Genome Resources at the EBI - Ensembl and Ensembl Genomes

Bert Overduin, Ph.D.

wellcome trust sanger institute

EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

Outline

- Introduction to Ensembl / Ensembl Genomes
- Highlights in 2011
- Demo 1: Browser basics
- Demo 2: Variant Effect Predictor
- Demo 3: Adding custom tracks
- Demo 4: BioMart
- Future plans for 2012
- Help & Workshops
- Acknowledgements

Goal

To provide access to genome-scale data from completely sequenced species of scientific interest from across the taxonomy

History

- 1999: Start of Ensembl project for the Human Genome Project
- 2000: First release of data and web interface
- 2009: First release of Ensembl Genomes
- 2011: Ensembl v65: 63 genomes
- 2011: Ensembl Genomes v12: 335 genomes

- Ensembl: EBI & Wellcome Trust Sanger Institute
- Ensembl Genomes: EBI

Species Ensembl

Primates

Rodents etc.

Laurasiatheria

Afrotheria

Xenartha

Other mammals

Birds & reptiles

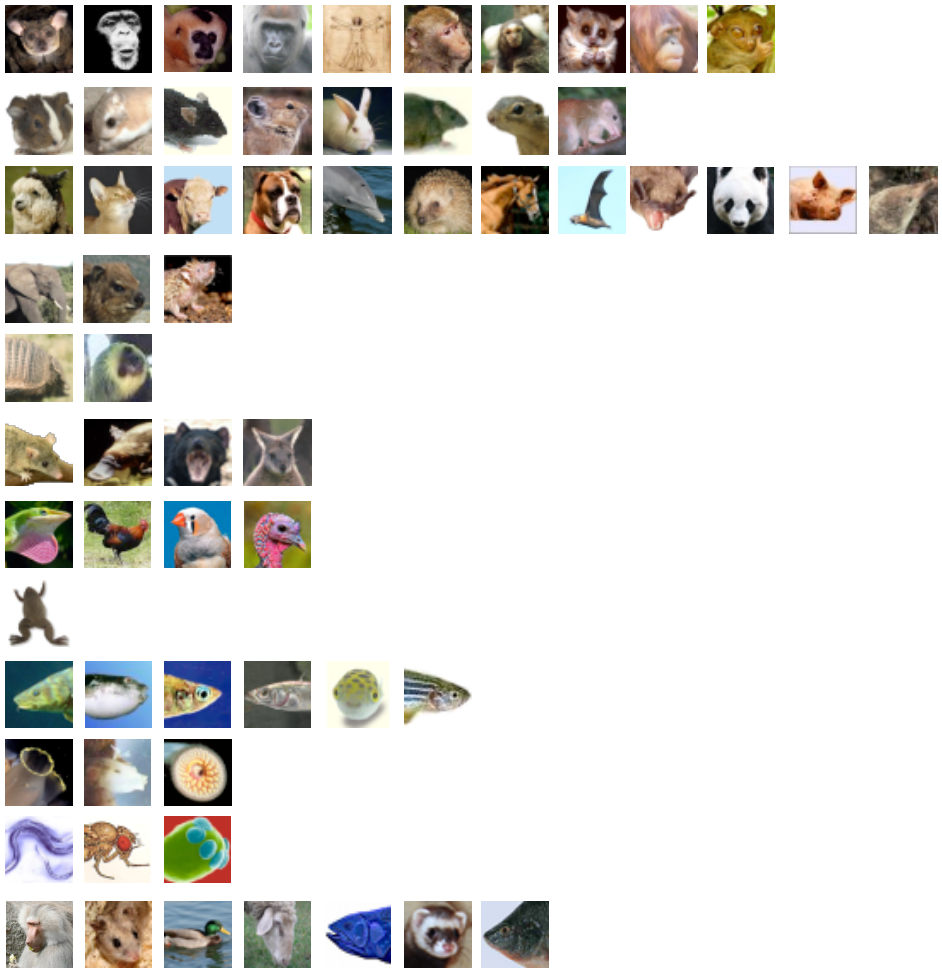
Amphibians

Fish

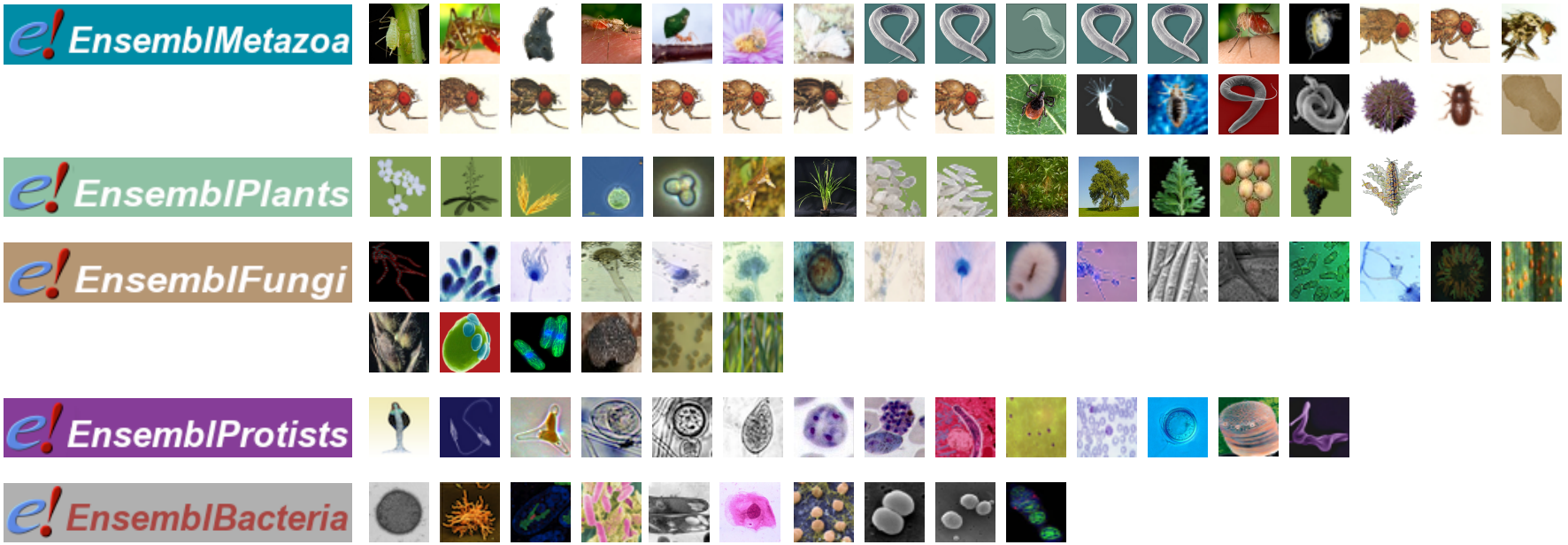
Other chordates

Other eukaryotes

On *Pre!* Ensembl



Species Ensembl Genomes



Annotation

