Management of digital sequence information

Guy Cochrane

EMBL-EBI

CCTAACCAAGTACTACTTAAAATCAGTAAACTGCCACTCCCTCCCCACTACCCTATATGA
AAATAATTAAAAAAGCTCCTACGCTTCTTTTTTTGACCGCCCTCTACCTCTTTTCACAAA
GGCTATCAACTCGAATTTTCTATTGCTTTATTTTATCAAAAAAAACATAAAAATTTTTGT
ATCAAAAAATATCATTCAAGAGGGGGGGATTTTCACCCCCACCGCCGGCACCAAAGCCAA
AATTCTTGGAATCAAACTACCTCTTGTCCTGGTTTTCCTAATGTACGAGTAACGTGGCAA
CATATTATGCCTATCGTGCATCCAACTATCTGCCACACGACTATTTTTTAGTACTCTTCG
GAGTGTAAGCCGAACACTTAGGGCGAGAAACCACCAACCCGCTCCTGACGATACGATGAC
CAGATCTGAGGACTTTTATTGTAGAGTGCCTTACTTCCCTTGAGGCGCCACTGGTTAAAA
TCTATGGACACGACTCGAAGATTCATTCATCAATTGGATCGAACGGGTACCTGGCGGCTG
CTTGATAACTAATCAGCCCATGATCCCTCAGCCTCCTCTAAGCACATCTGGTAATTTTTT
ATTTTTCTGTGTGGTCAACCAACATTACGGTAATATGTCTGGTACTACACGATCTAAAGC
TGAACATAACATGCAATTGTTTTATTTGGACCAAATAGAATGAGTAAATTATATGAATGA
TTAT

EMBL-EBI

A sequence in isolation…

```
CCTAACCAAGTACTACTTAAAATCAGTAAACTGCCACTCCCTCCCCACTACCCTATATGA
AAATAATTAAAAAAGCTCCTACGCTTCTTTTTTTGACCGCCCTCTACCTCTTTTCACAAA
GGCTATCAACTCGAATTTTCTATTGCTTTATTTTATCAAAAAAAACATAAAAATTTTTGT
ATCAAAAAATATCATTCAAGAGGGGGGGATTTTCACCCCCACCGCCGGCACCAAAGCCAA
AATTCTTGGAATCAAACTACCTCTTGTCCTGGTTTTCCTAATGTACGAGTAACGTGGCAA
CATATTATGCCTATCGTGCATCCAACTATCTGCCACACGACTATTTTTTAGTACTCTTCG
GAGTGTAAGCCGAACACTTAGGGCGAGAAACCACCAACCCGCTCCTGACGATACGATGAC
CAGATCTGAGGACTTTTATTGTAGAGTGCCTTACTTCCCTTGAGGCGCCACTGGTTAAAA
TCTATGGACACGACTCGAAGATTCATTCATCAATTGGATCGAACGGGTACCTGGCGGCTG
CTTGATAACTAATCAGCCCATGATCCCTCAGCCTCCTCTAAGCACATCTGGTAATTTTTT
ATTTTTCTGTGTGGTCAACCAACATTACGGTAATATGTCTGGTACTACACGATCTAAAGC
TGAACATAACATGCAATTGTTTTATTTGGACCAAATAGAATGAGTAAATTATATGAATGA
TTAT
```

.. cannot be interpreted

.. is unlikely to be informative

EMBL-EBI

A collection of sequences..

260 million sequences

.. can be interpreted

.. drives discovery

EMBL-EBI

International Nucleotide Sequence Database Collaboration (INSDC)

DDBJ
DNA Data Bank of Japan
National Institute of Genetics

INSDC

GenBank
NCBI

ENA
European Nucleotide Archive
EMBL-EBI

Values

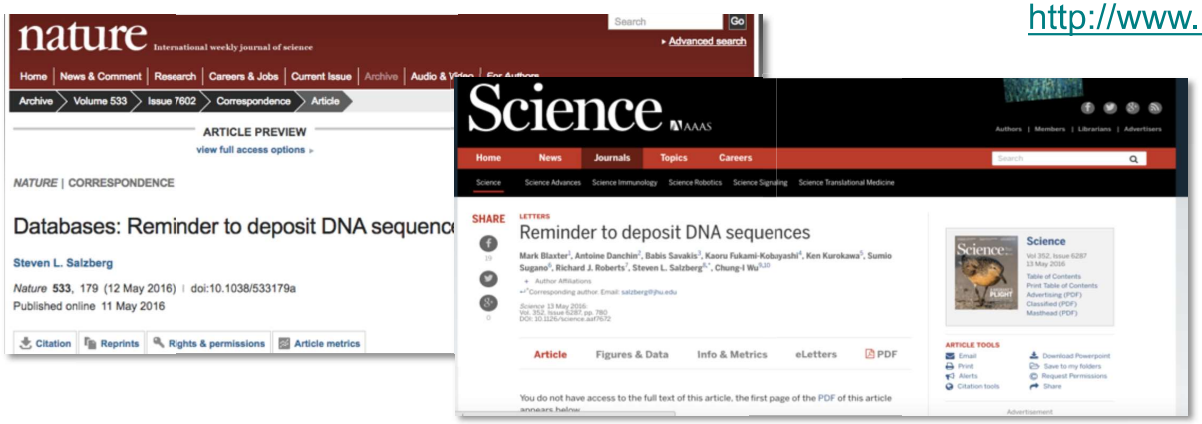
- open access for all
- globally comprehensive
- spanning life science domains
- permanent database of record
- public forum for the scientific process

