Ensembl

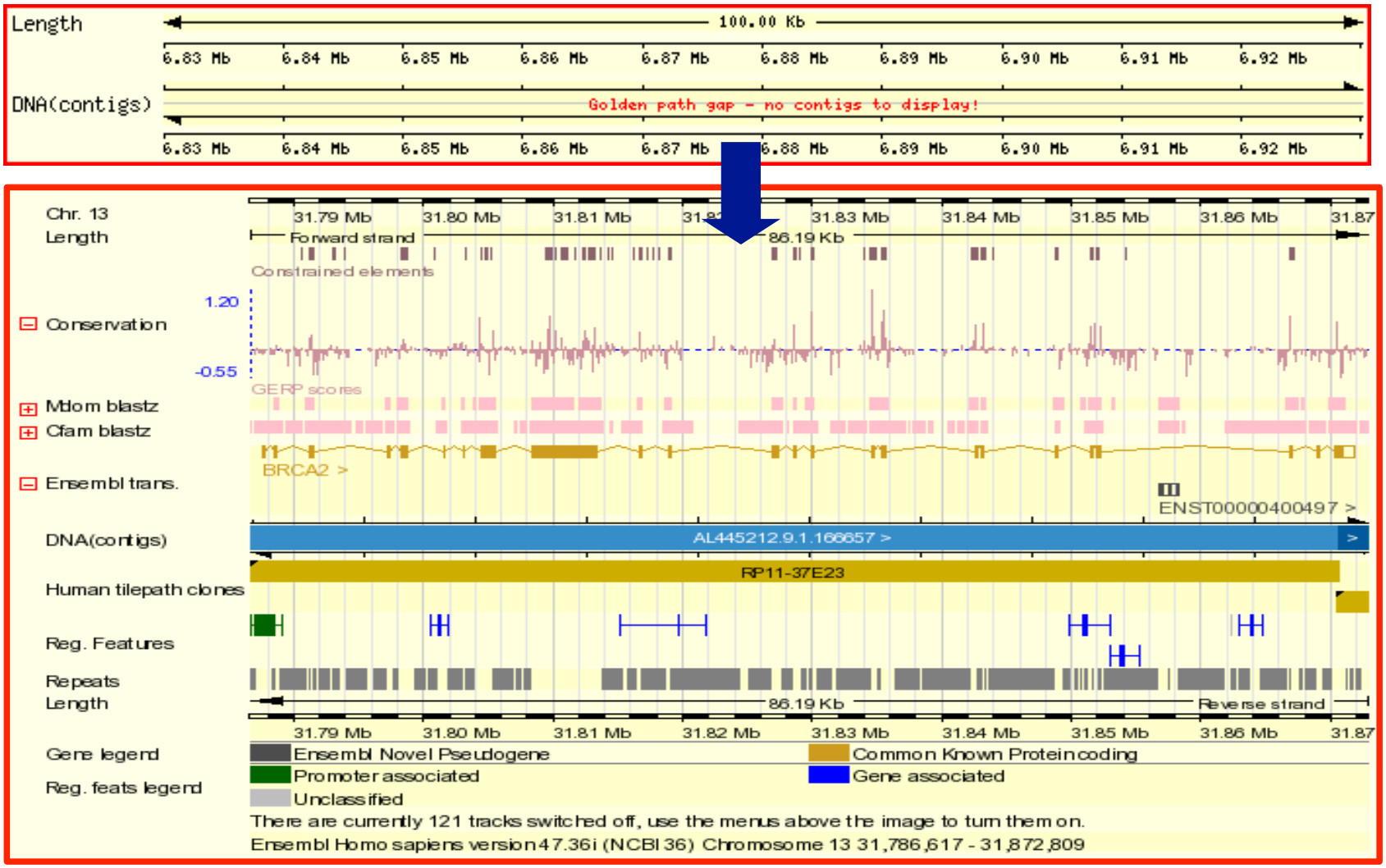
Javier Herrero *(Former Ensembl member)*
Comparative Genomics Project Leader
The Genome Analysis Center (TGAC)
Norwich Research Park
Norwich, UK

