The International Nucleotide Sequence Database Collaboration

Guy Cochrane, EMBL – European Bioinformatics Institute
Nucleotide sequencing: the ubiquitous method

Credits:
Applied Biosystems
Oxford Nanopore Technologies
https://universe-review.ca/R11-16-DNAsequencing.htm
History

EMBL Data Library & GenBank collaborate

DNA Databank of Japan (DDBJ) joins collaboration

First mandated submissions

Expansion to cover next generation platforms

Continuation as foundational data resource between partners:
- European Nucleotide Archive (ENA) at EMBL-EBI
- DDBJ at NIG
- GenBank and SRA at NCBI
Scope

- regular data exchange
- data standards
- open and unrestricted access

- globally comprehensive coverage
- scientific database of record
- public forum for the scientific process

INSDC

International Nucleotide Sequence Database Collaboration
Collaborative instruments

- Feature Table Definitions ([http://www.insdc.org/documents/feature-table](http://www.insdc.org/documents/feature-table))
- Unified accessioning system
- Data model for raw data
- Status conventions
- Various controlled vocabularies
Scale

• Data volume
  • Several petabytes on disk
  • 1.3 petabase pairs
  • >1 million taxa
  • 1 submission every 6 minutes

• Scale by record type
  • 800,000,000 assembled/annotated sequences
  • 87,000 assemblies
  • 230,000,000 coding genes
  • 760,000 non-coding genes

• Literature references
  • 327,000 total publications
  • 57,000 ‘re-use’ publications

• Growth
  • Doubling times as low as few months for raw data

• Counting minimally, 60 staff
Governance & advisory structure

International INSDC Advisory Committee

DDBJ Advisors
- Sumio Sugano, Institute of Medical Science, The University of Tokyo
- Ken Kurokawa, Earth-Life Science Institute, Tokyo Institute of Technology
- Kaoru Fukami-Kobayashi, Bioresource Information Division, RIKEN BioResource Center

EMBL Advisors
- Antoine Danchin, AMAbiotics SAS, Evry, France and Scientific Advisor to the CEA, France
- Babas Savakis, University of Crete and IMBB-FORTH, Heraklion
- Jean Weissenbach, Genoscope, Evry
- Mark Baxter, GenPool Genomics Facility and Institute of Evolutionary Biology, University of Edinburgh

GenBank Advisors
- Steven Salzberg, Johns Hopkins University School of Medicine, Baltimore, MD, US
- Rich Roberts, New England Biolabs, Beverly, MA, US
- Chung-I Wu, University of Chicago, US and Beijing Institute of Genomics, China
Services: submissions

Start >> Sample >> Finish

Please create new samples by uploading a spreadsheet or by following the instructions below.

1. Please select the checklist that you wish to use for your sample submission.
2. If you already have a spreadsheet containing your data, upload it here.

(Example views from ENA services)
Services: data discovery

temperature>=10 AND temperature<=25 AND geo_box1(42, 17, 43, 18)

tax_tree(10090) AND library_source="GENOMIC" AND instrument_platform="ILLUMINA" AND library_strategy="ChIP-Seq"

(Application views from ENA services)
Reach & impact

• Mandatory submissions in all major journals that publish sequence-based science
• Mature infrastructure - ‘post-citation’ phase
• INSDC accessions referenced routinely in cases of data reuse
• Data providers
  • 2,000-5,000 active data submitters
  • 11,000 ‘centres’
• Data consumers
  • Direct users: scale of 10s-100s of thousands per month
  • Many more via secondary resources, data mirrors, etc.
ENA outward interactions
Funding

• INSDC partner institutions are independent with no shared funding
  • Local context defines technical and other operational configurations
  • Provides resilience and neutrality
• DDBJ and DRA
  • Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) via a management Expense grant for Inter-University Research Institute Corporation to DDBJ
• NCBI GenBank and SRA
  • GenBank and SRA are funded under the Intramural Research Program of the National Institutes of Health, National Library of Medicine
• EMBL-EBI ENA
  • Supported by European Molecular Biology Laboratory and a variety of funding agencies, including the European Commission; the UK Biotechnology and Biological Sciences Research Council; the Wellcome Trust; the Gordon and Betty Moore Foundation
Layers of funding

- **Operation**
  - Technical and service delivery
    - hardware, network, databases, network, service desks, training
  - Development required to ‘stand still’ in the face of growth
    - data compression, performant search systems, data exchange
- **Innovation**
  - New functions, data types
    - new platforms, new scientific applications, new taxonomic groups
  - New communities and projects
    - standards groups, data coordination, curation, model organism databases
Acknowledgements

• Ilene Karsch-Mizrachi, NCBI
• Yaskaz Nakamura, DDBJ
• Members of the INSDC teams across partners institutes
• INSDC International Advisory Committee