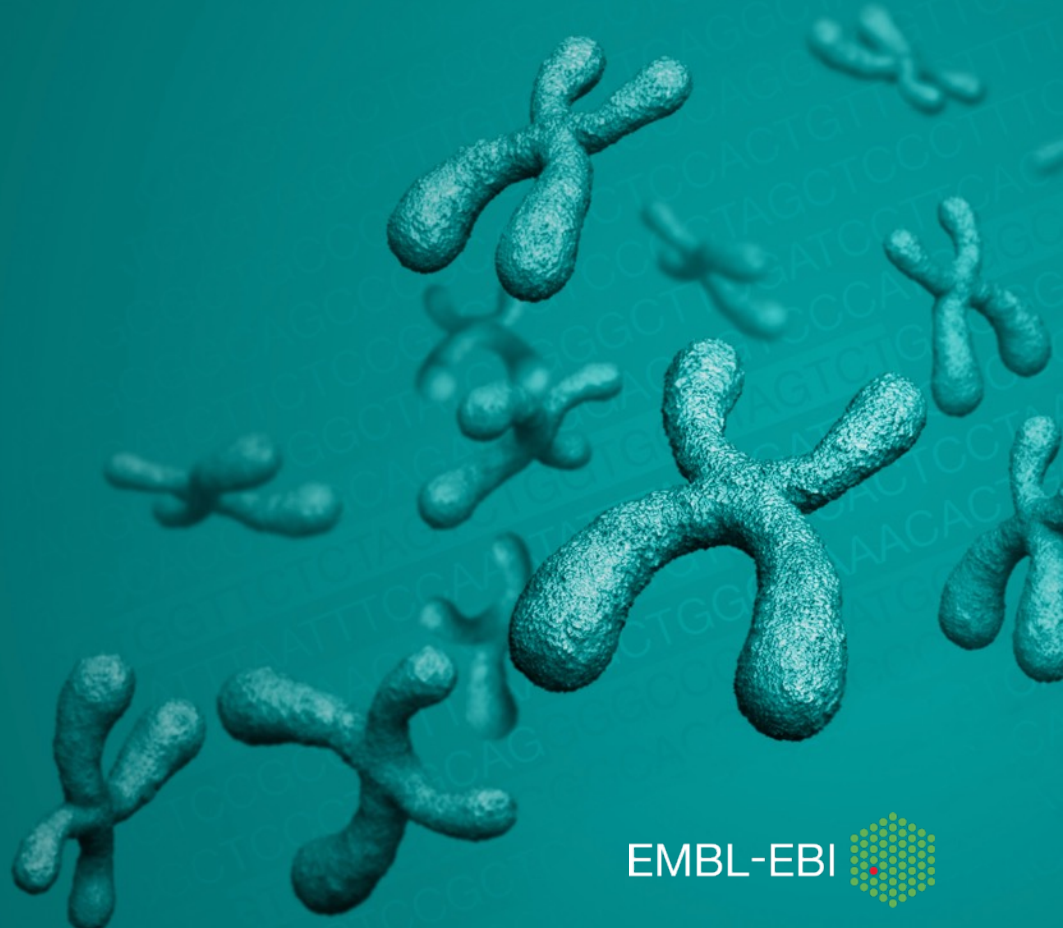
Open access & the International Nucleotide Sequence Database Collaboration

Guy Cochrane



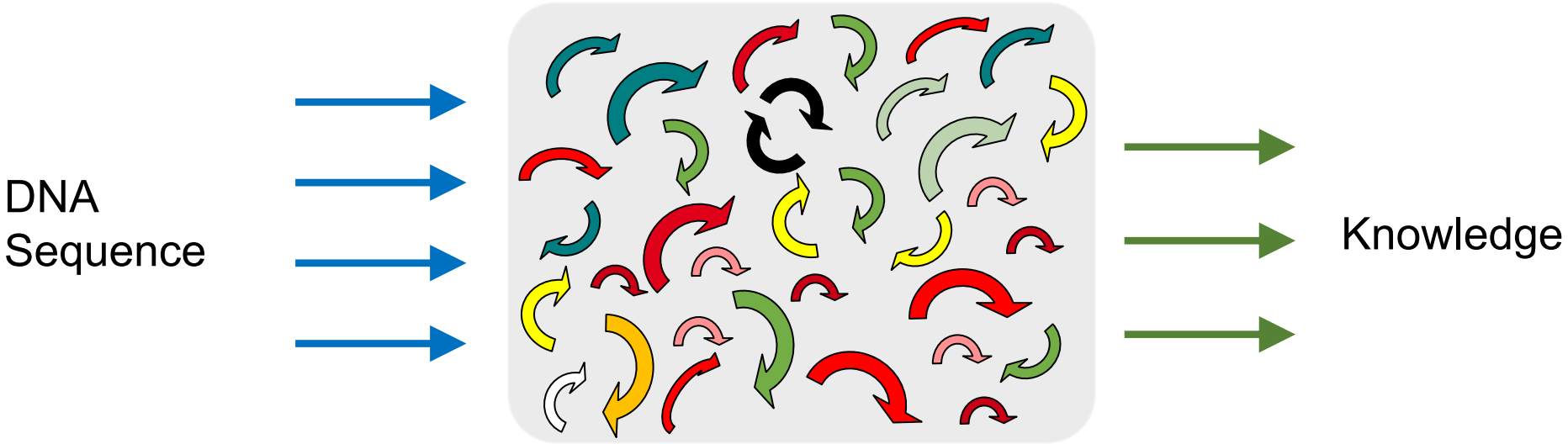
EMBL-EBI

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

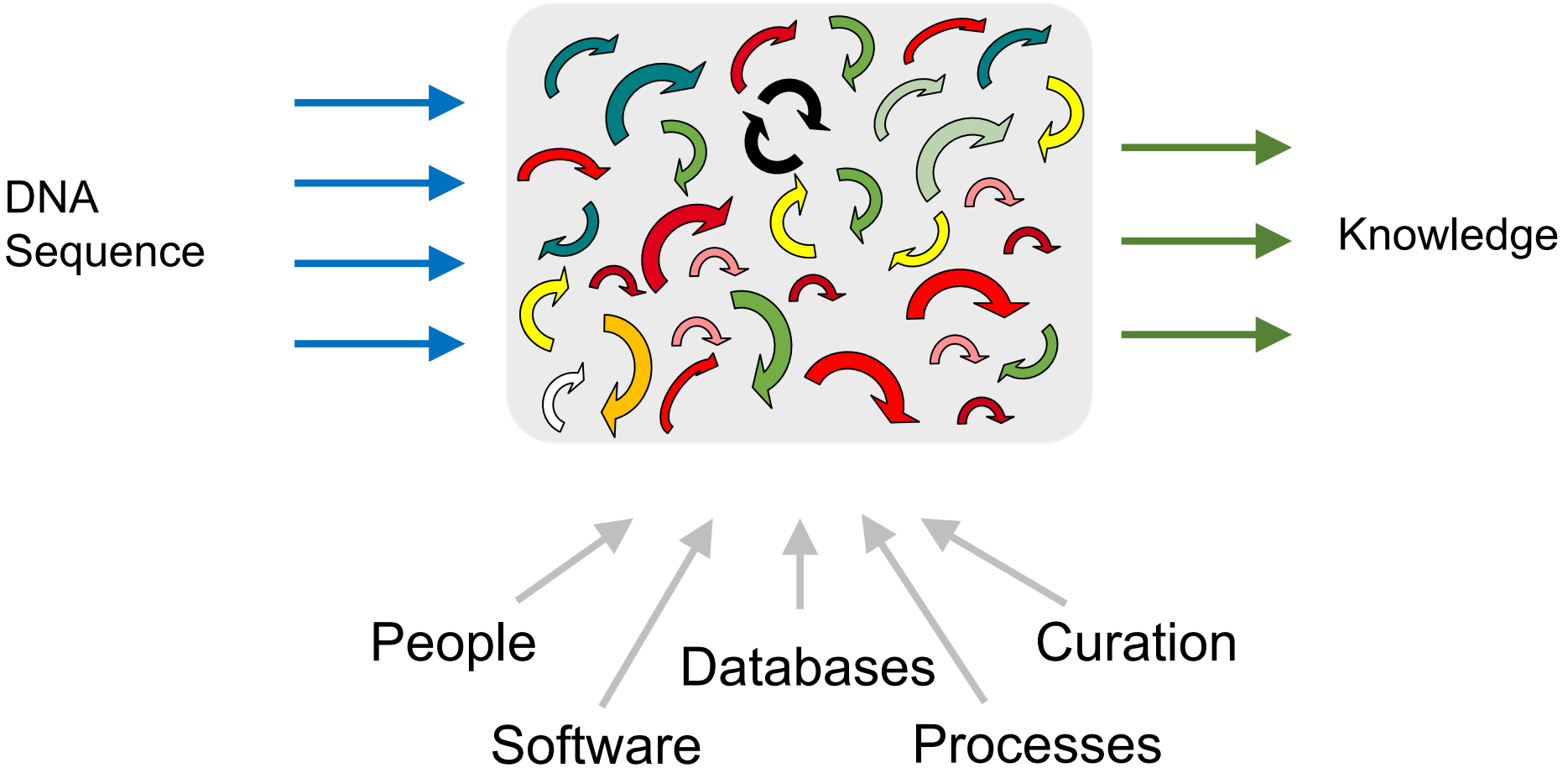
Isaac Newton

EMBL-EBI

The life science “machine”



The life science “machine”

