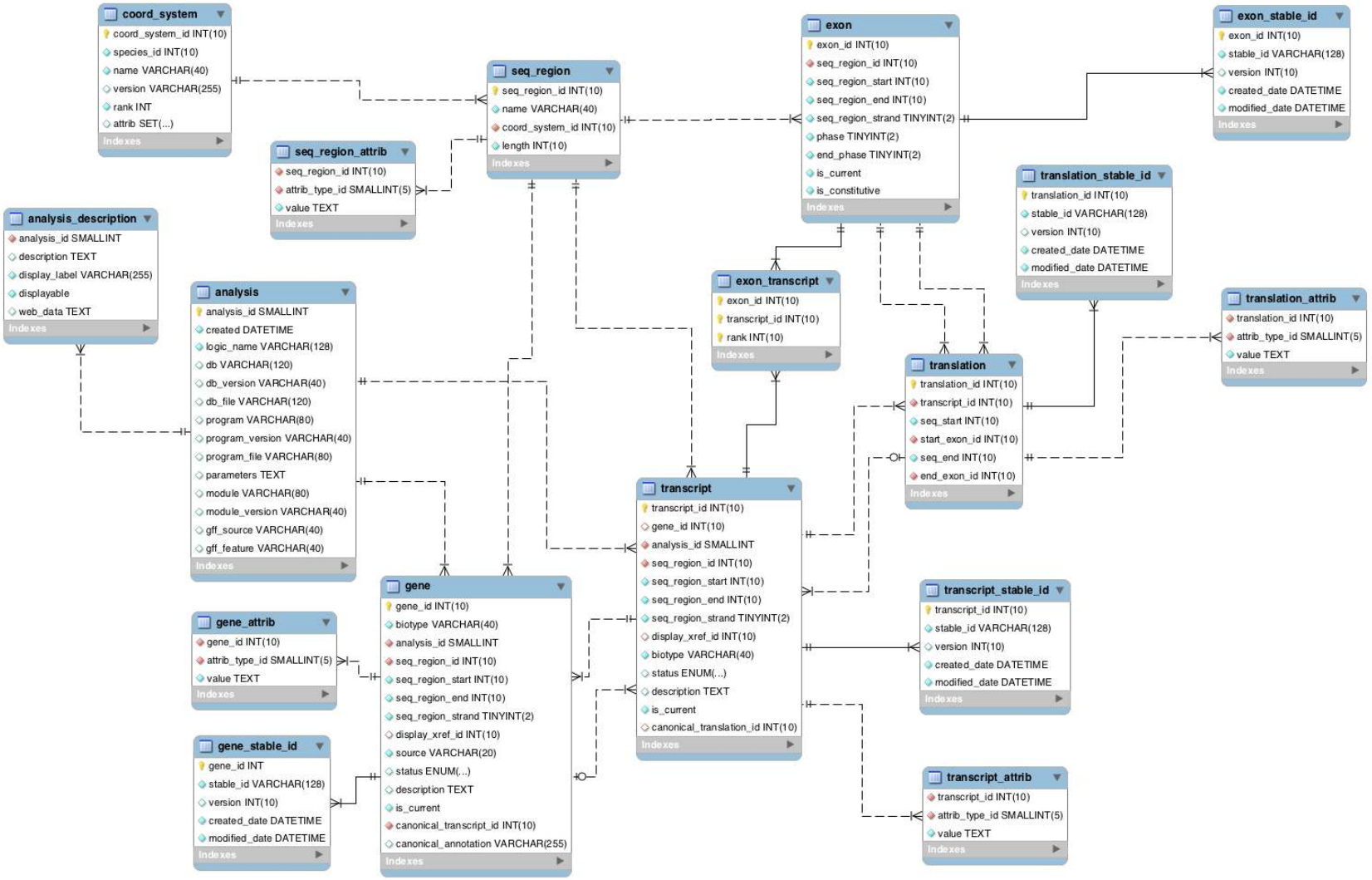
Advanced Access to Genome Annotation

- FTP site
  - ftp://ftp.ensembl.org/
  - FASTA files for mRNAs, ncRNAs, and proteins
  - EMBL and GenBank files of annotated genome sequence slices
  - SQL table dumps
  - Gene Transfer Format (GTF) annotated gene sets
- MySQL interface
  - http://www.ensembl.org/info/data/mysql.html
  - ensembldb.ensembl.org
- Perl API
  - http://www.ensembl.org/info/docs/api/
- Amazon Web Services
  - http://aws.amazon.com/publicdatasets/

