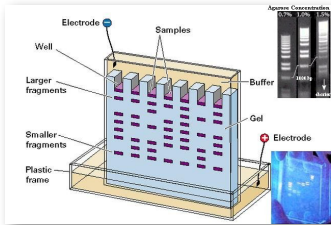
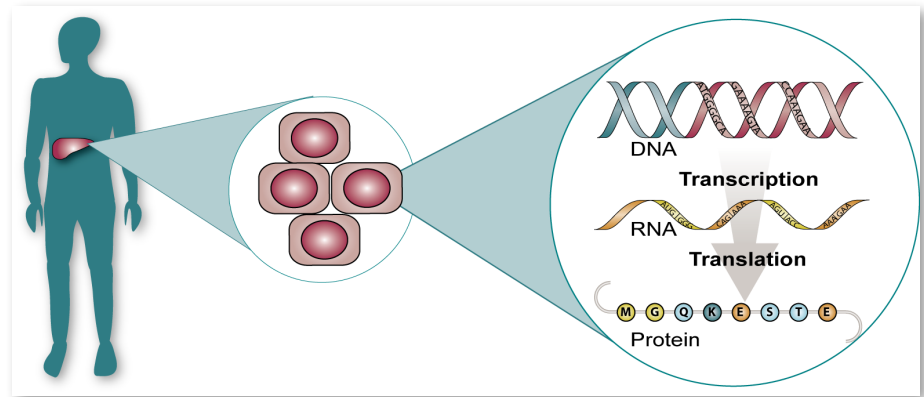


# The International Nucleotide Sequence Database Collaboration

Guy Cochrane, EMBL – European Bioinformatics Institute



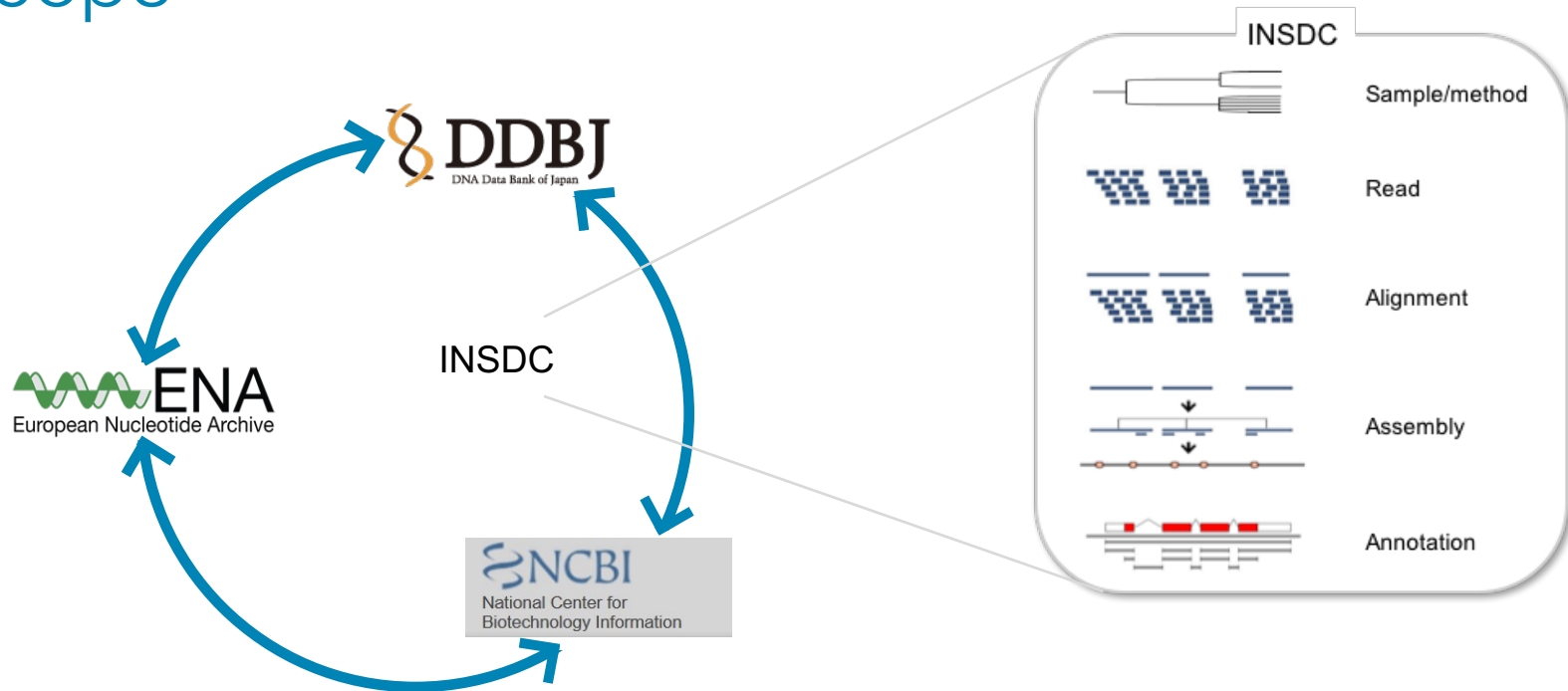
# Nucleotide sequencing: the ubiquitous method