Organisation

- established early 1980s
- major ongoing investment
- structure and governance
- model for scientific collaboration

INSDC

http://www.insdc.org/

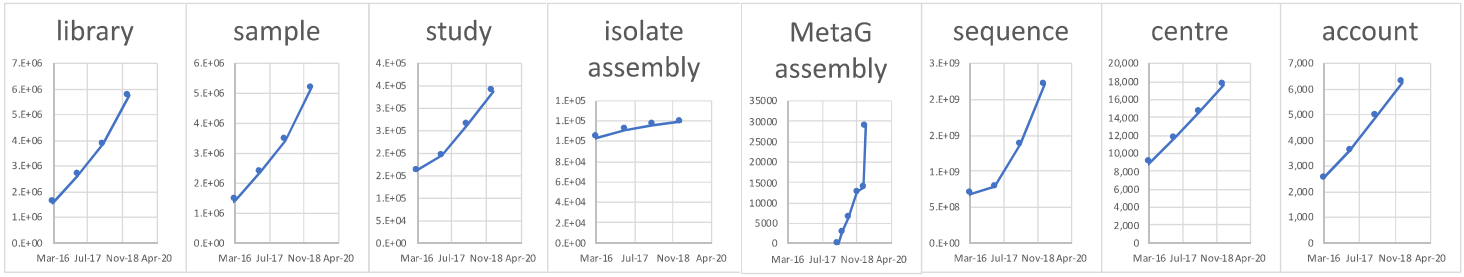


Instruments

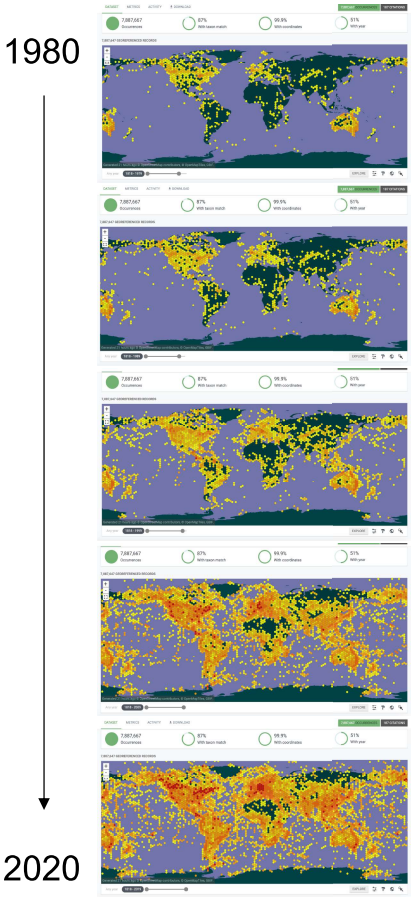
- regular data exchange
- accession scheme
- data standards
- mandatory submission agreement
- services and software (node-level)

Scale

- **Rate**: 1 new dataset every 6 minutes
- **Data**: $2x10^9$ sequences and $1x10^{16}$ base pairs of read data across $2x10^6$ taxa
- **Usage**: 2,000 submitters; 10x thousands monthly consumers; 10x millions of monthly hits, many times this globally
- **Support**: 46 tickets per day and in-person training delivered to more than 350 users per annum
- **Infrastructure:** includes 100s of petabytes of storage, ~$50 USD mil./year
- **Growth:**



- **E.G. biodiversity data**
  - 7.2 million records point to natural history museums, culture collections and biobanks
  - 7.9 million sequences with geographical coordinates
  - 2.1 million sequences have place name annotation
  - 1 million raw sequencing data sets have coordinates