EMBL-EBI

wellcome trust sanger institute

Ensembl mission

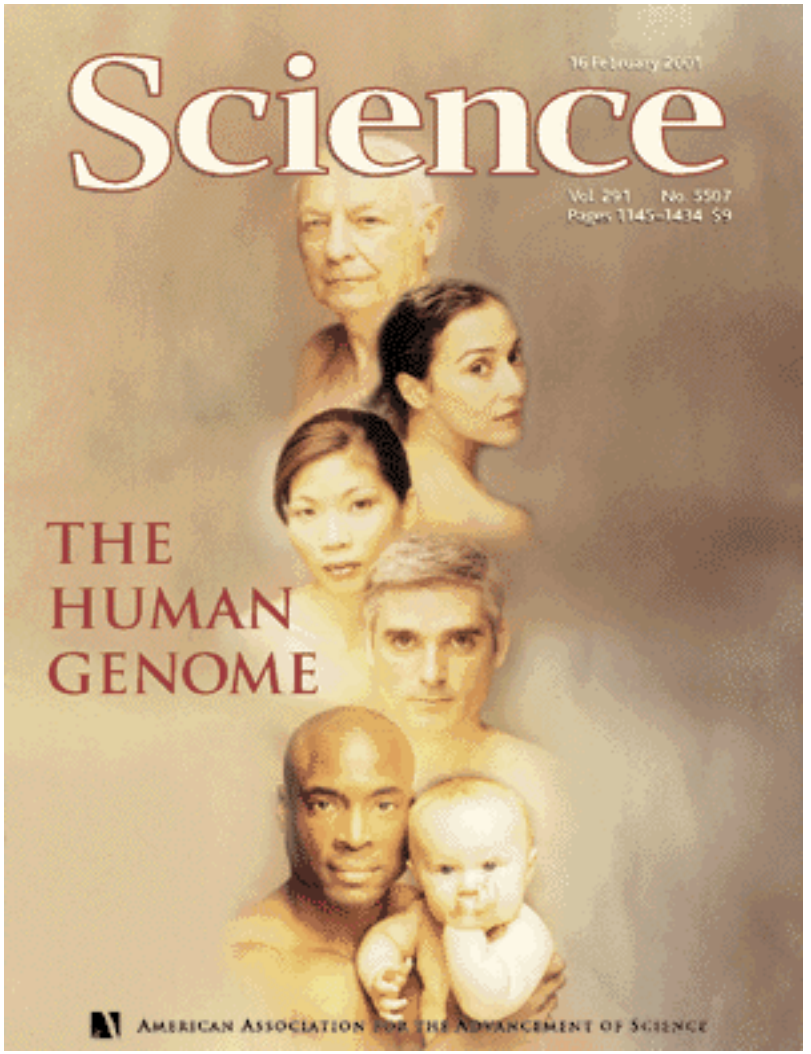
To enable genomic science by providing high-quality, integrated annotation on vertebrate genomes within a consistent and accessible infrastructure.



Ensembl Concept

- Collaborative project of the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute
- Provides annotation and analysis of chordate genomes
- Open by design
  - Code is BSD, not GNU
  - All data is freely available
- Continuously developed and comprehensively updated 5 times a year
- Diverse skills across the project
- Technology adopted and used by many other projects