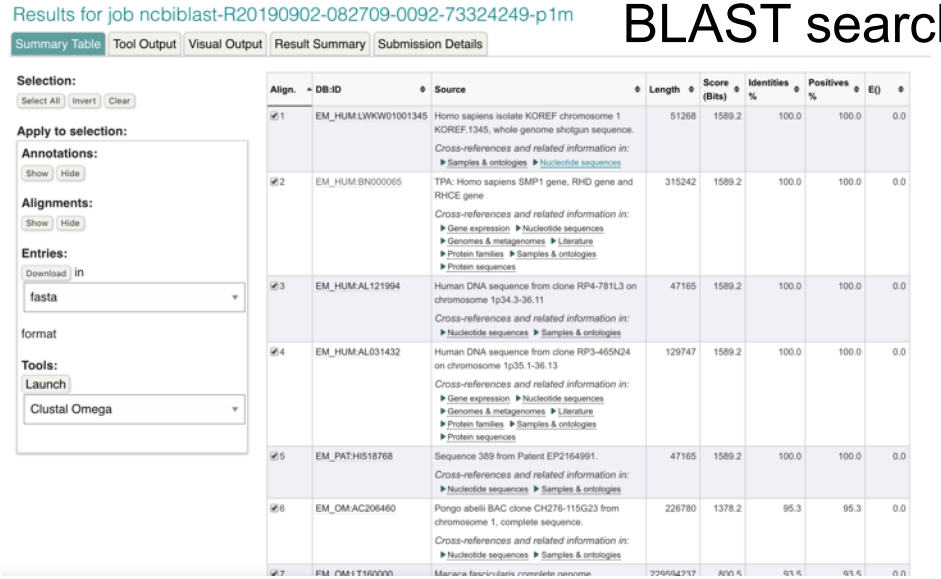


Open data are essential

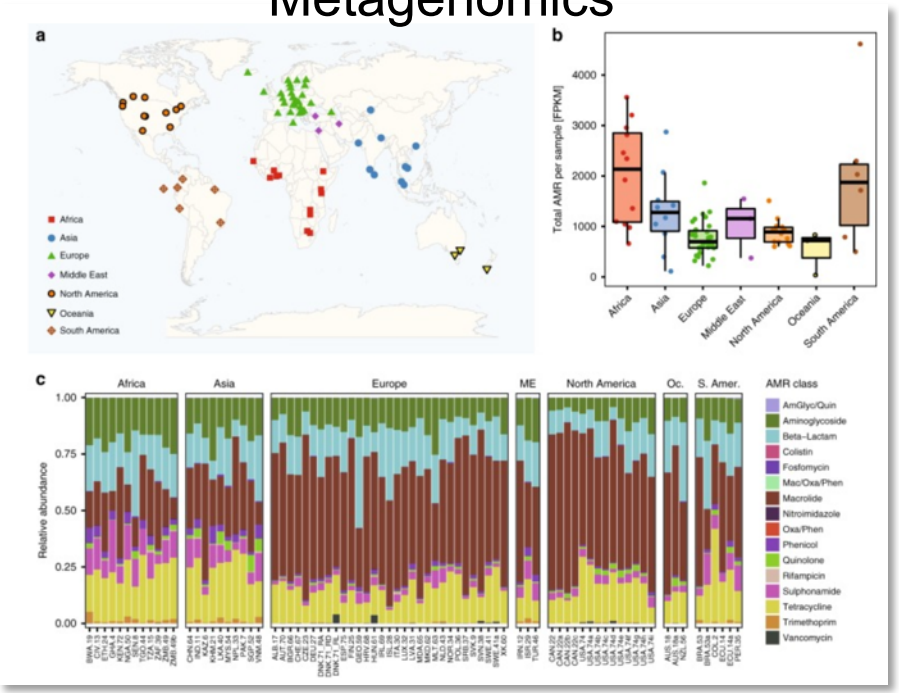
Tree building

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

BLAST search

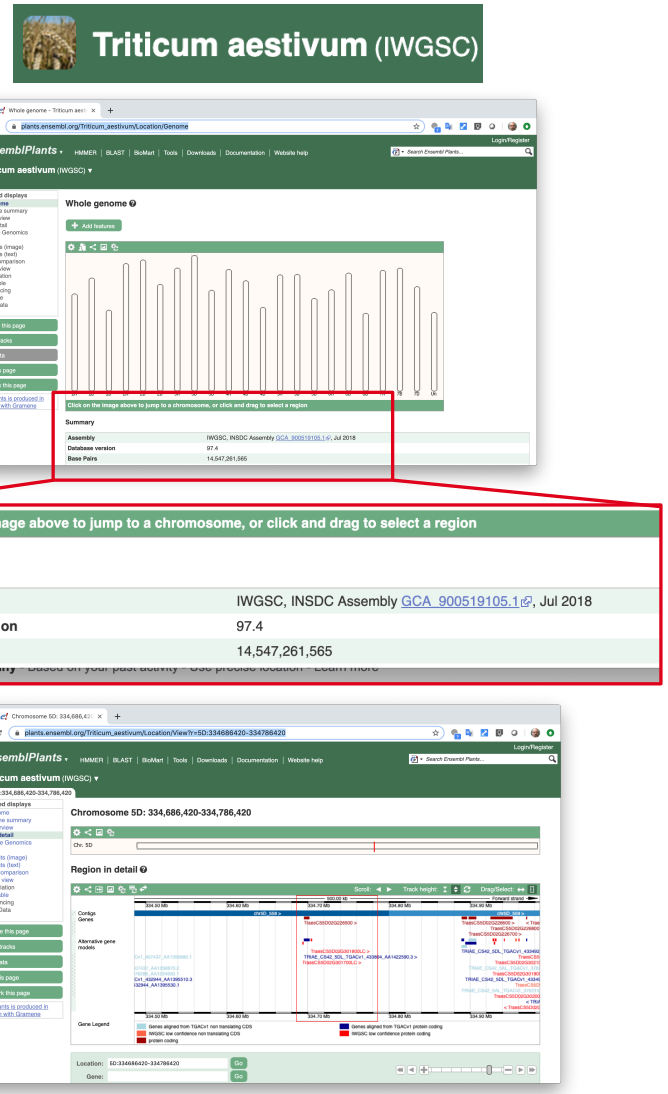


Metagenomics



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

Annotation by reference



https://plants.ensembl.org/Triticum_aestivum/Info/Index?db=core

EMBL-EBI

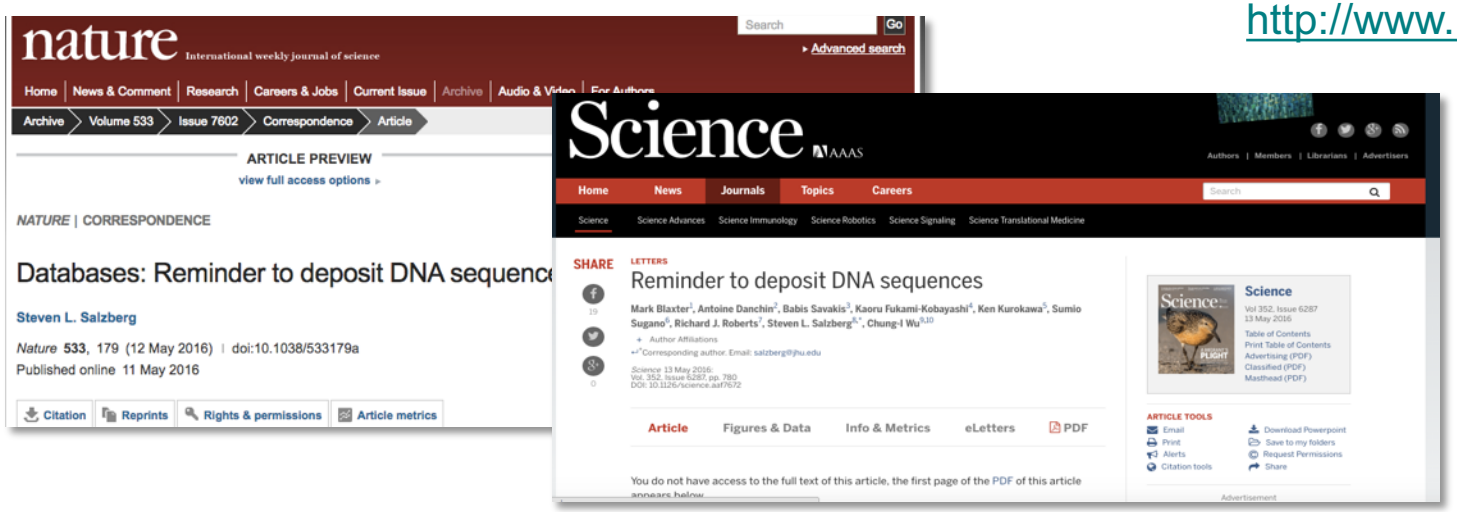
International Nucleotide Sequence Database Collaboration (INSDC)

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)

```
TATACTTAATTTTCGGTGCCTGAGCAGGAATAGTAGGAACCGGCTTAAGTTTATTAATCC
GAACTGAACTTAGTCAACCAGGCACTCTATTGGGCGATGACCAAATTTATAATGTAATCG
TGACTGCTCATGCATTCGTTATAATCTTCTTTATAGTTATACCAATTATAATCGGCGGGT
TTGGAAACTGATTGATACCCCTCATAATTGGAGCCCCAGACATAGCCTTTCCACGAATAA
ATAATATAAGCTTCTGACTGTTACCACCATCTTTCCTTCTTCTATTAGCTTCAGCAGGAG
TAGAGGCCGGAGCAGGCACAGGATGAACAGTTTATCCGCCCCTTGCTGGAAATATTGCTC
ATGCCGGAGCATCAGTCGACTTAACTATCTTCTCACTGCACTTAGCAGGGGCCTCTTCAA
TCCTAGCCTCTATTAATTTTATCACCACAATTATTAACATAAAATCACCATCAATTACAC
AATACCAAACACCACTTTTTGTGTGATCATTACTTATCACTGCAATTCTTCTACTTCTAT
CATTACCAGTACTAGCAGCAGGAATTACAATACTTCTGACAGACCGAAATCTTAACACCA
CATTCTTCGACCCAGCTGGAGGAGGTGACCCAATTCTCTATCAACATTTATTCTGATTCT
TTGGCCACCCAGAAGTCTA
```