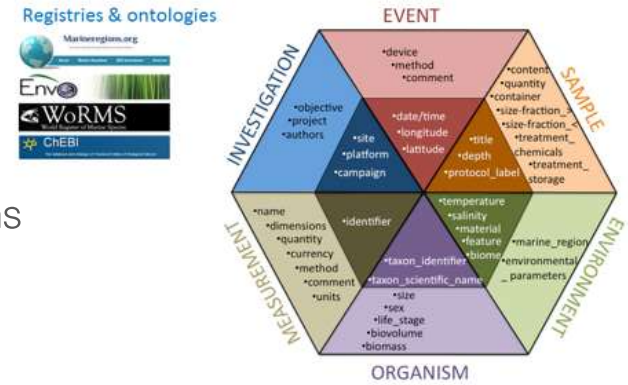


https://doi.org/10.15468/cndomv, courtesy of GBIF

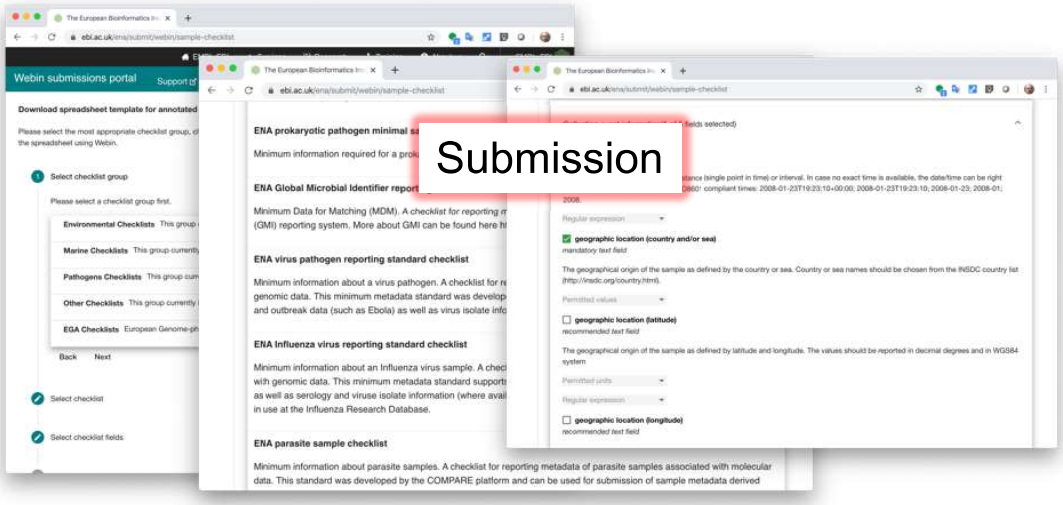
EMBL-EBI

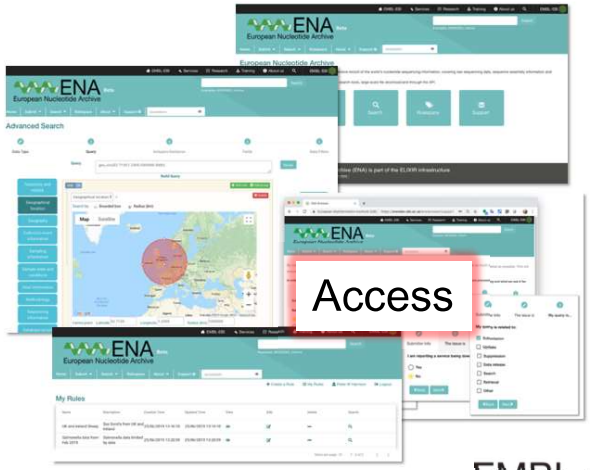
The INSDC data cycle

- Data provider
  - Validation, organization, adding structure and curation
  - Compliance with standards and established conventions
  - Links to external data (e.g. academic publications)
- Database
  - Curation and integration with the overall corpus
  - Indexing for discovery and reuse
  - Open services, freely available to the world



Ten Hoopen P., Pesant S., Kottmann R., Kopf A., Bicak M., Claus S., Deneudt K., Borremans C., Thijsse P., Dekeyzer S., Schaap D., **Bowler C.**, Glöckner F.O., **Cochrane G.** Data standards for Marine Microbial Biodiversity, Bioinformatics and Biotechnology (M2B3) *Standards in Genomic Sciences* 10:20 doi:10.1186/s40793-015-0001-5 (2015).
http://www.standardsingenomics.com/content/10/1/20

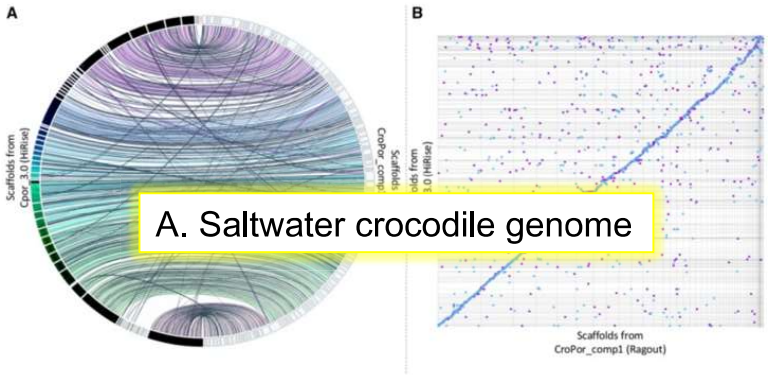




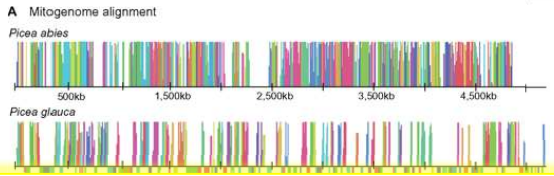
EMBL-EBI

Direct use

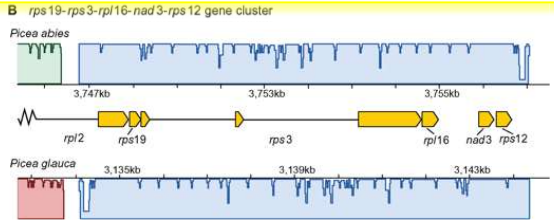
- 1,600 publications citing sequence accessions per month in 2020