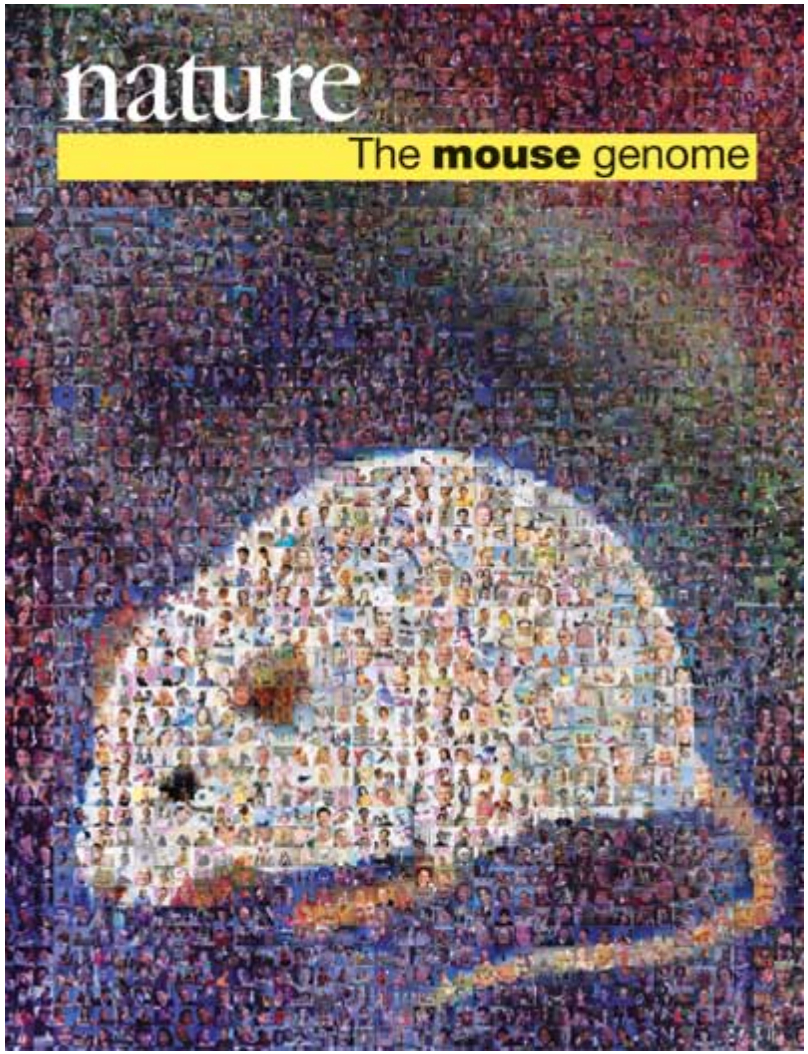
Human genome (Feb 2001)



Nature, 15th Feb 2001

Science, 15th Feb 2001

Mouse genome (Dec 2002)



- 2nd mammal genome
- model organism in lab
- 14% smaller than human
- 40% can be aligned to human
- 5% under purifying selection
- 0.5 substitutions per site, twice as many in the mouse lineage

Nature 420, 520-562

Rat genome (Apr 2004)

- 3rd mammal, 2nd rodent
- Similar number of genes in all 3 species
- 40% eutherian specific seq.
- 30% rodent specific seq., mostly repeats
- At least half of unaligned seq. is rat-specific repeats

