

Combined study on digital sequence information in public and private databases and traceability



Dr. Amber Hartman Scholz, Deputy to the Director



Our terms of reference

1. study on ongoing developments in the field of traceability of digital information, including how traceability is addressed by databases, and how these could inform discussions on digital sequence information on genetic resources (decision [14/20](#), para. 11 (c))
2. study on public and, to the extent possible, private databases of digital sequence information on genetic resources, including the terms and conditions on which access is granted or controlled, the biological scope and the size of the databases, numbers of accessions and their origin, governing policies, and the providers and users of the digital sequence information on genetic resources and encourages the owners of private databases to provide the necessary information (decision 14/20, para. (d))

We had to report quantifiable information on DSI, so we used NSD as a proxy

DSI AHTEG 2018:

Nucleotide Sequence Data = NSD

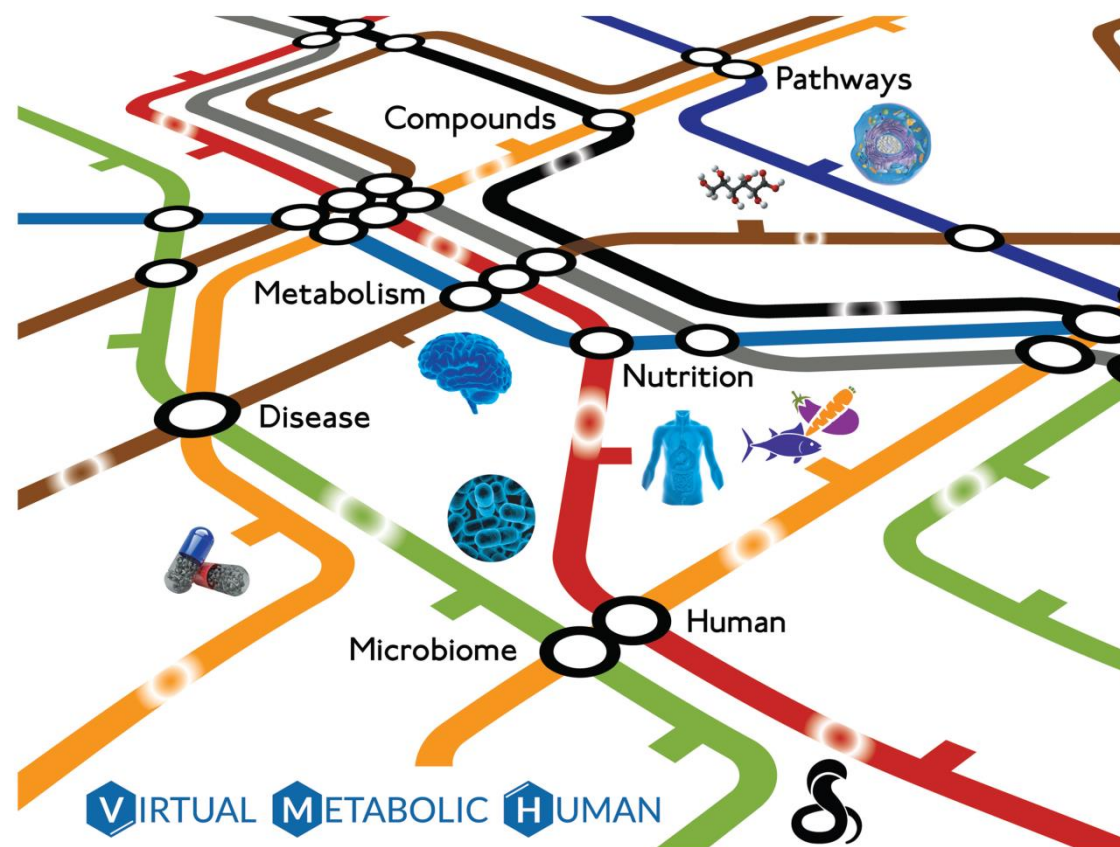
- The nucleic acid sequence reads and the associated data
- Information on the sequence assembly, its annotation and genetic mapping. This information may describe whole genomes, individual genes or fragments thereof, barcodes, organelle genomes or single nucleotide polymorphisms
- Information on gene expression
- Data on macromolecules and cellular metabolites
- Information on ecological relationships, and abiotic factors of the environment
- Function, such as behavioural data
- Structure, including morphological data and phenotype
- Information related to taxonomy
- Modalities of use

Where to start our database analysis? The 1-stop annual database issue on biological databases

Nucleic Acids Research

VOLUME 47 **DATABASE ISSUE** JANUARY 8, 2019

<https://academic.oup.com/nar>



Leibniz-Institut • DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen

OXFORD
UNIVERSITY PRESS

Open Access

No barriers to access – all articles freely available online



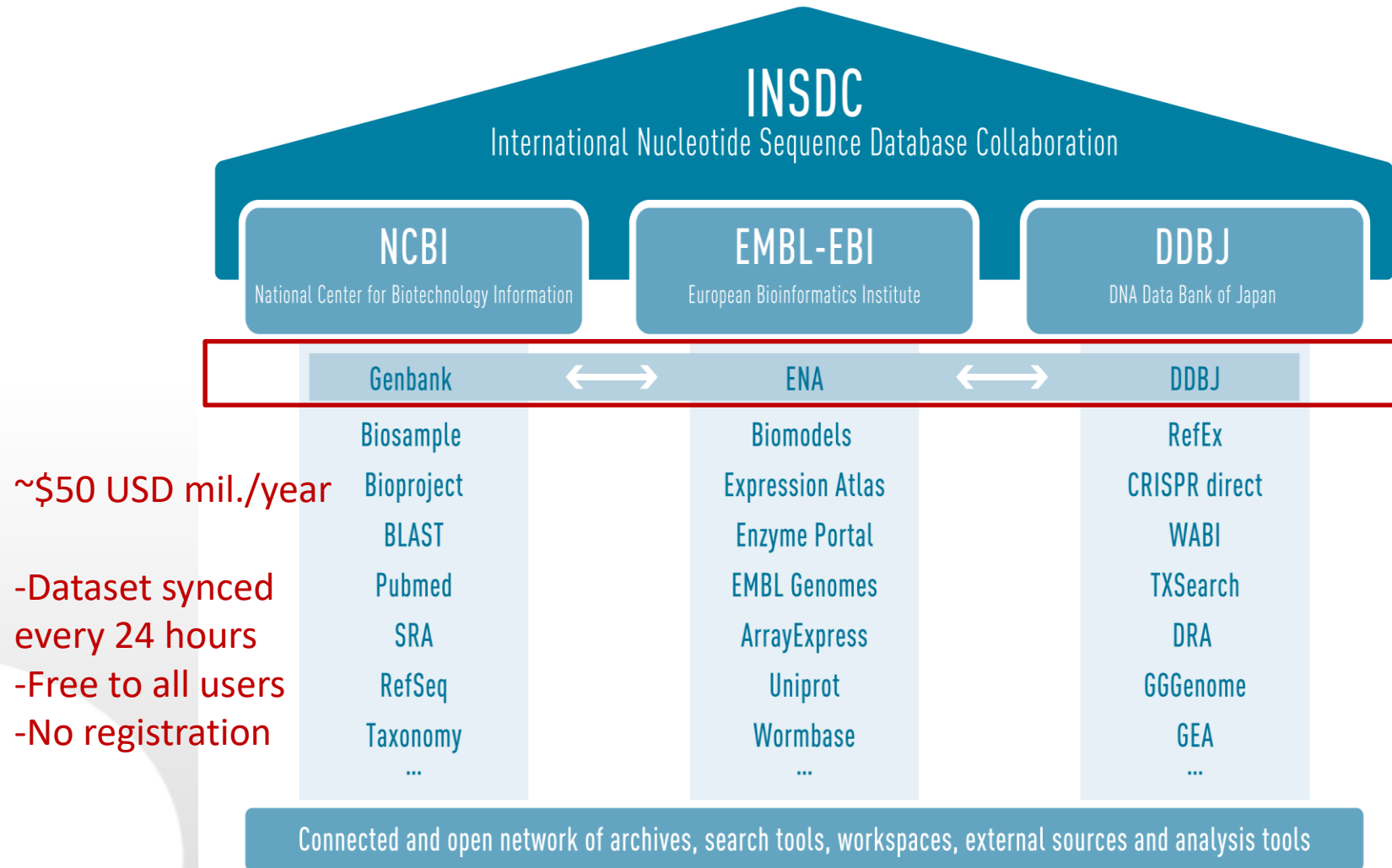
How many public NSD databases are there?

