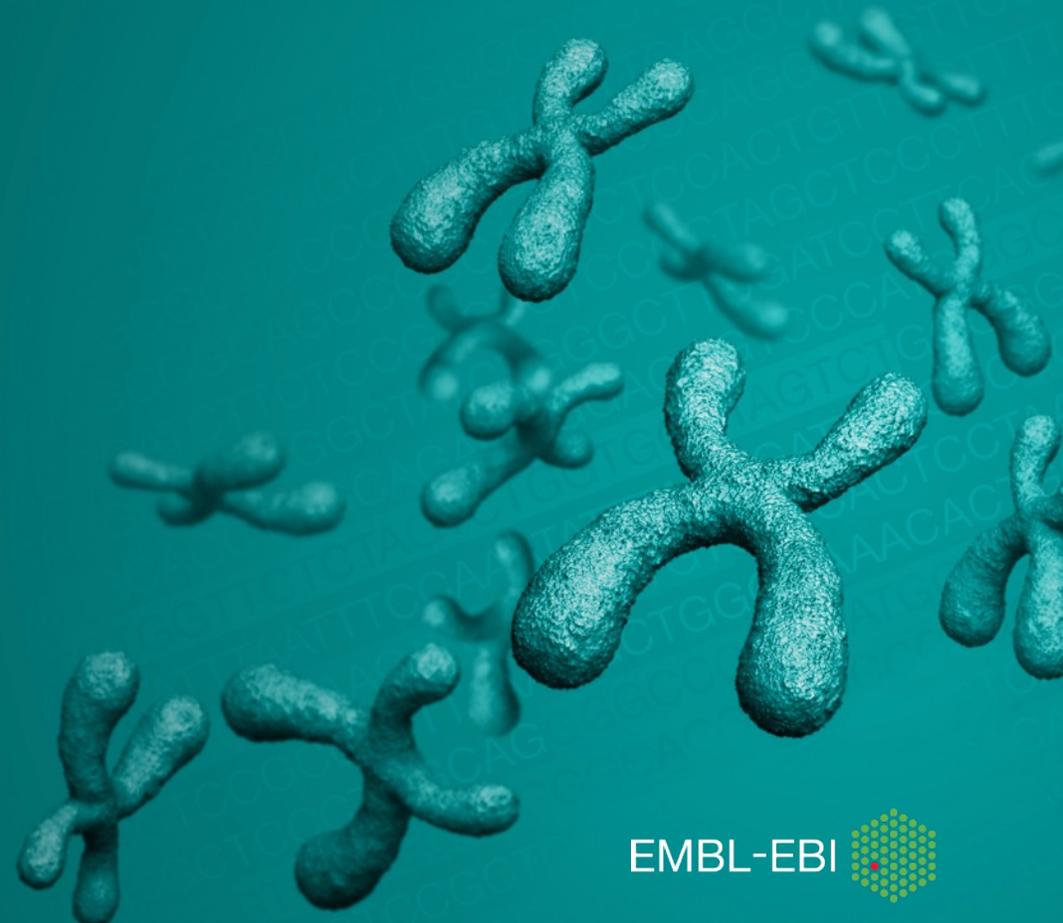


# Open access & the International Nucleotide Sequence Database Collaboration

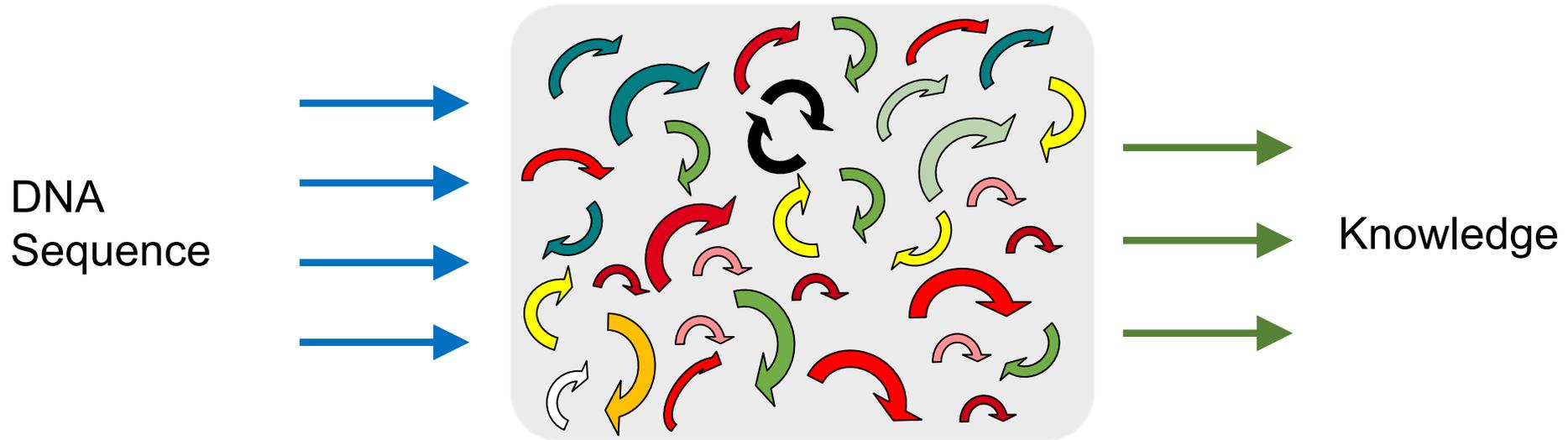
Guy Cochrane



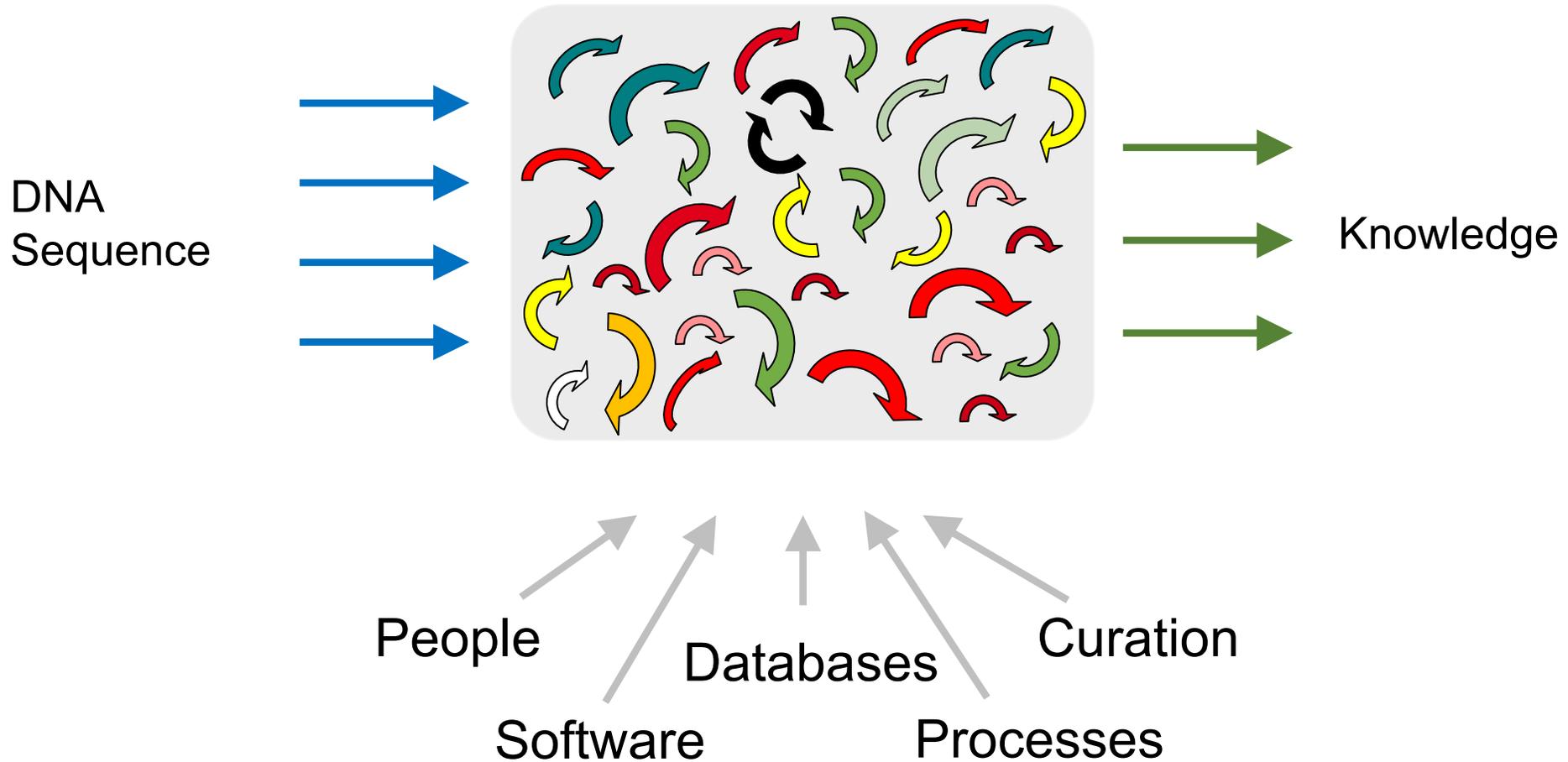
“If I have seen further than others, then it is by standing on the shoulders of giants.”