- Inclusion of species depends on various criteria (model organism? community interest / demand? funding? completeness / quality of genome assembly?)
- A broad taxonomic coverage is aimed for

e!Ensembl

- Annotation in-house by the Ensembl team

e!EnsemblGenomes

- Annotation preferably by or in collaboration with the scientific community for the species in question

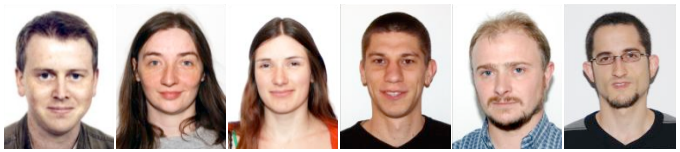
Ensembl genebuild

Genome assembly

+

Experimental evidence (cDNAs & proteins)

Genebuild pipeline



Ensembl Genes

Data

- Genomic sequence
- Gene/transcript/protein models
- External references
- Mapped cDNAs, proteins, microarray probes, BACs, cytogenetic bands, markers, repeats etc.
- Comparative data: orthologs and paralogs, protein families, whole genome alignments, syntenic regions
- Variation data: sequence variants, structural variants
- Regulatory data: “best guess” set of regulatory elements

Access to data

• Web browser	http://www.ensembl.org (with US West, US East and Asia mirrors and Pre! and Archive! sites) http://www.ensemblgenomes.org
• BioMart	http://www.biomart.org
• FTP	ftp.ensembl.org/pub ftp.ensemblgenomes.org/pub
• Public MySQL server	ensembldb.ensembl.org:5306:anonymous mysql.ebi.ac.uk:4157:anonymous
• Ensembl API	http://www.ensembl.org/info/docs/api