Sequence

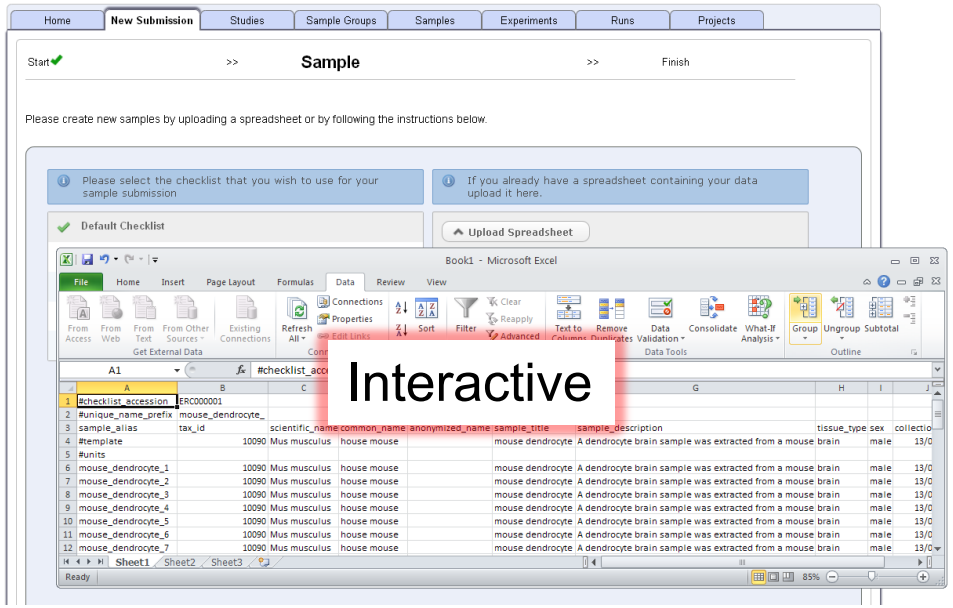
```
ID   MK791304; SV 1; linear; genomic DNA; STD; VRT; 679 BP.
XX
AC   MK791304;
XX
DT   04-NOV-2019 (Rel. 142, Created)
DT   04-NOV-2019 (Rel. 142, Last updated, Version 1)
XX
DE   Narcine maculata isolate CAS-AM-SR cytochrome c oxidase subunit I (COI)
DE   gene, partial cds; mitochondrial.
XX
KW   .
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
DR   MD5; f83e14e5913fb53c7d28dc9e9960a043.
XX
CC   ##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                   tcaaccaaccacaaagacattggcac, rev_name: coi r, rev_seq:
FT                   tagacttctgggtggccaaagaatca"
FT                   /db_xref="taxon:1455688"
FT   CDS             <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="YLIFGAWAGMVGTGLSLLIRTELSQPGTLLGDDQIYNVIVTAHAF
FT                   VMIFFMVMPIMIGGFGNWLMPLMIGAPDMAFPRMNNMSFWLLPPSFLLLLASAGVEAGA
FT                   GTGWTVYPPLAGNIAHAGASVDLTIFSLHLAGASSILASINFITTIINMKSPSITQYQT
FT                   PLFVWSLLITAILLLLSLPVLAAGITMLLTDRNLNTTFFDPAGGGDPILYQHLFWFFGH
FT                   PEV"
XX
SQ   Sequence 679 BP; 197 A; 166 C; 109 G; 207 T; 0 other;
     tatacttaat tttcggtgcc tgagcaggaa tagtaggaac cggcttaagt ttattaatcc        60
     gaactgaact tagtcaacca ggcactctat tgggcgatga ccaaatttat aatgtaatcg       120
     tgactgctca tgcattcgtt ataatcttct ttatagttat accaattata atcggcgggt       180
     ttggaaactg attgataccc ctcataattg gagccccaga catagccttt ccacgaataa       240
     ataatataag cttctgactg ttaccaccat ctttccttct tctattagct tcagcaggag       300
     tagaggccgg agcaggcaca ggatgaacag tttatccgcc ccttgctgga aatattgctc       360
     atgccggagc atcagtcgac ttaactatct tctcactgca cttagcaggg gcctcttcaa       420
     tcctagcctc tattaatttt atcaccacaa ttattaacat aaaatcacca tcaattacac       480
     aataccaaac accacttttt gtgtgatcat tacttatcac tgcaattctt ctacttctat       540
     cattaccagt actagcagca ggaattacaa tacttctgac agaccgaaat cttaacacca       600
     cattcttcga cccagctgga ggaggtgacc caattctcta tcaacattta ttctgattct       660
     ttggccaccc agaagtcta                                                    679
//
```

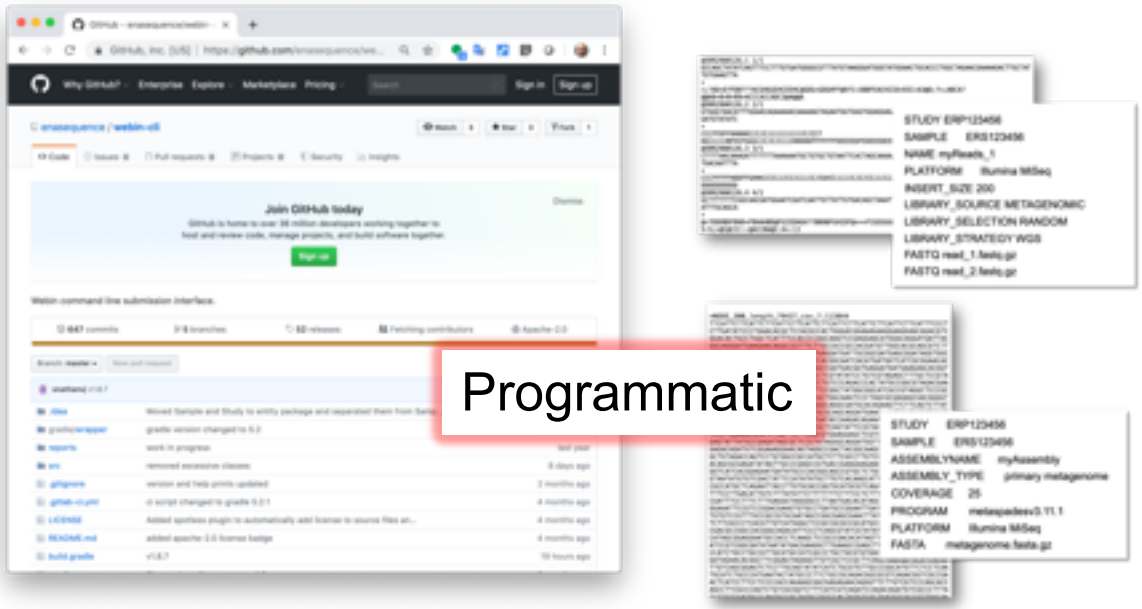
- Clerical information
- Taxonomy
- Literature
- Cross-references
- Source information
- Biological features

http://www.insdc.org/

The data submission process

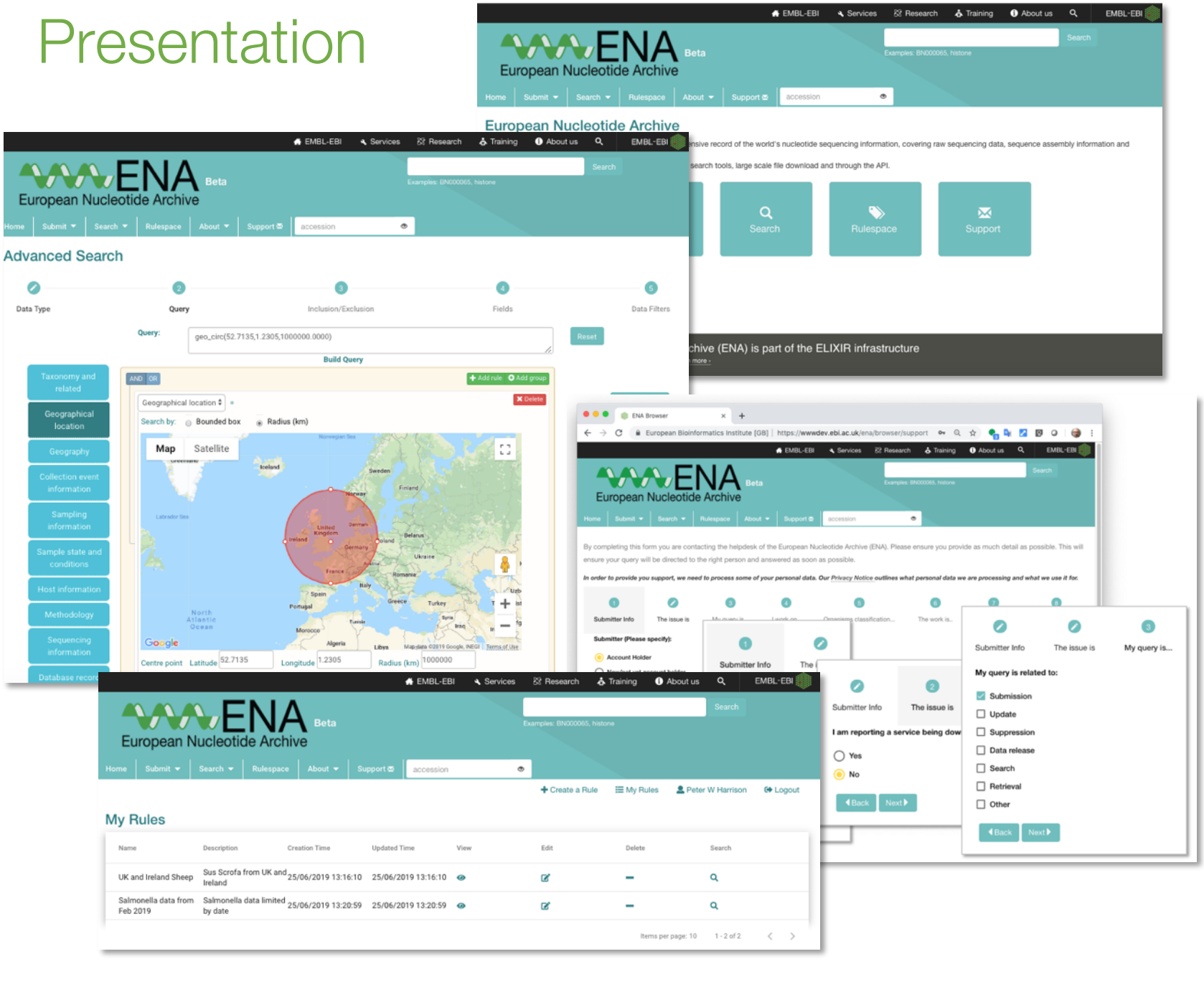
- Data provider
  - Validation, organization, adding structure and curation
  - Compliance with standards and established conventions
  - Links to external data (e.g. academic publications)
- Database
  - Curation and integration with the overall corpus
  - Indexing for discovery and reuse
  - Open services,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