*Isaac Newton*

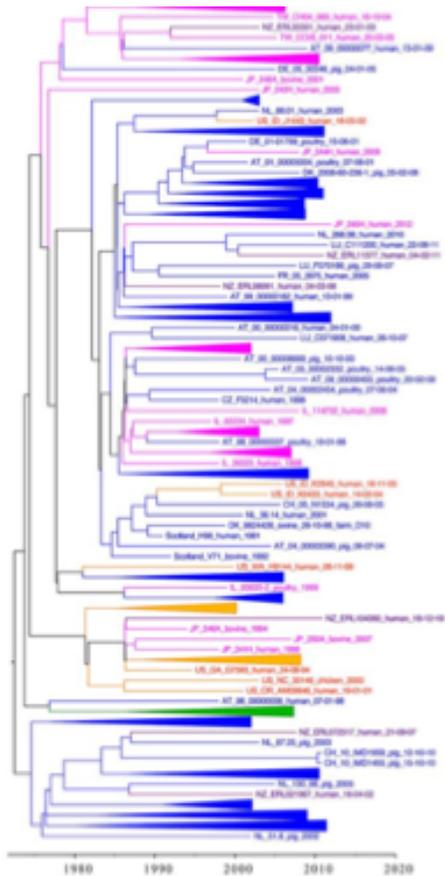
# The life science “machine”



# The life science “machine”



Open data are essential



Tree building

Pimplapas Leekitcharoenphon *et al.* Appl. Environ. Microbiol. 2016; doi:10.1128/AEM.03821-15

Results for job ncbiblast-R20190902-082709-0092-73324249-p1m **BLAST search**

Summary Table | Tool Output | Visual Output | Result Summary | Submission Details

**Selection:** Select All | Invert | Clear

**Apply to selection:**

**Annotations:** Show | Hide

**Alignments:** Show | Hide

**Entries:** Download | In

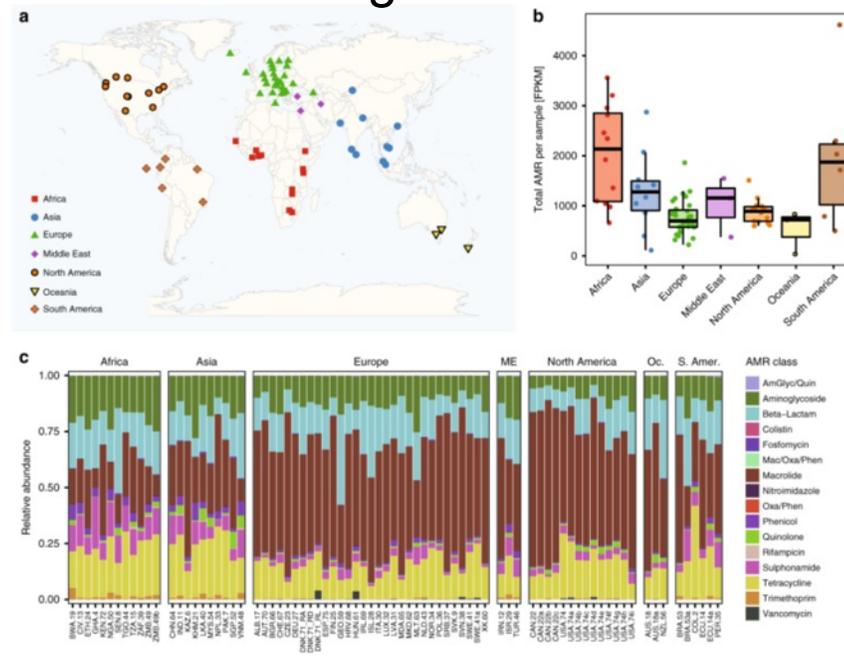
fasta

**Tools:** Launch

Clustal Omega

Align.	DB-ID	Source	Length	Score (Bits)	Identities %	Positives %	E <sub>0</sub>
✓1	EM_HUMLWKW01001345	Homo sapiens isolate KOREF chromosome 1 KOREF.1345, whole genome shotgun sequence.	51268	1589.2	100.0	100.0	0.0
✓2	EM_HUM.BN000005	TPA: Homo sapiens SMP1 gene, RHD gene and RHCE gene	315242	1589.2	100.0	100.0	0.0
✓3	EM_HUM.AL121994	Human DNA sequence from clone RP4-781L3 on chromosome 1p34.3-36.11	47165	1589.2	100.0	100.0	0.0
✓4	EM_HUM.AL031432	Human DNA sequence from clone RP3-465N24 on chromosome 1p35.1-36.13	129747	1589.2	100.0	100.0	0.0
✓5	EM_PAT.H518768	Sequence 389 from Patent EP2164991.	47165	1589.2	100.0	100.0	0.0
✓6	EM_OM.AC206460	Pongo abelli BAC clone CH276-115G23 from chromosome 1, complete sequence.	226780	1378.2	95.3	95.3	0.0
✓7	FM_011169000	Marinae faecal/urine complete genome	770662717	801.4	84.4	84.4	0.0

## Metagenomics



Hendriksen RS, *et al.* Nat Commun. 2019; 10: 1124.

## Annotation by reference



plants.ensembl.org/Triticum\_aestivum/Location/Genome

EnsemblPlants - HMMER | BLAST | BioMart | Tools | Downloads | Documentation | Website help

**Triticum aestivum (IWGSC)**

Whole genome

Click on the image above to jump to a chromosome, or click and drag to select a region

**Summary**

Assembly	IWGSC, INSDC Assembly <a href="#">GCA_900519105.1</a> , Jul 2018
Database version	97.4
Base Pairs	14,547,261,565

Click on the image above to jump to a chromosome, or click and drag to select a region

**Summary**

Assembly	IWGSC, INSDC Assembly <a href="#">GCA_900519105.1</a> , Jul 2018
Database version	97.4
Base Pairs	14,547,261,565

plants.ensembl.org/Triticum\_aestivum/Location/View?tr=SD:334686420-334786420

EnsemblPlants - HMMER | BLAST | BioMart | Tools | Downloads | Documentation | Website help

**Triticum aestivum (IWGSC)**

Chromosome SD: 334,686,420-334,786,420

Region in detail

Gene Legend

- Gene aligned from TSLC11 on transcripting cDNA
- Gene with evidence from transcripting cDNA
- Gene coding

[https://plants.ensembl.org/Triticum\\_aestivum/Info/Index?db=core](https://plants.ensembl.org/Triticum_aestivum/Info/Index?db=core)

# International Nucleotide Sequence Database Collaboration (INSDC)



## Values

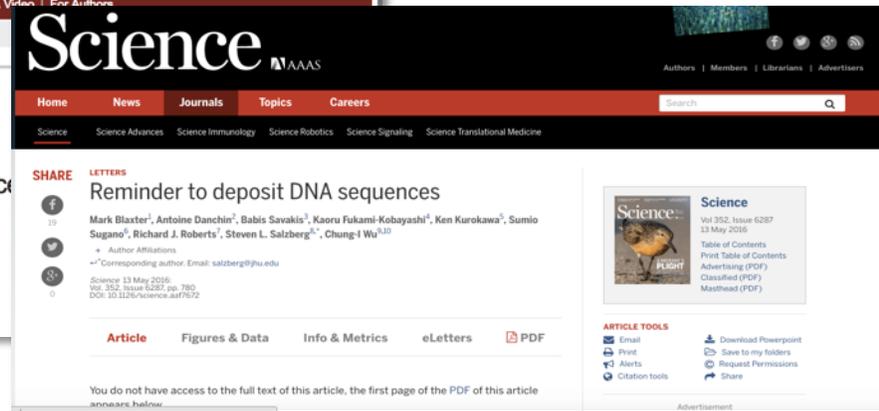
- open access for all
- globally comprehensive
- spanning life science domains
- permanent database of record
- public forum for the scientific process

