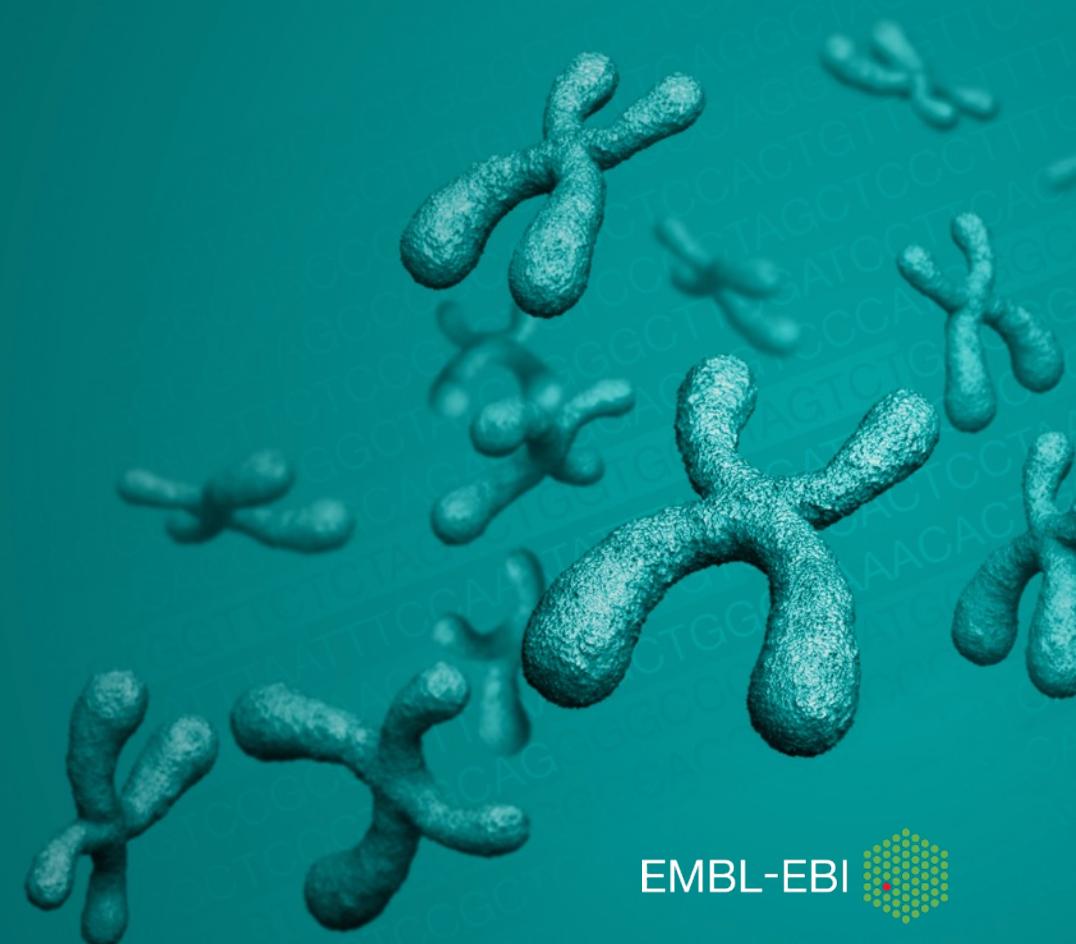


Open access & the International Nucleotide Sequence Database Collaboration

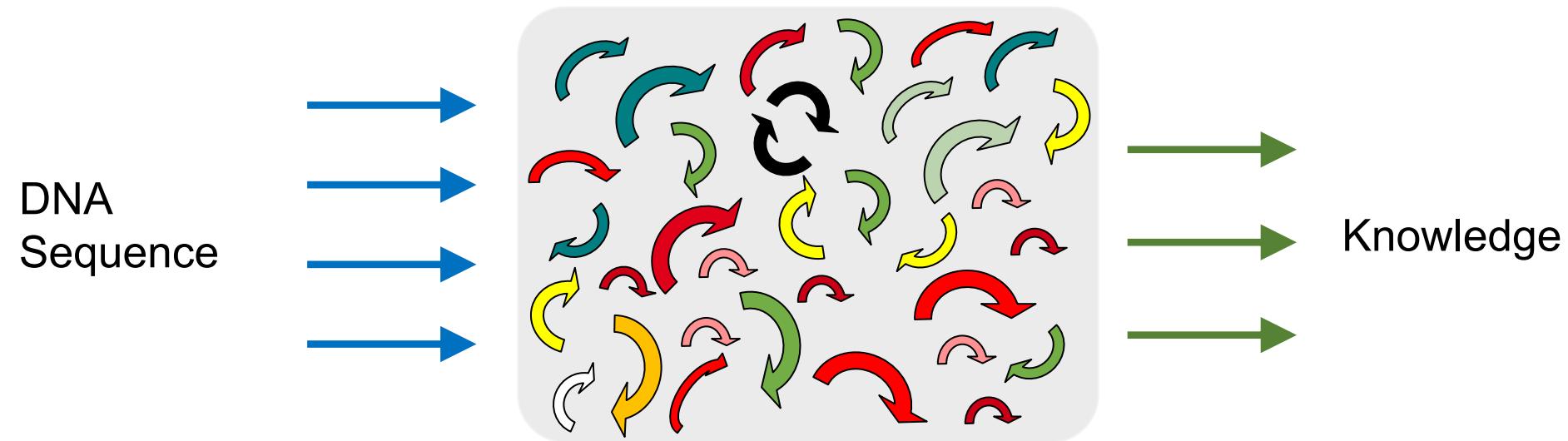
Guy Cochrane



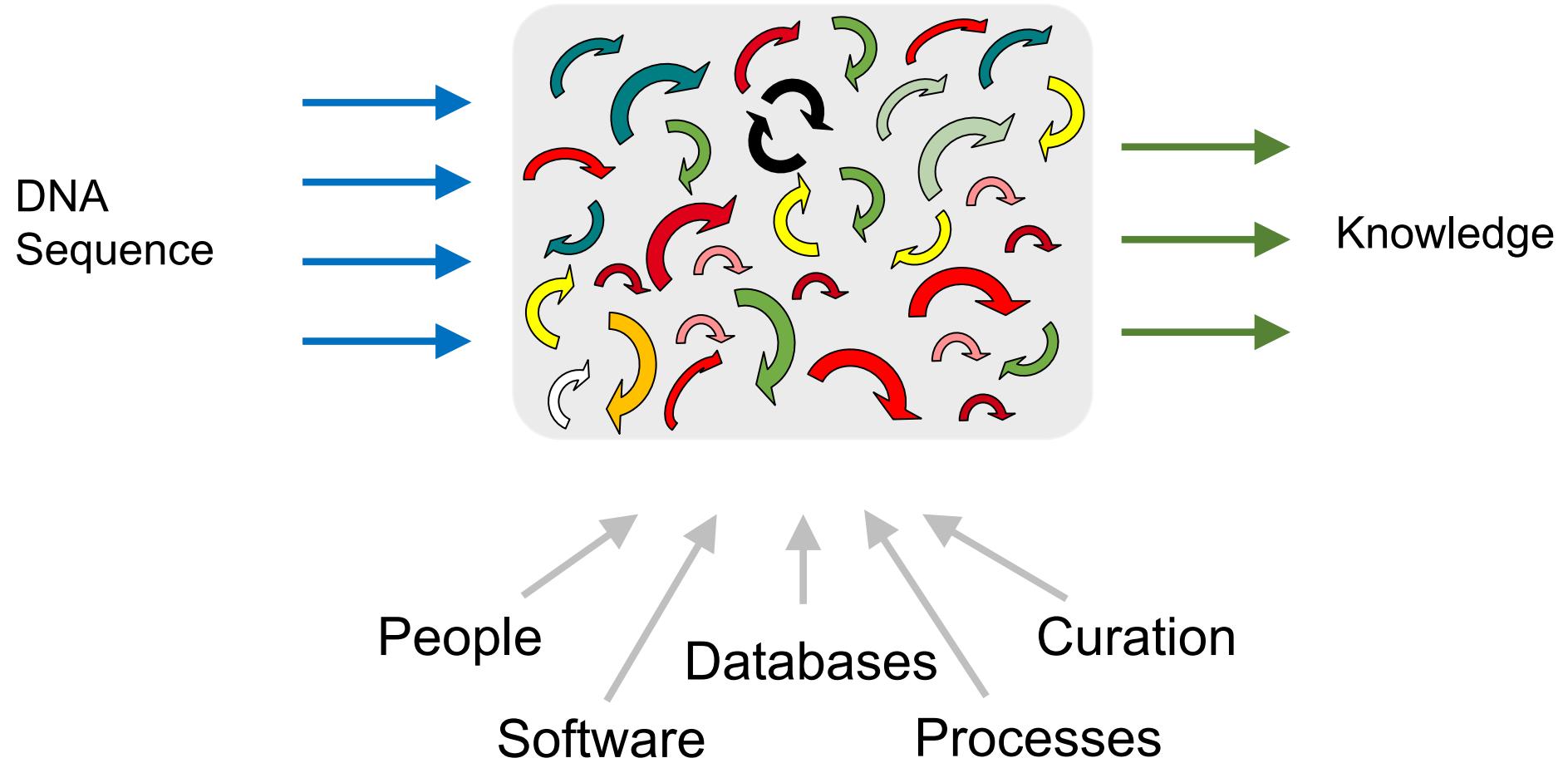
“If I have seen further than others, then it
is by standing on the shoulders of giants.”

Isaac Newton

The life science “machine”



The life science “machine”



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



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

The image displays two screenshots of scientific journals. The top screenshot is from the journal 'nature', showing a DNA sequence with the text "Databases: Reminder to deposit DNA sequences" and the author's name "Steven L. Salzberg". The bottom screenshot is from the journal 'Science', showing a bird in flight with the text "Reminder to deposit DNA sequences" and the author's name "Mark Blaxter¹, Antoine Danchin², Babis Savakis³, Kaoru Fukami-Kobayashi⁴, Ken Kurokawa⁵, Sumio Sugano⁶, Richard J. Roberts⁷, Steven L. Salzberg^{8,9}, Chung-I Wu¹⁰". Both screenshots include navigation menus and article metadata.

Organisation

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



Instruments

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

International Nucleotide Sequence Database Collaboration (INSDC)

TATACTTAATTTCGGTGCGCTGGAGCAGGAATAGTAGGAACCCGGCTTAAGTAACTTATTAACTCG
GAACTGAACCTAGTCACCCAGGCACTCTATTGGGCGATGACCAAATTATAATGTAATCG
TGACTGCTCATGCATTGTTATAATCTTCTTATAGTTATACCAATTATAATCGGGGGT
TTGGAAACTGATTGATACCCCTCATATTGGAGCCCCAGACATAGCCTTCCACGAATAA
ATAATATAAGCTCTGACTGTTACCACCATTTCTCTTCTATTAGCTTCAGCAGGAG