EMBL-EBI

Presentation



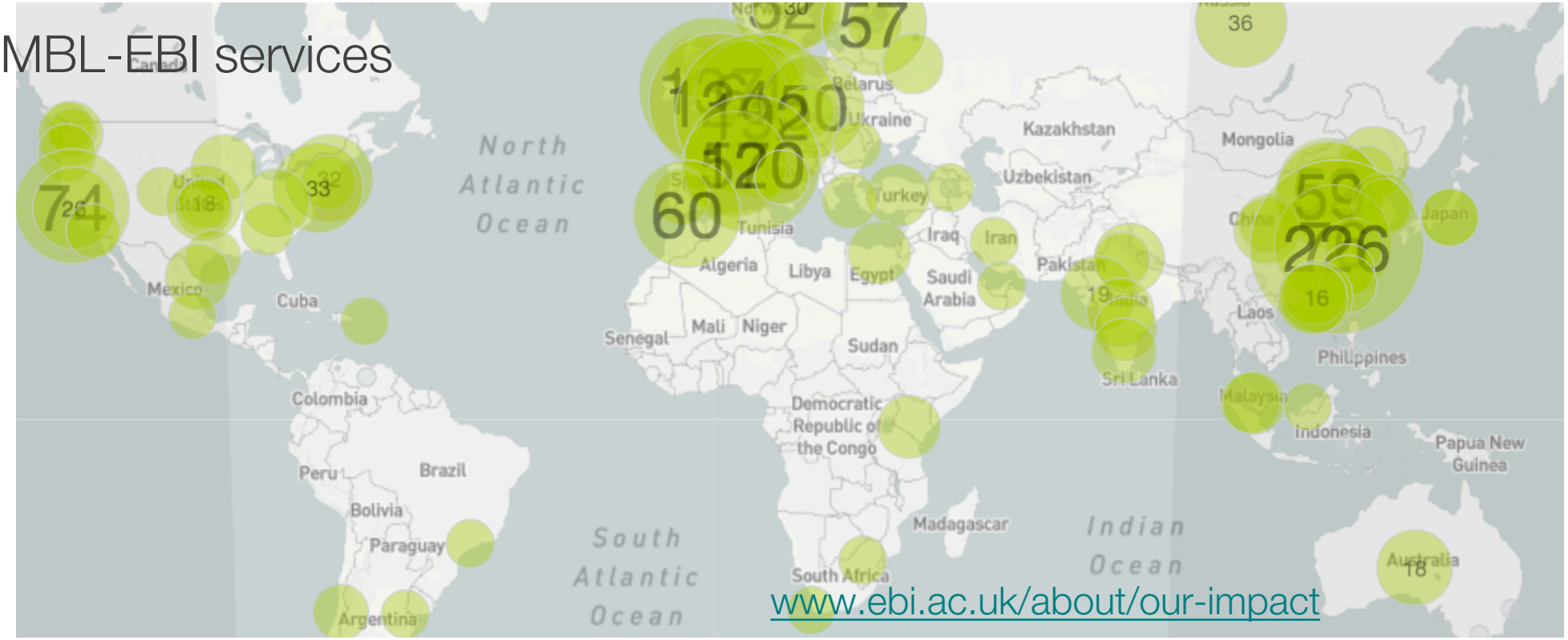
- Search
  - Sample characteristics
  - Function
  - Sequence similarity
- Browse
  - Study
  - Taxonomy
  - Sequencing method
  - Historic versions
- Filter
  - Reusable/shareable rules
  - Synchronisation with latest data
- Retrieve
  - Download tools

EMBL-EBI

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
- **Adoption**: *the sequence database of record for the broadest scientific community*