How many rely on the INSDC?

Public database inventory

	NAR database summary	–	1778 (100%)
	DBs focussing on NSD	–	808 (45%)
95% directly link to or download NSD from the INSDC	Non-Human DBs	–	743 (42%)
Only 1 of these 38 databases has an “access policy”	DBs allowing upload	–	38 (<2.4%)
99.9% of public NSD databases rely on the INSDC	Operating outside INSDC	–	3 (0.16%)

The INSDC is the core infrastructure but there are dozens of databases & tools in these organizations



INSDC access and use policy (published 2002)

1. ...uniform policy of **free and unrestricted access to all of the data** records their databases contain.
2. The INSD **will not attach statements to records that restrict access to the data**, limit the use of the information in these records, or prohibit certain types of publications based on these records. Specifically, **no use restrictions or licensing requirements** will be included in any sequence data records...
3. All **database records submitted to the INSD will remain permanently accessible** as part of the scientific record...
4. ..information displayed on the Web sites maintained by the INSD is **fully disclosed to the public...**

This access policy is „inherited“ by the downstream databases.

<http://www.insdc.org/policy.html>

Science 298 (5597): 1333 15 Nov 2002

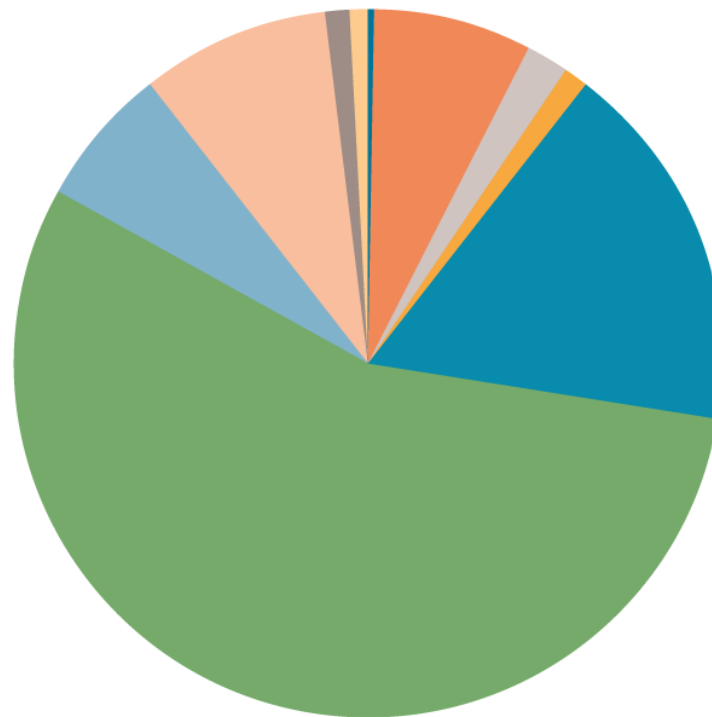
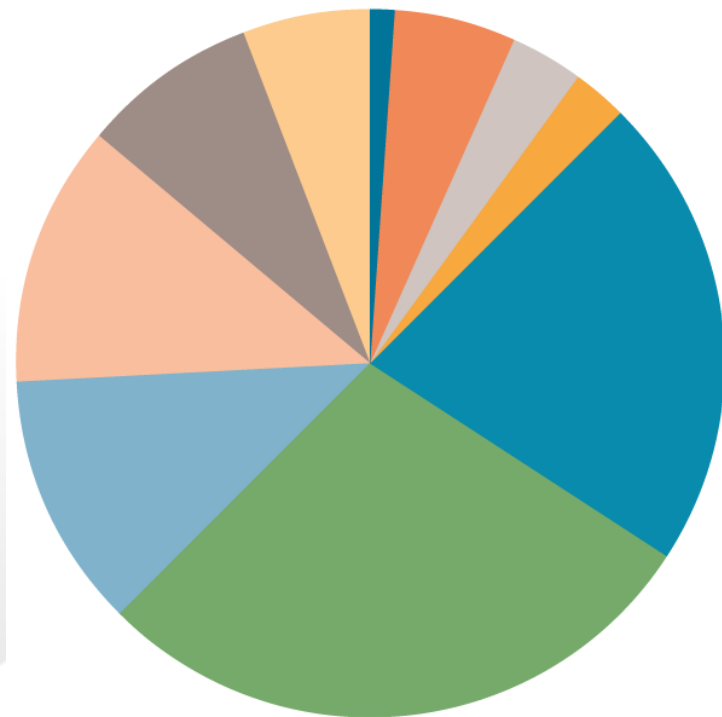
Why do scientists put NSD in a public database for everyone to see and use?

1. Publishing: Journals *will not publish* your paper unless the data is in the INSDC.
2. Funding: Grant agencies very often *require* it.
3. Ethics: Scientific reproducibility, integrity, data security
4. Comparison: ACGTACGT means nothing without context!