TAGAGGCCGGAGCAGGCACAGGATGAACAGTTATCGCCCTTGCTGAAATATTGCTC
ATGCCGGAGCATCAGTCGACTTAACCTATCTTCACTGCACTTAGCAGGGGCCTCTTCAA
TCCTAGCCTCTATTAACTTATCACCACAATTATTAACATAAAATCACCATCAATTACAC
AATACCAACACCACTTTTGTGTGATCATTACTTACACTGCAATTCTCTACTTCTAT
CATTACCACTAGCAGCAGGAATTACAATACTTCTGACAGACCGAAATCTAACACCA
CATTCTCGACCCAGCTGGAGGAGGTGACCCAATTCTCTATCAACATTATTCTGATTCT
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:
tcaaccaaccacaaagacatggcac, rev_name: coi r, rev_seq:
tagacttcgttgcggccaaagaatca"
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="YLIFGAWAGMVGTLGLSLLIRTELSPQPGTLGGDDQIYNNVITAHAF
VMIFPMVMPIMIGFGGNLMLPLMIGAPDMAPPSSFLLLASAGVEAGA
GTGWTVPVPPLAGNIHAGASVDLTFSLHLAGASSILASINFITIINMKSPSITQYQT
FT PLFVWSLLITAILLSSLVLAAGITMLLTDRNLNTFFDPAGGDPILYQHFLWFQGH
PEV"
XX
SQ Sequence 679 BP; 197 A; 166 C; 109 G; 207 T; 0 other;
ttacttaat ttccgtgccc ttagcaggaa tagtaggaac cggcttaatg ttataatcc 60
gaactgaat tagtcaccaa gggacttat tggggatca cccaaattat ataatgtatcg 120
ttgactgtca tcatttcgtt ataaatcttc tttatgtatt accaaattat atccgggggt 180
ttggaaactg atttgatcccc ctccataatg gggcccccaga catagccccc ccacgaataa 240
ataatccatg ctctgtactg ttaccacat cttcccttc ttatcttagt tcagcaggag 300
tagaggcccg acggccggaa ggatcgaatc ttatccggcc cttttgttcc aaatattgtcc 360
atggccggcc atccggccggaa ttatctatctt ttccatgtca ctttgcgggg gcctttccaa 420
tccttagccctt ttaatattttt atccaccaatc ttatataatc cccaaatccaca tcataatcc 480
aataccaaac accatctttt gtgtgtatcat tactttatccat tgcacccatcc ttactttat 540
cattacccatg actatccatc ggaatttacaa tactttgtcc agacccaaat cttaaacacca 600
cattttccatc cccatccatg ggatgttccca caatttttca tcaacatcca ttcttgatttc 660
ttggccaccc agaactcta 679