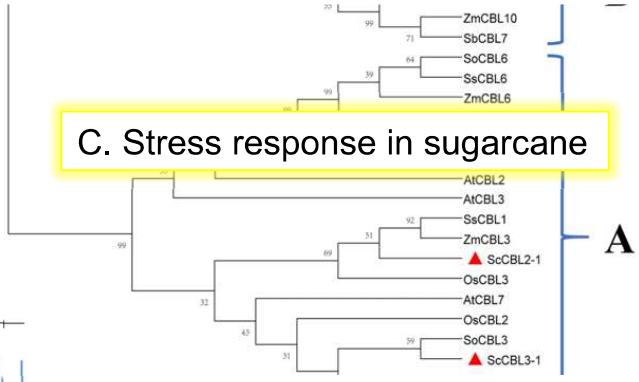


A. Saltwater crocodile genome

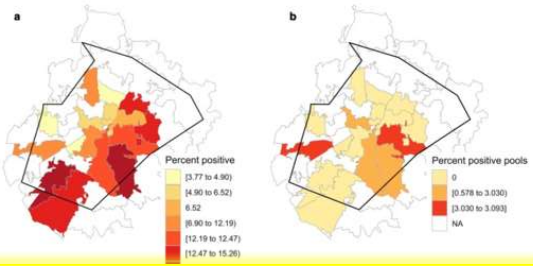


B. Norway spruce evolutionary biology

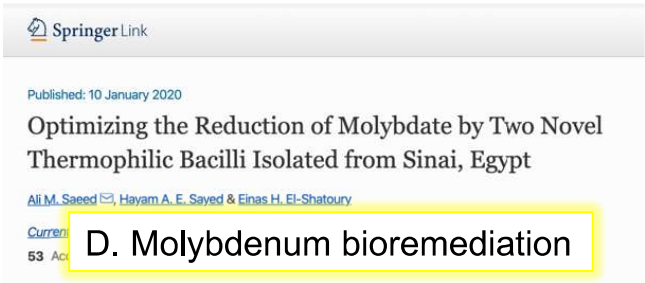




C. Stress response in sugarcane



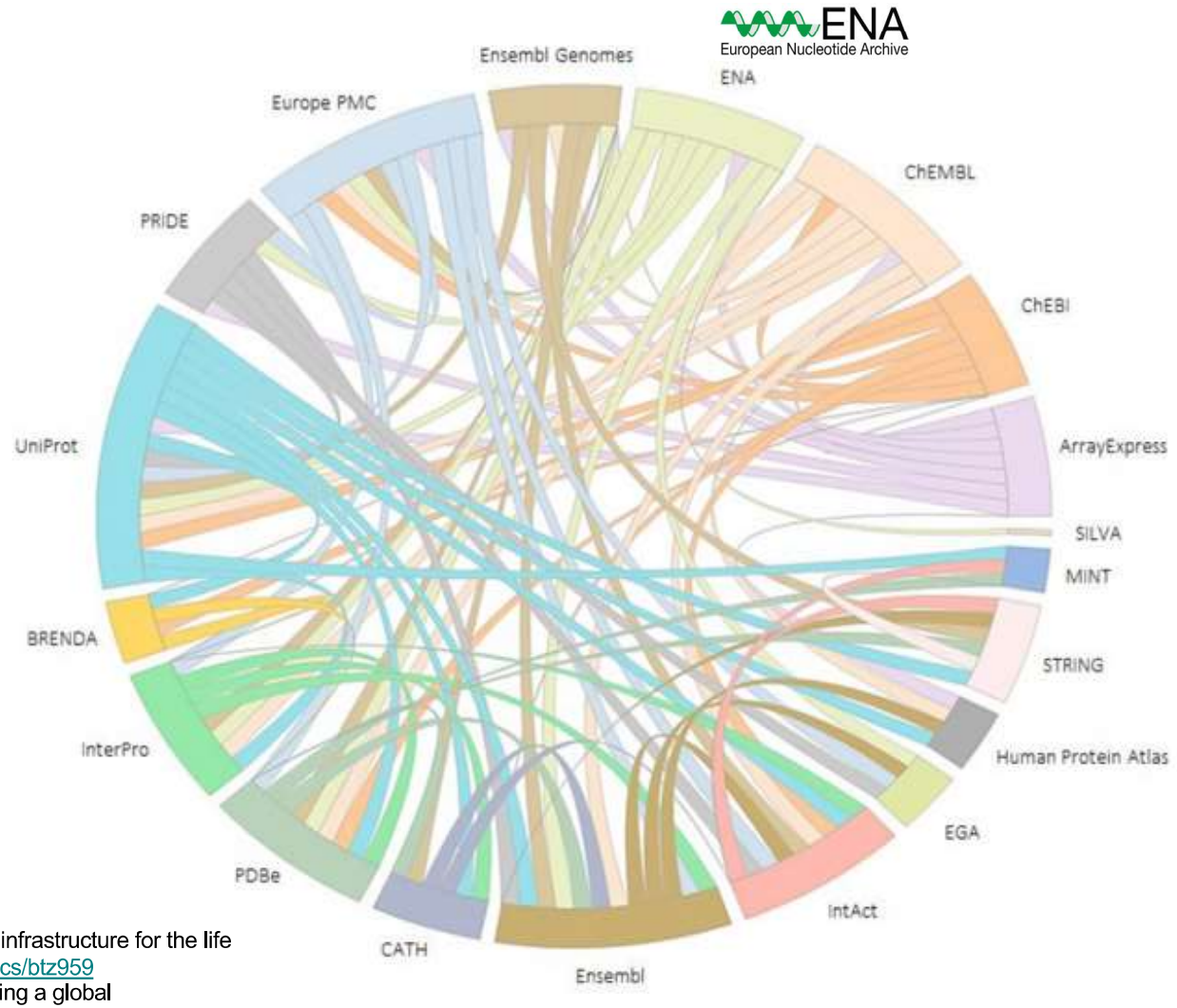
E. Mosquito-borne heartworm in dogs



D. Molybdenum bioremediation

A. Ghosh A et al., Genome Biol Evol. 2020 Jan;12(1) 3635-3646. doi:10.1093/gbe/evz269. PMID: 31821505; © The Author(s) 2020; http://creativecommons.org/licenses/by/4.0/
B. Sullivan AR et al., Genome Biol Evol. 2020 Jan;12(1) 3586-3598. doi:10.1093/gbe/evz263. PMID: 31774499; © The Author(s) 2020; http://creativecommons.org/licenses/by-nc/4.0/
C. Su W et al., Sci Rep. 2020 Jan;10(1) 167. doi:10.1038/s41598-019-57058-7. PMID: 31932662; http://creativecommons.org/licenses/by/4.0/
D. Saeed AM et al., Curr Microbiol. 2020 Jan . doi:10.1007/s00284-020-01874-y. PMID: 31925514.
E. Spence Beaulieu MR et al., Parasit Vectors. 2020 Jan;13(1) 12. doi:10.1186/s13071-019-3874-0. PMID: 31924253; PMCID: PMC6953185. http://creativecommons.org/licenses/by/4.0/; http://creativecommons.org/publicdomain/zero/1.0/