NSD comes from all branches of the tree of life & organisms outside of CBD scope

ENTRIES

BASE PAIRS

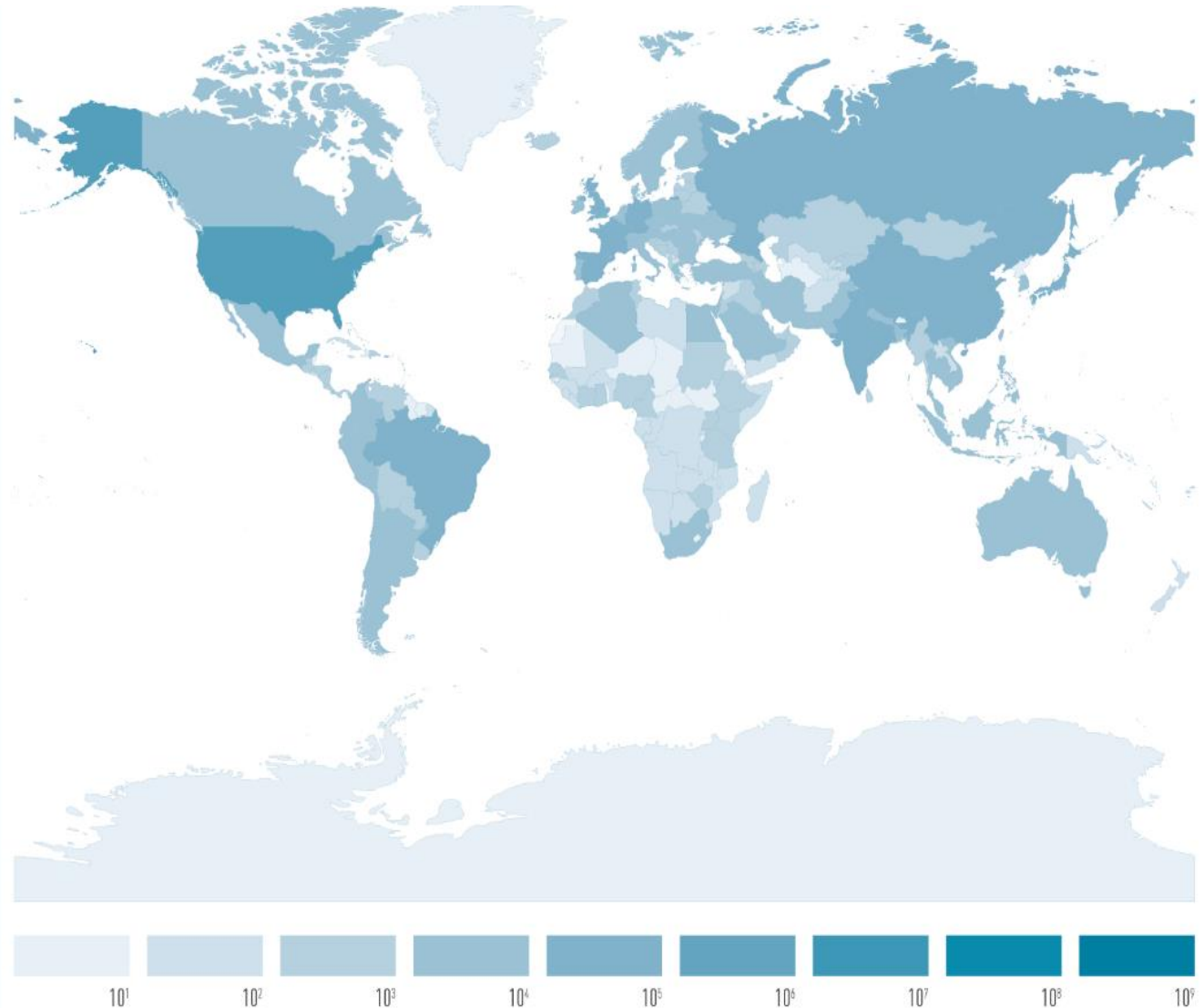


TAXONOMIC GROUP	ENTRIES	BASE PAIRS
Viruses	1,21%	0,32%
Bacteria/Archaea	5,60%	7,37%
Fungi	3,29%	1,95%
Single-celled Eukaryotes	2,56%	0,91%
Plants & Algae	21,61%	17,12%
Animals	28,45%	55,62%
Human	11,52%	6,25%
Model Organisms	11,93%	8,67%
Other/synthetic (microorganisms)	8,03%	1,16%
Unidentified (no tax. ID)	5,82%	0,61%

There are 10-15 million total users of INSDC. They live in every country in the world.

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 %

Costs: \$3-5 per user
50% of users live in
countries that do not
contribute to NSD
infrastructure costs



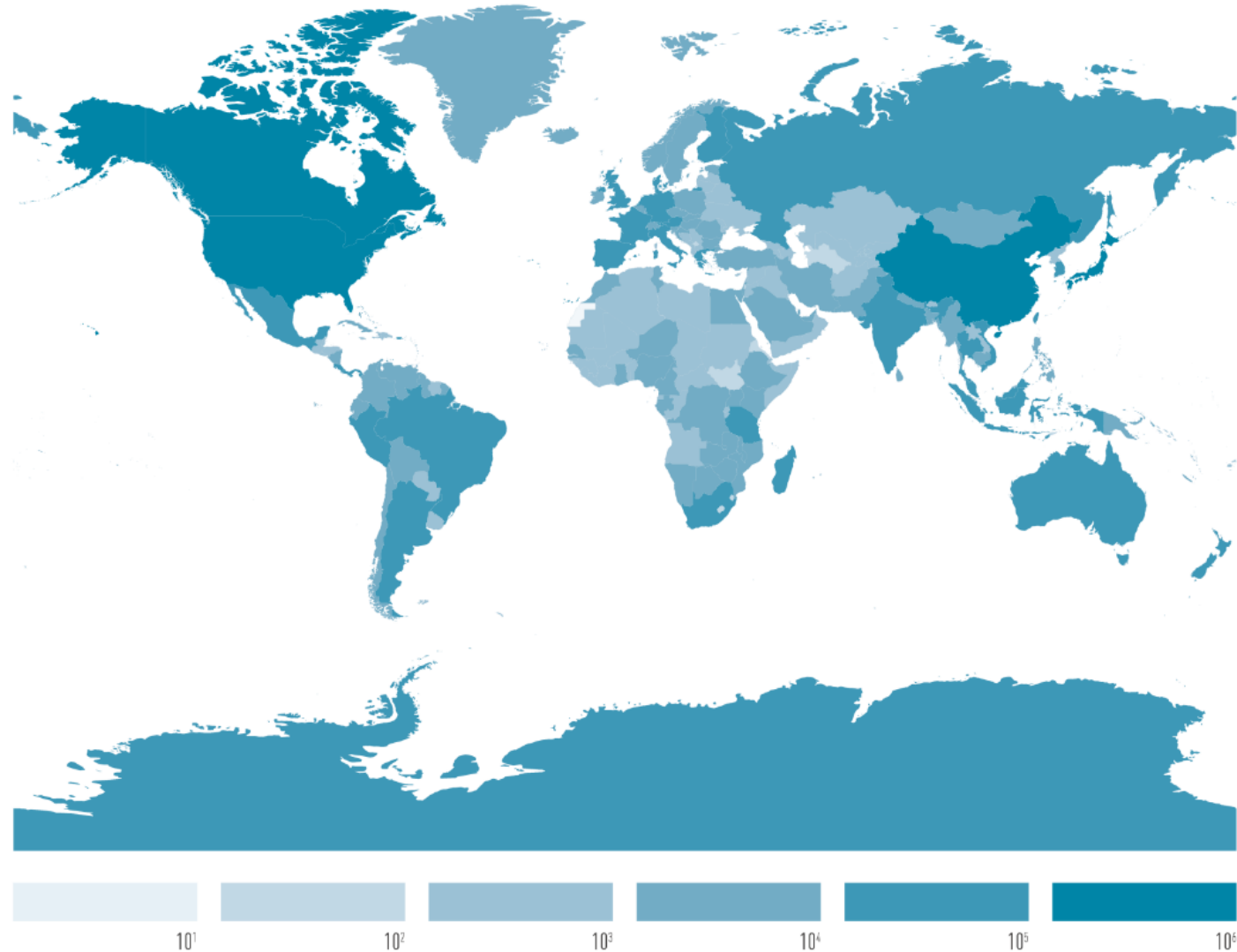
Where does the original GR for these NSD come from?