Highlights in 2011

Ensembl

- Genebuilds for turkey and cod
- Genebuild on new cow assembly (UMD 3.1)
- Added rabbit to whole-genome multiple alignments
- 3-way avian whole-genome alignment and constrained elements (chicken, turkey, zebra finch)
- Variation db for cat (dbSNP127)
- Updated variation data for cow (dbSNP133), dog (DGVa), pig (Illumina PorcineSNP60 Bead Chip, DGVa)
- Improved Variant Effect Predictor (VEP) and failed variation pipeline
- Sortable tracks, saving of configurations and configuration sets
- Support for large file formats (BAM, BigWig, VCF)

Highlights in 2011

e! EnsemblGenomes

- 31 new species
- Plants: *Chlamydomonas reinhardtii, Cyanidioschyzon merolae, Glycine max, Oryza glaberrima, Selaginella moellendorffii*
- Fungal plant pathogens: *Ashbya gossypii, Fusarium oxysporum, Gibberella moniliformis, Gibberella zeae, Mycosphaerella graminicola, Nectria haematococca, Phaeosphaeria nodorum, Puccinia triticina, Ustilago maydis*
- Oomycete plant pathogens: *Phytophthora infestans, Phytophthora ramorum, Phytophthora sojae, Pythium ultimum*
- Active collaborations within PhytoPath (http://www.phytopathdb.org/) and PomBase projects
- Variation db for *Arabidopsis thaliana* contains over 14 million variants from over 1600 strains

Demo 1 - Browser basics

Background:

The *CYN* gene encodes cyanate hydratase, an enzyme found in bacteria and plants that catalyses the reaction of cyanate with bicarbonate to produce ammonia and carbon dioxide:

$$NCO^- + HCO_3^- + 2H^+ = NH_3 + 2CO_2$$

Task:

Explore the *CYN* gene of *Vitis vinifera* (grape).

Variant Effect Predictor (VEP)

- Predicts functional consequences of variants on Ensembl genes
- Web interface, standalone Perl script and Perl API
- Accepts tab-delimited, VCF and pileup format as input

Demo 2 - Variant Effect Predictor

Background:

Variants in the bestrophin 1 (*BEST1*) gene are associated with various retinal disorders in man. Dog is used as a model to study these. The following are a number of new variants discovered in the *BEST1* gene of a Lapponian Herder:

chr	start	end	alleles	strand
18	57500034	57500034	A/G	+
18	57500028	57500028	G/T	+
18	57500027	57500027	G/T	+
18	57499959	57499958	-/C	+
18	57499929	57499929	G/T	+
18	57499981	57499981	G/T	+
18	57499834	57499834	A/T	+
18	57449754	57449754	C/T	+

Task:

Determine the effect of the variants on dog *BEST1*.

Adding custom tracks

- Upload data to Ensembl (5 MB size limit) or attach file on web-accessible server (http or ftp) to Ensembl (no size limit)
- Possible formats:

BAM	sequence alignments (no upload)
BED	genes / features
BedGraph	continuous-valued data
BigWig	continuous-valued data (no upload)
GBrowse	genes / features
GFF	genes / features
GTF	genes / features
PSL	sequence alignments
VCF	variants (no upload)
WIG	continuous-valued data

Demo 3 - Adding custom tracks

Background:

The file SRR070570.bam contains alignments of Illumina RNAseq reads from a wildtype *Arabidopsis thaliana* strain.

The bam file and its bam.bai index file are located at http://www.ebi.ac.uk/~bert/.

Task:

Attach SRR070570.bam to Ensembl Genomes.

Check the expression of a constitutive and a non-constitutive *Arabidopsis* gene, e.g. *RBCS1A* (ribulose bisphosphate carboxylase small chain 1A) and *PR1* (pathogenesis-related protein 1).

BioMart

- Data retrieval tool
- Originally developed for Ensembl (EnsMart)
- Now used by many large data resources
- Integrated with several widely used software packages
- Joint project between the European Bioinformatics Institute (EBI) and the Ontario Institute for Cancer Research (OICR)
- Website : http://www.biomart.org

Principle

- Step 1 – Dataset
  Choose your dataset

- Step 2 – Filters
  Limit your dataset

- Step 3 – Attributes
  Specify what information you want to output

- Step 4 – Results
  Preview and output your results

Demo 4 - BioMart

Background:

"Lactation" (GO:0007595) is the Gene Ontology (GO) term for the biological process of "the secretion of milk by the mammary gland".

Task:

Retrieve all cow genes that are annotated with the GO term "lactation".

