The European Nucleotide Archive (ENA)

Bert Overduin, Ph.D.

EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

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

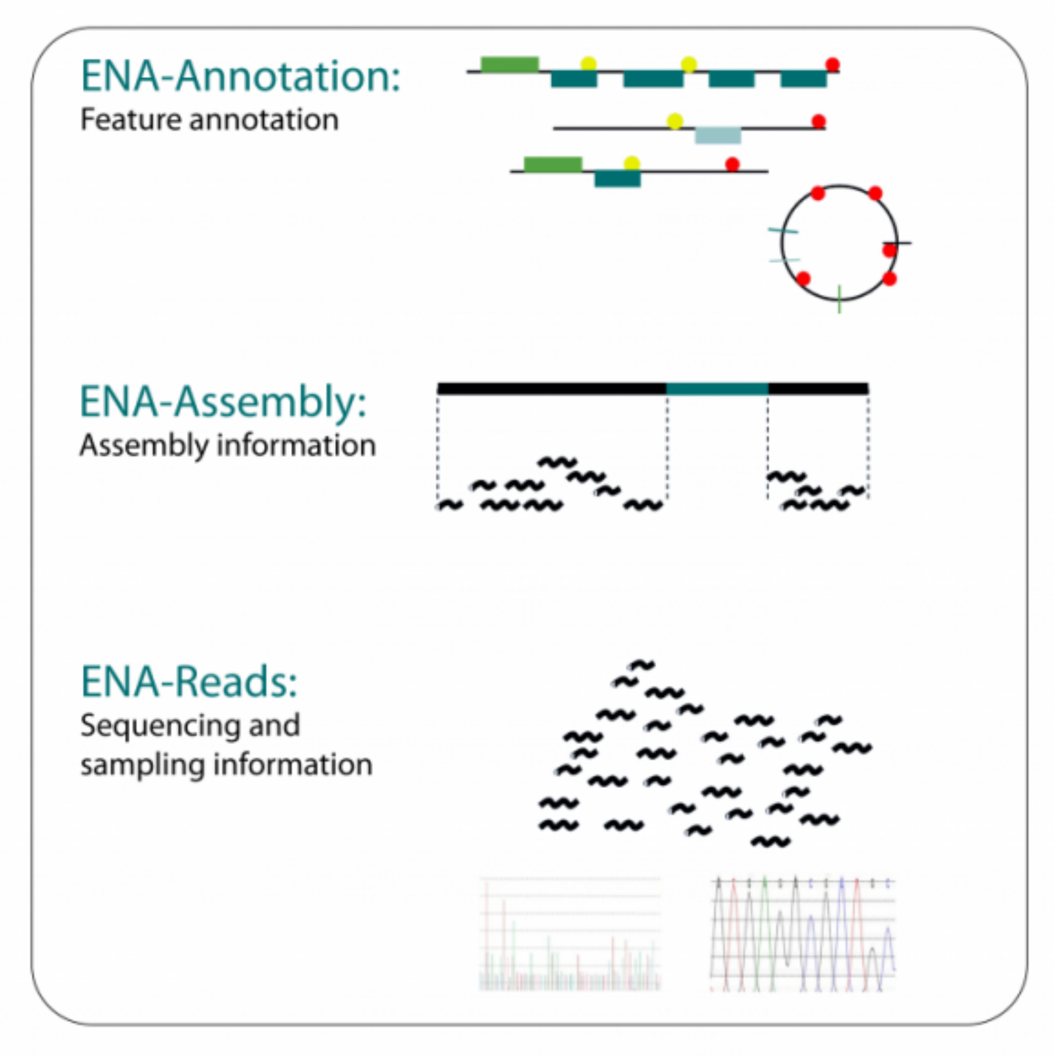
- International Nucleotide Sequence Database Collaboration
- Consists of ENA, NCBI GenBank and DNA Data Bank of Japan
- Databases are synchronized on a daily basis
- http://www.insdc.org