Low-Level Exporting

- Direct database access at “ensembldb.ensembl.org”
- MySQL client program
  - UNIX® and Linux®
    - Client and server software
      http://dev.mysql.com/downloads/mysql/
  - Windows®
    - WinMySQL 1.1 (graphical user interface)
      http://www.winmysql.com/download/wmysr11.zip

Database Schema



Example Query

Retrieve Ensembl Transcript and Peptide IDs for ENSG00000010704

```
mysql -u anonymous -h ensembldb.ensembl.org
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 1699364 to server version: 4.1.20
standard-log

Type 'help;' or '\h' for help. Type '\c' to clear the buffer.

mysql> use homo_sapiens_core_41_36c;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> SELECT gene_stable_id.stable_id AS gene, transcript_stable_id.stable_id
AS transcript, translation_stable_id.stable_id AS peptide FROM gene,
transcript, translation, gene_stable_id, transcript_stable_id,
translation_stable_id WHERE gene.gene_id = transcript.gene_id AND
transcript.transcript_id = translation.transcript_id AND gene_stable_id.gene_id
= gene.gene_id AND transcript_stable_id.transcript_id =
transcript.transcript_id AND translation_stable_id.translation_id =
translation.translation_id AND gene_stable_id.stable_id = 'ENSG00000010704';
```

Query Result

Result:

gene	transcript	peptide
ENSG00000010704	ENST00000309234	ENSP00000311698
ENSG00000010704	ENST00000349999	ENSP00000259699
ENSG00000010704	ENST00000317896	ENSP00000313776
ENSG00000010704	ENST00000353147	ENSP00000312342
ENSG00000010704	ENST00000352392	ENSP00000315936
ENSG00000010704	ENST00000336625	ENSP00000337819
ENSG00000010704	ENST00000345823	ENSP00000344033
ENSG00000010704	ENST00000357618	ENSP00000350238
ENSG00000010704	ENST00000317880	ENSP00000313489

Application Programme Interfaces (APIs)

- Encapsulate Ensembl “technology”
- Stand-alone “products”
  - Object-oriented Perl
  - Partly based on BioPerl
  - Modular organisation
  - Ensembl core, comparative genomics, pipeline, …
  - Used for web, command line and application programme interfaces
- Freely-available, open-source
  - Community development

Example Perl Program

Retrieve Ensembl Transcript and Peptide IDs for ENSG00000010704

```
#!/software/bin/perl

use strict;
use warnings;

use Bio::EnsEMBL::Registry;
my $reg = "Bio::EnsEMBL::Registry";

$reg->load_registry_from_db(
        -host => 'ensembldb.ensembl.org',
        -user => 'anonymous');

my $gene_adaptor = $reg->get_adaptor ("human", "core", "Gene");
my $gene = $gene_adaptor->fetch_by_stable_id('ENSG00000010704');
my @transcripts = @{$gene->get_all_Transcripts()};

print "Gene\t\tTranscript\tPeptide\n";
foreach my $transcript(@transcripts){
        print $gene->stable_id, "\t", $transcript->stable_id, "\t",
        $transcript->translation->stable_id, "\n";
}
```