Future plans for 2012

Ensembl

- Genebuilds for duck (?), salmon (?), sheep (?), tilapia
- Genebuilds on new assemblies for cat (Felis_catus-6.2), chicken (Gallus_gallus-4.0), dog (CanFam3.1), pig (Sscrofa10.2)
- Include RNAseq data in genebuild
- VEP support for structural variants
- New BLAST/BLAT interface
- http://www.ensembl.info/roadmap

Future plans for 2012

e! EnsemblGenomes

- New species: barley, *Brassica* (from BrassEnsembl), foxtail millet, *Oryza brachyanta*, potato, tomato, *Gaeumannomyces graminis*, *Magnaporthe oryzae*, *Magnaporthe poae*, tsetse fly
- New assemblies: maize (B73_RefGen_v3), *Oryza sativa* ssp. japonica cu. Nipponbare (Os-Nipponbare-Reference-IRGSP-1.0; IRGSP1.0), poplar
- Variation db and new gene annotation for wheat stem rust pathogen
- New query interface for data re plant-fungal pathogen interactions (PhytoPath; http://www.phytopathdb.org/)
- Widened development of community annotation pipelines

Help

- Helpdesk:

  helpdesk@ensembl.org
  helpdesk@ensemblgenomes.org

- Mailing lists:

  http://www.ensembl.org/info/about/contact/mailing.html
  http://plants.ensembl.org/info/about/contact/mailing.html

- Ensembl YouTube and YouKu (优酷网) channels:

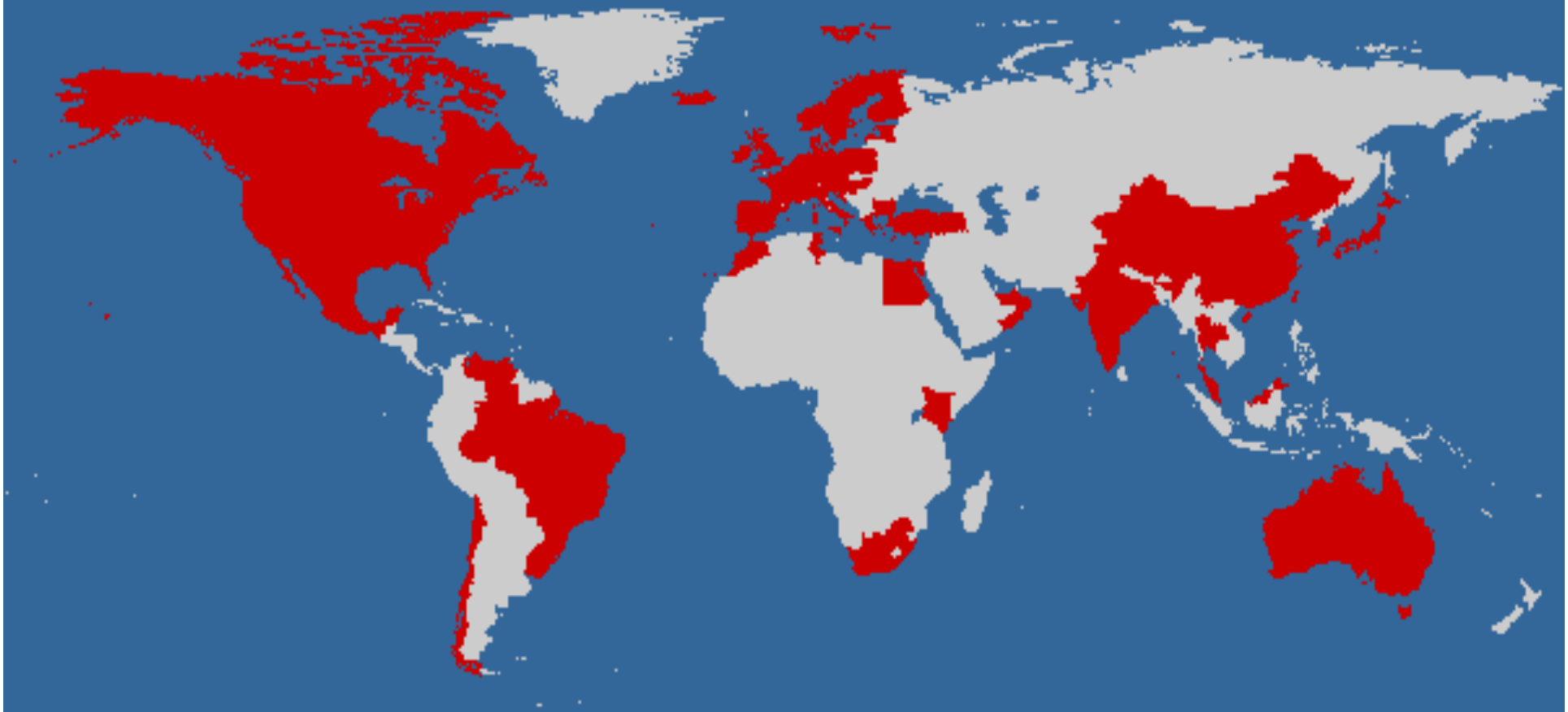
  http://www.youtube.com/user/EnsemblHelpdesk
  http://u.youku.com/user_show/uid_Ensemblhelpdesk