Every country in the world.

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 %

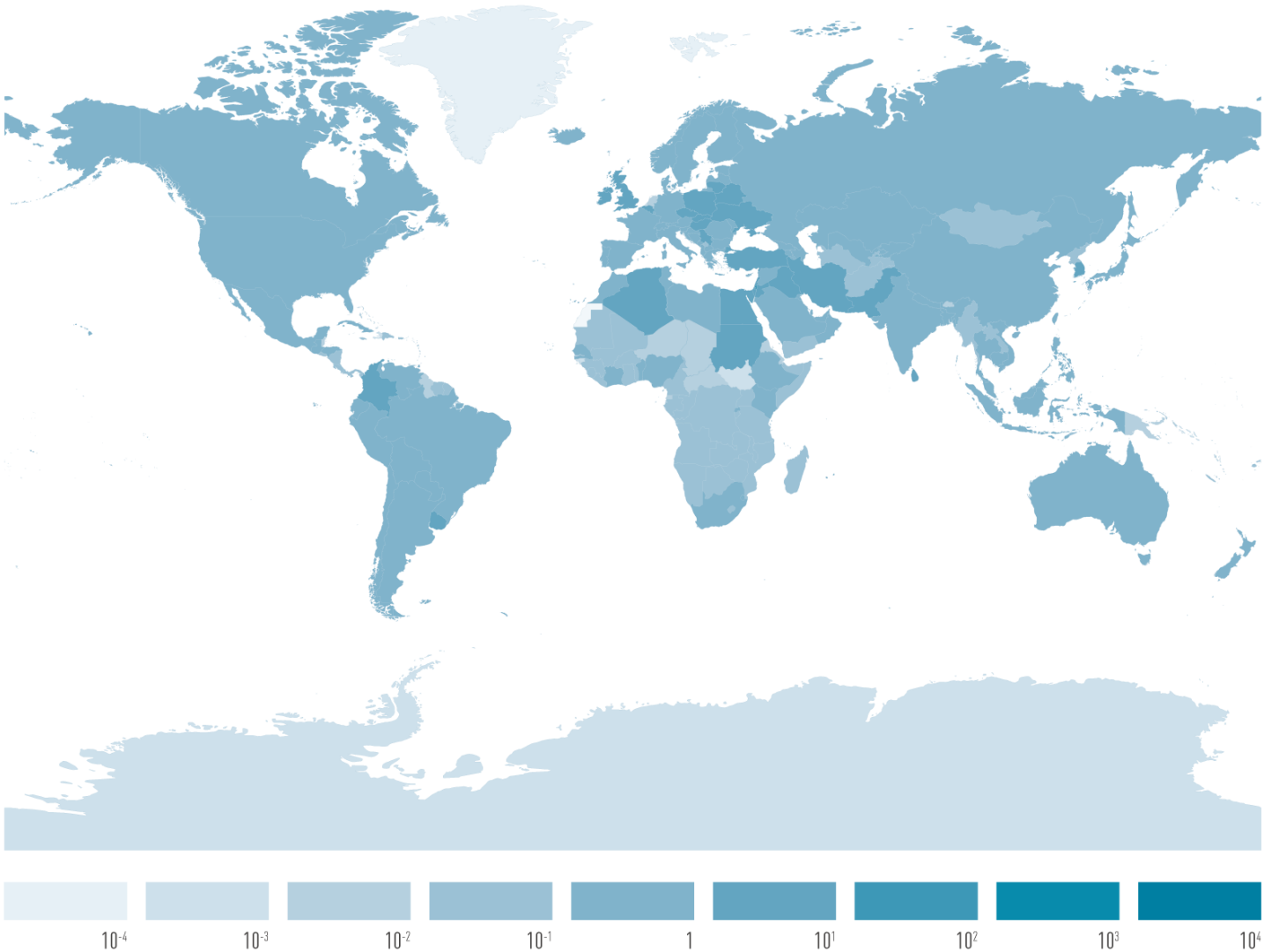
- 16% of NSD has a country tag
- 52% of tagged-NSD comes from 4 countries



For each country: How many users are there per sequence entry in the INSDC?

How does database usage compare to provided sequences?

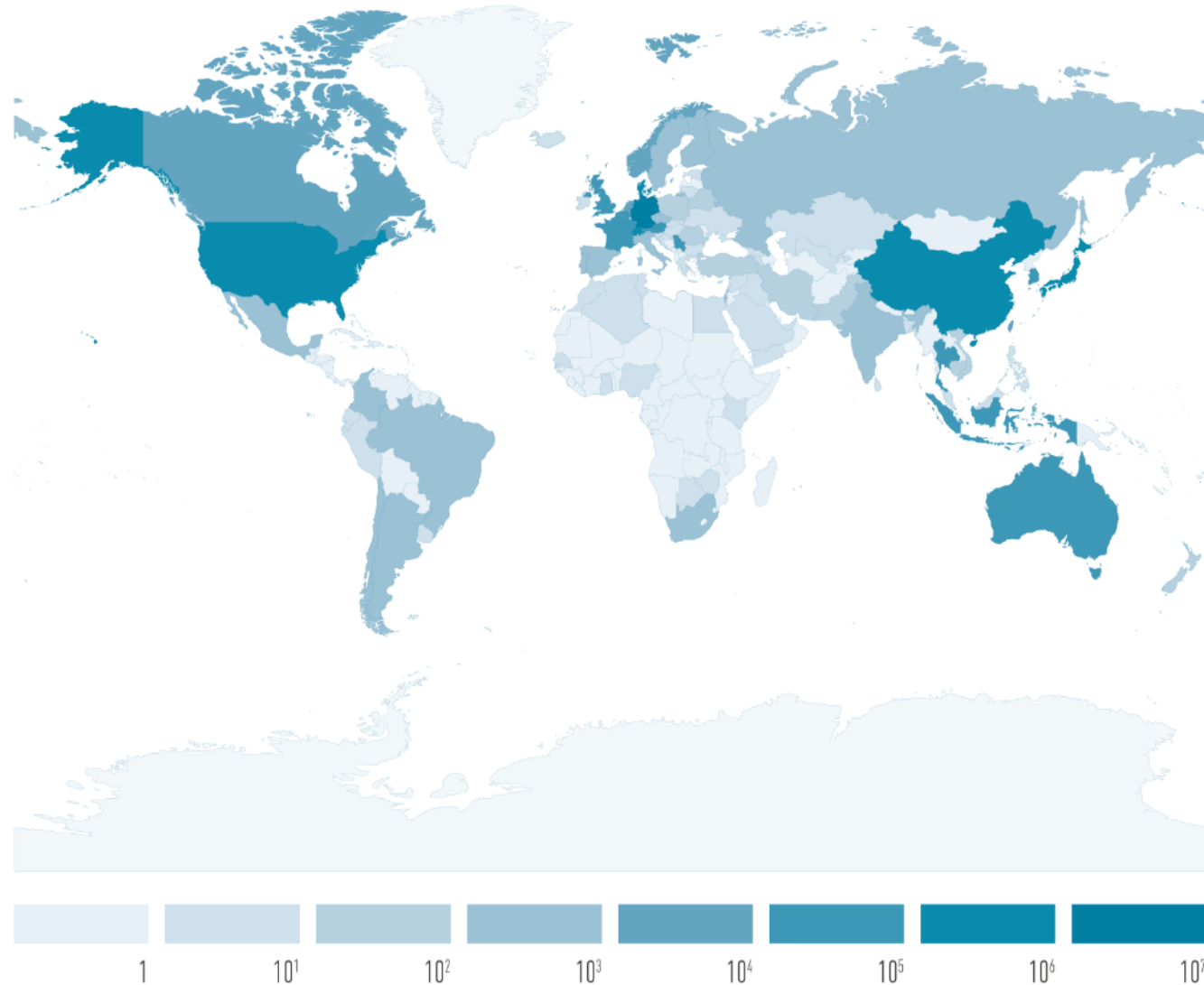
1. Lebanon	27.68
2. Ukraine	23.84
3. Belarus	22.88
4. Iraq	21.41
5. Colombia	19.79
6. Algeria	19.54
7. Pakistan	19.37
8. Republic of Korea	18.81
9. Belgium	17.83
10. Poland	17.47