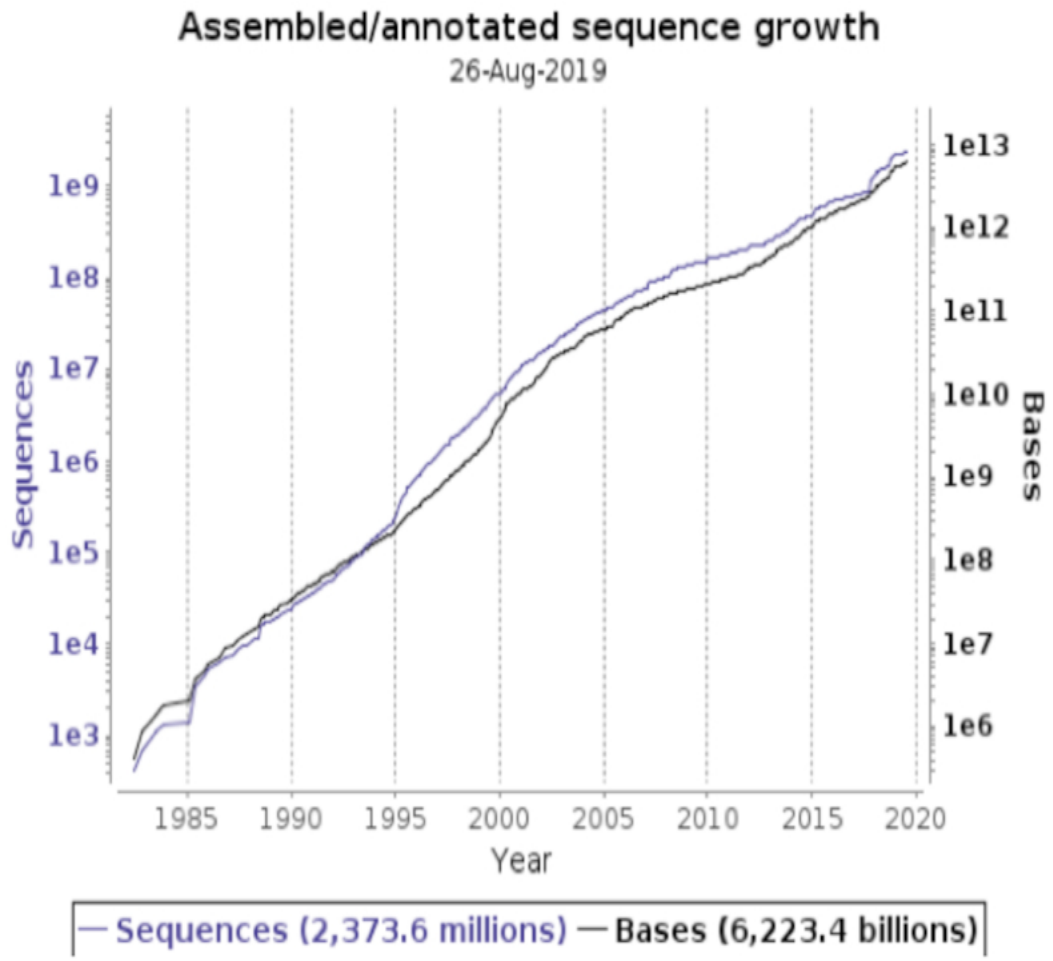
EMBL-EBI services

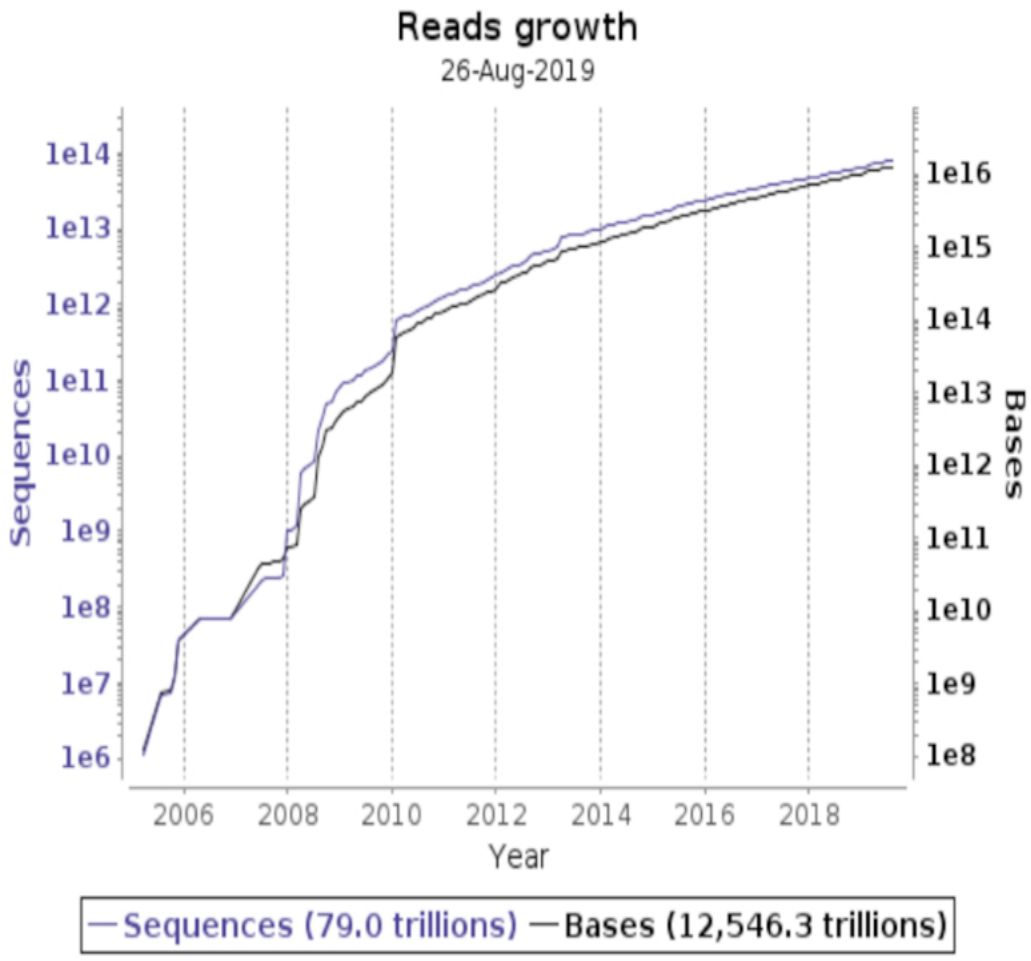


www.ebi.ac.uk/about/our-impact

EMBL-EBI

Data growth



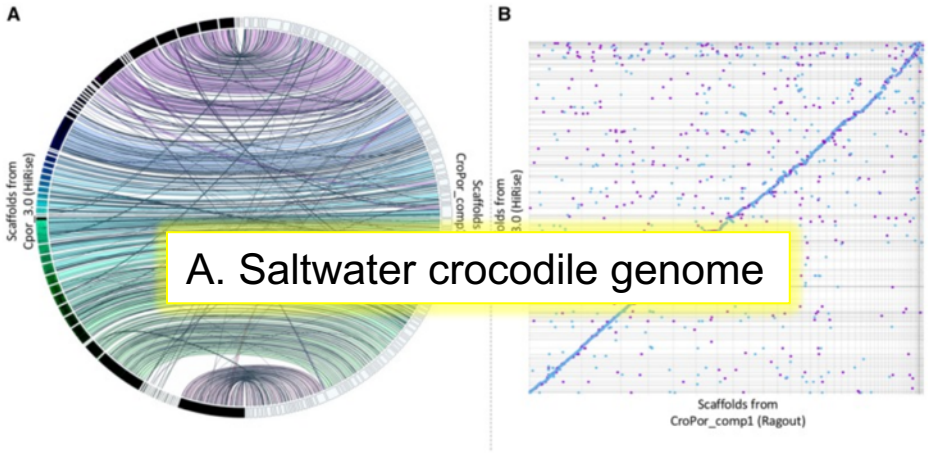


Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D70–D76, https://doi.org/10.1093/nar/gkz1063

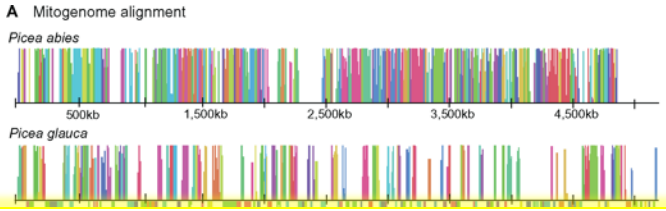
OXFORD
UNIVERSITY PRESS

Direct use

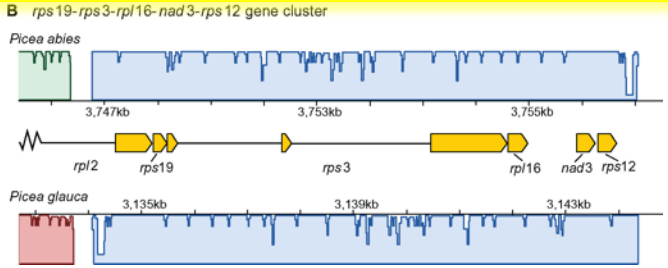
- 1,306 publications citing sequence accessions since the 1st Jan. 2020