ENA
European Nucleotide Archive

NCBI

DDBJ
DNA Data Bank of Japan

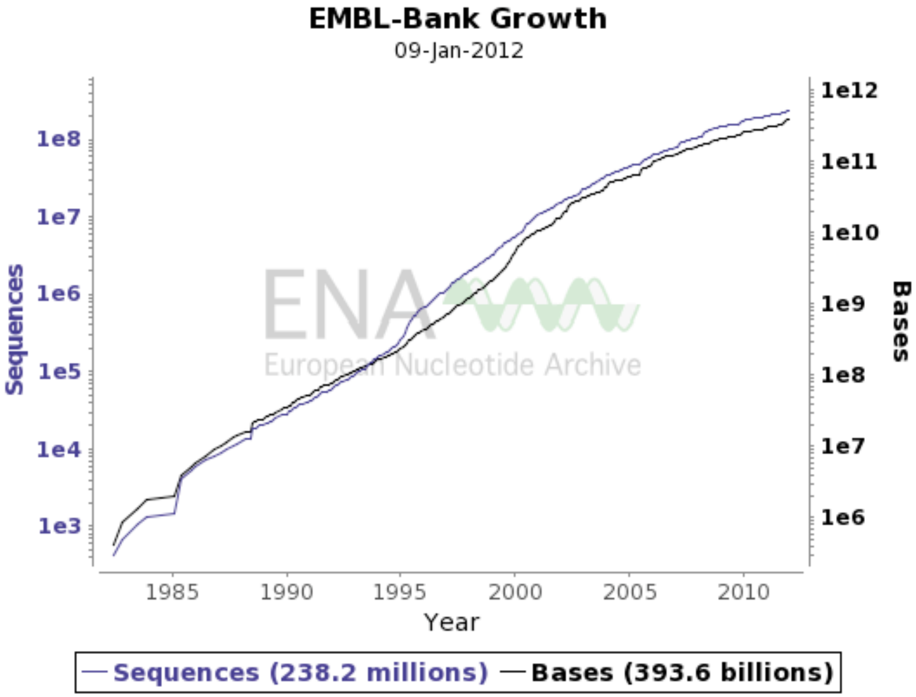
Data architecture

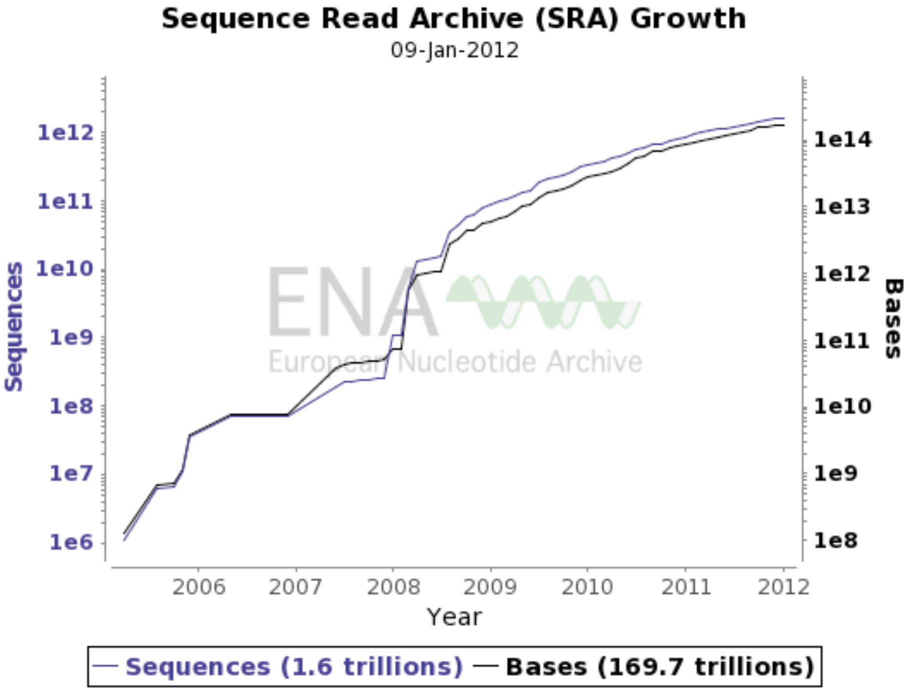


EMBL-Bank

SRA

Content


EMBL-Bank Growth
09-Jan-2012
Sequences
1e3
1e4
1e5
1e6
1e7
1e8
Bases
1e6
1e7
1e8
1e9
1e10
1e11
1e12
1985
1990
1995
2000
2005
2010
Year
ENA
European Nucleotide Archive
— Sequences (238.2 millions) — Bases (393.6 billions)


Sequence Read Archive (SRA) Growth
09-Jan-2012
Sequences
1e6
1e7
1e8
1e9
1e10
1e11
1e12
Bases
1e8
1e9
1e10
1e11
1e12
1e13
1e14
2006
2007
2008
2009
2010
2011
2012
Year
ENA
European Nucleotide Archive
— 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