A central collage of images and text labels representing various fields of study where nucleotide sequencing is applied. The labels are highlighted in yellow boxes and include: **Genomics**, **Epigenomics**, **Metagenomics**, **Clinical medicine**, **Medical research**, **Transcriptomics**, **Variation**, **Evolutionary & population biology**, **Agriculture**, **Forensics**, **Genetic counselling**, **Biotechnology**, **Border control**, **Biodiversity & taxonomy**, **Environmental sciences**, **Ecology**, and **Food science**. Each label is accompanied by a small image related to that field.



# Scope



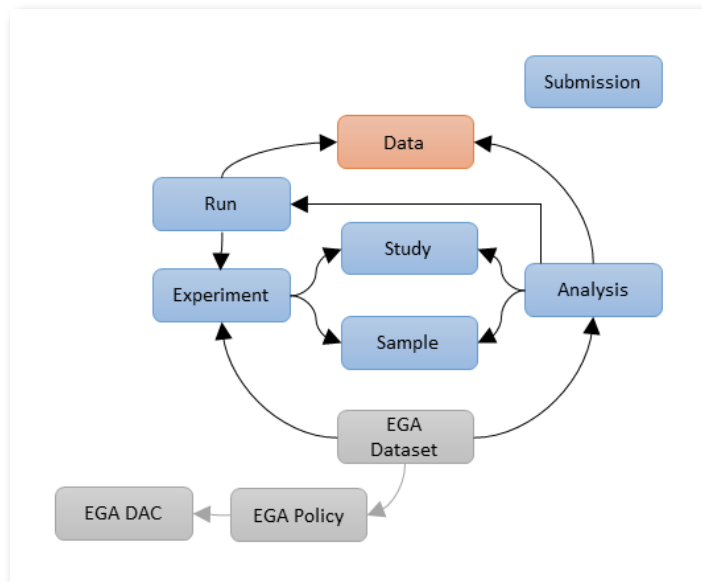
- regular data exchange
- data standards
- open and unrestricted access
- globally comprehensive coverage
- scientific database of record
- public forum for the scientific process



# Collaborative instruments

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

Feature	gene
<b>Definition</b>	region of biological interest identified as a gene and for which a name has been assigned;
<b>Optional Qualifiers</b>	<pre> allele="text" /citation=[number] /db_xref="&lt;database&gt;&lt;identifier&gt;" /experiment=" CATEGORY: text" /function="text" /gene="text" /gene_synonym="text" /inference=" CATEGORY: TYPE (same species) :EVIDENCE_BASIS" /locus_tag="text" (single token) /map="text" /nc="text" /old_locus_tag="text" (single token) /opereon="text" /product="text" /pseudo /pseudogene="TYPE" /phenotype="text" /standard_name="text" /trans_splicing                     </pre>
<b>Comments</b>	the gene feature describes the interval of DNA that corresponds to a genetic trait or phenotype; the feature is, by definition, not strictly bound to its positions at the ends; it is meant to represent a region where the gene is located.
<b>Last Updated</b>	Wed Nov 11 15:36:03 2015
<b>Qualifier</b>	product
<b>Definition</b>	name of the product associated with the feature, e.g. the mRNA of an mRNA feature, the polypeptide of a CDS, the peptide of a mat_peptide, etc.
<b>Value Format</b>	"text"
<b>Example</b>	<pre> /product="trypsinogen" (when qualifier appears in CDS feature) /product="trypsin" (when qualifier appears in mat_peptide feature) /product="XYZ neural-specific transcript" (when qualifier appears in mRNA feature)                     </pre>



INSDC Status Document | INSDC

[www.insdc.org/documents/insdc-status-document](http://www.insdc.org/documents/insdc-status-document)