Where do ftp requests (downloads of entire INSDC dataset) come from?

Where do ftp requests come from?

1. Germany	11451001
2. United States	6333363
3. China	5397918
4. Switzerland	2999900
5. Japan	2180902
6. Denmark	1207616
7. Singapore	1064037
8. Serbia	666395
9. Republic of Korea	643485
10. France	426430



Private databases: 100% of interviewees download INSDC via ftp, generate NSD, submit NSD to INSDC

CASE STUDY	EMPLOYEES	FOCUS FOR NSD + SI	% OF PUBLIC DATA	SUBMIT DATA TO PUBLIC	P&P PARTNERSHIPS	USE PATENT DATABASES
1: Novozymes	> 6,000	Enzymes	~ 50%	yes	yes	yes
2: Company X	> 20,000	Health, materials and nutrition	~ 95%	yes	yes	yes
3: Company Y	> 2,000	Plant breeding and seed production	~ 50–80%	yes	yes	yes
4: TraitGenetics	> 20	Molecular markers and genotypes in plants	?	yes	yes	no
5: BASF SE	> 122,000	Various areas	~ 50–90%	yes	yes	yes
6: Company Z	> 350	Enzymes for DNA handling	?	yes	yes	yes

Patent NSD

1. 20% of NSD in INSDC comes from patent-associated NSD directly deposited by a patent office
2. NSD itself is not patented
3. No geographical information associated with this NSD
4. Patent NSD often receives a new AN

The NSD databases are connected and data flows in a predictable way.

What can be learned from other traceability technologies?

Where do scientists put their NSD?

How do The Complete Genome of *Haloferax volcanii* DS2 plasmid pHV3, complete sequence of DS2, a Model Archaeon

GenBank: CP001953.1

[FASTA](#) [Graphics](#)

Amber L. Hartman, et al.

Category

Display Name (*)

NCBI Taxon ID (*)

Domain (*)

Phylum (*)

Class (*)

Order

Family

Genus (*)

Species (*)

Strain (*)

Culture Collection

Biosafety Level

Project Information

Project description

Accession #s

Project Type (*)

Project Status (*)

Contact Name (*)

Contact Email (*)

GC Percent

Sequencing Center

Sequencing Center

Funding Agency

Funding Agency

Publication Journal

Publication Volume

Publication Link

Sequencing Information

Sequencing Status

D13378.1.1470, *Haloferax gibbonsii*, 1470 nuc
EU308225.1.1438, *Haloferax* sp. FB247_8, 1438 nuc
AB081732.1.1473, *Haloferax lucentense*, 1473 nuc
KY827082.1.1265, *Haloferax* sp., 1265 nuc
DQ458842.1.1429, *Haloferax* sp. YT226, 1429 nuc
CP011947.2862162.2863642, *Haloferax gibbonsii*,
AB037474.1.1472, *Haloferax alexandrinus*, 1472 nuc
AOLL01000014.72909.74350, *Haloferax alexandrinus*,
AOLJ01000012.98386.99827, *Haloferax gibbonsii*,
AB663378.1.1473, *Haloferax lucentense*, 1473 nuc
HQ438275.1.925, *Haloferax* sp. PR13, 925 nuc
GQ478070.1.1448, *Haloferax* sp. CT4-2, 1448 nuc
GQ478076.1.1448, *Haloferax* sp. CS1-9, 1448 nuc
FJ746722.1.1473, *Haloferax* sp. H4, 1473 nuc
AY425724.1.1404, *Haloferax volcanii*, 1404 nuc
CP001956.2770163.2771635, *Haloferax volcanii* DS2
CP001956.1598192.1599664, *Haloferax volcanii* DS2
AOHU01000104.40516.41975, *Haloferax volcanii* DS2
AB074566.1.1472, *Haloferax volcanii*, 1472 nuc
KC354391.1.933, *Haloferax* sp. B
KF650666.1.925, *Haloferax volcanii*, 925 nuc
KC354398.1.902, *Haloferax volcanii*, 902 nuc
KC354382.1.927, *Haloferax lucentense*, 927 nuc
KC354384.1.931, *Haloferax volcanii*, 931 nuc
KC354396.1.934, *Haloferax* sp. BAB-2
KC354394.1.946, *Haloferax* sp. BAB-2
DQ915828.1.901, *Haloferax volcanii*, 901 nuc
KC354392.1.928, *Haloferax volcanii*, 928 nuc
KX197500.1.1415, *Haloferax* sp., 1415 nuc
KF321798.1.1379, *Haloferax* sp. A2FP3cs, 1379 nuc
JX067386.1.1435, *Haloferax* sp. C27, 1435 nuc
JF781313.1.903, *Haloferax* sp. KPS1, 903 nuc

silva
high quality ribosomal RNA databases

[Home](#) [SILVAngs](#) [Browser](#) [Search](#) [ACT](#) [Download](#) [Do](#)

SILVA

Welcome to the SILVA rRNA database project

A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.

SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (*Bacteria*, *Archaea* and *Eukarya*).

SILVA are the official databases of the software package ARB.

For more background information → [Click here](#)

SILVAngs

silva
angs

Check out our service for Next Generation Amplicon data

SILVA Alignment, Classification and Tree (ACT) Service

The SILVA ACT service combines alignment, search and classify as well as reconstruction of trees in a single web application.