## Organisation

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



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



## Instruments

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

# International Nucleotide Sequence Database Collaboration (INSDC)

```
ID MK791304; SV 1; linear; genomic DNA; STD; VRT; 679 BP.
XX
AC MK791304;
XX
DT 04-NOV-2019 (Rel. 142, Created)
DT 04-NOV-2019 (Rel. 142, Last updated, Version 1)
XX
DE Narcine maculata isolate CAS-AM-SR cytochrome c oxidase subunit I (COI)
DE gene, partial cds; mitochondrial.
XX
KW .
XX
OS Narcine maculata
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Chondrichthyes;
OC Elasmobranchii; Batoidea; Torpediniformes; Narcinidae; Narcine.
OG Mitochondrion
XX
RN [1]
RP 1-679
RA Arulmoorthy M.P., Sureandiran B.;
RT ;
RL Submitted (13-APR-2019) to the INSDC.
RL Cas in Marine Biology, Annamalai University, Anna Kovil, Parangipettai,
RL Tamil Nadu 608502, India
XX
MD5; f83e14e5913fb53c7d28dc9e9960a043.
XX
##Assembly-Data-START##
CC Sequencing Technology :: Sanger dideoxy sequencing
CC ##Assembly-Data-END##
XX
FH Key Location/Qualifiers
FH
FT source 1..679
FT /organism="Narcine maculata"
FT /organelle="mitochondrion"
FT /map="11 29' N; 79 46' E"
FT /isolate="CAS-AM-SR"
FT /mol_type="genomic DNA"
FT /country="India"
FT /isolation_source="Mudasal Odai landing centre"
FT /collected_by="Arulmoorthy"
FT /collection_date="24-Jan-2019"
FT /sex="female"
FT /tissue_type="fin"
FT /PCR_primers="fwd name: coi f, fwd_seq:
FT tcaaccaaccacaagaacattggcac, rev_name: coi r, rev_seq:
FT tagacttctgggtggcacaagaatca"
FT /db_xref="taxon:1455688"
FT
FT CDS
FT <1..>679
FT /codon_start=3
FT /transl_table=2
FT /product="cytochrome c oxidase subunit I (COI)"
FT /protein_id="QFU19390.1"
FT /translation="YLIFGAWAGMVGTLGSLLRTELSQPGTLLGDDQIYNVIVTAHAF
FT VMIFFMVPMIMIGGFGNWLMLMIGAPDMAPFRMNMMSFWLLPPSFLLLASAGVEAGA
FT GTGWTVPPLAGNIAHAGASVDLTFSLHLGASSILASINFTITINMKSPSITQYQT
FT PLFVWSLITAILLLSLPVLAAIGITMLLTDRLNLTFFDPAGGGDPILYQHLFWFFGH
FT PEV"
XX
SQ Sequence 679 BP; 197 A; 166 C; 109 G; 207 T; 0 other;
tatacttaat tttcgggtgcc tgagcaggaa tagtaggaac cggcttaagt ttattaatcc 60
gaactgaact tagtcaacca ggcactctat tggggcgtga ccaatttat aatgtaatcg 120
tgactgctca tgcattcgtt ataattctct ttatagttat accaattata atcggcgggt 180
ttggaaactg attgataacc ctcataattg gagcccaga catagccttt ccaagataaa 240
ataatataag cttctgactg ttaccaccat ctttctctct tctattagct tcaagcaggag 300
tagaggccgg agcaggcaca ggatgaacag tttatccccc ccttgcctga aatattgctc 360
atgccggagc atcagtcgac ttaactatct tctcactgca cttagcaggg gctcttcaaa 420
tctagcctc tattaatctt atcaccacaa ttattaacat aaaatcacca caaattacac 480
aataccaaac accacttttt gtgtgatcat tacttaacac tgcaattctt ctactcttat 540
cattaccagt actagcagca ggaattacaa tacttctgac agaccgaaat cttaacacca 600
cattcttga cccagctgga ggaggtgacc caattctcta tcaacattta tcttgattct 660
ttggccacc cagaagtcta
//
```

```
TATACTTAATTTTCGGTGCCTGAGCAGGAATAGTAGGAACCGGCTTAAGTTTATTAATCC
GAAGTGAACCTTAGTCAACCAGGCACCTCTATTGGGCGATGACCAAATTTATAATGTAATCG
TGACTGCTCATGCATTCGTTATAATCTTCTTTATAGTTATACCAATTATAATCGGCGGGT
TTGGAAACTGATTGATACCCCTCATAATTGGAGCCCCAGACATAGCCTTTCCACGAATAA
ATAATATAAGCTTCTGACTGTTACCACCATCTTTCCTTCTTCTATTAGCTTCAGCAGGAG
TAGAGGCCGGAGCAGGCACAGGATGAACAGTTTATCCGCCCTTGCTGGAAATATTGCTC
ATGCCGGAGCATCAGTCGACTTAACTATCTTCTCACTGCACTTAGCAGGGGCCCTTTCAA
TCCTAGCCTCTATTAATTTTATCACCACAATTATTAACATAAAAATCACCATCAATTACAC
AATACCAAACACCAC'TTTTGTGTGATCATTACTTATCACTGCAATTCCTTCTACTTCTAT
CATTACCAGTACTAGCAGCAGGAATTACAATAC'TTCTGACAGACCGAAATCTTAAACCA
CATTCTTCGACCCAGCTGGAGGAGGTGACCCAATCTCTATCAACATTTATCTGATTCT
TTGGCCACCCAGAAGTCTA
```

Sequence

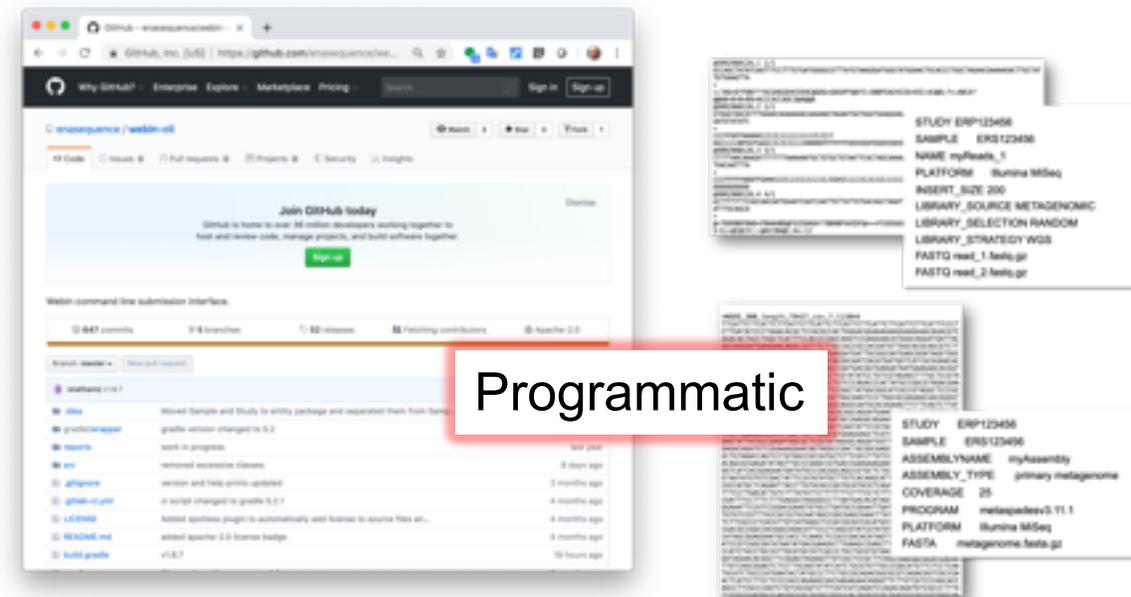
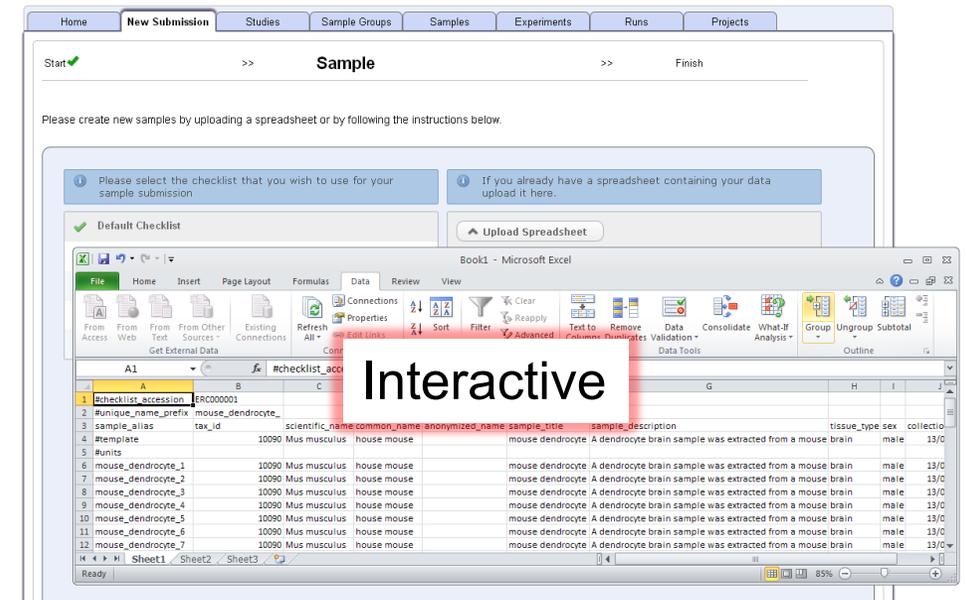
- Clerical information
- Taxonomy
- Literature
- Cross-references
- Source information
- Biological features



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

# The data submission process

- 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



# Presentation

The collage displays various parts of the ENA website:

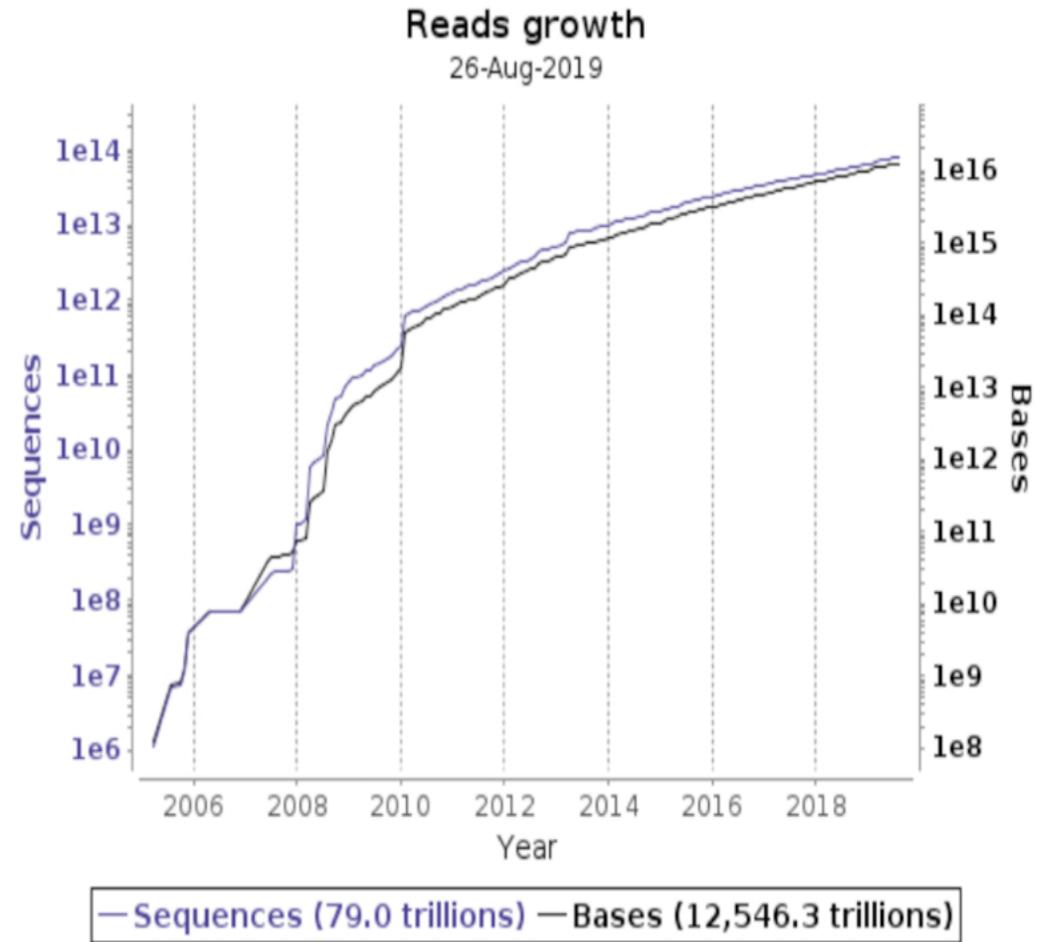
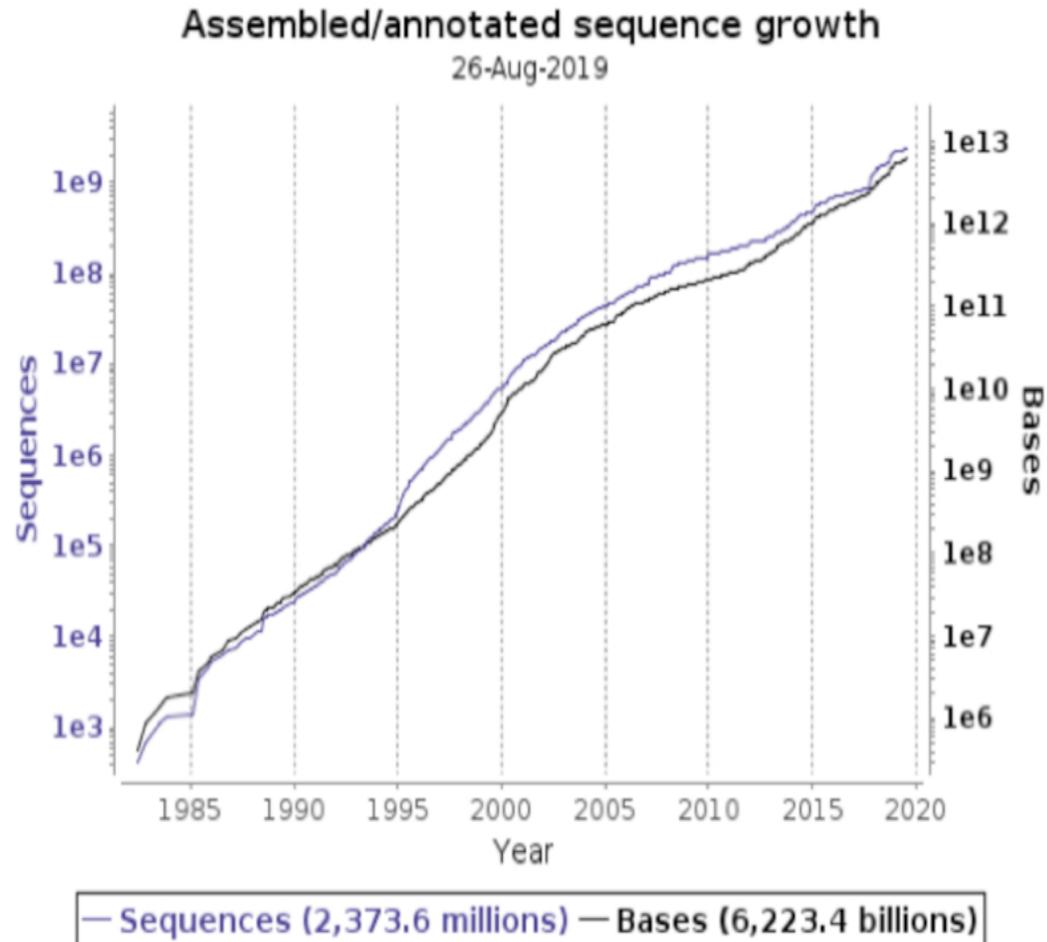
- Home Page:** Features the ENA logo, navigation menu (Home, Submit, Search, Rulespace, About, Support), and a search bar with examples like "BN000065, histone".
- Advanced Search:** Shows a query input field with "geo\_circ(52.7135,1.2305,1000000.0000)" and a "Build Query" section with a map of Europe. The map shows a red circle centered on the UK/Ireland region.
- Submission Form:** A multi-step form for reporting a service issue, with steps numbered 1 through 6. It includes fields for "Submitter info", "The issue is", and "My query is...".
- My Rules Table:** A table listing user-defined search rules.

Name	Description	Creation Time	Updated Time	View	Edit	Delete	Search
UK and Ireland Sheep	Sus Scrofa from UK and Ireland	25/06/2019 13:16:10	25/06/2019 13:16:10				
Salmonella data from Feb 2019	Salmonella data limited by date	25/06/2019 13:20:59	25/06/2019 13:20:59				

- Search
  - Sample characteristics
  - Function
  - Sequence similarity
- Browse
  - Study
  - Taxonomy
  - Sequencing method
  - Historic versions
- Filter
  - Reusable/shareable rules
  - Synchronisation with latest data
- Retrieve
  - Download tools



# Data growth

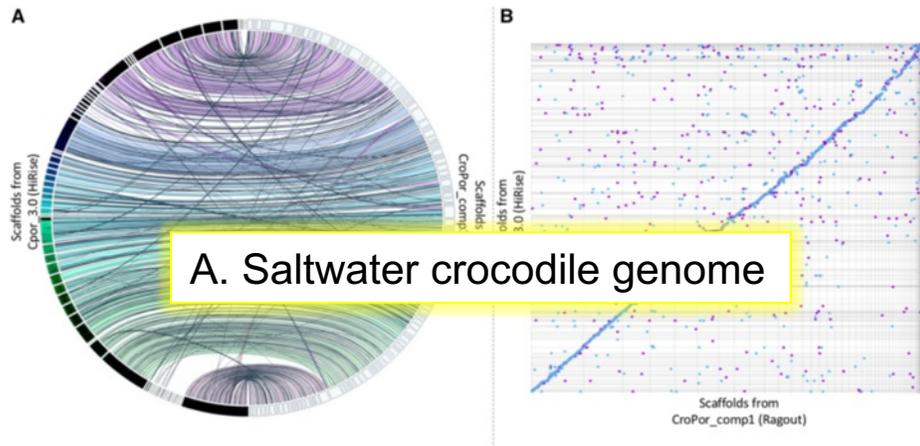


*Nucleic Acids Research*, Volume 48, Issue D1, 08 January 2020, Pages D70–D76, <https://doi.org/10.1093/nar/gkz1063>

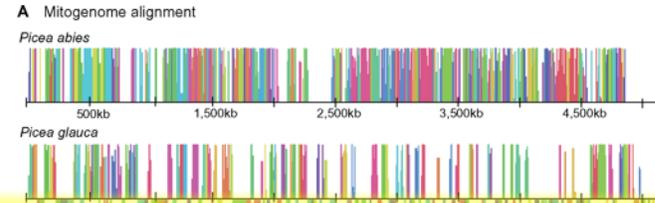
© The Author(s) 2019. Published by Oxford University Press on behalf of *Nucleic Acids Research*. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

# Direct use

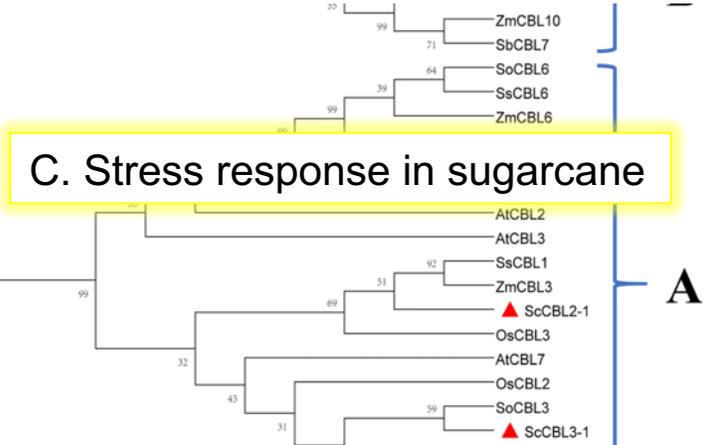
- 1,306 publications citing sequence accessions since the 1<sup>st</sup> Jan. 2020



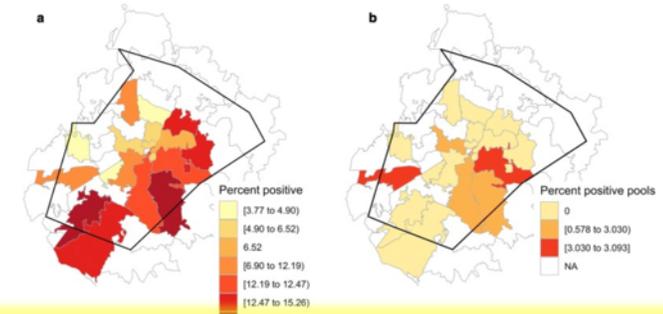
**A. Saltwater crocodile genome**



**B. Norway spruce evolutionary biology**



**C. Stress response in sugarcane**



**E. Mosquito-borne heartworm in dogs**

SpringerLink

Published: 10 January 2020

Optimizing the Reduction of Molybdate by Two Novel Thermophilic Bacilli Isolated from Sinai, Egypt