INSDC International Nucleotide Sequence Database Collaboration

ABOUT INSDC POLICY ADVISORS DOCUMENTS

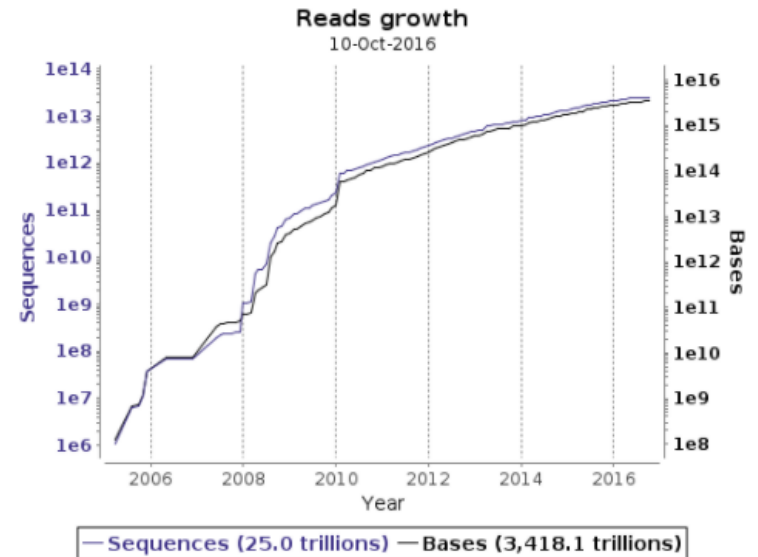
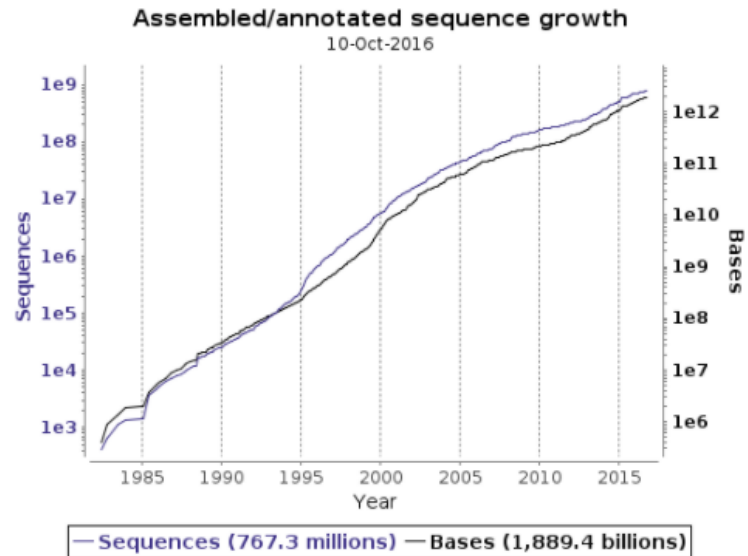
INSDC Status Document

Status name	Causes	Implications
<b>Public</b>	Data are submitted with no request for confidential hold prior to publication or have reached an owner-agreed public release date.	Data are fully available.
<b>Confidential</b>	Data owner requires and indicates to INSDC staff that confidentiality is required until a release date or publication in the literature, whichever comes earlier.	Data are not available publicly through any means. A data release date is recorded for the data, which are subsequently and automatically released as <b>Public</b> on reaching this date or being cited in a publication prior to this date. In the event that a release date must be extended, data owners are required to contact the INSDC partner responsible for the submission with sufficient notice*.
<b>Suppressed</b>	(1) Data are found by the owner to be incorrectly annotated or contaminated with no immediate opportunity on the part of the owner to be updated.  (2) Data owners realise after sequences have been released that they failed to request a confidential status, either at the time of submission, or within the period between completion of submission processing and the date on which the submission is normally made available to the public (this time period can vary among the INSDC members).	Data are removed where possible from direct search tools (such as text and sequence similarity search) but remain available by accession number.
<b>Replaced</b>	Data owners generate new data under new accession identifiers that directly replace existing data; this is expected to be rare since	Data are removed where possible from direct search tools (such as text and sequence similarity search) but remain available by accession number. Where possible,



# 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

www.insdc.org/advisors

ABOUT INSDC POLICY ADVISORY COMMITTEE



**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
- Babis Savakis, University of Crete and IMBB-FORTH, Heraklion
- Jean Weissenbach, Genoscope, Evry
- Mark Blaxter, GenePool 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

**INSDC**  
International Nucleotide Sequence Database Collaboration

Site maintained by the External Services team at [EMBL-EBI](#) | [Terms of Use](#) | [Privacy](#) | [Cookies](#)

**nature** International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 533 > Issue 7602 > Correspondence > Article

ARTICLE PREVIEW  
view full access options >

NATURE | CORRESPONDENCE

## Databases: Reminder to deposit DNA sequences

Steven L. Salzberg

Nature 533, 179 (12 May 2016)  
Published online 11 May 2016

Citation | Reprints | Rights

As members of the Advisory Committee for the International Nucleotide Sequence Database Collaboration (INSDC), which includes the European Nucleotide Archive (ENA) and GenBank, we remind you of the importance of depositing complete DNA sequences in the INSDC databases.