Clerical information

Taxonomy

Literature

Cross-references

Source information

Biological features



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

Please select the checklist that you wish to use for your sample submission
If you already have a spreadsheet containing your data upload it here.
Default Checklist
Book1 - Microsoft Excel

#checklist_accession	ER000003	sample_name	common_name	anonymized_name	sample_title	sample_description	tissue_type	sex	collection_date
1	ER000003	mouse_dendrocyte_1	Mus musculus	house mouse	mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain	male	13/0	
2		mouse_dendrocyte_2	Mus musculus	house mouse	mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain	male	13/0	
3		mouse_dendrocyte_3	Mus musculus	house mouse	mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain	male	13/0	
4		mouse_dendrocyte_4	Mus musculus	house mouse	mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain	male	13/0	
5		mouse_dendrocyte_5	Mus musculus	house mouse	mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain	male	13/0	
6		mouse_dendrocyte_6	Mus musculus	house mouse	mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain	male	13/0	
7		mouse_dendrocyte_7	Mus musculus	house mouse	mouse dendrocyte	A dendrocyte brain sample was extracted from a mouse brain	male	13/0	

Join GitHub today
GitHub is home to over 30 million developers working together to host and review code, manage projects, and build software together.

Main command-line submission interface:

commit	branch	release	Pending contributions	Apache 2.0
username	main			
grc				last year
grc_repair				10 days ago
reports				2 months ago
etc				4 months ago
gitignore				4 months ago
grc_vcf				4 months ago
LICENSE				4 months ago
README.md				4 months ago
build_profile				10 hours ago

STUDY ERP123456
SAMPLE ER5123456
NAME myReads_1
PLATFORM Illumina MiSeq
INSERT_SIZE 200
LIBRARY_SOURCE METADENOMIC
LIBRARY_SELECTION RANDOM
LIBRARY_STRATEGY WGS
FASTQ read_1.fastq.gz
FASTQ read_2.fastq.gz

STUDY ERP123456
SAMPLE ER5123456
ASSEMBLYNAME myAssembly
ASSEMBLY_TYPE primary metagenome
COVERAGE 25
PROGRAM metaspades v3.1.1
PLATFORM Illumina MiSeq
FASTA metagenome.fasta.gz

Presentation

The collage of screenshots illustrates the ENA platform's features:

- Search:** The main ENA homepage with a search bar and navigation menu.
- Advanced Search:** A detailed view of the search interface, showing a query builder with a map for geographical location filtering.
- Support:** A form for contacting the helpdesk, asking for submission-related support.
- Filter:** A screenshot of the "My Rules" section, showing a table of saved search rules with columns for Name, Description, Creation Time, Updated Time, View, Edit, Delete, and Search.

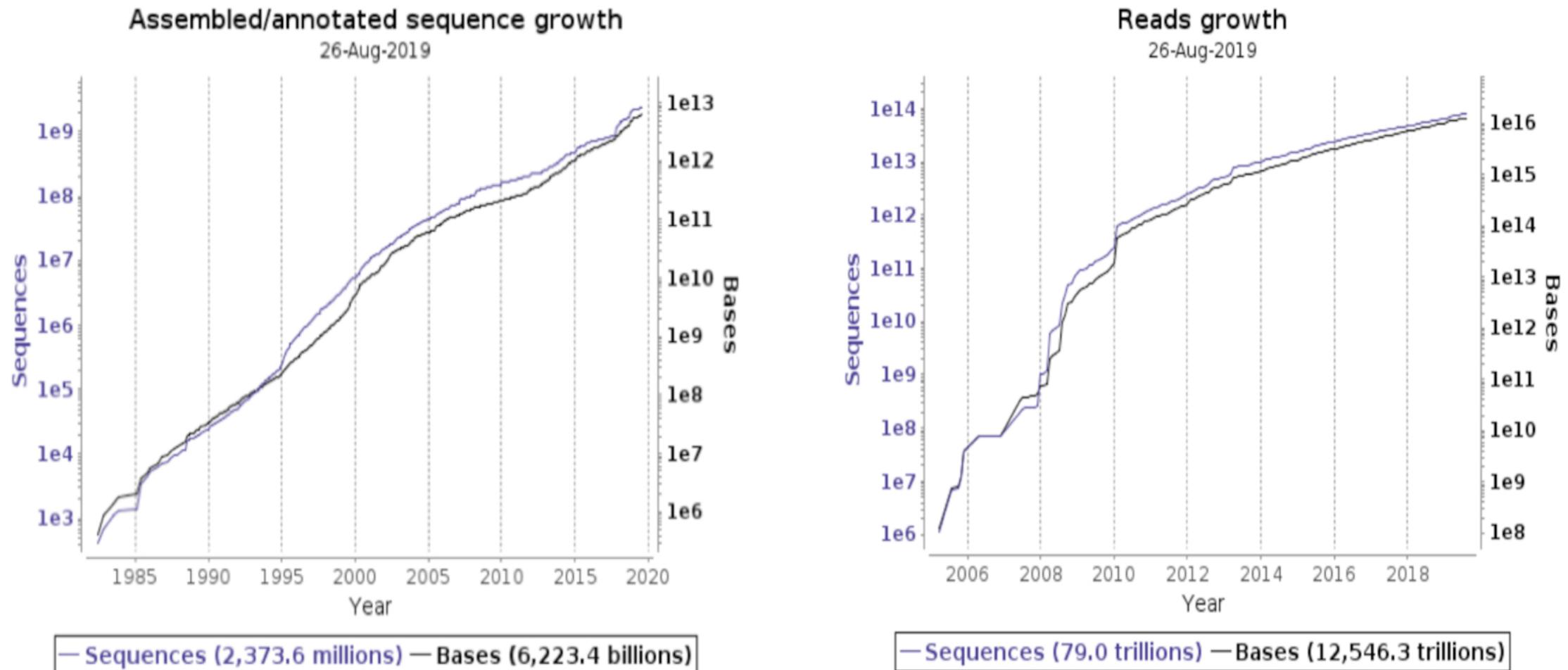
- Search
 - Sample characteristics
 - Function
 - Sequence similarity
- Browse
 - Study
 - Taxonomy
 - Sequencing method
 - Historic versions
- Filter
 - Reusable/shareable rules
 - Synchronisation with latest data
- Retrieve
 - Download tools

Scale

- Rate: 1 new dataset every 6 minutes
- Data: 2×10^9 sequences and 1×10^{16} base pairs of read data across 2×10^6 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*