Program Output

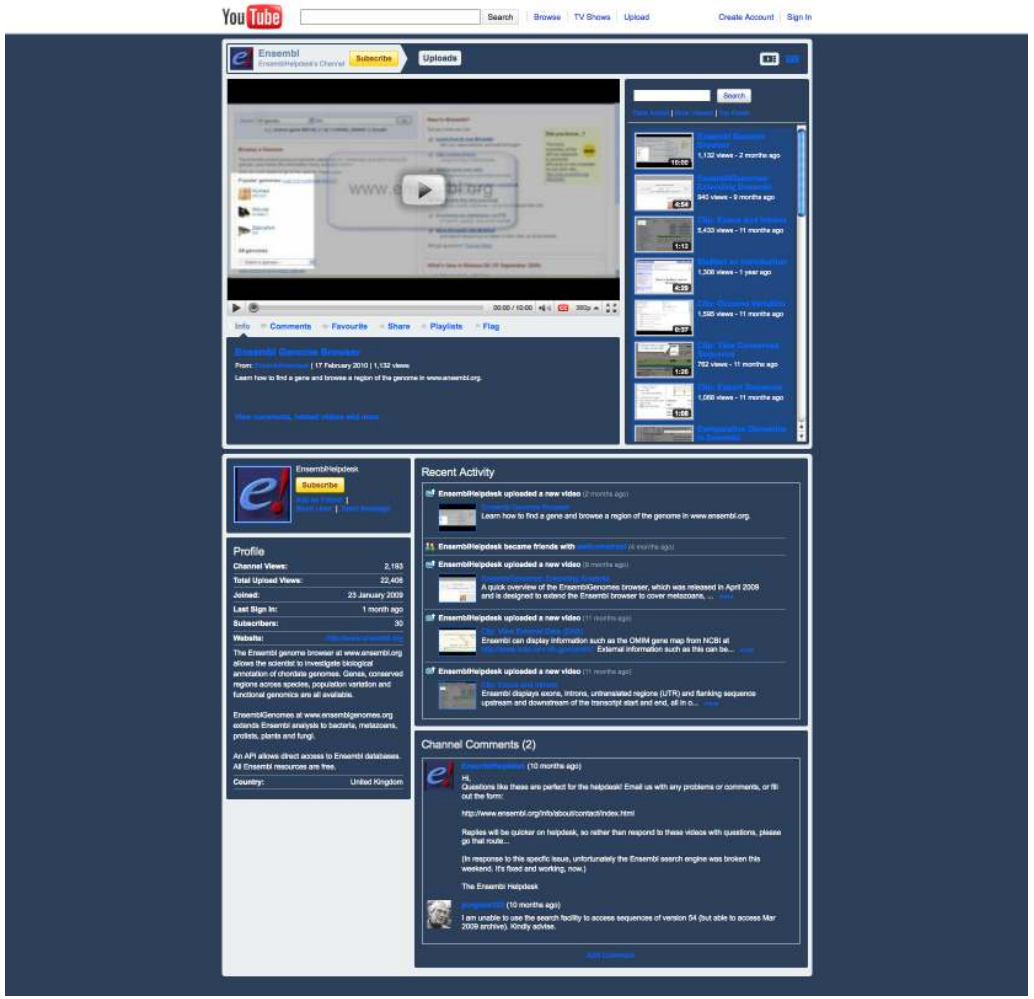
Result:

```
Gene             Transcript       Peptide
ENSG00000010704 ENST00000309234 ENSP00000311698
ENSG00000010704 ENST00000349999 ENSP00000259699
ENSG00000010704 ENST00000317896 ENSP00000313776
ENSG00000010704 ENST00000353147 ENSP00000312342
ENSG00000010704 ENST00000352392 ENSP00000315936
ENSG00000010704 ENST00000336625 ENSP00000337819
ENSG00000010704 ENST00000345823 ENSP00000344033
ENSG00000010704 ENST00000357618 ENSP00000350238
ENSG00000010704 ENST00000317880 ENSP00000313489
```

Ensembl Support

- Ensembl helpdesk
  - Private mailing list
  - General enquiries, feedback and support
    helpdesk@ensembl.org
- Ensembl announcements
  - Public mailing list
  - Low-volume, announcement of new releases
    ensembl-announce@ebi.ac.uk
- Ensembl developers
  - Public mailing list
  - Good for technical support
    ensembl-dev@ebi.ac.uk
- Ensembl Blog ensembl.blogspot.com

Ensembl Helpdesk YouTube Channel



http://www.youtube.com/user/EnsemblHelpdesk

QUESTIONS

ANSWERS