Ali M. Saeed, Hayam A. E. Sayed & Elnas H. El-Shatoury

**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/>

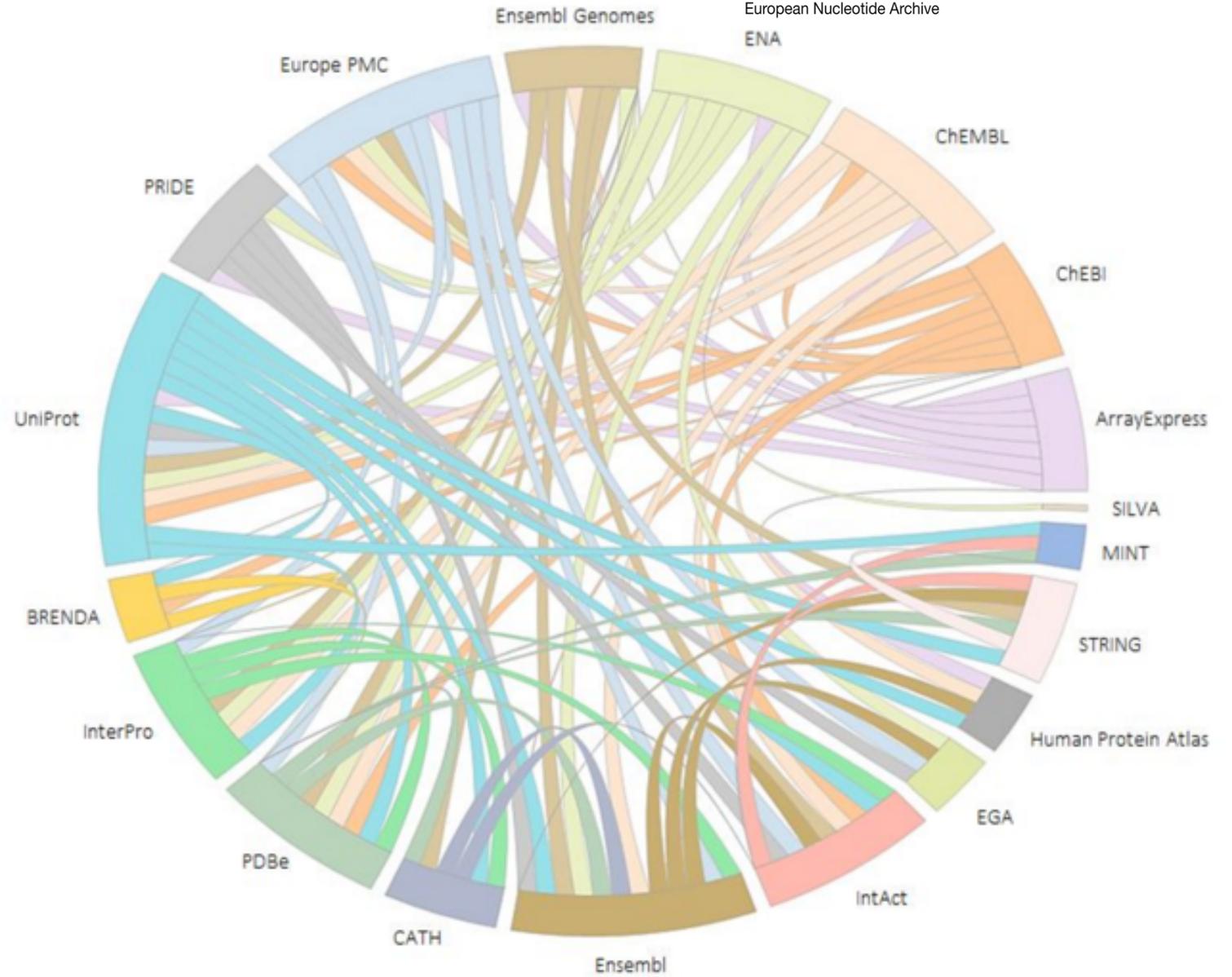
**B.** Sullivan AR et al., Genome Biol Evol. 2020 Jan;12(1) 3586-3598. doi:10.1093/gbe/evz263. PMID: 31774499; © The Author(s) 2020; <http://creativecommons.org/licenses/by-nc/4.0/>

**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/>

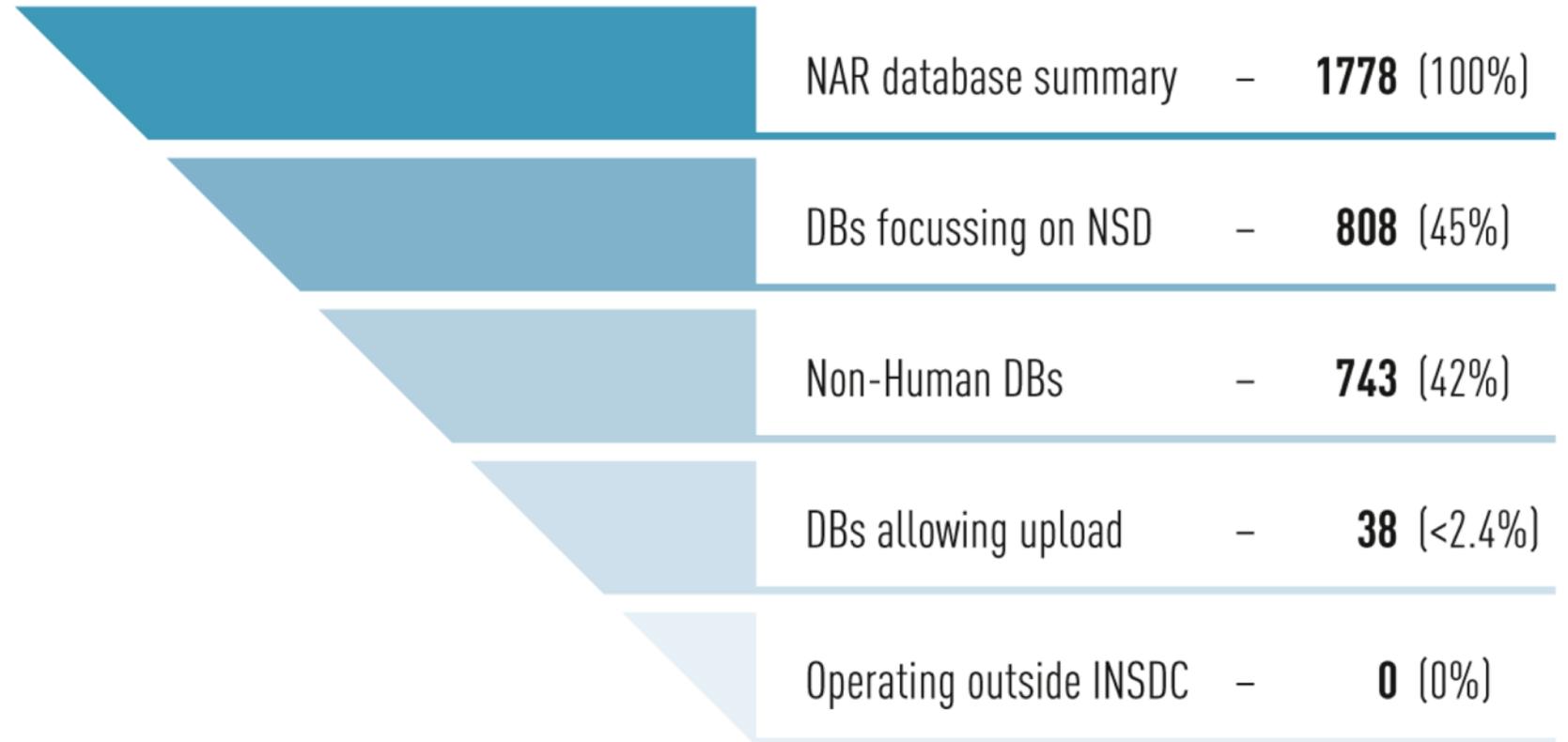
# Data reach



# Further reach



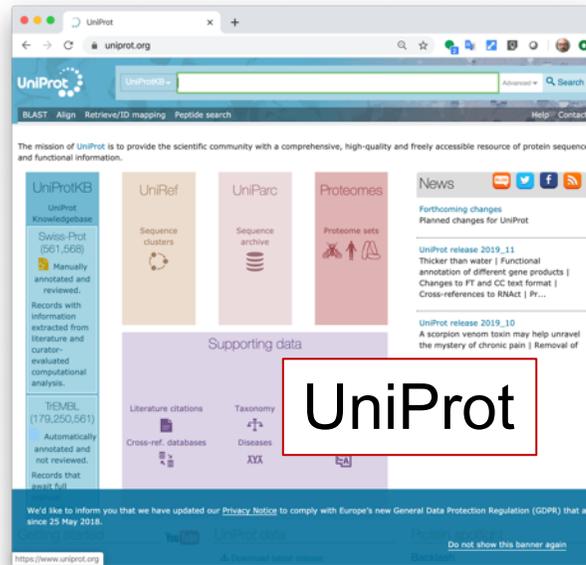
## Public database inventory



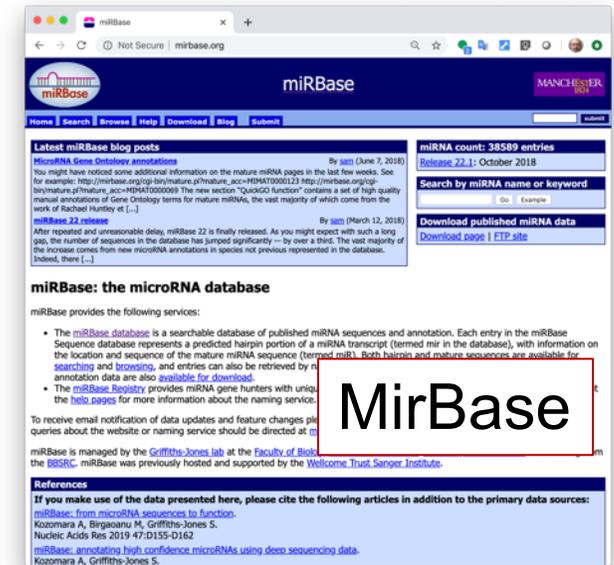
Rohden F, Huang S, Dröge G, Hartman Scholz A, and contributing authors (2019). Combined study in DSI in public and private databases and DSI traceability. <https://www.cbd.int/abs/DSI-peer/Study-Traceability-databases.pdf>