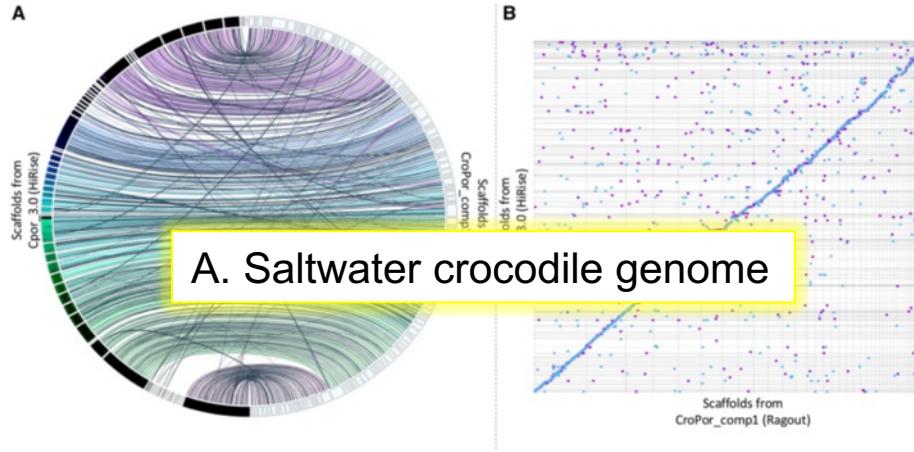


Data growth



Direct use

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



SpringerLink

Published: 10 January 2020

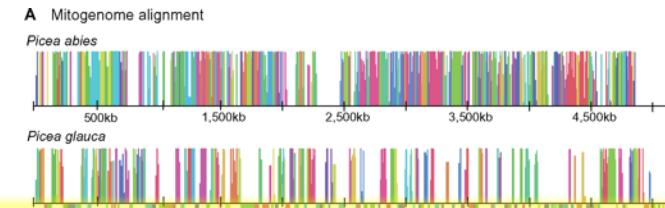
Optimizing the Reduction of Molybdate by Two Novel Thermophilic Bacilli Isolated from Sinai, Egypt

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

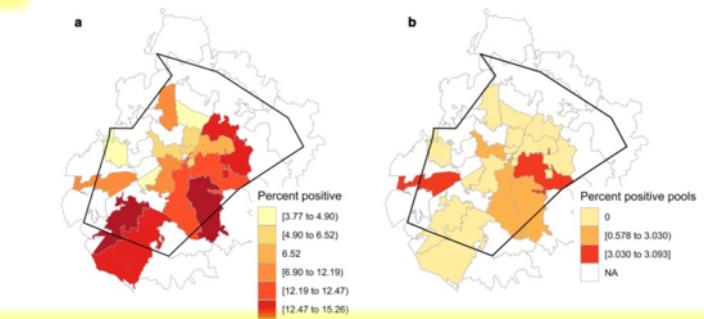
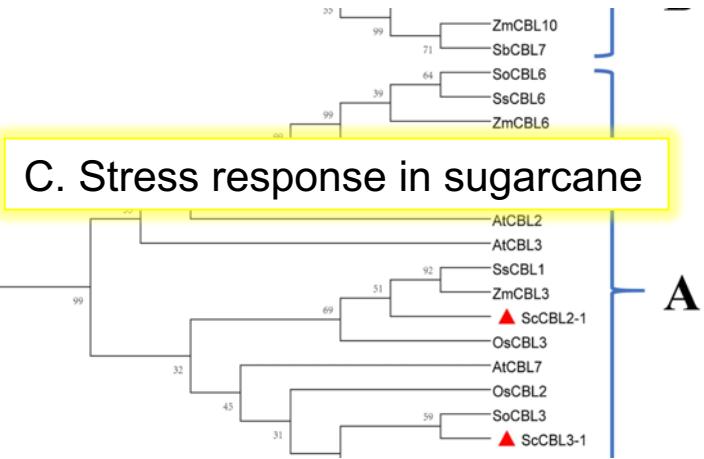
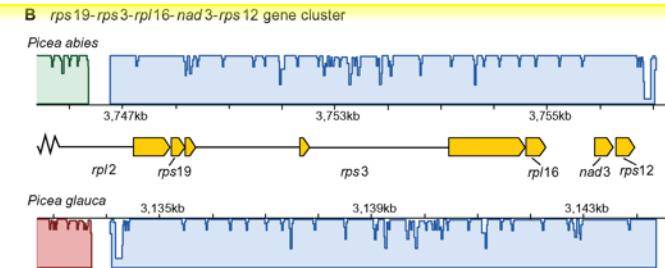
Curren-

