The European Nucleotide Archive (ENA)

Bert Overduin, Ph.D.
Outline

• Introduction
• Highlights in 2011
• Demo 1: EMBL-Bank
• Demo 2: SRA
• Future plans for 2012
• Help
• Acknowledgements
Goal

To provide a comprehensive record of the world’s nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation
History

• 1980: EMBL Data Library (EMBL Heidelberg, Germany)
  • World’s first public database of nucleotide sequences
• 1995: EMBL-Bank (EBI Hinxton, UK)
• 2008: Trace Archive
  • Capillary Sequencing reads
• 2008: Sequence (formerly: Short) Read Archive (SRA)
  • Next Generation Sequencing reads (454, Solexa/Illumina, SOLiD etc.)
INSDC

- International Nucleotide Sequence Database Collaboration
- Consists of ENA, NCBI GenBank and DNA Data Bank of Japan
- Databases are synchronized on a daily basis
- [http://www.insdc.org](http://www.insdc.org)
Data architecture

ENA-Annotation: Feature annotation

ENA-Assembly: Assembly information

ENA-Reads: Sequencing and sampling information

EMBL-Bank

SRA
Content

EMBL-Bank Growth
09-Jan-2012

Sequence Read Archive (SRA) Growth
09-Jan-2012

--- Sequences (238.2 millions) --- Bases (393.6 billions)

--- Sequences (1.6 trillions) --- Bases (169.7 trillions)
What can you do with ENA?

- Permanently archive your sequence data and disseminate them to the global research community
- Share your pre-publication data with collaborators
- Reduce your local hardware requirements for archiving NGS data
- Report novel annotation relating to existing sequence data
- Locate, retrieve and aggregate existing sequence data for analysis and meta-analysis
- Browse existing sequence and annotation referred to in literature
- Find all sequences and annotation available for your gene of interest
- Find out what is known about your sequence of interest
- Link through from nucleotide data to a host of integrated resources
Submitting data

- Many journals and funders require authors to submit their sequence to an INSDC database prior to publication
- You should only submit to one INSDC database (ENA, GenBank or DDBJ)
- Unique accession numbers are assigned to all submitted data
- Submitted data can be made public immediately or kept private until the associated work has been published
- Once public, submitted data will be exchanged with NCBI and DDBJ
- Data belong to the submitter and can only be updated with submitter consent
Submitting data

• Preferred: Webin interactive web submission system
• Other tools for e.g. genome projects and large sequencing centers
• http://www.ebi.ac.uk/ena/about/submit_and_update
Retrieving data

- ENA Browser
  - Free text search: ENA homepage, EB-eye
  - Sequence similarity search: ENA homepage, ENA Sequence Search
  - Programmatic data access using REST URLs
  - Formats: FASTA, FASTQ, flat file, HTML, XML

- Bulk data download: using FTP or Aspera

- [http://www.ebi.ac.uk/ena/about/search_and_browse](http://www.ebi.ac.uk/ena/about/search_and_browse)
Highlights in 2011

• SRA Webin release and extensive improvement with upload tools and other new functionality
• EMBL-Bank Webin new templates
• Genome collections design and prototyping
• Darwin Core presentation of taxonomy
• Publication of reference-based compression proof of principle (Genome Res. 2011 May;21(5):734-40.)
• Release of several version of CRAM toolkit for reference-based compression
• NGS data from SRA available as data source in Galaxy (http://main.g2 bx.psu.edu/)
Retrieve and browse the mitochondrial genome of the cave bear \textit{(Ursus spelaeus)}.
Demo 2 - SRA

Whole-genome resequencing reveals loci under selection during chicken domestication.


Department of Medical Biochemistry and Microbiology, Uppsala University, Box 582, SE-75123 Uppsala, Sweden.

Author Information The SOLiD reads have been submitted to the NCBI Sequence Read Archive under the study accession number SRP001870. SNPs reported in this

Retrieve and browse the (meta) data for this study.
Future plans for 2012

- Project and genome collections support in Webin
- Integration of EMBL-Bank and SRA Webin into a single system
- Genome collections
- Taxonomy services
- Further integration with other databases through enhanced cross-referencing
- Full launch of ENA Sequence Search, web and SOAP services
- Improved data consumer interfaces and services
- [http://www.ebi.ac.uk/ena/about/forthcoming_changes](http://www.ebi.ac.uk/ena/about/forthcoming_changes)
Help

• Data submissions, helpdesk, enquiries:
  datasubs@ebi.ac.uk

• Updates, publication notifications:
  update@ebi.ac.uk
### EBI Train online

<table>
<thead>
<tr>
<th><strong>The European Nucleotide Archive: Quick tour</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td><strong>Topic</strong></td>
</tr>
<tr>
<td><strong>Data resources used</strong></td>
</tr>
<tr>
<td><strong>Level</strong></td>
</tr>
<tr>
<td><strong>Duration</strong></td>
</tr>
<tr>
<td><strong>Target Audience</strong></td>
</tr>
<tr>
<td><strong>Background knowledge required</strong></td>
</tr>
<tr>
<td><strong>Author</strong></td>
</tr>
</tbody>
</table>

http://www.ebi.ac.uk/training/online/course/european-nucleotide-archive-quick-tour
Acknowledgements


EMBL, EU, Wellcome Trust, BBSRC
EBI is an Outstation of the European Molecular Biology Laboratory.

Come and see us at booth 302!

Training courses
Meet the experts
PhD and post doc opportunities
Visitor’s programme
Research and services

Careers
Brochures and factsheets
Industry programme

EBI is an Outstation of the European Molecular Biology Laboratory.
ENA User Survey 2011


PDF of this presentation

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