EBI Train online

Ensembl: Browsing chordate genomes	
Description	This course focuses on *chordate* (mostly vertebrate) genomes on the *Ensembl* website at www.ensembl.org. It provides a quick beginner's guide to the overall structure of the Ensembl genome browser.
Topic	Genes and Genomes
Data resources used	Ensembl
Level	Beginner
Duration	3hours
Target Audience	Bioinformaticians; Biologists; Medical researchers; Molecular biologists
Background knowledge required	A knowledge of some genomics is required, and some bioinformatics knowledge would be useful but is not essential. For more information on how to complete the courses in Train online please see 'About the courses'.
Author	Giulietta Spudich

http://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes

Workshops



until now:
49 countries on 5 continents

in 2011:
~ 90 workshops

Workshops

- Browser (0.5-2 days) and API (1-3 days) workshops
- Combination of lectures and hands-on exercises
- Advertised on http://www.ensembl.info/workshops/calendar/

- You can host your own workshop!
- For academic institutions there is, apart from the instructor's expenses, no fee
- You only need a computer room and participants
- You can get more info from me (bert@ebi.ac.uk) or at the EBI booth (302)

Stay in touch

- Blog:

  http://www.ensembl.info

- Facebook:

  http://www.facebook.com/Ensembl.org

- Twitter:

  http://twitter.com/Ensembl

Acknowledgements

Ensembl

- WTSI

- Wellcome Trust
- NIH-NHGRI
- EMBL
- EU

EnsemblGenomes

- CADRE
- Gramene
- VectorBase
- WormBase
- PomBase

- EMBL
- BBSRC
- Wellcome Trust
- Bill and Melinda Gates Foundation
- EU

biomart

- OICR

Acknowledgements

Paul Flicek, Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Clapham P, Guy Coates, Susan Fairley, Stephen Fitzgerald, Laurent Gil, Leo Gordon, Maurice Hendrix, Thibaut Hourlier, Nathan Johnson, Andreas Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Monika Komorowska, Gautier Koscielny, Eugene Kulesha, Pontus Larsson, Ian Longden, Will McLaren, Matthieu Muffato, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Harpreet Riat, Graham Ritchie, Magali Ruffier, Michael Schuster, Daniel Sobral, Amy Tang, Kieron Taylor, Stephen Trevanion, Jana Vandrovcova, Simon White, Mark Wilson, Steven Wilder, Bronwen Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé Fernández-Suarez, Jennifer Harrow, Javier Herrero, Tim Hubbard, Anne Parker, Glenn Proctor, Giulietta Spudich, Jan Vogel, Andy Yates, Amonida Zadissa, Steve Searle

Paul Kersey, Dan Staines, Dan Lawson, Eugene Kulesha, Paul Derwent, Jay Humphrey, Daniel Hughes, Stephen Keenan, Arnaud Kerhornou, Gautier Koscielny, Nick Langridge, Mark McDowall , Karyn Megy, Uma Maheswari, Michael Nuhn, Michael Paulini, Helder Pedro, Iliana Toneva, Derek Wilson, Andy Yates, Ewan Birney

Posters

P941
Genome Annotation in Ensembl
Susan Fairley

P942
Ensembl Plants: An Integrating Resource for Plant Genomics and Variation
Paul Kersey

Training courses

Careers

Meet the experts

Brochures and factsheets

Come and see us at booth 302!

PhD and post doc opportunities

Industry programme

Research and services

Visitor's programme

EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

PDF of this presentation

http://www.ebi.ac.uk/~bert/past_workshops.html