SILVA ACT is available at: → www.arb-silva.de/act

SILVA Tree Viewer

The SILVA Tree Viewer is a web application to browse and query the SILVA guide trees.

A technical preview is available at: → www.arb-silva.de/treeviewer



<https://today.uconn.edu/school-stories/papke-archaea/#>

Traceability lessons from other technologies?

- **Music & video industry? passive use**
 - Bioinformatics analysis = highly interactive
 - Types of analysis = unlimited
- **Cloud genomics: closed system, defined analysis**
 - Perhaps useful for specialized fields, but requires prior knowledge of what user wants to do with NSD

Blockchain: what did we learn?

Requirements	CBD challenges
Significant costs to build	Funding? Users?
Probably not compatible with INSDC	Re-invent the wheel, lack of connectivity, unable to publish
Would NSD would stay in the blockchain? Useres perform their analyses here?	CBD users are very diverse with thousands of analyses tools. Very difficult to know what do users want to do with data.
Size limits of the blockchain	Novel NSD cannot be simplified → Global dataset too big for Blockchain
Cost-benefit analysis	Human NSD = known value CBD NSD = unknown demand
Users need to be motivated to use blockchain system	Bitcoin has no value outside of blockchain NSD is still NSD outside of blockchain. Value is not lost.

Technical observations

- 16% of entries reported country information
- 44% of entries did not report a country but should have
- 4% of NSD entries linked to GR
- Digital objects (like IRCCs) could theoretically be linked to NSD but not PIC/MAT (unpublished docs)
- Information in patent application is not fully transferred to INSDC (e.g., country information)

Key lessons learned

- Majority of NSD in INSDC is out of CBD scope
- INSDC is the core infrastructure for sequence data (99.9%)
- Database ecosystem is complex and growth of NSD is exponential... BIG data
- Traceability in databases is possible and already being done. After download from databases, traceability is a LOT harder.
- If NSD from INSDC is used in a patent application this is not traceable
- Blockchain cannot handle size of the INSDC dataset

Acknowledgments



Fabian Rohden
Germany / Canada



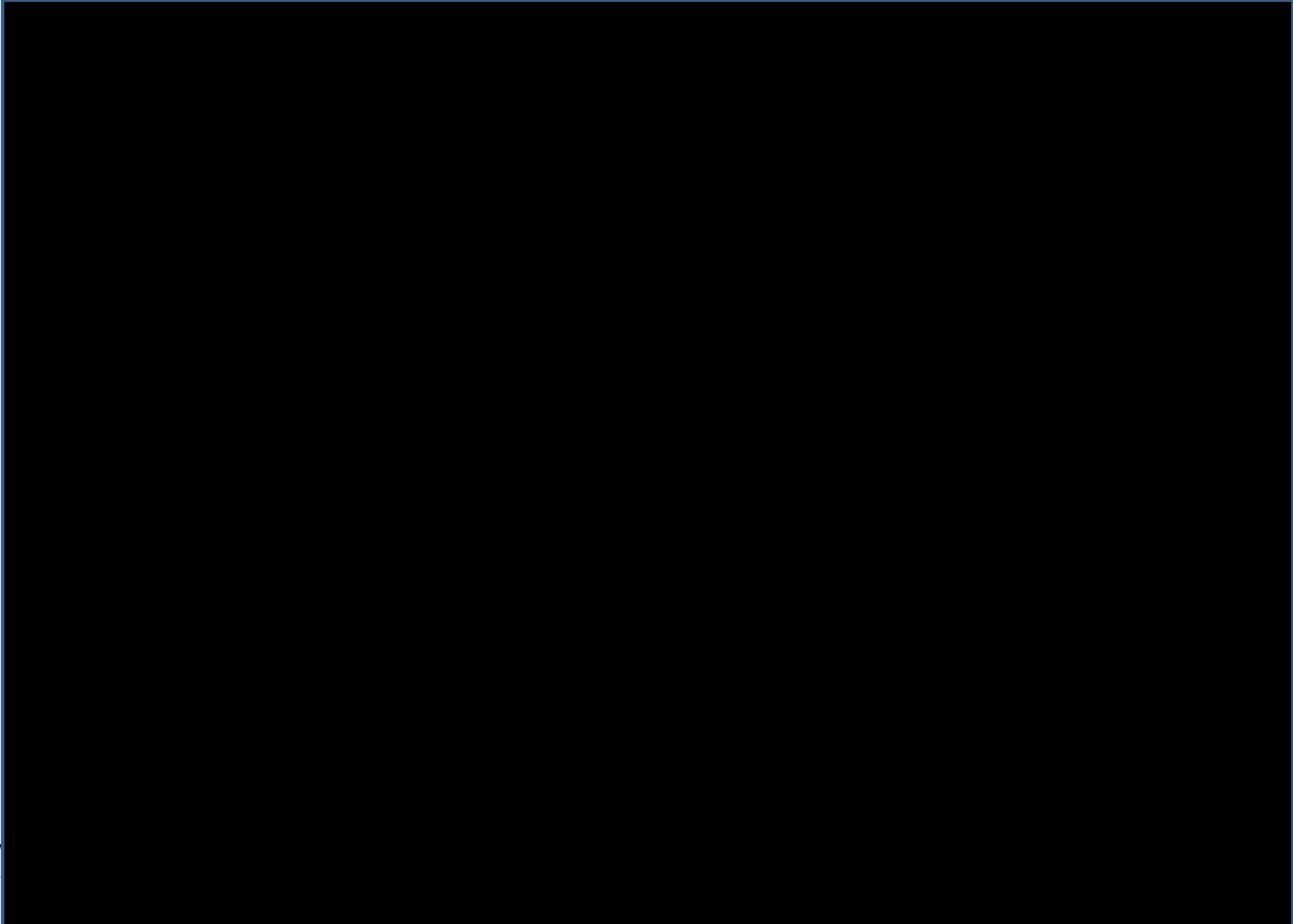
Sixing Huang
China / Germany



Gabriele Dröge
Germany

Katharine Barker, USA
Walter G. Berendsohn, Germany
Jonathan A. Coddington, USA
Manuela da Silva, Brazil
Jörg Overmann, Germany
Ole Seberg, Denmark
Michelle van der Bank, South Africa
Xun Xu, China





DSR



Databases not in the NAR issue: GISAID example

- Users must register, identify themselves and affiliation
- No data sharing with non-GISAID users
- Use/re-use requires consent and usually collaboration
- No traceability outside of database
- Thousands of users (not millions)
- Enforcement of policy is unclear
- Data silo (data analysis and comparisons limited, no data flow to other databases)

Why do scientists worry about “data silos”? Because you never know when you need to connect new types of data....