# Secondary databases

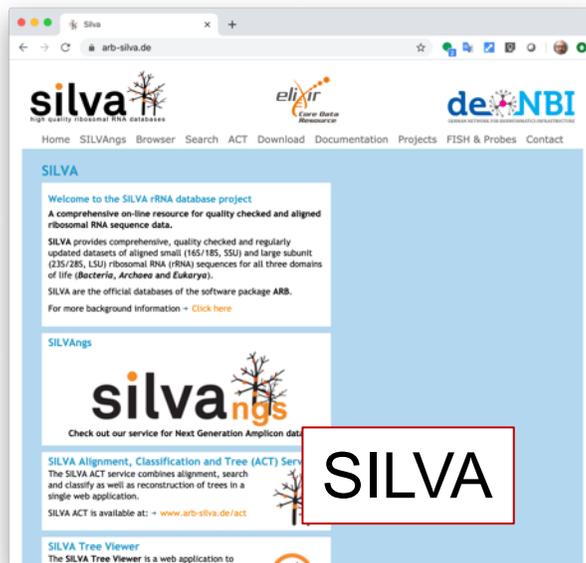
- Connected to INSDC services and content
- More specialist services to user communities



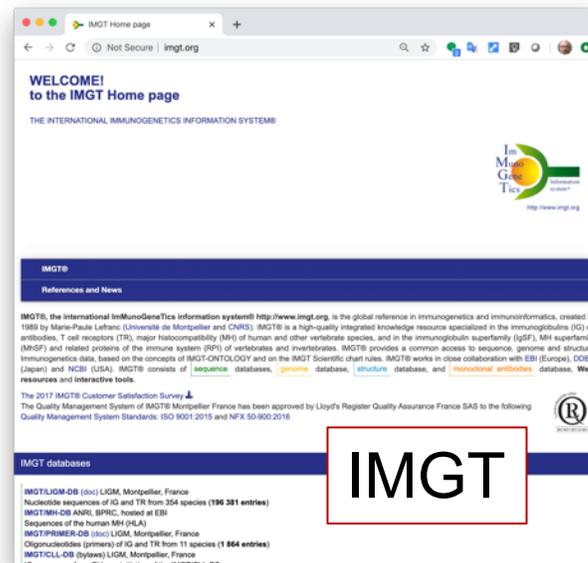
<https://www.uniprot.org/>



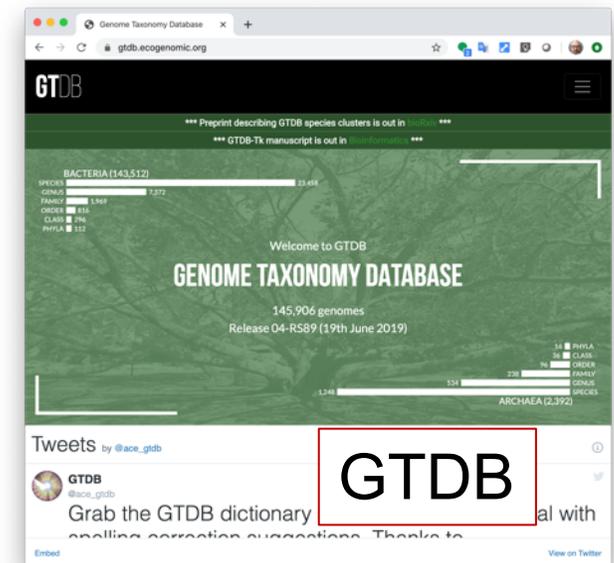
<http://www.mirbase.org/>



<https://www.arb-silva.de/>

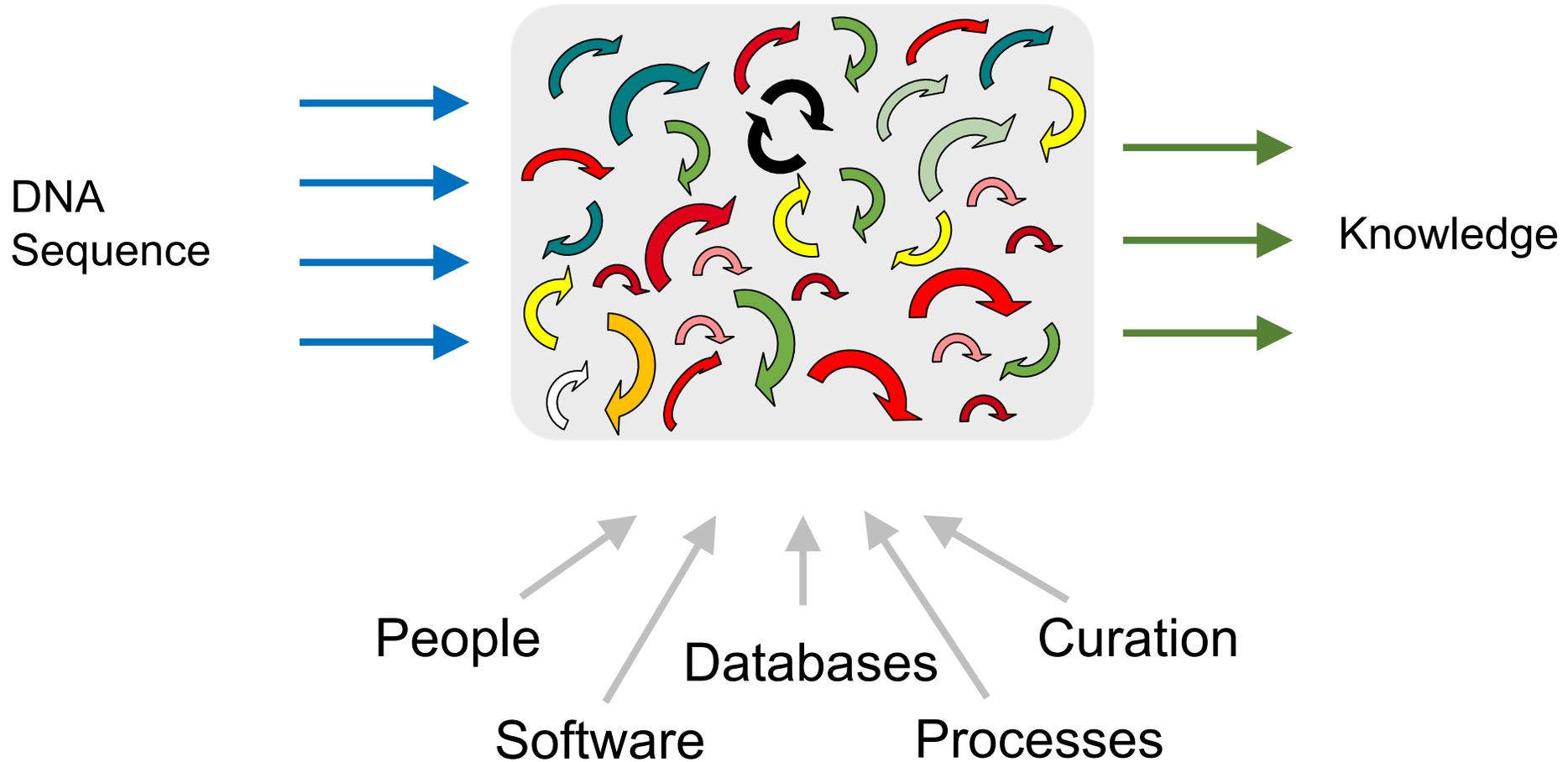


<http://www.imgt.org/>



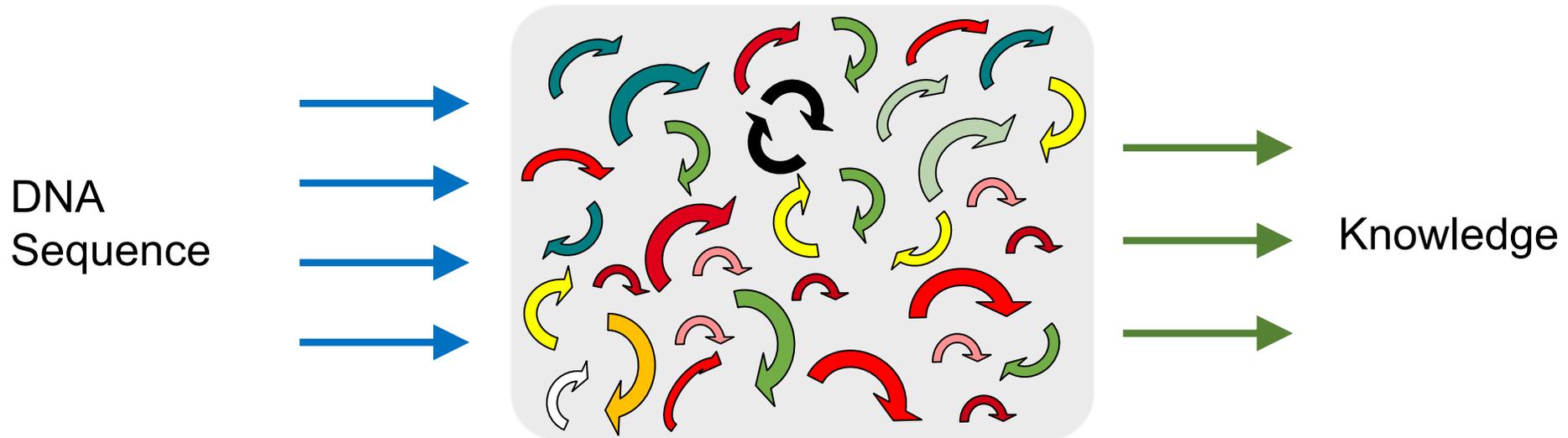
<https://gtdb.ecogenomic.org/>

# The life science “machine”



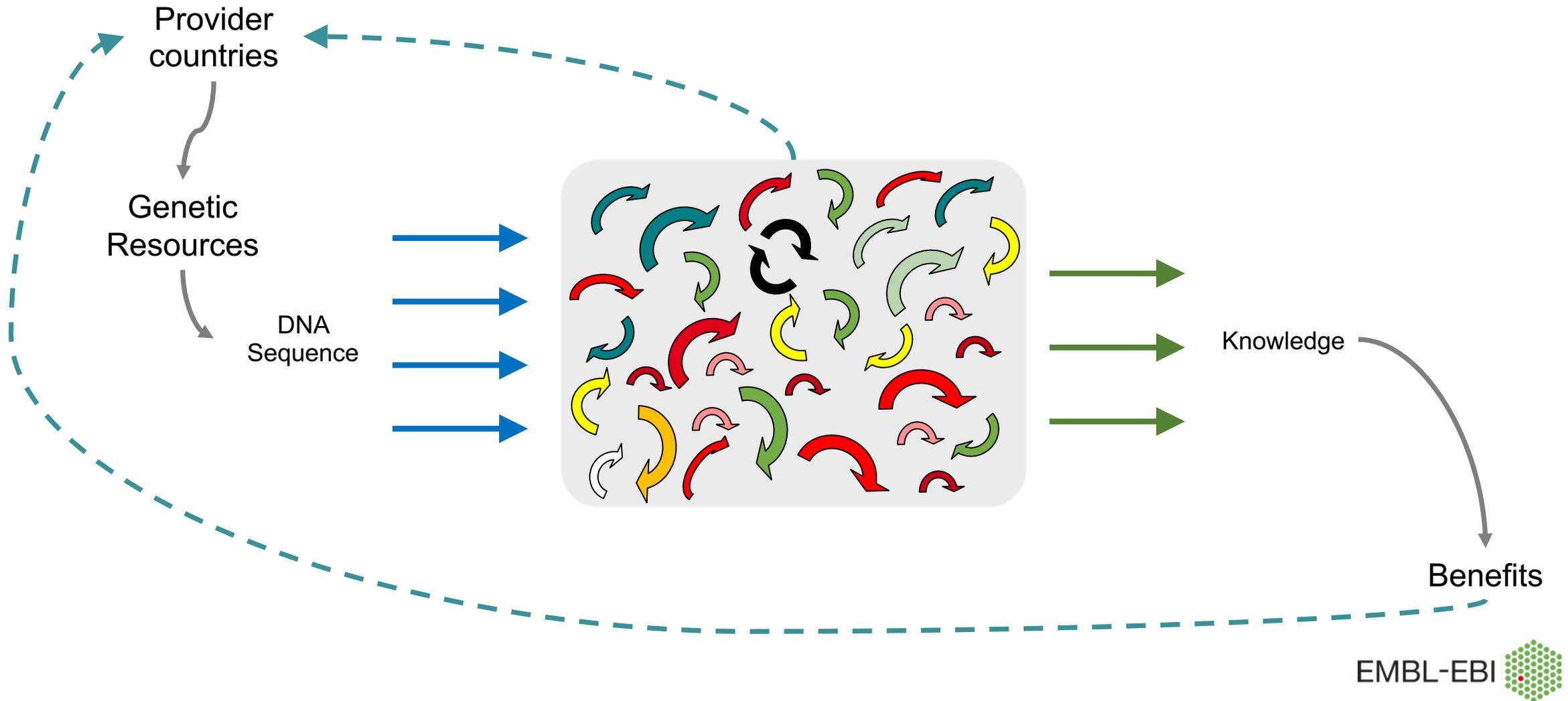
# The life science “machine”

- High productivity drives scientific progress



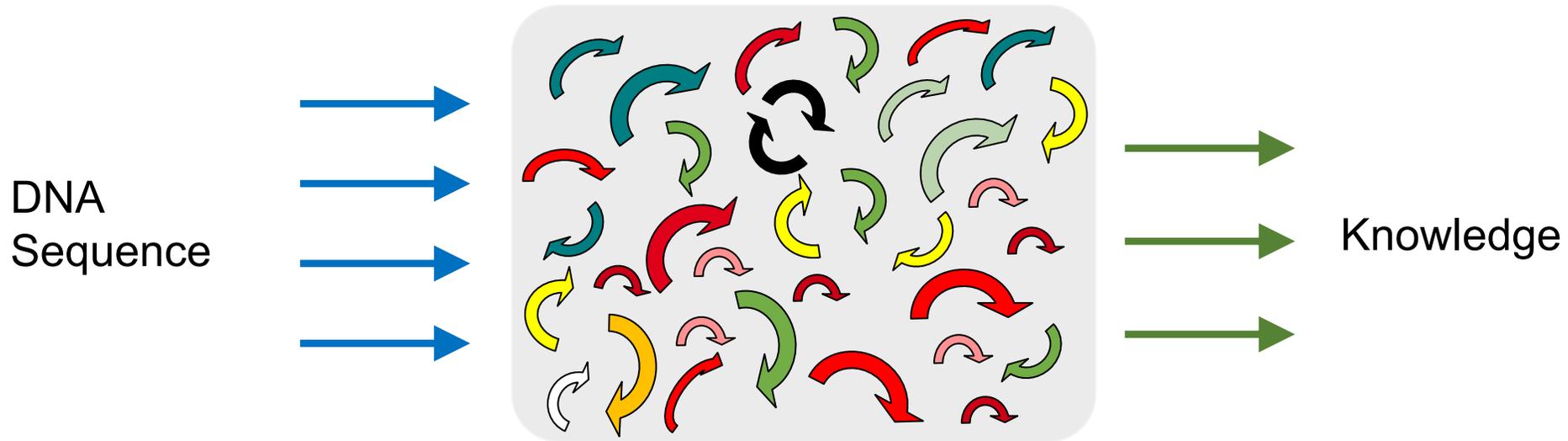
# Productivity in the machine

- High productivity drives ABS



# Productivity in the machine

- Understand risks
- Assert core values
- Identify opportunities



# What if.. visibility of sequence were reduced?

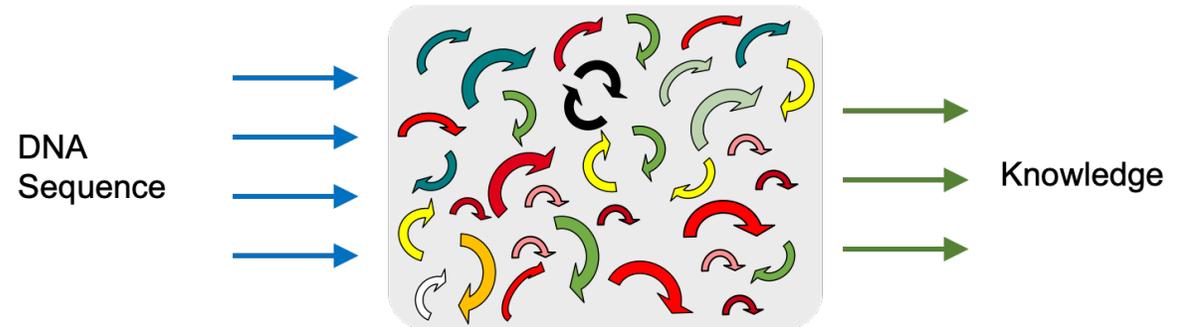