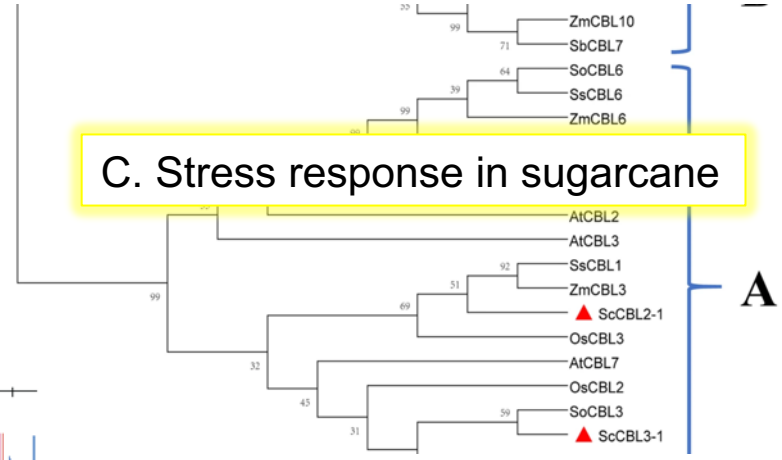


A. Saltwater crocodile genome

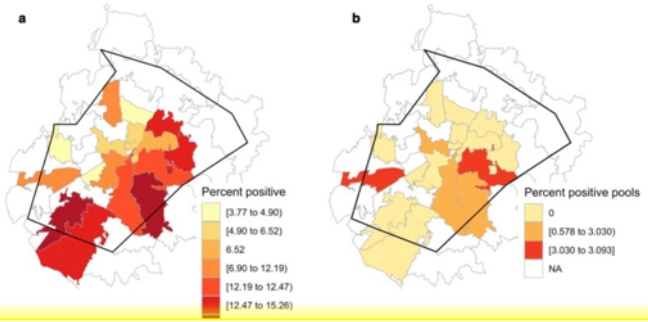


B. Norway spruce evolutionary biology

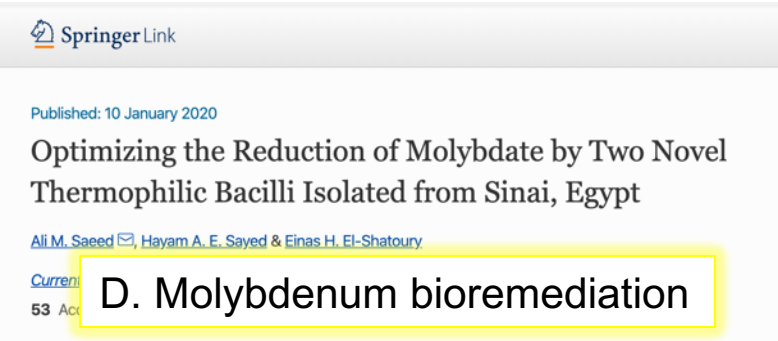




C. Stress response in sugarcane



E. Mosquito-borne heartworm in dogs

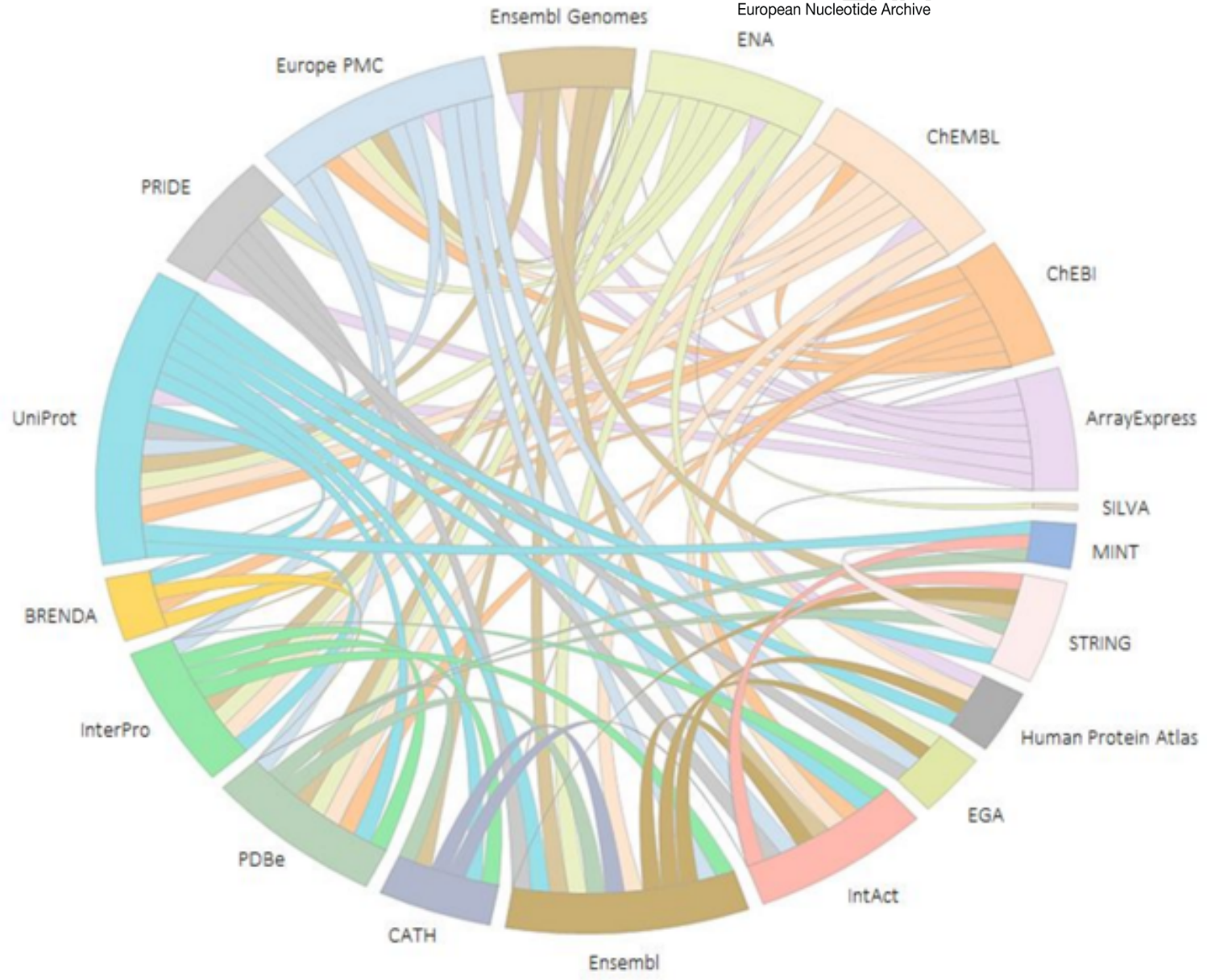


D. Molybdenum bioremediation

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

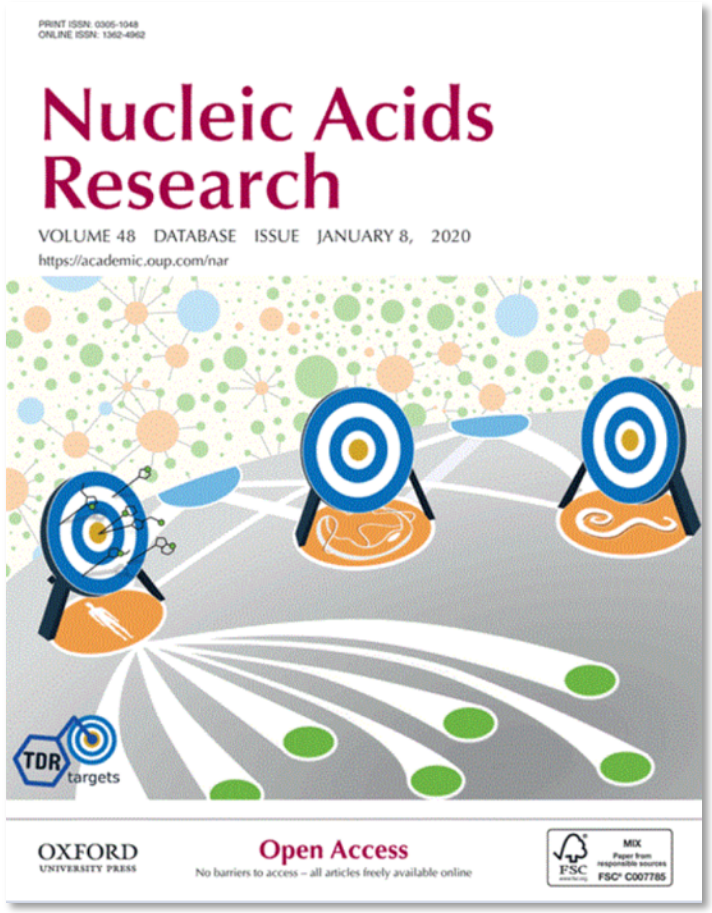
ENA
European Nucleotide Archive



elixir
Core Data Resource

Core resource connectivity figure from Drysdale *et al.*, https://www.biorxiv.org/content/10.1101/598318v1

Further reach



Public database inventory

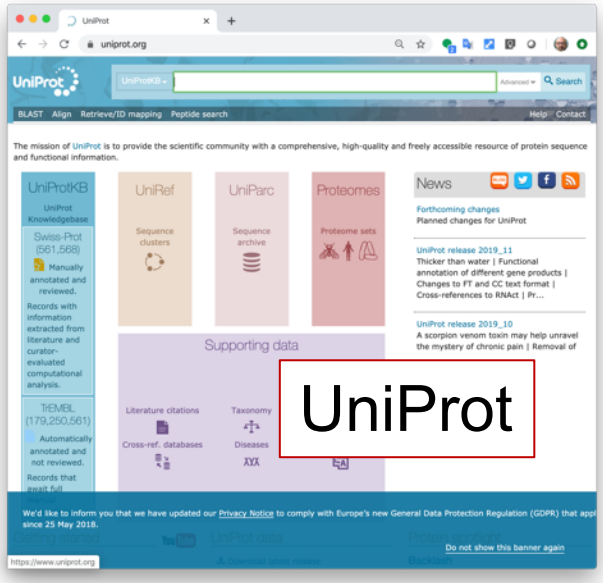
NAR database summary	–	**1778** (100%)
DBs focussing on NSD	–	**808** (45%)
Non-Human DBs	–	**743** (42%)
DBs allowing upload	–	**38** (<2.4%)
Operating outside INSDC	–	**0** (0%)

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

Secondary databases

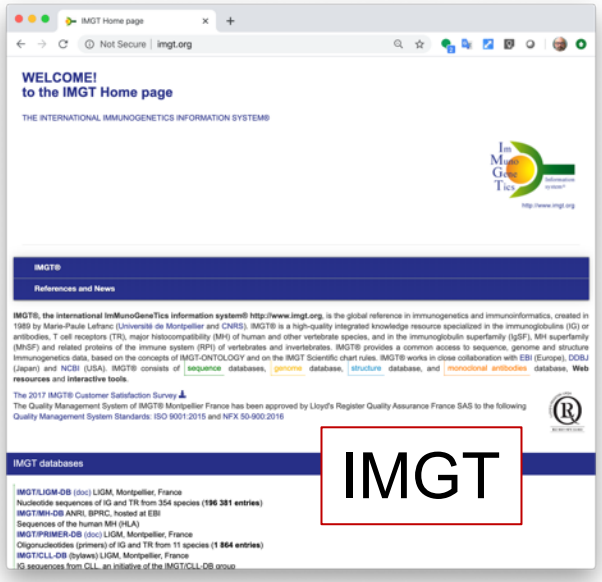
- Connected to INSDC services and content
- More specialist services to user communities