Viral Sequences

All (338,993)
Sequences (41,237)
Raw reads (162,767)
Sequenced samples (122,092)
Studies (183)
Genes (22)
Browser (1)
Variants (12,691)

Host Sequences

All (4,784)
Human studies (controlled access) (4)
Human reads (consented for full access) (3,237)
Other species reads (1,526)
Association studies (17)

Expression

All (84)
Gene expression (4)
Gene expression experiments (23)
Single cell expression (4)
Single cell expression experiments (19)
Protein expression experiments (34)

Proteins

All (969)
Protein sequences (79)
Protein families (315)
Protein structures - Knowledge Base (7)
Protein structures (260)
Electron microscopy density maps (294)
Electron microscopy public image archive (14)

Biochemistry

All (2,559)
Pathways (16)
Interactions (2,052)
Complexes (29)
Compound document (0)
Drug targets (453)
Metabolomics experiments (1)

Imaging

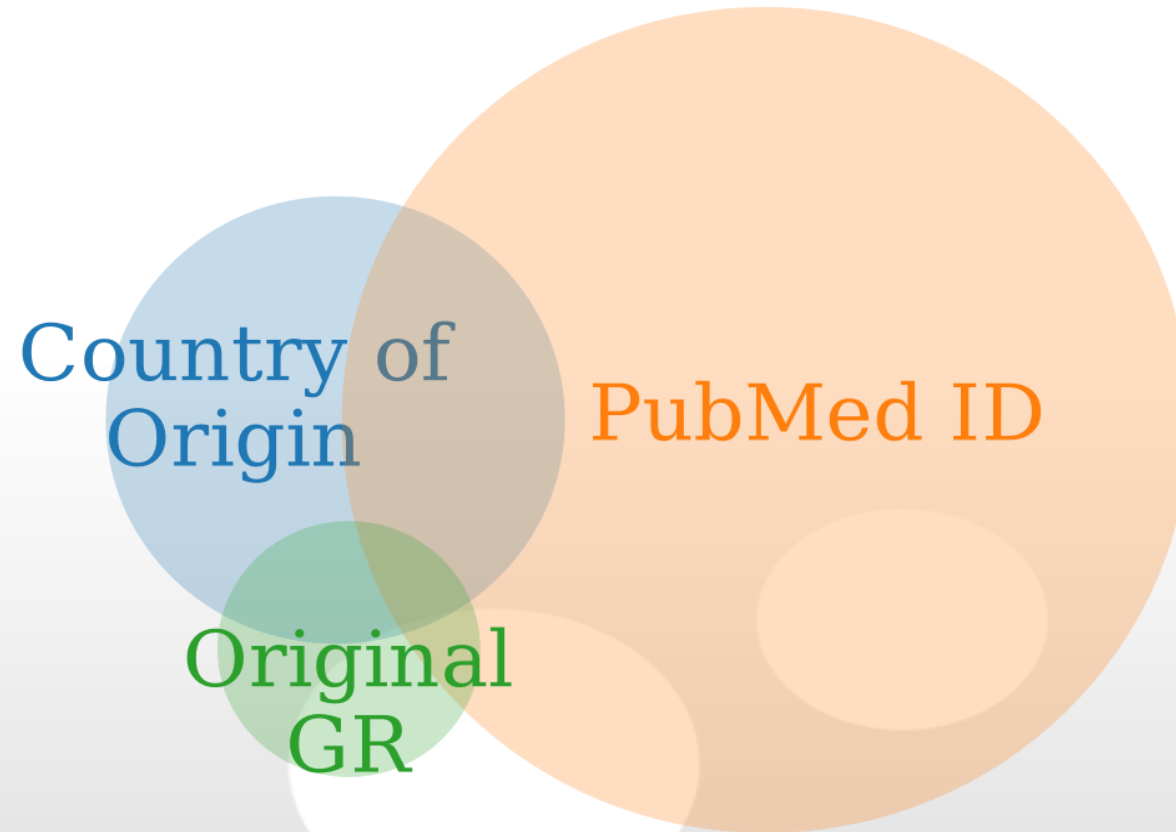
All (16)
Images (2)
Electron microscopy public image archive (14)

Literature

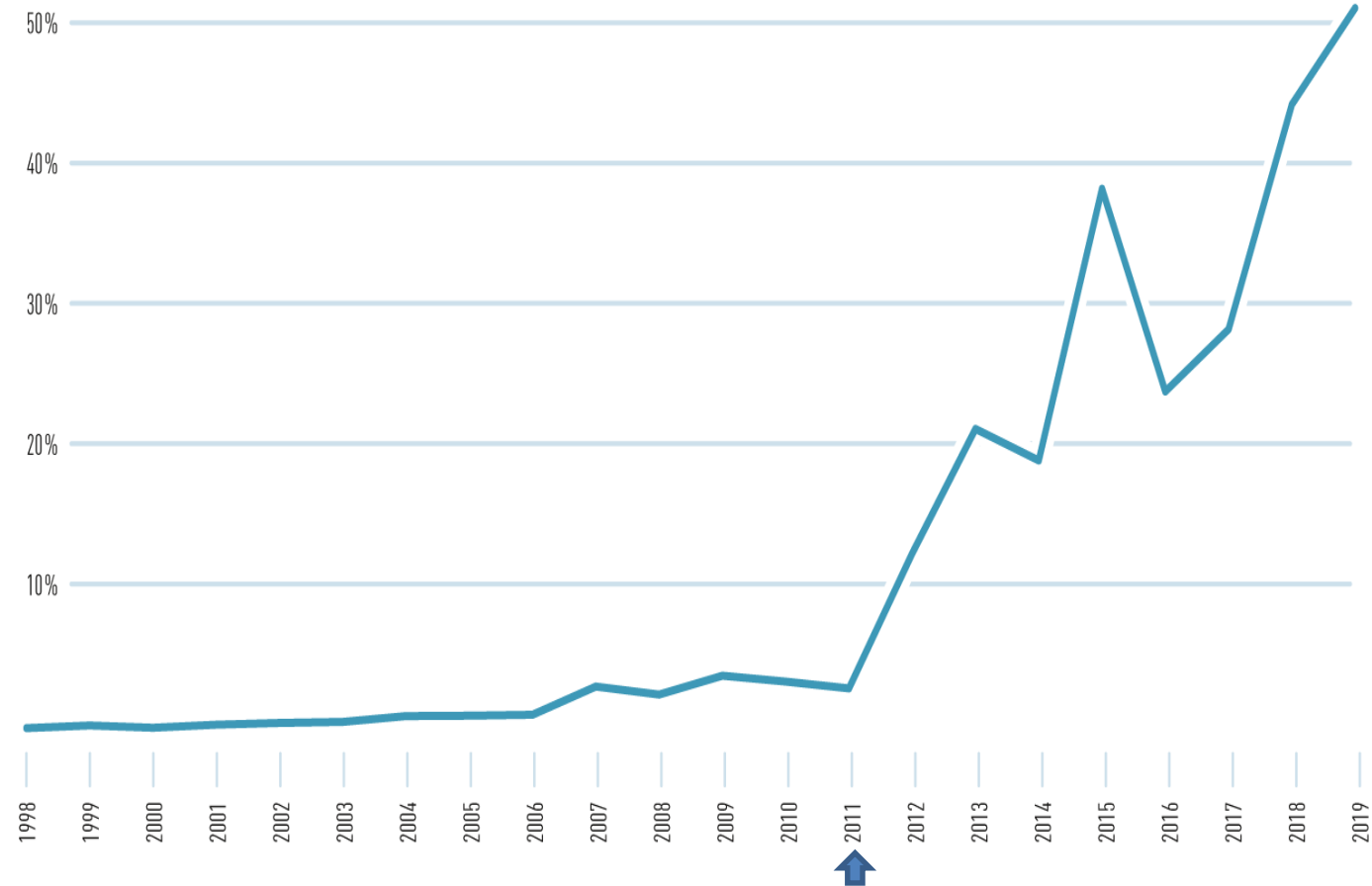
All (234,955)
Coronaviruses (120,779)
Diseases (106,991)
Related viruses and diseases (2,159)
Genes, receptors and antibodies (5,022)
Supplementary material (4)

<https://www.covid19dataportal.org/>

What kinds of data can be associated with NSD?