53 Acc

D. Molybdenum bioremediation



B. Norway spruce evolutionary biology



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/>

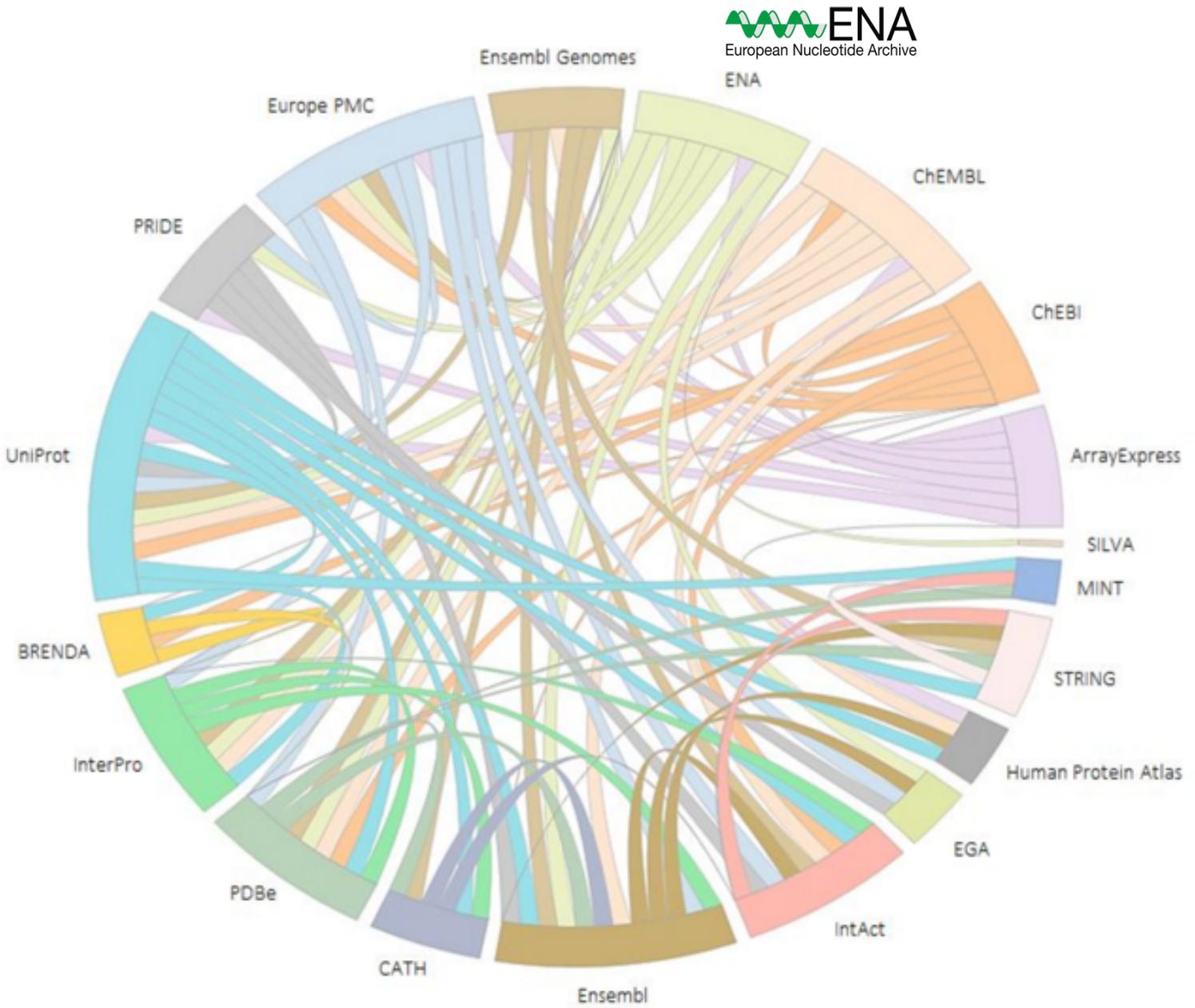
B. Sullivan AR et al., *Genome Biol Evol*. 2020 Jan;12(1):3586-3598. doi:10.1093/gbe/evz263. PMID: 31774499; © The Author(s) 2020; <http://creativecommons.org/licenses/by-nc/4.0/>

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/>

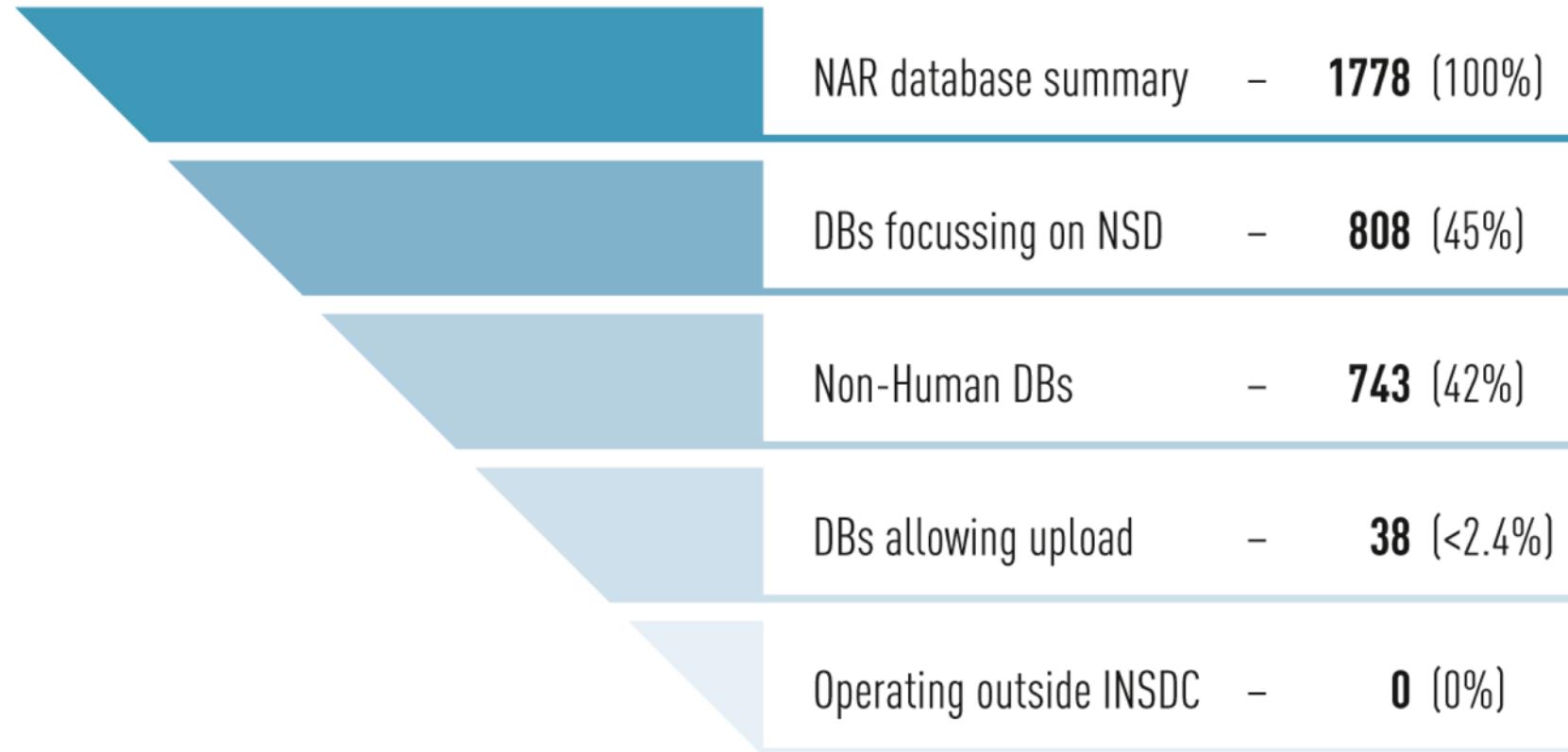
Data reach



Further reach



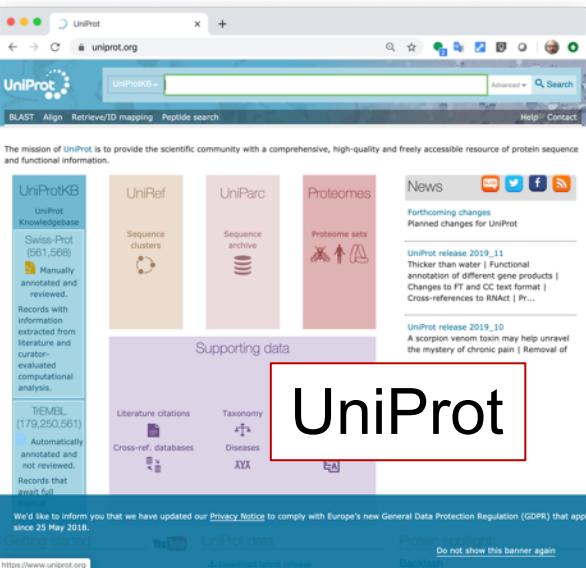
Public database inventory



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.ncbi.nlm.nih.gov/pmc/articles/PMC6907333/>