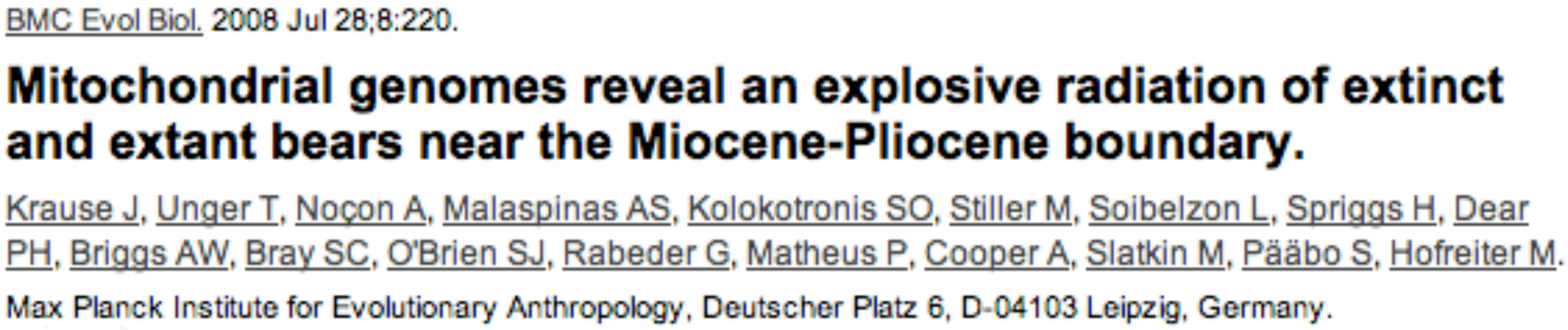
Retrieving data

- 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

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/)

Demo 1 - EMBL-Bank

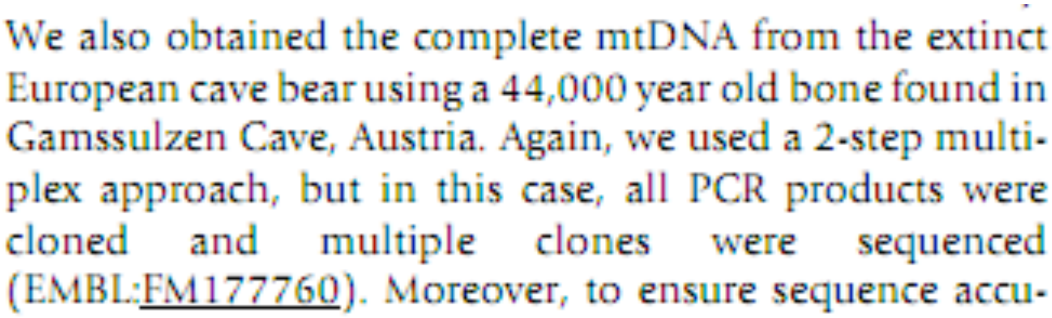


BMC Evol Biol. 2008 Jul 28;8:220.

Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary.

Krause J, Unger T, Noçon A, Malaspinas AS, Kolokotronis SO, Stiller M, Soibelzon L, Spriggs H, Dear PH, Briggs AW, Bray SC, O'Brien SJ, Rabeder G, Matheus P, Cooper A, Slatkin M, Pääbo S, Hofreiter M.

Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany.