Editors' pick  
Image credit: L. Bourouiba/The Fluid Dynamics of Disease Transmission Laboratory/MIT

**Science** AAAS

Home | News | Journals | Topics | Careers

Science | Science Advances | Science Immunology | Science Robotics | Science Signaling | Science Translational Medicine

SHARE LETTERS

## Reminder to deposit DNA sequences

Mark Blaxter<sup>1</sup>, Antoine Danchin<sup>2</sup>, Babis Savakis<sup>3</sup>, Kaoru Fukami-Kobayashi<sup>4</sup>, Ken Kurokawa<sup>5</sup>, Sumio Sugano<sup>6</sup>, Richard J. Roberts<sup>7</sup>, Steven L. Salzberg<sup>8\*</sup>, Chung-I Wu<sup>3,10</sup>

\*Corresponding author. Email: [salzberg@jh.edu](mailto:salzberg@jh.edu)

Science 13 May 2016  
Vol. 352, Issue 6287, pp. 780  
DOI: 10.1126/science.1257672

Article | Figures & Data | Info & Metrics | eLetters | PDF

You do not have access to the full text of this article, the first page of the PDF of this article appears below.

ARTICLE TOOLS

- Email
- Print
- Alerts
- Citation tools
- Download Powerpoint
- Save to my folders
- Request Permissions
- Share

Advertisement

# Services: submissions

The screenshot shows the 'New Submission' page in the ENA submission system. The navigation tabs include Home, New Submission, Studies, Sample Groups, Samples, Experiments, Runs, and Projects. The 'Sample' section is active, showing a progress bar from 'Start' to 'Finish'. Below the progress bar, there is a prompt: 'Please create new samples by uploading a spreadsheet or by following the instructions below.' Two informational boxes provide instructions: 'Please select the checklist that you wish to use for your sample submission' and 'If you already have a spreadsheet containing your data upload it here.' An 'Upload Spreadsheet' button is visible. A Microsoft Excel spreadsheet is overlaid on the page, showing a checklist for sample submission. The spreadsheet has columns for various fields and rows of sample data.

Default Checklist

Upload Spreadsheet

Book1 - Microsoft Excel

#	A	B	C	D	E	F	G	H	I	J
1	#checklist_accession	ERC000001								
2	#unique_name_prefix	mouse_dendrocyte_								
3	sample_alias	tax_id	scientific_name	common_name	anonymized_name	sample_title	sample_description	tissue_type	sex	collectio
4	#template	10090	Mus musculus	house mouse		mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain		male	13/0
5	#units									
6	mouse_dendrocyte_1	10090	Mus musculus	house mouse		mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain		male	13/0
7	mouse_dendrocyte_2	10090	Mus musculus	house mouse		mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain		male	13/0
8	mouse_dendrocyte_3	10090	Mus musculus	house mouse		mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain		male	13/0
9	mouse_dendrocyte_4	10090	Mus musculus	house mouse		mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain		male	13/0
10	mouse_dendrocyte_5	10090	Mus musculus	house mouse		mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain		male	13/0
11	mouse_dendrocyte_6	10090	Mus musculus	house mouse		mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain		male	13/0
12	mouse_dendrocyte_7	10090	Mus musculus	house mouse		mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain		male	13/0

(Example views from ENA services)