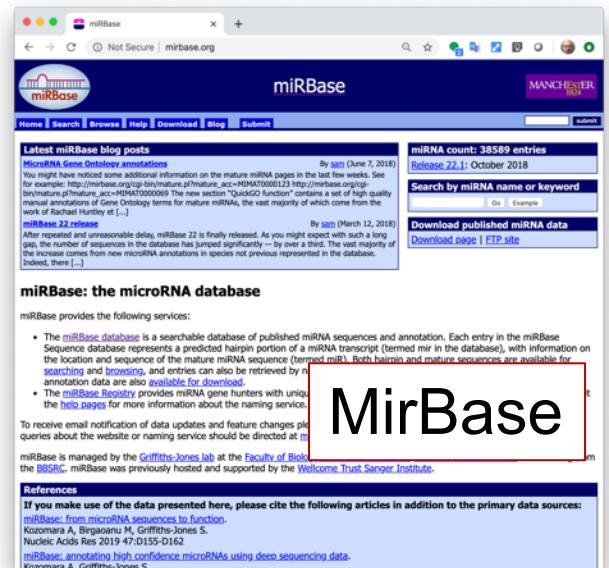
Secondary databases

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



The screenshot shows the UniProt homepage. It features a sidebar with sections for UniProtKB (Swiss-Prot and TrEMBL), UniRef, UniParc, Proteomes, News, and Forthcoming changes. The main content area displays a grid of supporting data categories: Sequence clusters, Sequence archive, Proteome sets, Literature citations, Taxonomy, Diseases, and ATOM. A large central banner says "UniProt". Below the banner, a message informs users about the update of the Privacy Notice to comply with the General Data Protection Regulation (GDPR) from May 25, 2018.

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



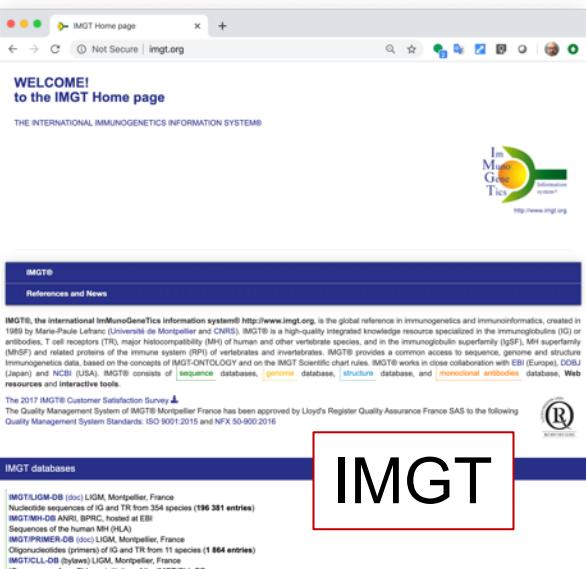
The screenshot shows the miRBase homepage. It features a header with the miRBase logo and navigation links. The main content area displays a banner for "Latest miRBase blog posts" and a section for "miRBase 22 release". A large central banner says "MirBase". Below the banner, a message indicates that after a reasonable delay, miRBase 22 is finally released. The page also includes a search bar and download links.

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



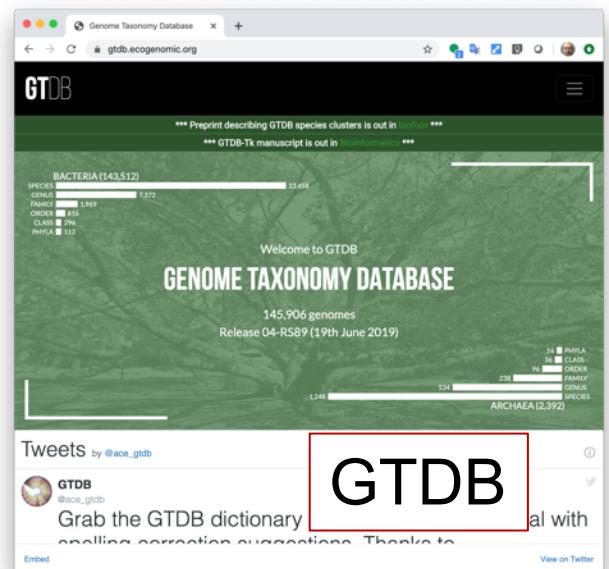
The screenshot shows the SILVA homepage. It features a header with logos for SILVA, ELIXIR, and deNBI. The main content area is divided into sections for SILVA (Welcome to the SILVA rRNA database project, SILVAngs, SILVA alignment, classification and tree (ACT) service, and SILVA tree viewer), IMGT (WELCOME to the IMGT Home page, IMGT databases, and IMGT GATC Bioinformatics center), and GTDB (Welcome to GTDB, Genome Taxonomy Database, and Tweets). A large central banner says "SILVA".

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



The screenshot shows the IMGT homepage. It features a header with the IMGT logo. The main content area displays a banner for "WELCOME to the IMGT Home page" and a section for "IMGT databases". A large central banner says "IMGT".