The amount of country-tagged-NSD has increased over time



Tell me more: 16% of NSD has country information?

Is the reported country information accurate?

Number of samples	Incorrect country tag	Correct country tag	No information
150	0	86 57%	64 43%

If country information was missing, should it have been reported?

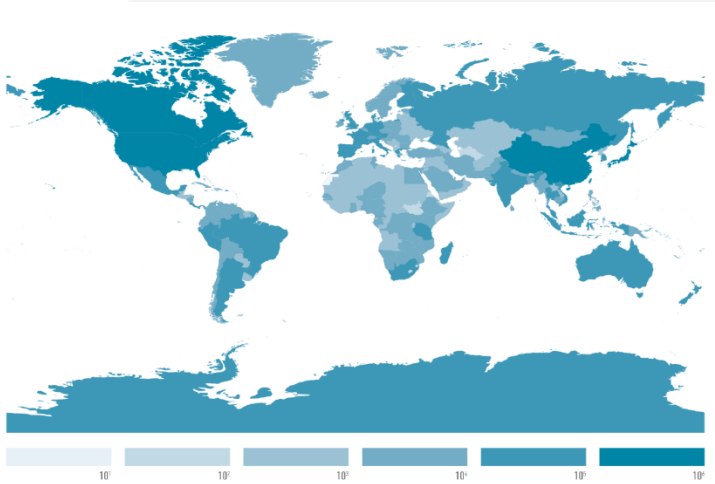
Number of samples	Publication accessible	Country could have been reported?	Origin of these 123 entries
660	282 43%	Yes: 124 44% No: 158 56%	48% other 46% environment 6% human microbial

For these 124 non-country-tagged entries:
28% US, 12% China, 3% Canada, 2% Japan,
3% India, 5% Brazil, 5% other

→ Looks like carelessness not deception

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

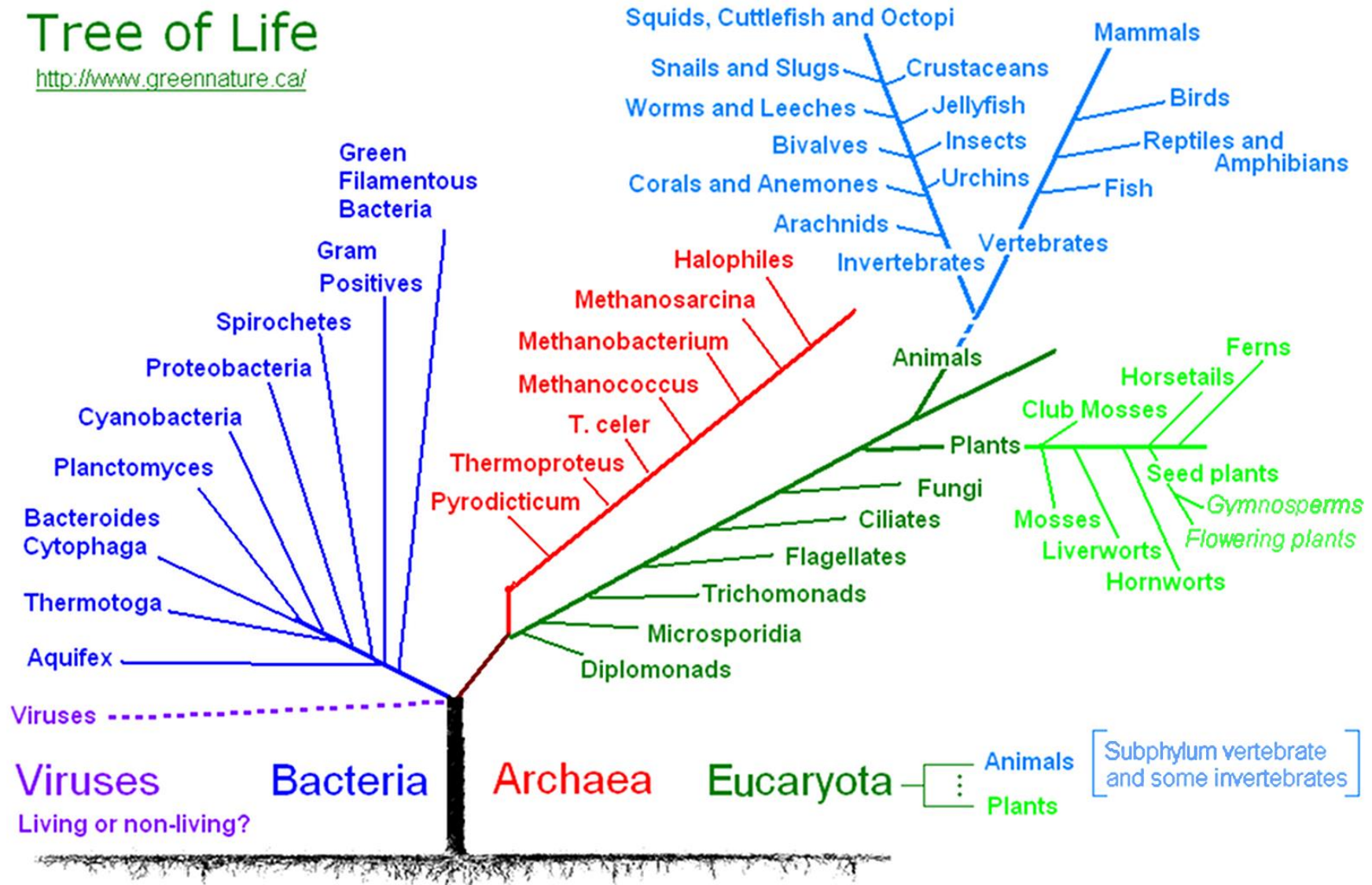
1. China	18.23 %
2. United States	17.39 %
3. Canada	9.10 %
4. Japan	7.24 %
5. India	3.46 %
6. Australia	2.66 %
7. Mexico	2.54 %
8. Brazil	2.30 %
9. Germany	1.83 %
10. Spain	1.58 %



DSI from every branch on the tree of life

Tree of Life

<http://www.greennature.ca/>



Entries vs. Information content

212,000,000 entries

321,000,000,000 bases