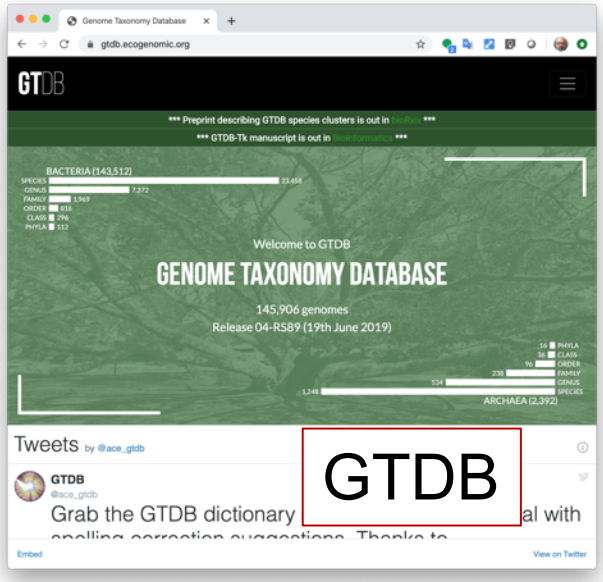
https://www.uniprot.org/

http://www.mirbase.org/

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

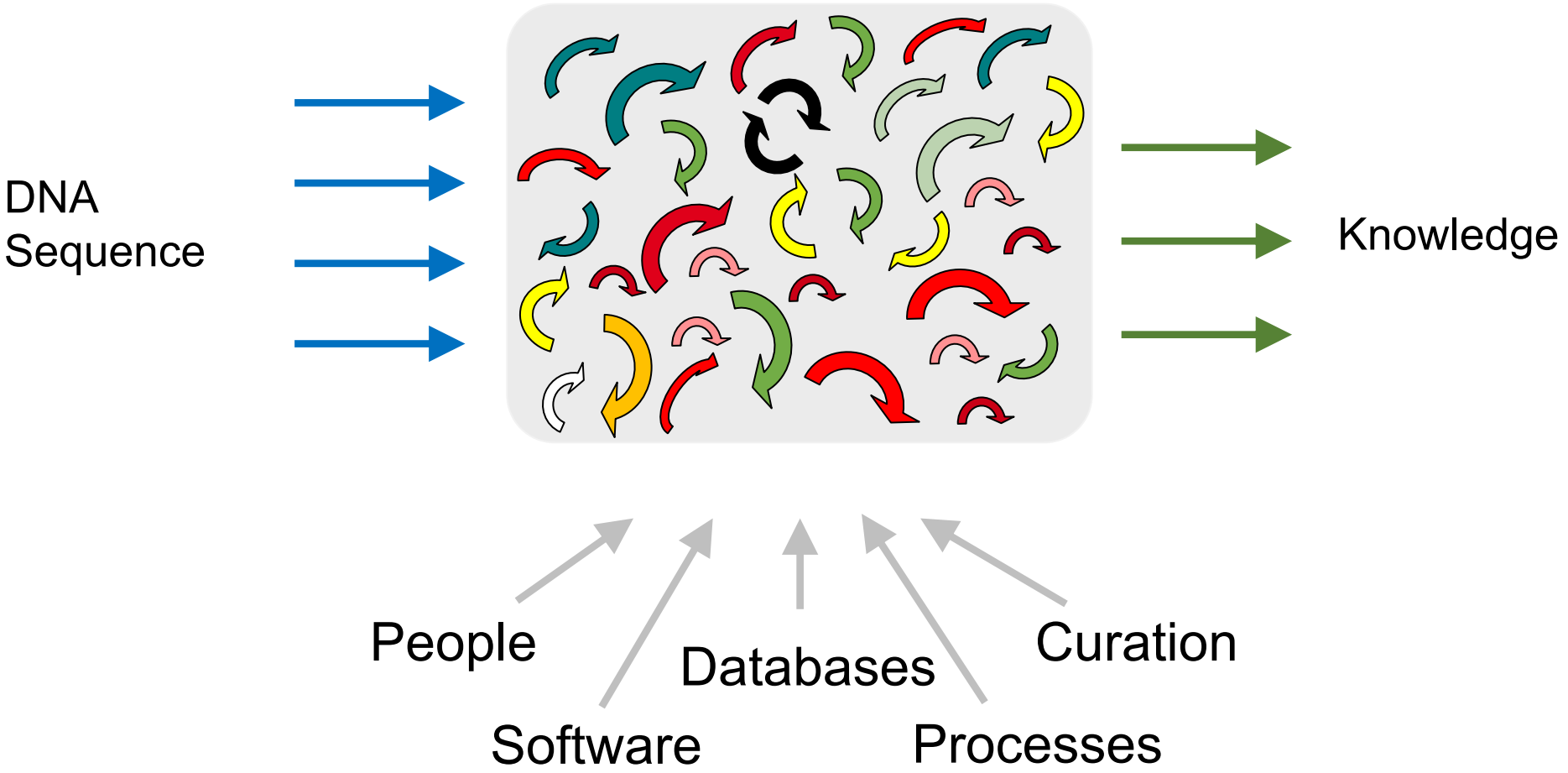


http://www.imgt.org/



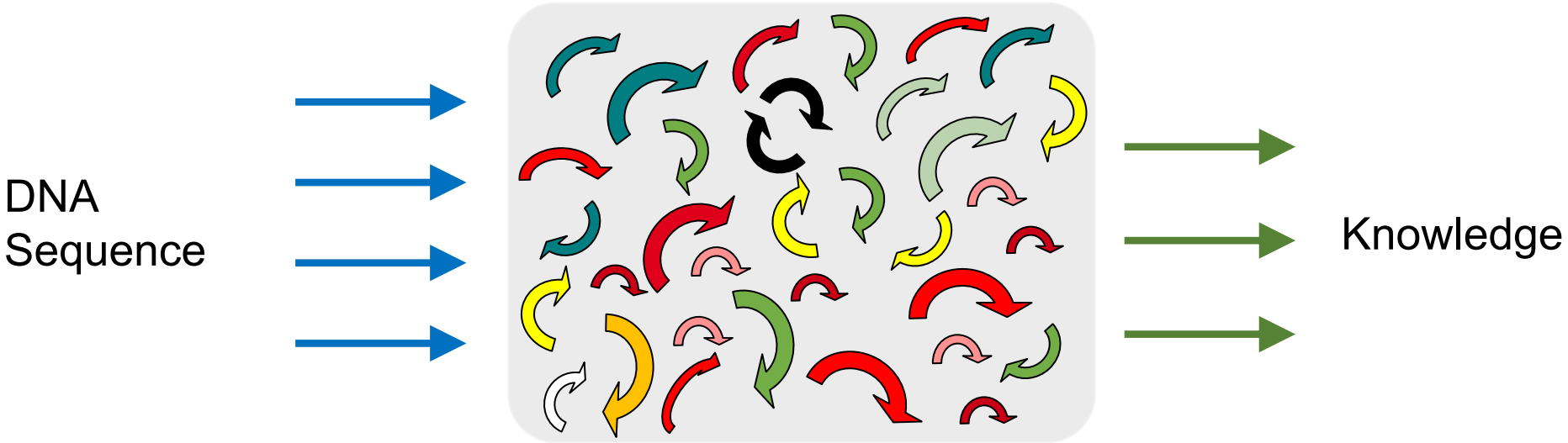
https://gtdb.ecogenomic.org/

The life science “machine”



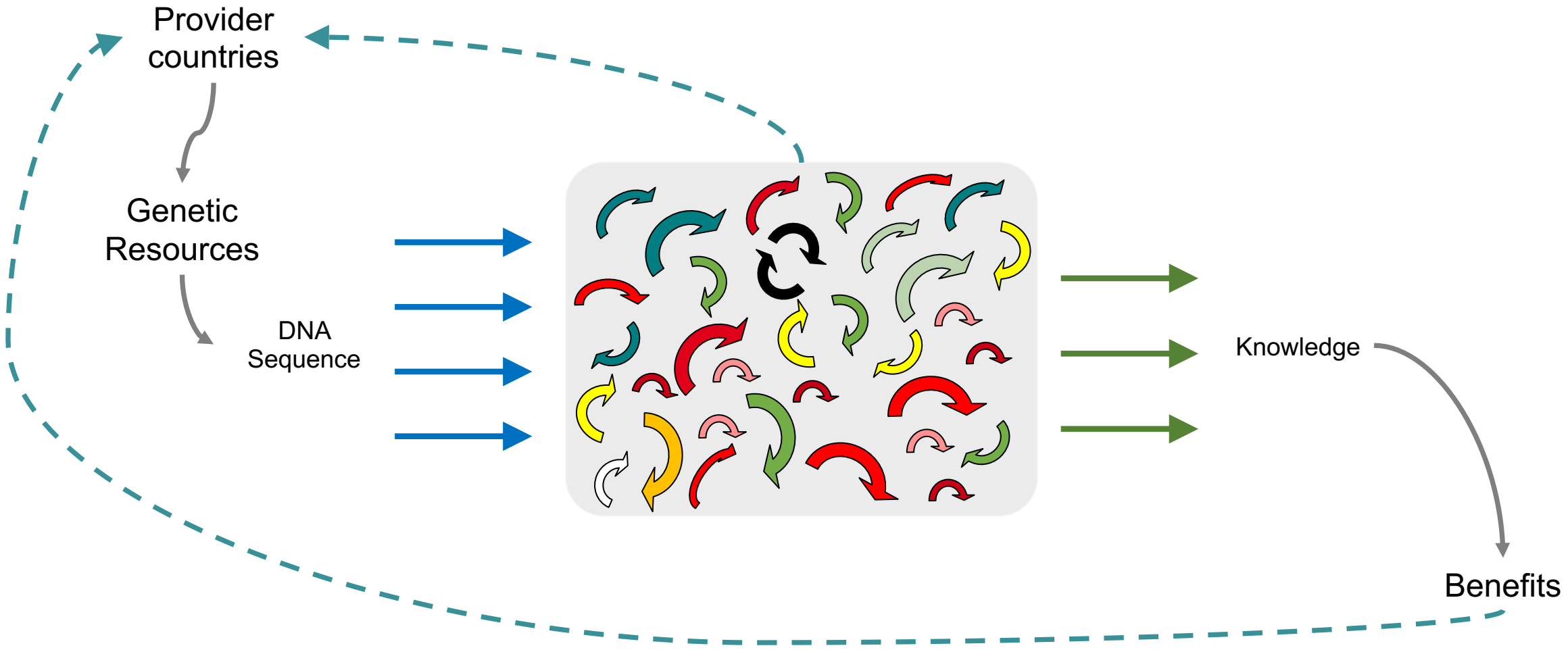
The life science “machine”

- High productivity drives scientific progress



Productivity in the machine

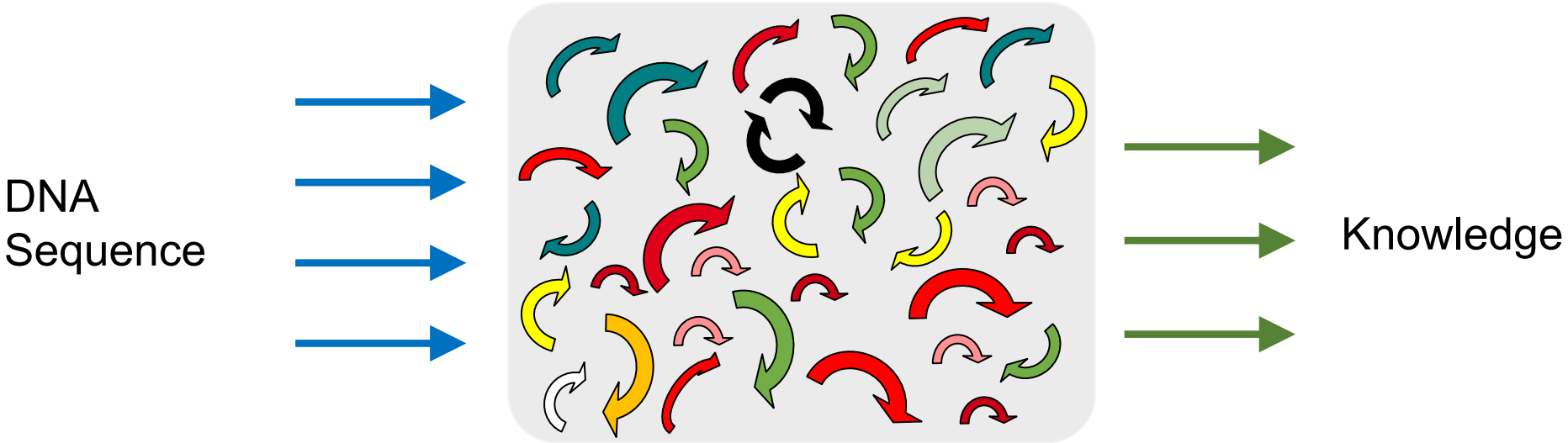
- High productivity drives ABS



EMBL-EBI

Productivity in the machine

- Understand risks
- Assert core values
- Identify opportunities

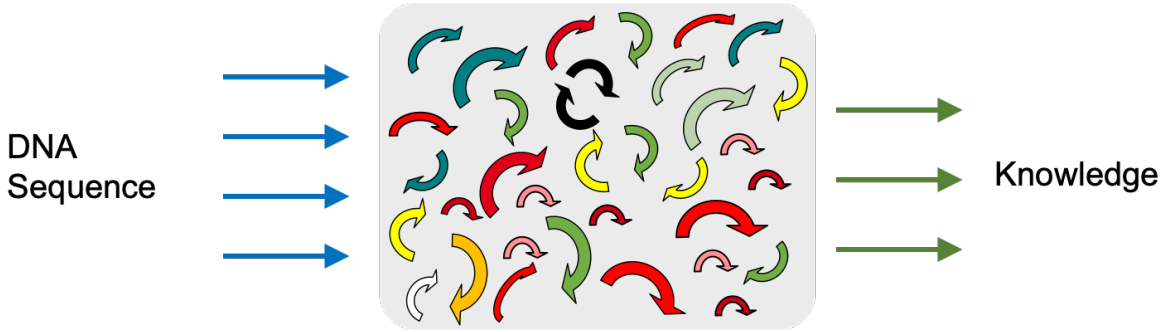


What if.. visibility of sequence were reduced?

- 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

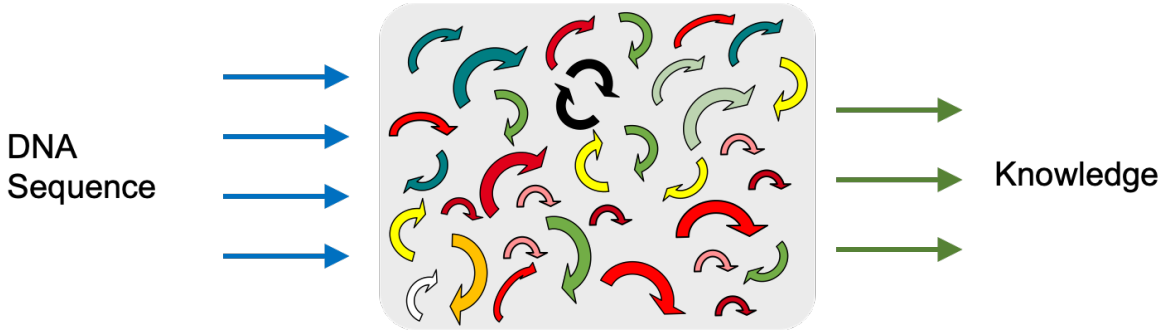
EMBL-EBI

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:

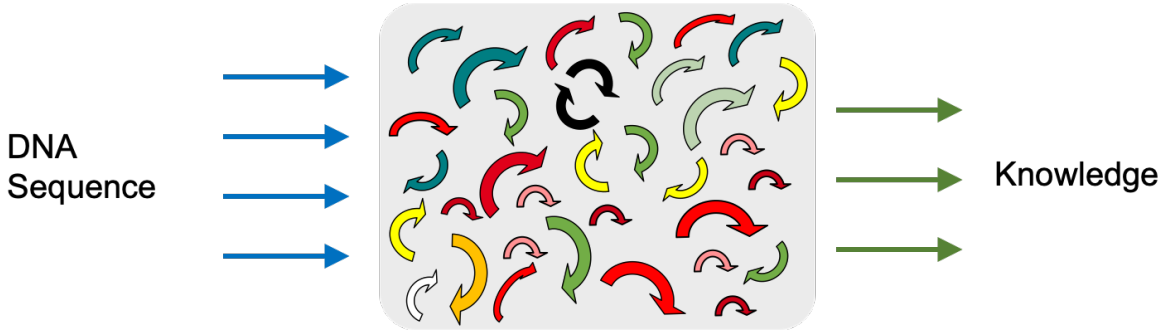
Terms of use must be uniform across sequences

EMBL-EBI

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:

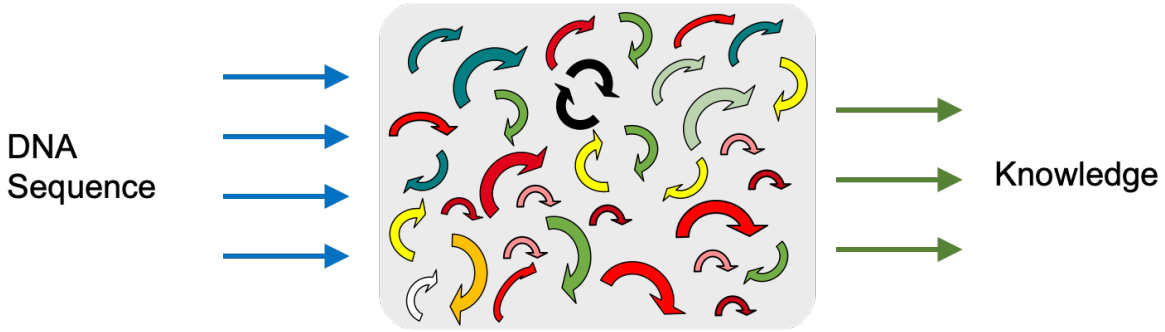
Location information is scientifically informative

EMBL-EBI

What if.. collaboration networks were tracked?

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

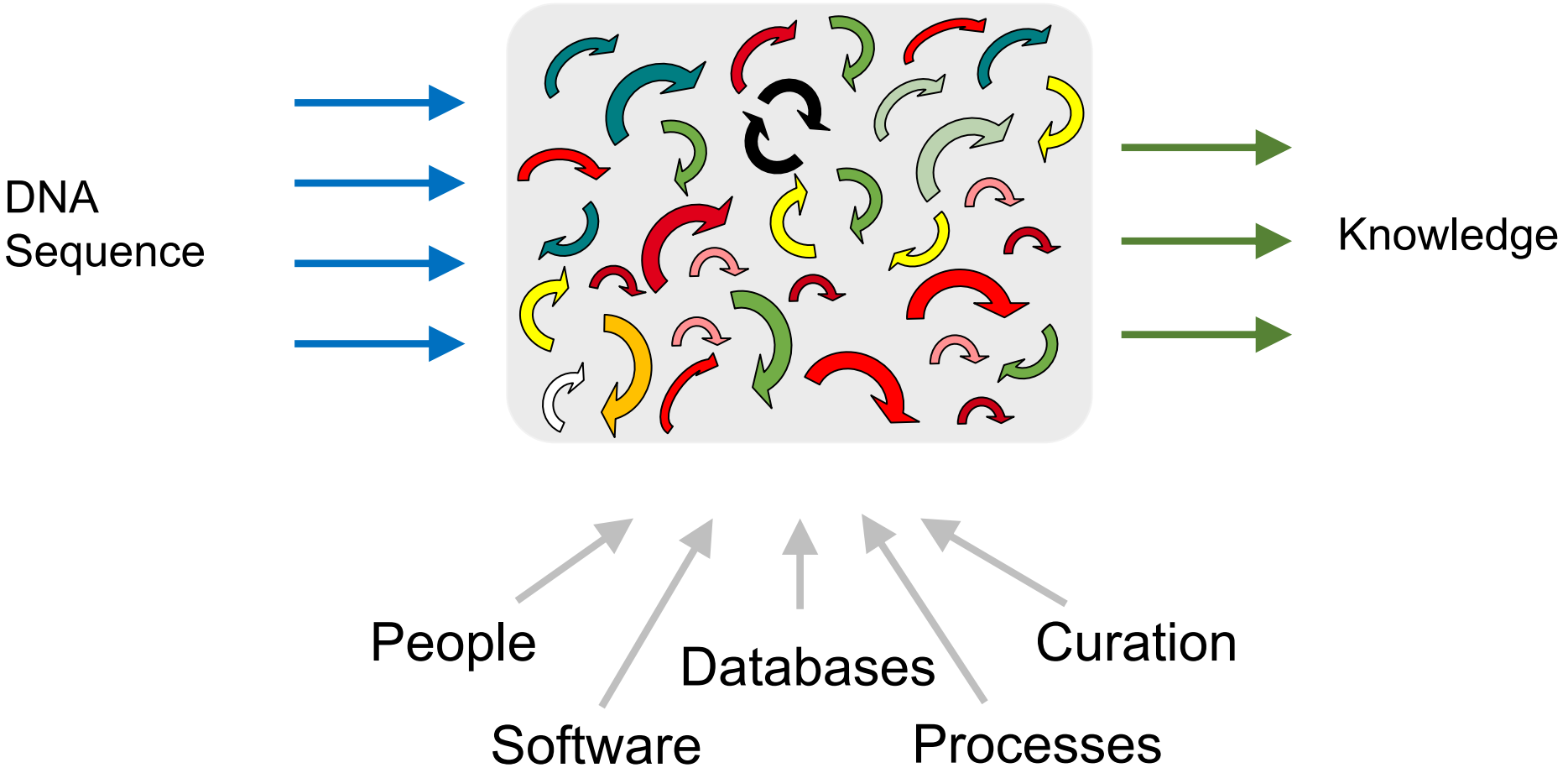
- Scientists would be able to identify potential collaborators
- Co-owned data or co-publishing authors may reveal data links

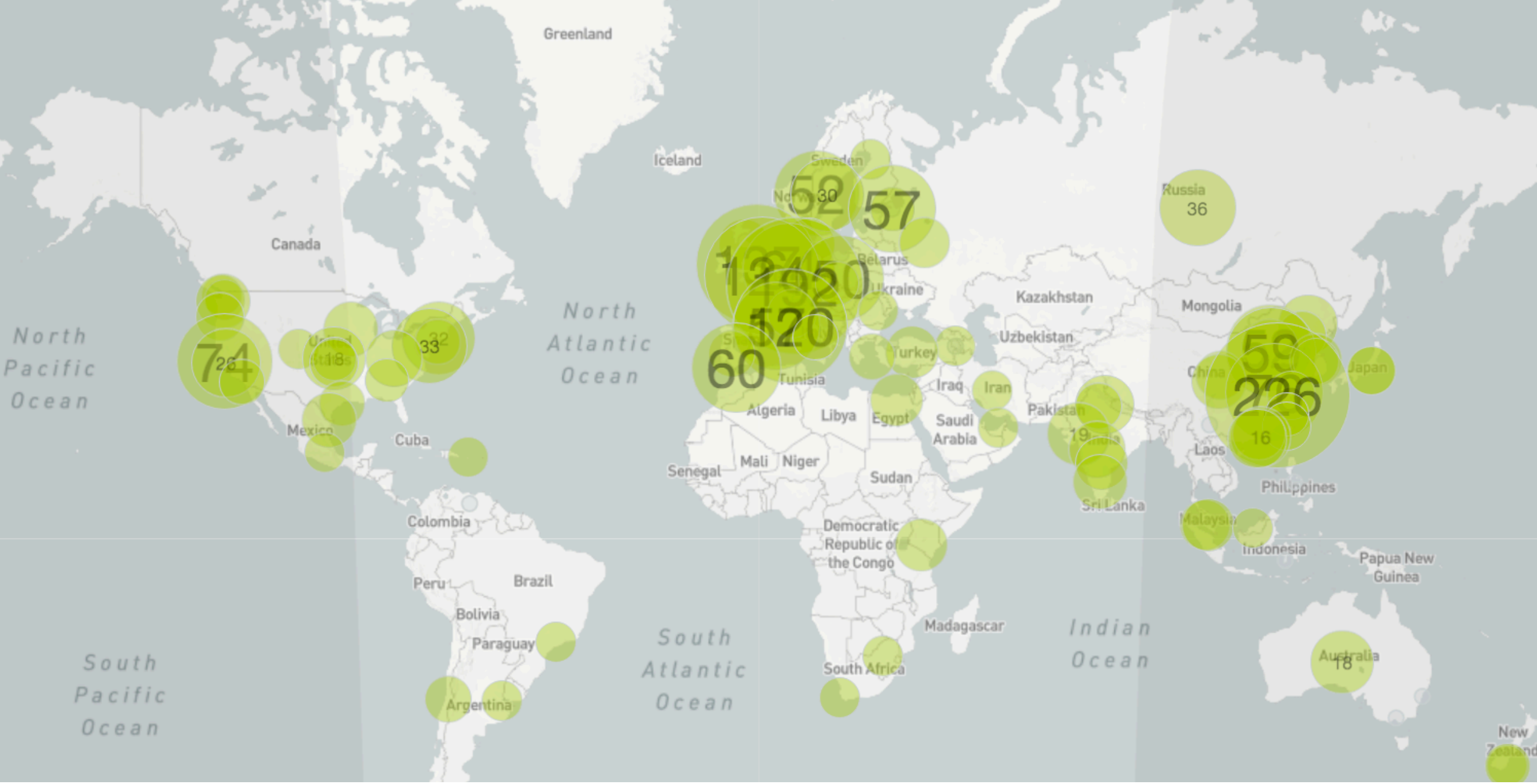


Opportunity:

Ownership of, and interest in, sequence can add productivity

The life science “machine”





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