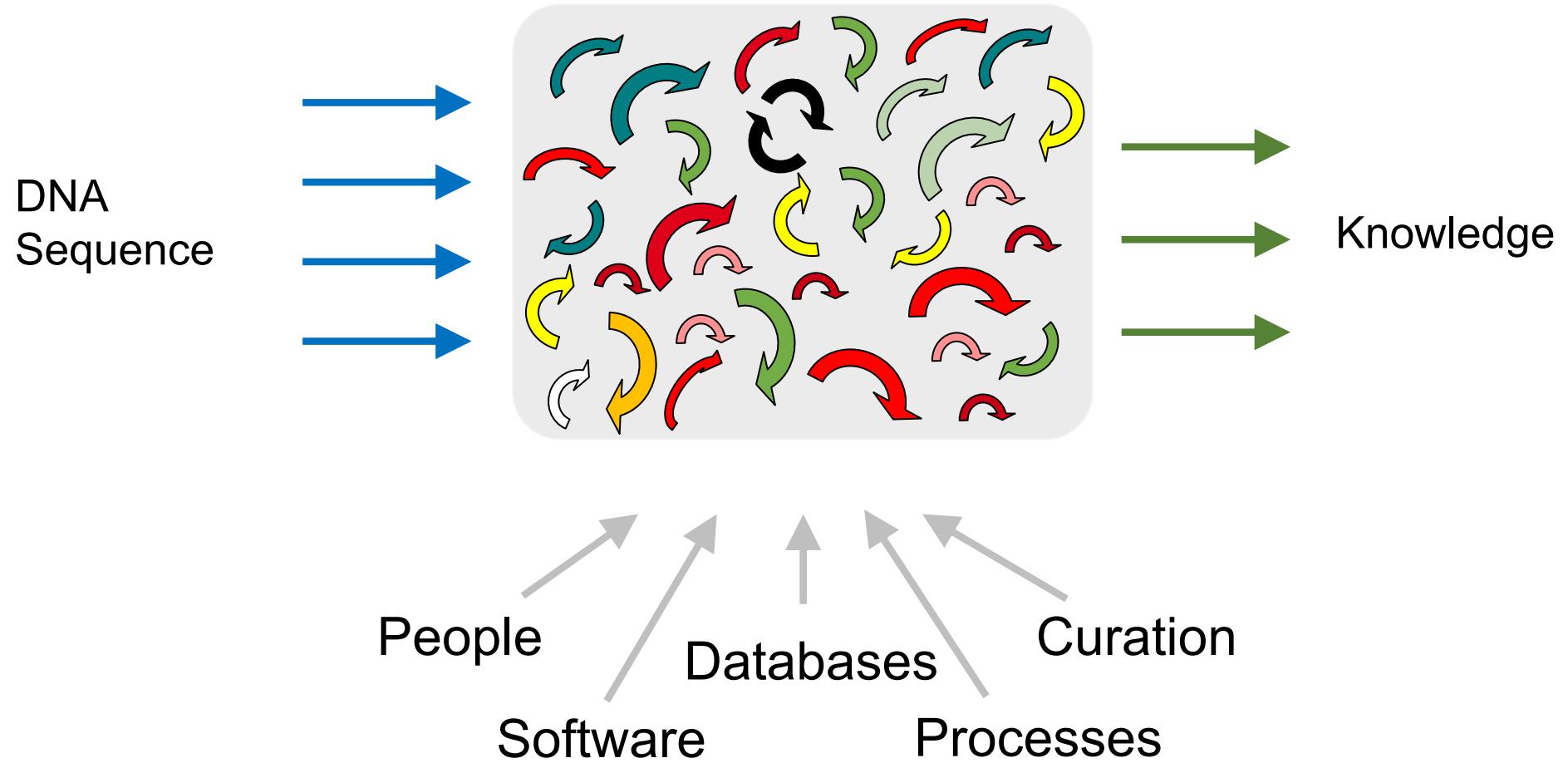
<http://www.imgt.org/>



The screenshot shows the GTDB homepage. It features a header with the GTDB logo. The main content area displays a banner for "Welcome to GTDB" and a section for "GENOME TAXONOMY DATABASE". A large central banner says "GTDB". Below the banner, there is a bar chart showing the distribution of species across different taxonomic ranks: BACTERIA (143,512), ARCHAEA (2,392), PHYLUM (54), CLASS (36), ORDER (14), FAMILY (534), GENUS (1,348), and SPECIES (33,454).

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

The life science “machine”



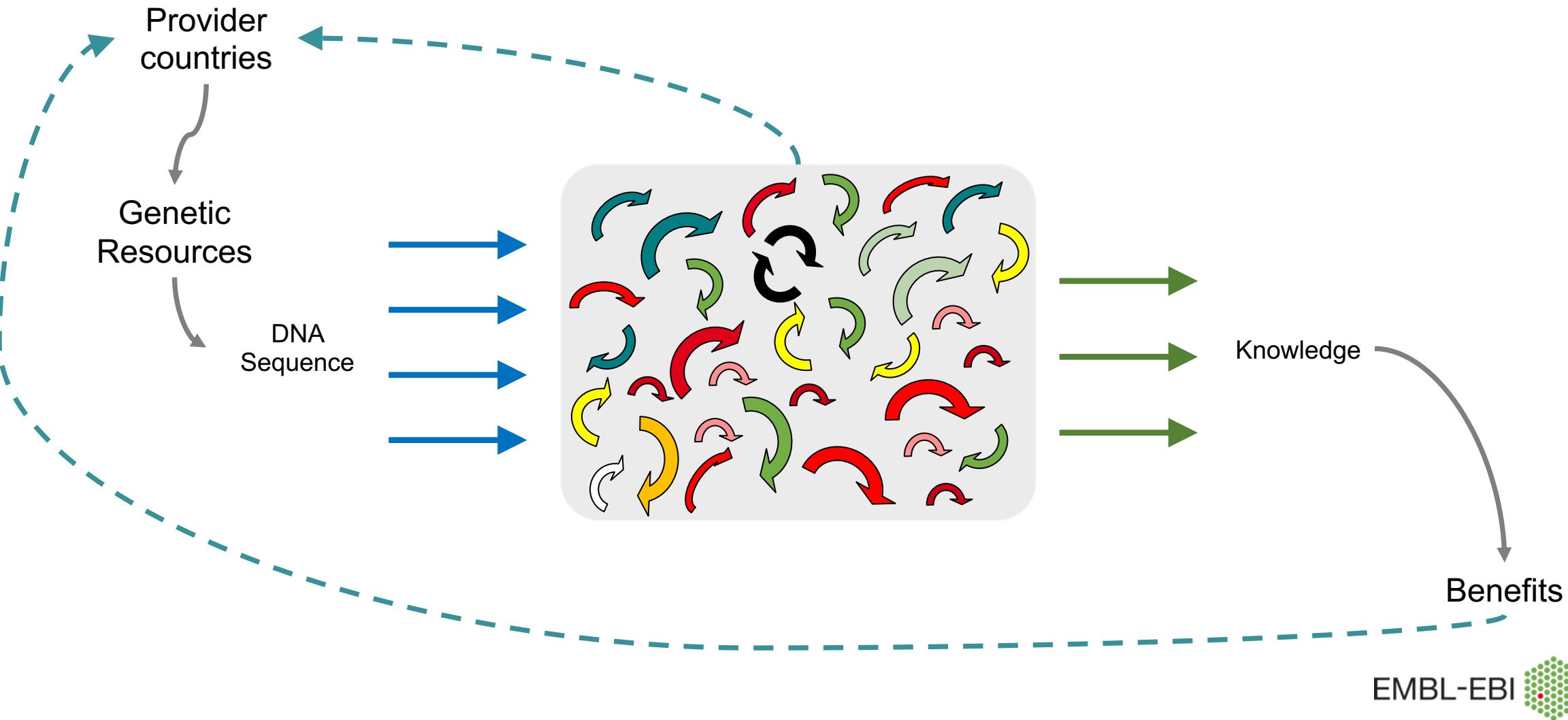
The life science “machine”