Data reach

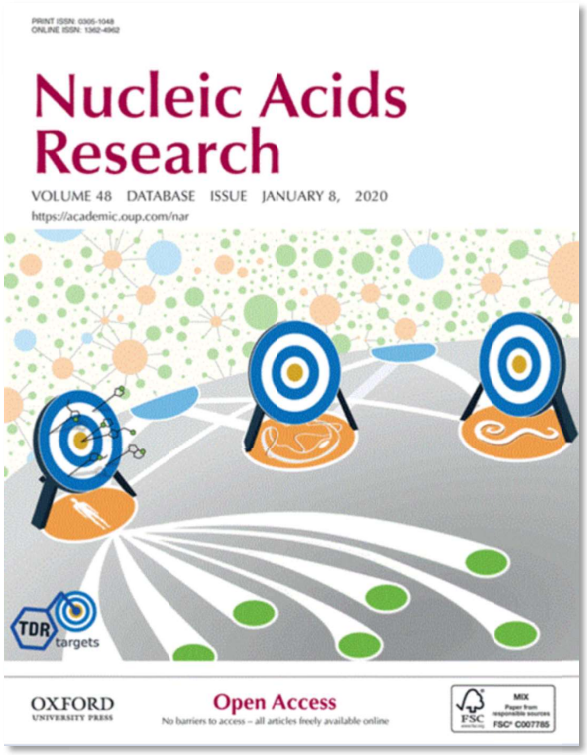


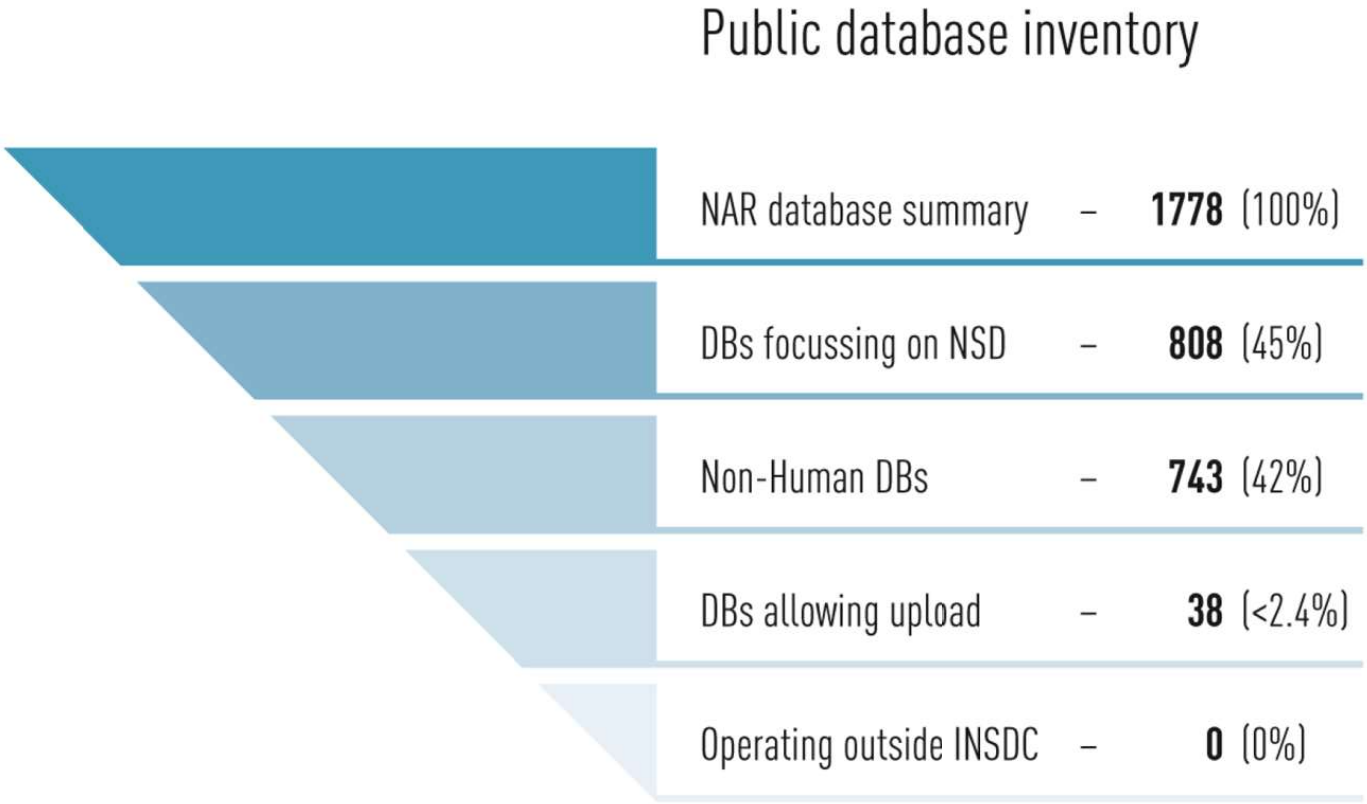
elixir Core Data Resource

Drysdale *et al.* (2020) The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences. *Bioinformatics*, 2020, 1–7; http://doi.org/10.1093/bioinformatics/btz959

Cook *et al.* (2020) The European Bioinformatics Institute in 2020: building a global infrastructure of interconnected data resources for the life sciences. *Nucleic Acids Research* 48:D17-D23; http://doi.org/10.1093/nar/gkz1033

Further reach





Rohden F, Huang S, Dröge G, Hartman Scholz A, and contributing authors (2019). Combined study in DSI in public and private databases and DSI traceability. https://www.cbd.int/abs/DSI-peer/Study-Traceability-databases.pdf

EMBL-EBI

Terms of use

Data Services