- E.G. users had to register and log-in
- Agreement not to redistribute in any form



- Researchers couldn't provide evidence trails for their findings
- Secondary databases wouldn't be able to show data to users



- Productivity drops to near-zero



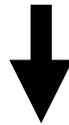
**Critical property of machine:  
Data must be visible**

# What if.. each sequence travelled with unique terms of use?

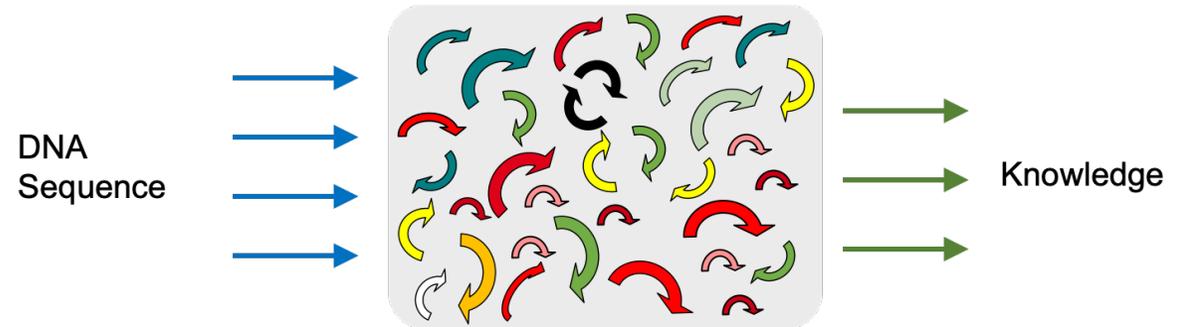
- E.G. license connected to each record, many different licenses



- Immediacy of reuse would be lost, researchers would have significant clerical work
- Secondary databases would either propagate license (impractical or impossible) or reject



- Productivity reduced substantially



**Critical property of machine:**

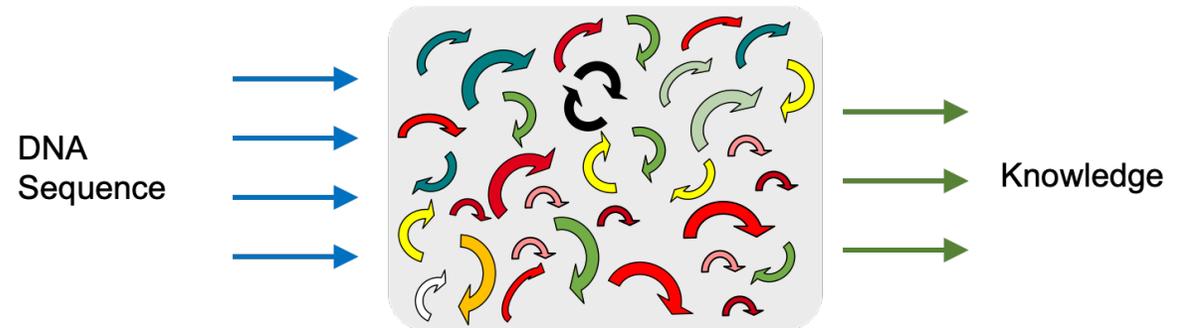
**Terms of use must be uniform across sequences**