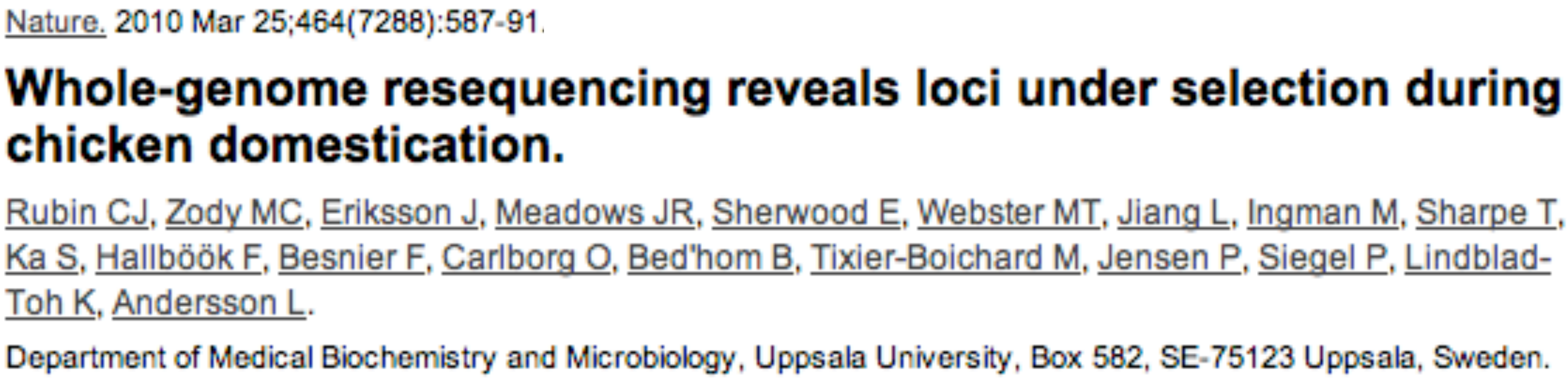


We also obtained the complete mtDNA from the extinct European cave bear using a 44,000 year old bone found in Gamssulzen Cave, Austria. Again, we used a 2-step multiplex approach, but in this case, all PCR products were cloned and multiple clones were sequenced (EMBL:FM177760). Moreover, to ensure sequence accu-

Retrieve and browse the mitochondrial genome of the cave bear (*Ursus spelaeus*).

© Mo Hassan

Demo 2 - SRA

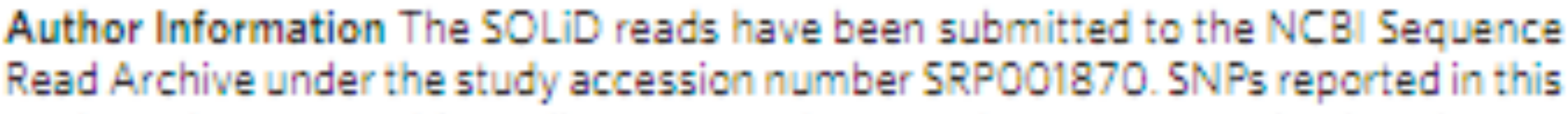


Nature. 2010 Mar 25;464(7288):587-91.

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

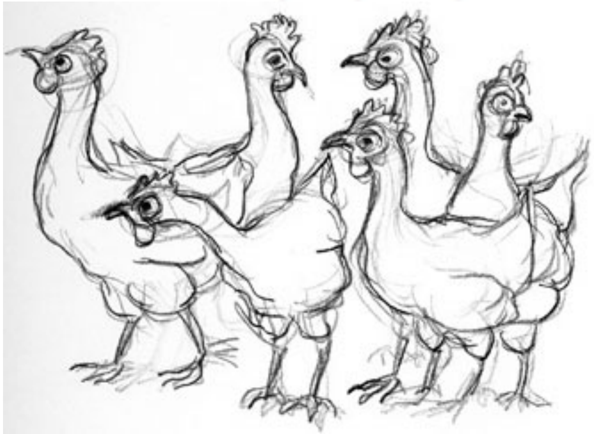
Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S, Hallböök F, Besnier F, Carlborg O, Bed'hom B, Tixier-Boichard M, Jensen P, Siegel P, Lindblad-Toh K, Andersson L.

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.