1. The online data services and databases of EMBL-EBI are generated in part from data contributed by the community who remain the data owners.

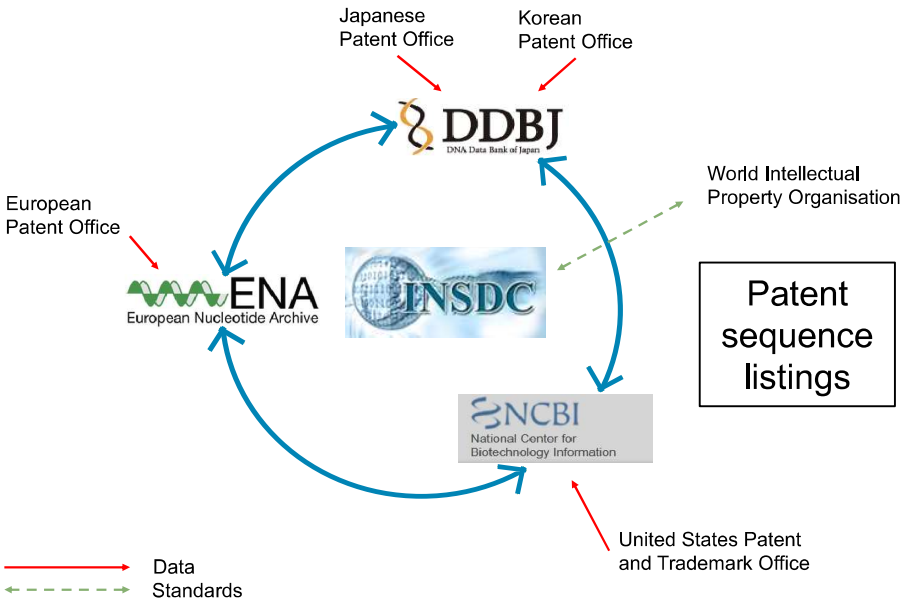
2. When you contribute scientific data to a database through our website or other submission tools this information will be released at a time and in a manner consistent with the scientific data and we may store it permanently.

3. EMBL-EBI itself places no additional restrictions on the use or redistribution of the data available via its online services other than those provided by the original data owners.

4. EMBL-EBI does not guarantee the accuracy of any provided data, generated database, software or online service nor the suitability of databases, software and online services for any purpose.