Nature 428, 493-521

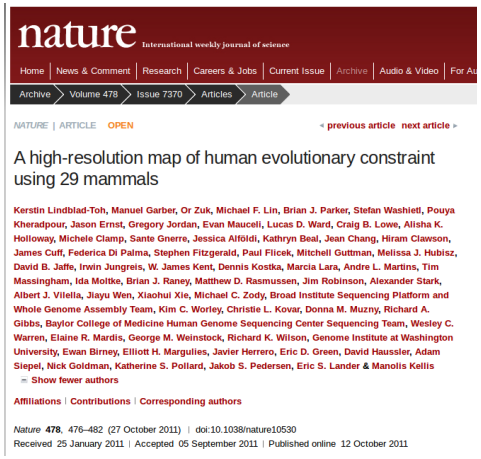
More and more genomes

2004

2005

2007

2008



2011

More than 50 vertebrate genomes have been “fully” sequenced

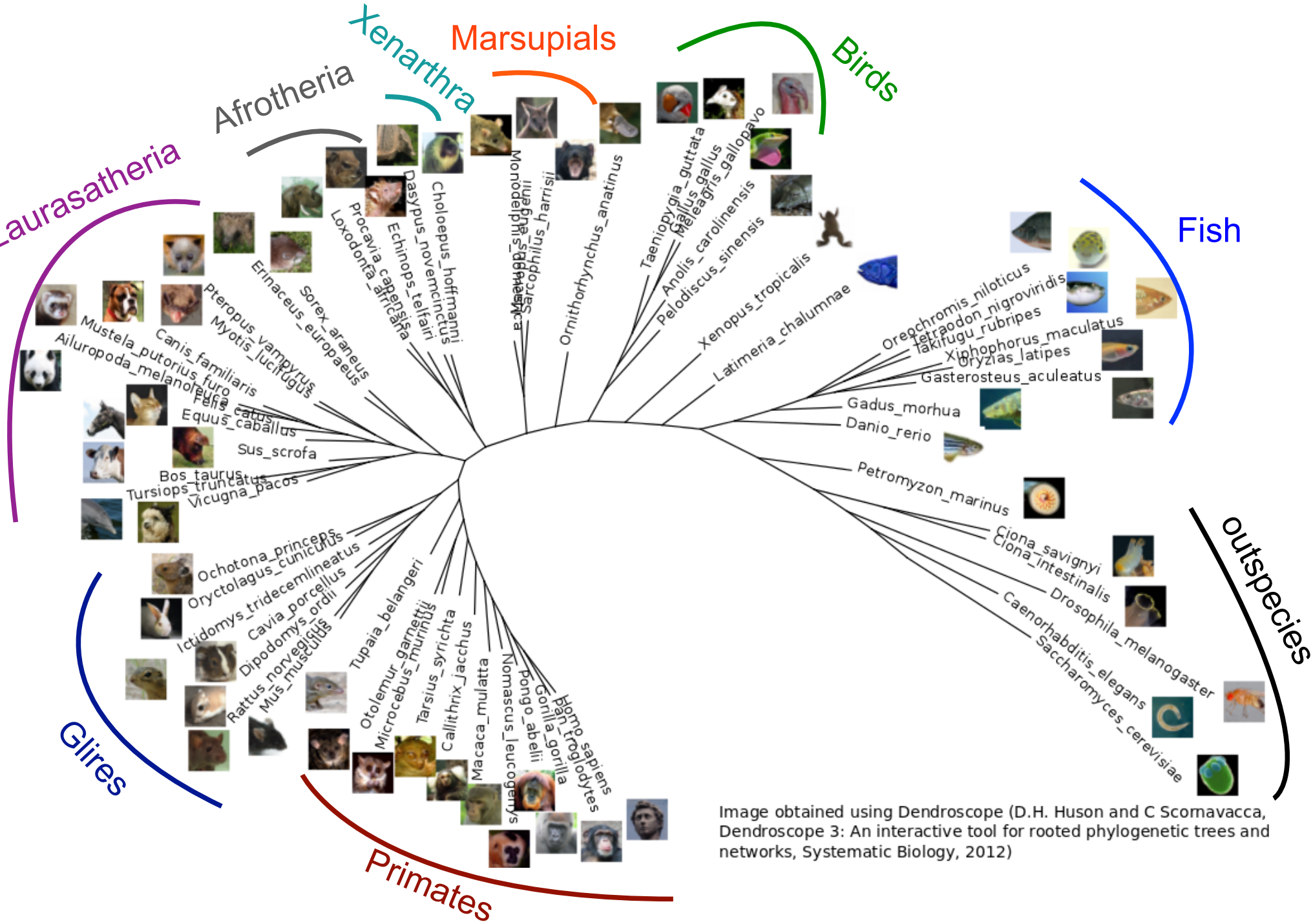
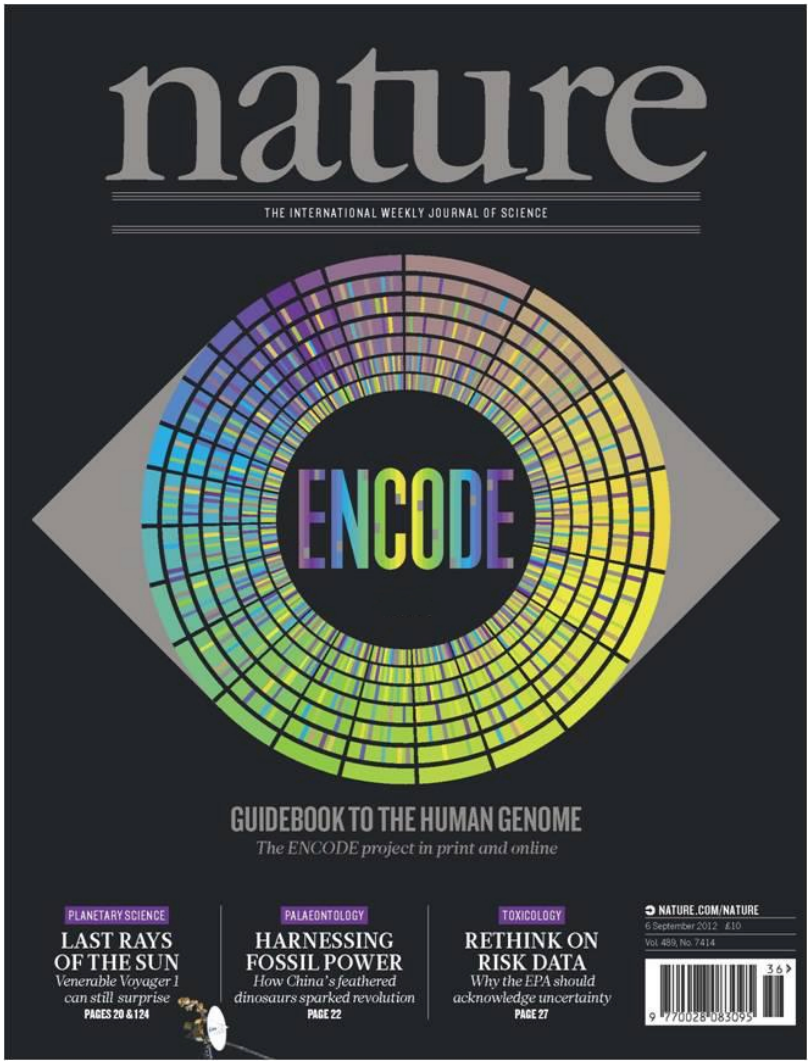