# What if.. location of sequenced material was mandatory?

- Submitters would be required to provide source location, such as country name or coordinates



- Submission blocked if this condition is not satisfied
- Data – particularly for biodiversity and ecology applications - are enriched

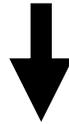


**Opportunity:**

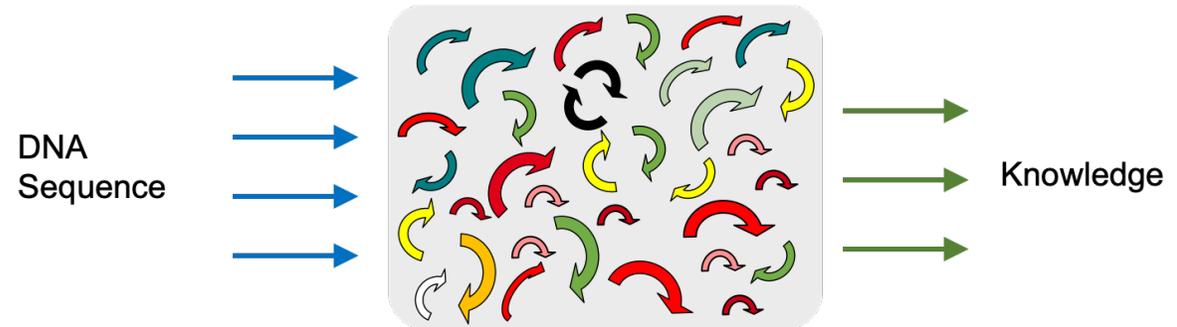
**Location information is scientifically informative**

# What if.. collaboration networks were tracked?

- E.G. INSDC would provide services to show networks of data owning institutions or networks of co-authors



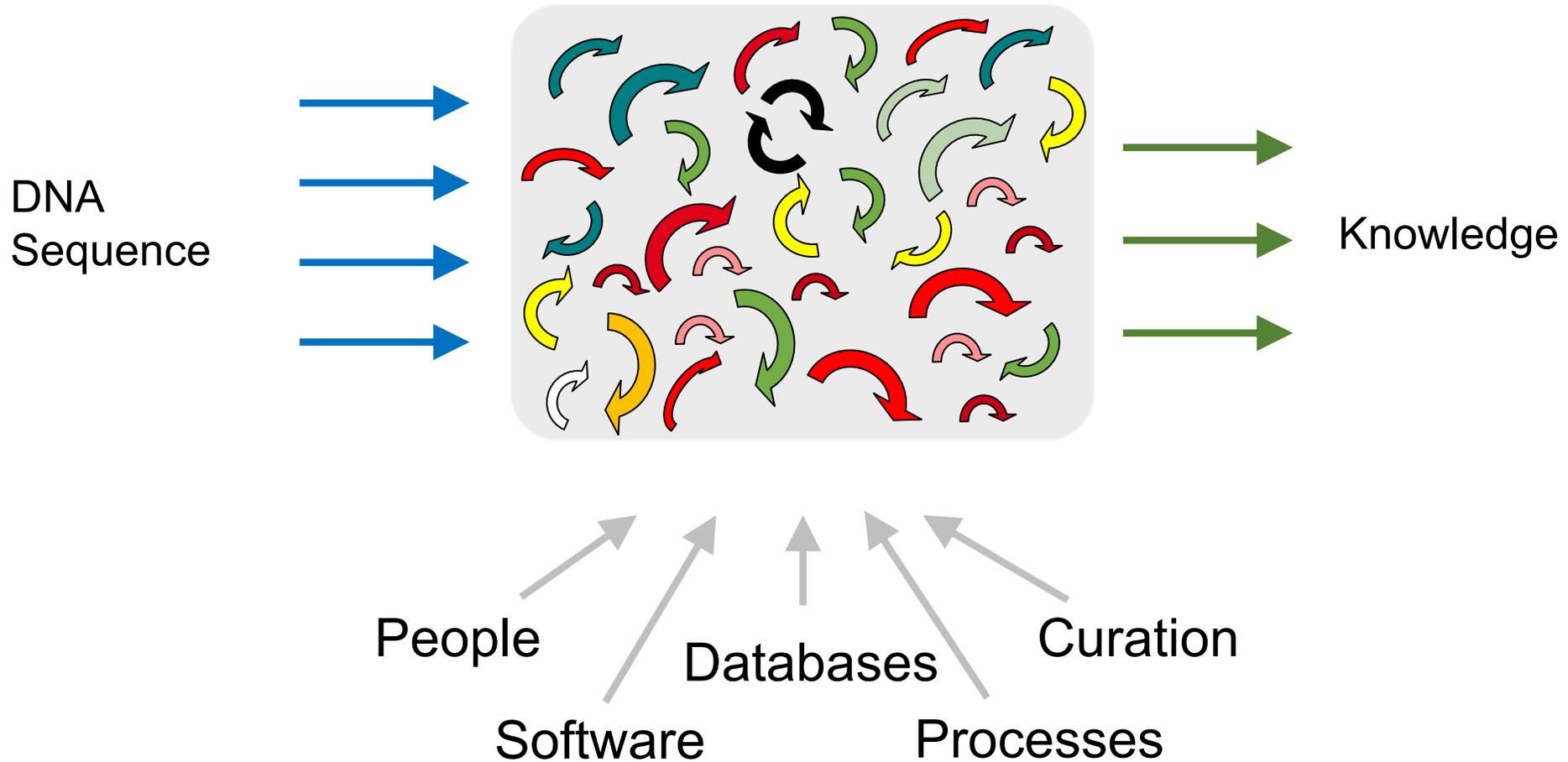
- Scientists would be able to identify potential collaborators
- Co-owned data or co-publishing authors may reveal data links

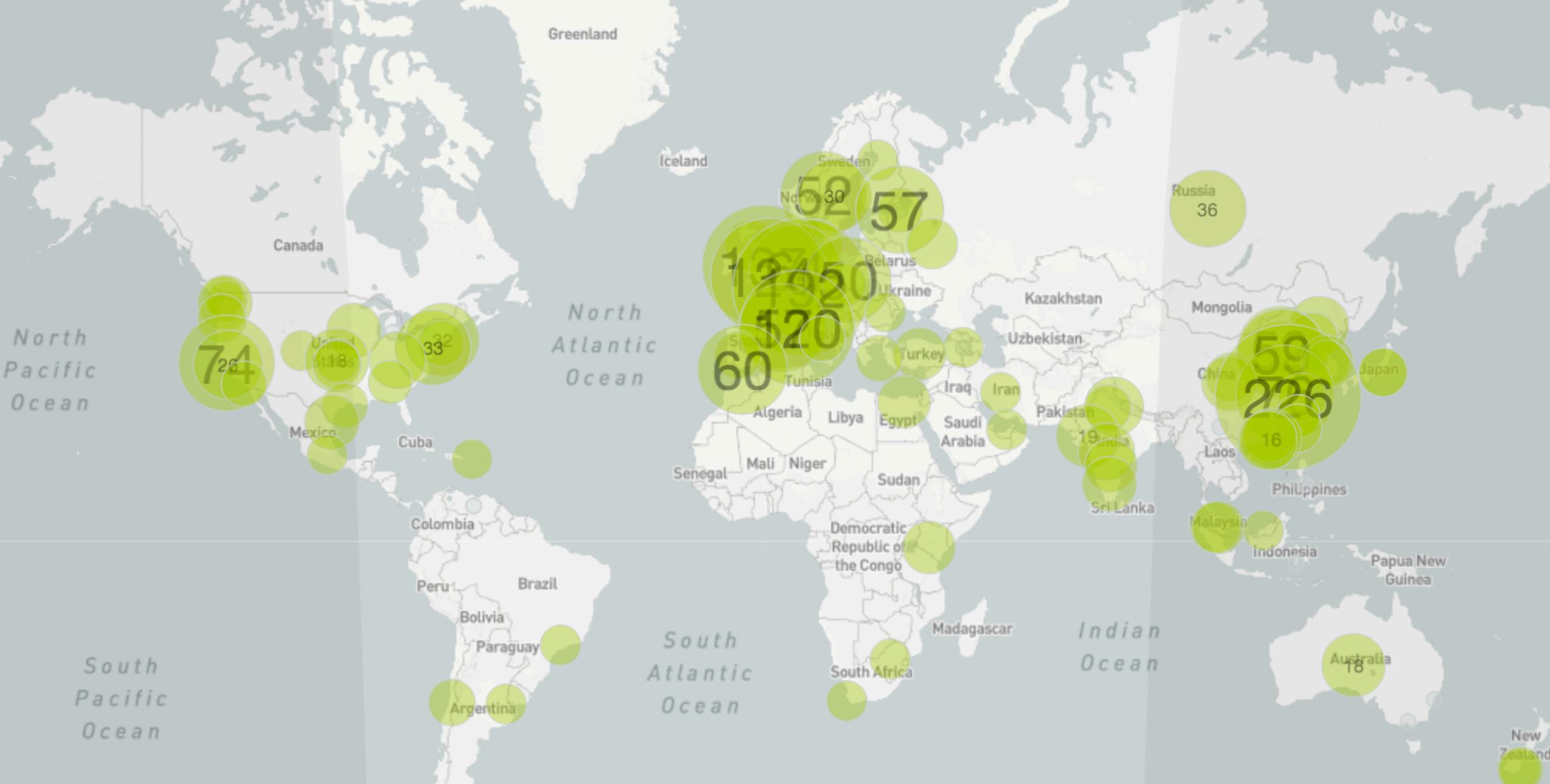


**Opportunity:**

**Ownership of, and interest in, sequence can add productivity**

# The life science “machine”





<https://www.ebi.ac.uk/web/livemap/live-data-map.html>