5. The original data may be subject to rights claimed by third parties, including but not limited to, patent, copyright, other intellectual property rights, biodiversity-related access and benefit-sharing rights. For the specific case of the EGA database and human data consented for biomedical research, these rights may be formalised in Data Access Agreements. It is the responsibility of users of EMBL-EBI services to ensure that their exploitation of the data does not infringe any of the rights of such third parties.

https://www.ebi.ac.uk/about/terms-of-use



Patent sequence listings

EXAMPLES OF PREPUBLICATION DATA-RELEASE GUIDELINES

Project type	Prepublication data release recommended	Prepublication data release optional
Genome sequencing	Whole-genome or mRNA sequence(s) of a reference organism or tissue	Sequences from a few loci for cross-species comparisons in a limited number of samples
Polymorphism discovery	Catalogue of variants from genomic and/or transcriptomic samples in one or more populations	Variants in a gene, a gene family or a genomic region in selected pedigrees or populations
Genetic association studies	Genomewide association analysis of thousands of samples	Genotyping of selected gene candidates
Somatic mutation discovery	Catalogue of somatic mutations in exomes or genomes of tumour and non-tumour samples	Somatic mutations of a specific locus or limited set of genomic regions
Microbiome studies	Whole-genome sequence of microbial	Sequencing of target locus in a limited me samples
RNA profiling		pression profiles of al system(s)
Proteomic studies	Mass spectrometry data sets from large panels of normal and disease tissues	Mass spectrometry data sets from a well-defined and limited set of tissues
Metabolomic studies	Catalogue of metabolites in one or more tissues of an organism	Analyses of metabolites induced in a perturbed biological system(s)
RNAi or chemical library screen	Large-scale screen of a cell line or organism analysed for standard phenotypes	Focused screens used to validate a hypothetical gene network
3D-structure elucidation	Large-scale cataloguing of 3D structures of proteins or compounds	3D structure of a synthetic protein or compound elucidated in the context of a focused project

Toronto Principles

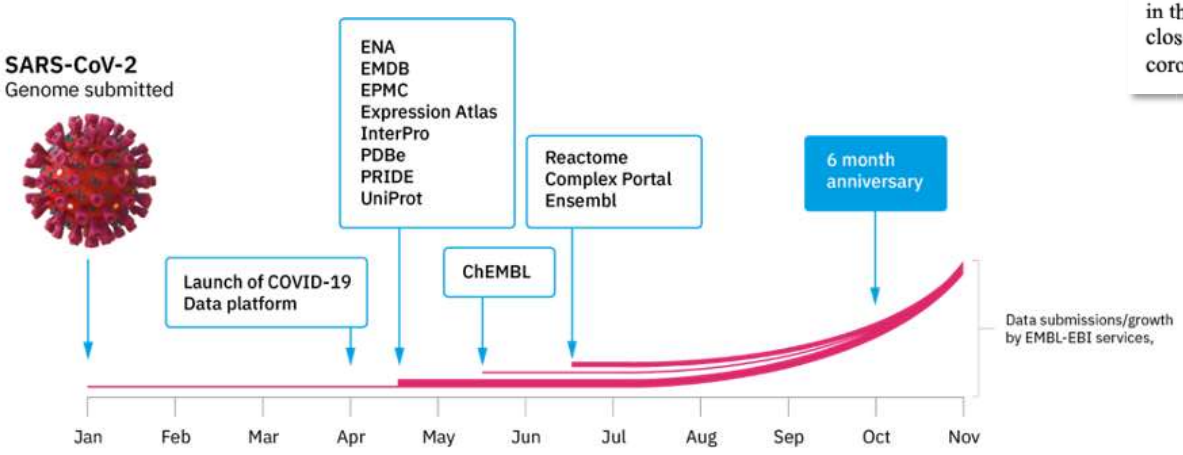
Nature **volume 461**, pages168–170(2009)

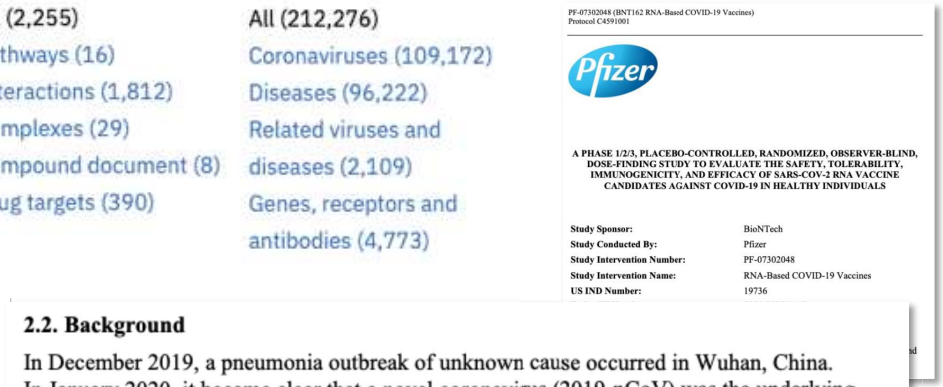
ABS-related obligations in existing agreements