© Robin Hoffman

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

Help

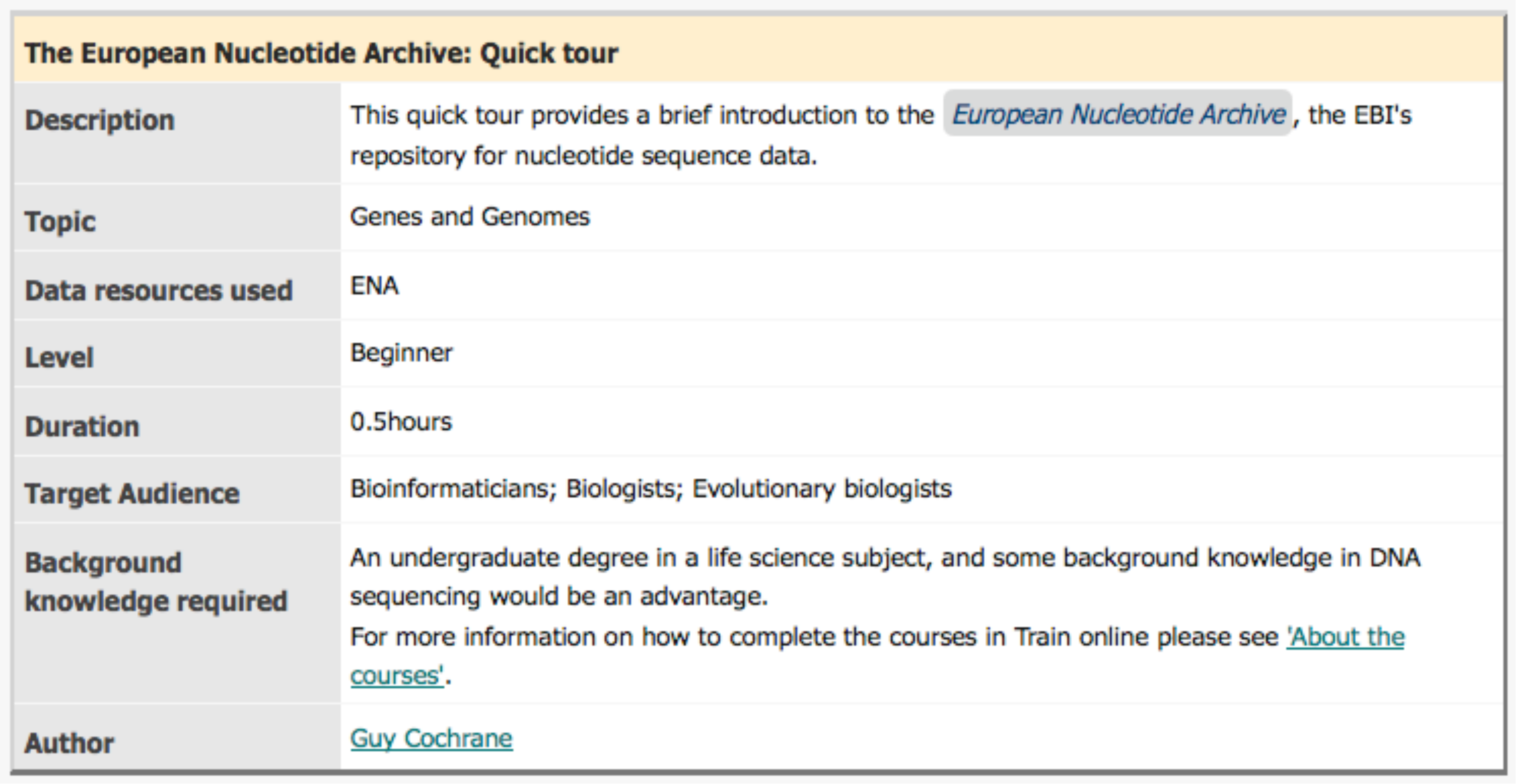
- Data submissions, helpdesk, enquiries:

  datasubs@ebi.ac.uk

- Updates, publication notifications:

  update@ebi.ac.uk

EBI Train online



The European Nucleotide Archive: Quick tour

Description	This quick tour provides a brief introduction to the *European Nucleotide Archive*, the EBI's repository for nucleotide sequence data.
Topic	Genes and Genomes
Data resources used	ENA
Level	Beginner
Duration	0.5hours
Target Audience	Bioinformaticians; Biologists; Evolutionary biologists
Background knowledge required	An undergraduate degree in a life science subject, and some background knowledge in DNA sequencing would be an advantage. For more information on how to complete the courses in Train online please see 'About the courses'.
Author	Guy Cochrane

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

Acknowledgements

Clara Amid, Ewan Birney, Lawrence Bower, Ana Cerdeño-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Richard Gibson, Neil Goodgame, Christopher Hunter, Mikyung Jang, Rasko Leinonen, Xin Liu, Arnaud Oisel, Nima Pakseresht, Sheila Plaister, Rajesh Radhakrishnan, Kethi Reddy, Stephane Rivière, Marc Rossello, Alexander Senf, Dimitriy Smirnov, Petra Ten Hoopen, Daniel Vaughan, Robert Vaughan, Vadim Zalunin and Guy Cochrane

EMBL, EU, Wellcome Trust, BBSRC

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.

ENA User Survey 2011

http://www.surveymonkey.com/s/ENA_User_Survey_2011

PDF of this presentation

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