Image obtained using Dendroscope (D.H. Huson and C Scornavacca, Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks, Systematic Biology, 2012)

1000 genomes pilot project (Oct 2010)

- 3 pilots:
  - 179 ind. low-cov
  - 2 trios high-cov
  - 679 ind. exon only
- Covers >95% of variants of any individual
- Each individual carries 250-300 loss-of-function variants; 50-100 implicated in inherited disorders

ENCODE project: ENCyclopedia Of DNA Elements



- 1640 data sets
- 147 different cell types
- 400+ authors

- Activity in 80% of the genome
- Evidence of negative selection (in aggregate) in primate-specific elements
- Could classify the genome into 7 different chromatin states
- Transcription ↔ Histone + TF

Nature 489, 57–74 (06 Sept 2012)

Ensembl: What do you get?

Genome Annotation

- Protein coding gene structure
  - Consistent with genome, predicted across all vertebrates
  - Manual annotations (human, mouse, zebrafish, MHC)
- RNA genes (including miRNA)
  - Consistent with genome, predicted in across mammals
- Additional identifiers per genes (Xref)
  - Affymetrix, EntrezGene, Uniprot…

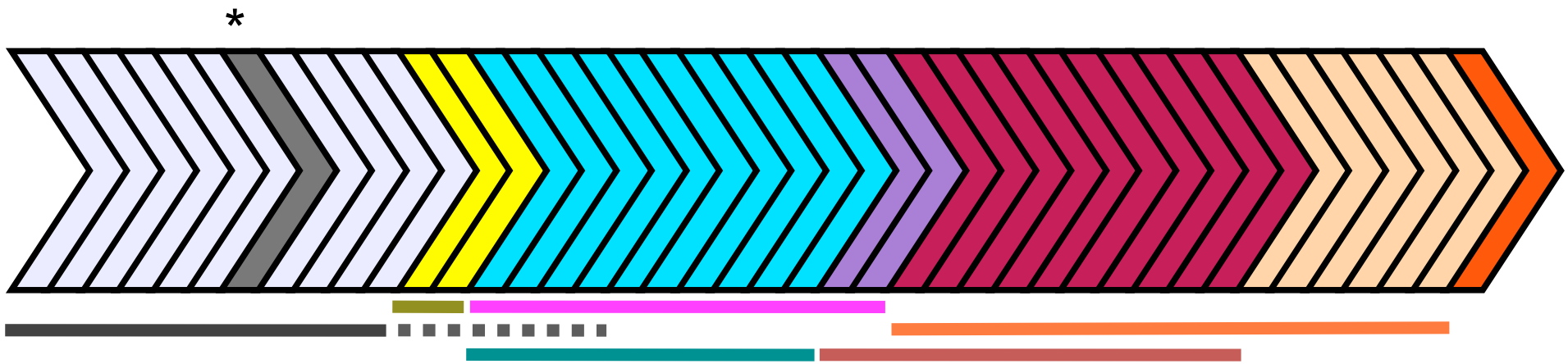
Variation, Comparative & Functional Genomics

- Genome alignments
  - Blastz, Blat, Pecan (multiple alignments), EPO
  - Homologues between genomes
  - Protein trees
- Variants (SNPs), CNVs, strains, genotypes
- ChIP-chip, ChIP-seq, segmentations

Infrastructure

- Website, Data mining tool, database and data dumps
- Portable, extendible, open source system with database, DAS, API, website, pipeline

Ensembl release cycle



Ensembl groups

- **GeneBuilders**: sequence masking, gene building
- **Core**: database schema, stable id mapping
- **Compara**: protein homology, genomic sequence alignments
- **Variation**: SNPs, CNVs, personal/strain genomes
- **Regulation**: probe mapping, functional data, segmentation
- **Web**: web site, new views for new data
- **Outreach**: help, workshops, tutorials
- **Production**: BioMart, coordination
- More people: Research, e! genomes, Zebrafish, Systems...

Acknowledgements

D48–D55 Nucleic Acids Research, 2013, Vol. 41, Database issue *Published online 30 November 2012*
doi:10.1093/nar/gks1236

Ensembl 2013

Paul Flicek[1,2,*], Ikhlak Ahmed[1], M. Ridwan Amode[2], Daniel Barrell[2], Kathryn Beal[1], Simon Brent[2], Denise Carvalho-Silva[1], Peter Clapham[2], Guy Coates[2], Susan Fairley[2], Stephen Fitzgerald[1], Laurent Gil[1], Carlos García-Girón[2], Leo Gordon[1], Thibaut Hourlier[2], Sarah Hunt[1], Thomas Juettemann[1], Andreas K. Kähäri[2], Stephen Keenan[1], Monika Komorowska[1], Eugene Kulesha[1], Ian Longden[1], Thomas Maurel[1], William M. McLaren[1], Matthieu Muffato[1], Rishi Nag[2], Bert Overduin[1], Miguel Pignatelli[1], Bethan Pritchard[2], Emily Pritchard[1], Harpreet Singh Riat[2], Graham R. S. Ritchie[1], Magali Ruffier[1], Michael Schuster[1], Daniel Sheppard[2], Daniel Sobral[1], Kieron Taylor[1], Anja Thormann[1], Stephen Trevanion[2], Simon White[2], Steven P. Wilder[1], Bronwen L. Aken[2], Ewan Birney[1], Fiona Cunningham[1], Ian Dunham[1], Jennifer Harrow[2], Javier Herrero[1], Tim J. P. Hubbard[2], Nathan Johnson[1], Rhoda Kinsella[1], Anne Parker[2], Giulietta Spudich[1], Andy Yates[1], Amonida Zadissa[2] and Stephen M. J. Searle[2]

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

wellcome trust

EMBL

National Human Genome Research Institute

BBSRC bioscience for the future

European Commission
Framework Programme 7

SEVENTH FRAMEWORK PROGRAMME

GEN2PHEN

BLUEPRINT epigenome

Quantomics
From Sequence to Consequence :
Tools for the Exploitation of Livestock Genomes