COVID-19 open data

Viral Sequences	Host Sequences	Expression	Proteins	Biochemistry	Literature
All (262,317)	All (1,046)	All (82)	All (712)	All (2,255)	All (212,276)
Sequences (27,649)	Human studies (controlled access) (4)	Gene expression (4)	Protein sequences (57)	Pathways (16)	Coronaviruses (109,172)
Raw reads (125,159)	Human reads (consented for full access) (784)	Gene expression experiments (23)	Protein families (126)	Interactions (1,812)	Diseases (96,222)
Sequenced samples (96,632)	Other species reads (250)	Single cell expression (4)	Protein structures - Knowledge Base (7)	Complexes (29)	Related viruses and diseases (2,109)
Studies (163)	Association studies (8)	Single cell expression experiments (19)	Protein structures (260)	Compound document (8)	Genes, receptors and antibodies (4,773)
Genes (22)		Protein expression experiments (32)	Electron microscopy density maps (250)	Drug targets (390)	
Browser (1)			Electron microscopy public image archive (12)		
Variants (12,691)					

https://www.covid19dataportal.org/





PF-07302048 (BNT162 RNA-Based COVID-19 Vaccines)
Protocol C4591001

Pfizer

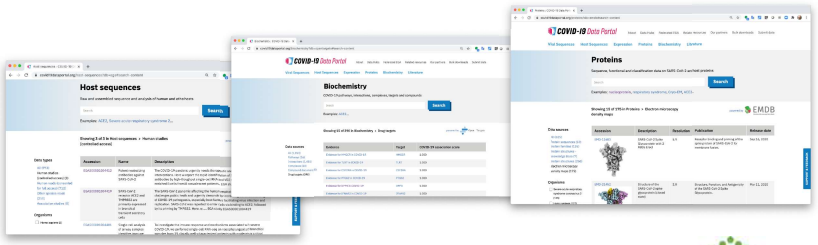
A PHASE 1/2/3, PLACEBO-CONTROLLED, RANDOMIZED, OBSERVER-BLIND, DOSE-FINDING STUDY TO EVALUATE THE SAFETY, TOLERABILITY, IMMUNOGENICITY, AND EFFICACY OF SARS-COV-2 RNA VACCINE CANDIDATES AGAINST COVID-19 IN HEALTHY INDIVIDUALS

Study Sponsor:	BioNTech
Study Conducted By:	Pfizer
Study Intervention Number:	PF-07302048
Study Intervention Name:	RNA-Based COVID-19 Vaccines
US IND Number:	19736

2.2. Background

In December 2019, a pneumonia outbreak of unknown cause occurred in Wuhan, China. In January 2020, *it became clear that a novel coronavirus* (2019-nCoV) was the underlying cause. Later in January, the genetic sequence of the 2019-nCoV became available to the World Health Organization (WHO) and public (MN908947.3), and the virus was categorized in the *Betacoronavirus* subfamily. By sequence analysis, the phylogenetic tree revealed a closer relationship to severe acute respiratory syndrome (SARS) virus isolates than to another coronavirus infecting humans, the Middle East respiratory syndrome (MERS) virus.

https://pfe-pfizercom-d8-prod.s3.amazonaws.com/2020-11/C4591001_Clinical_Protocol_Nov2020.pdf

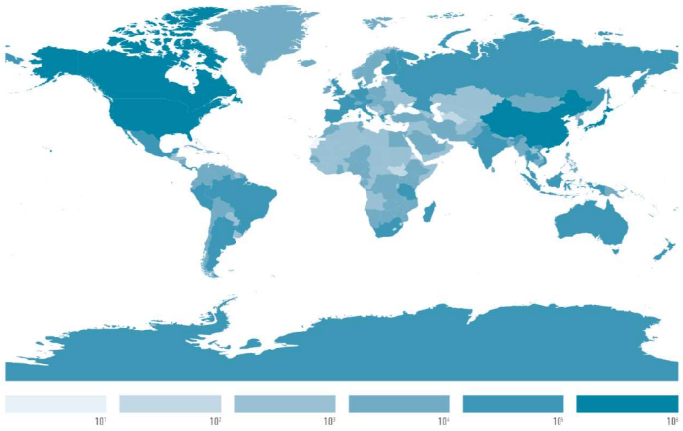


EMBL-EBI

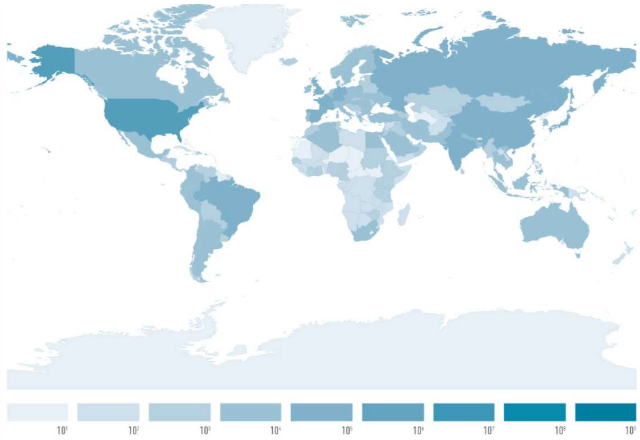
Direction of flow of non-monetary benefit

What is the country of origin for non-human NSD?

1.	China	18.23 %
2.	United States	17.39 %
3.	Canada	9.10 %
4.	Japan	7.24 %
5.	India	3.46 %
6.	Australia	2.66 %
7.	Mexico	2.54 %
8.	Brazil	2.30 %
9.	Germany	1.83 %
10.	Spain	1.58 %



Countries of DSI origin

DSI user countries

1.	United States	22.69 %
2.	China	15.42 %
3.	India	6.16 %
4.	Japan	3.97 %
5.	Germany	3.67 %
6.	United Kingdom	3.45 %
7.	France	2.84 %
8.	Brazil	2.83 %
9.	Spain	2.31 %
10.	Russian Federation	2.25 %