# Services: data discovery

*temperature>=10 AND temperature<=25 AND geo\_box1(42, 17, 43, 18)*

Accession	First public	Geographical location	Submitter's sample name	Depth (m)	Environment (Biome)	Temperature (C)	Sampling Site
<a href="#">SAMEA2591084</a>	2014-07-11	42.2038 N 17.715 E	TARA_E500000075	5.0	marine biome (ENVO:00000447)	17.32198	TARA_023
<a href="#">SAMEA2591093</a>	2014-06-23	42.2038 N 17.715 E	TARA_A100000551	5.0	marine biome (ENVO:00000447)	17.32198	TARA_023
<a href="#">SAMEA2591094</a>	2014-06-23	42.2038 N 17.715 E	TARA_A100000553	5.0	marine biome (ENVO:00000447)	17.32198	TARA_023
<a href="#">SAMEA2591095</a>	2014-06-26	42.2038 N 17.715 E	TARA_A100000552	5.0	marine biome (ENVO:00000447)	17.32198	TARA_023
<a href="#">SAMEA2591096</a>	2014-06-26	42.2038 N 17.715 E	TARA_A100000547	5.0	marine biome (ENVO:00000447)	17.32198	TARA_023
<a href="#">SAMEA2591097</a>	2014-07-18	42.2038 N 17.715 E	TARA_E500000056	5.0	marine biome (ENVO:00000447)	17.32198	TARA_023
<a href="#">SAMEA2591098</a>	2014-07-18	42.1735 N 17.7252 E	TARA_E500000081	55.0	marine biome (ENVO:00000447)	15.194062	TARA_023
<a href="#">SAMEA2591099</a>	2014-07-11	42.1735 N 17.7252 E	TARA_E500000080	55.0	marine biome (ENVO:00000447)	15.194062	TARA_023
<a href="#">SAMEA2591103</a>	2014-06-23	42.1735 N 17.7252 E	TARA_A100000558	55.0	marine biome (ENVO:00000447)	15.194062	TARA_023

*tax\_tree(10090) AND library\_source="GENOMIC" AND instrument\_platform="ILLUMINA" AND library\_strategy="ChIP-Seq"*

Study accession	Sample accession	Run accession	Scientific name	Fastq files (ftp)	Fastq files (galaxy)	Submitter's sample name
<a href="#">PRJEB6568</a>	<a href="#">SAMEA2604495</a>	<a href="#">ERR537823</a>	<a href="#">Mus musculus</a>	<a href="#">File 1</a> <a href="#">File 2</a>	<a href="#">File 1</a> <a href="#">File 2</a>	E-MTAB-2661:Exp2-Irf5-WT-120m-R2
<a href="#">PRJEB6568</a>	<a href="#">SAMEA2604492</a>	<a href="#">ERR537824</a>	<a href="#">Mus musculus</a>	<a href="#">File 1</a> <a href="#">File 2</a>	<a href="#">File 1</a> <a href="#">File 2</a>	E-MTAB-2661:Exp2-Irf5-KO-000m-R1
<a href="#">PRJEB6568</a>	<a href="#">SAMEA2604494</a>	<a href="#">ERR537825</a>	<a href="#">Mus musculus</a>	<a href="#">File 1</a> <a href="#">File 2</a>	<a href="#">File 1</a> <a href="#">File 2</a>	E-MTAB-2661:Exp2-Irf5-input-120m
<a href="#">PRJEB6568</a>	<a href="#">SAMEA2604493</a>	<a href="#">ERR537826</a>	<a href="#">Mus musculus</a>	<a href="#">File 1</a> <a href="#">File 2</a>	<a href="#">File 1</a> <a href="#">File 2</a>	E-MTAB-2661:Exp2-Irf5-WT-000m-R1
<a href="#">PRJEB6568</a>	<a href="#">SAMEA2604496</a>	<a href="#">ERR537827</a>	<a href="#">Mus musculus</a>	<a href="#">File 1</a> <a href="#">File 2</a>	<a href="#">File 1</a> <a href="#">File 2</a>	E-MTAB-2661:Exp2-Irf5-WT-120m-R1
<a href="#">PRJEB6568</a>	<a href="#">SAMEA2604491</a>	<a href="#">ERR537828</a>	<a href="#">Mus musculus</a>	<a href="#">File 1</a> <a href="#">File 2</a>	<a href="#">File 1</a> <a href="#">File 2</a>	E-MTAB-2661:Exp2-Irf5-KO-120m-R2
<a href="#">PRJEB6568</a>	<a href="#">SAMEA2604500</a>	<a href="#">ERR537829</a>	<a href="#">Mus musculus</a>	<a href="#">File 1</a> <a href="#">File 2</a>	<a href="#">File 1</a> <a href="#">File 2</a>	E-MTAB-2661:Exp2-Irf5-input-000m
<a href="#">PRJEB6568</a>	<a href="#">SAMEA2604498</a>	<a href="#">ERR537830</a>	<a href="#">Mus musculus</a>	<a href="#">File 1</a> <a href="#">File 2</a>	<a href="#">File 1</a> <a href="#">File 2</a>	E-MTAB-2661:Exp2-Irf5-KO-120m-R1
<a href="#">PRJEB6568</a>	<a href="#">SAMEA2604497</a>	<a href="#">ERR537831</a>	<a href="#">Mus musculus</a>	<a href="#">File 1</a>	<a href="#">File 1</a>	E-MTAB-2661:Exp2-Irf5-WT-000m-R1

(Example 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.





# 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