- High productivity drives scientific progress



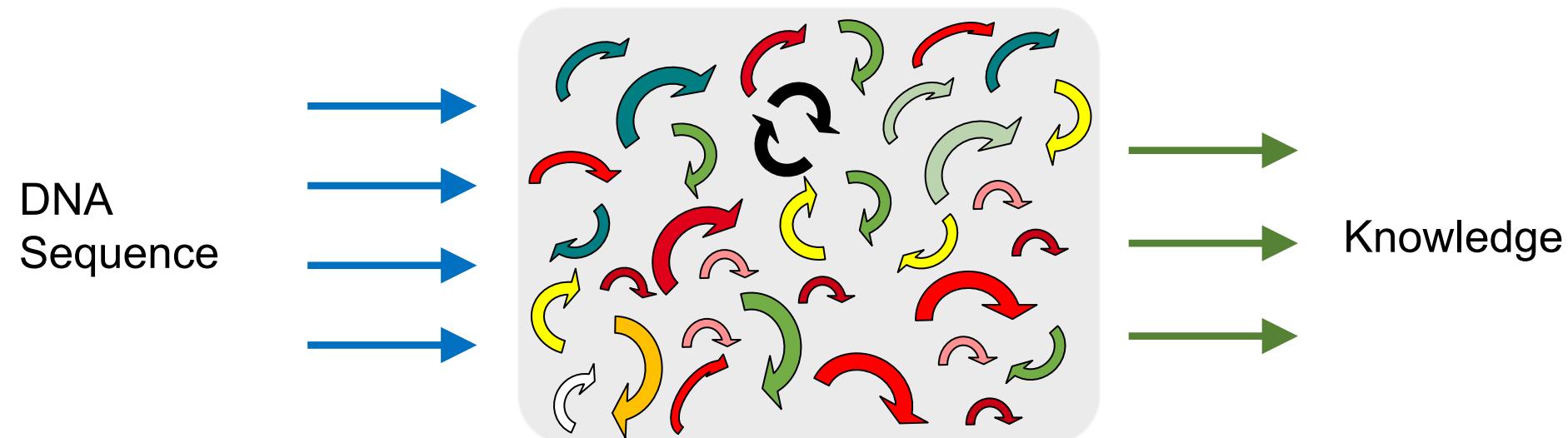
Productivity in the machine

- High productivity drives ABS



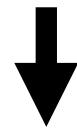
Productivity in the machine

- Understand risks
- Assert core values
- Identify opportunities

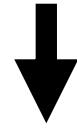


What if.. visibility of sequence were reduced?

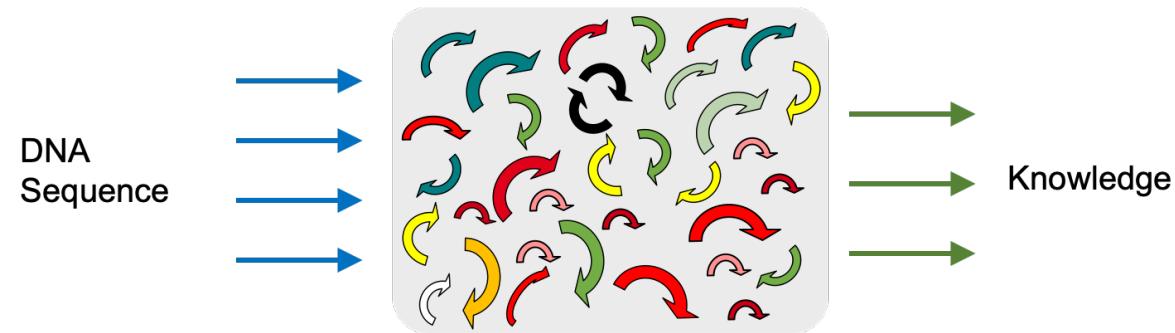
- E.G. users had to register and log-in
- Agreement not to redistribute in any form



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



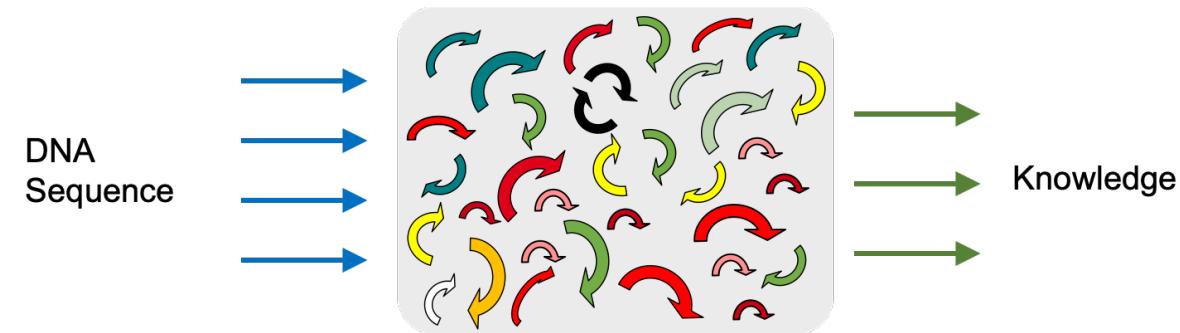
- Productivity drops to near-zero



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

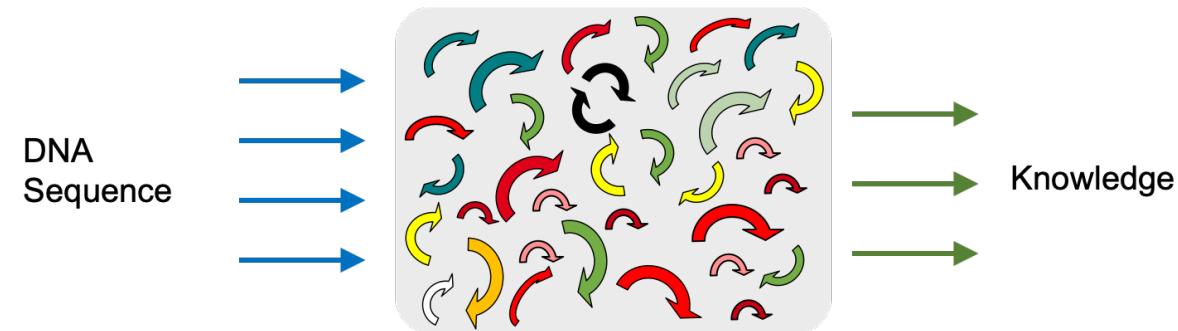
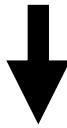


Critical property of machine:

**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

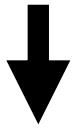


- Submission blocked if this condition is not satisfied
- Data – particularly for biodiversity and ecology applications - are enriched

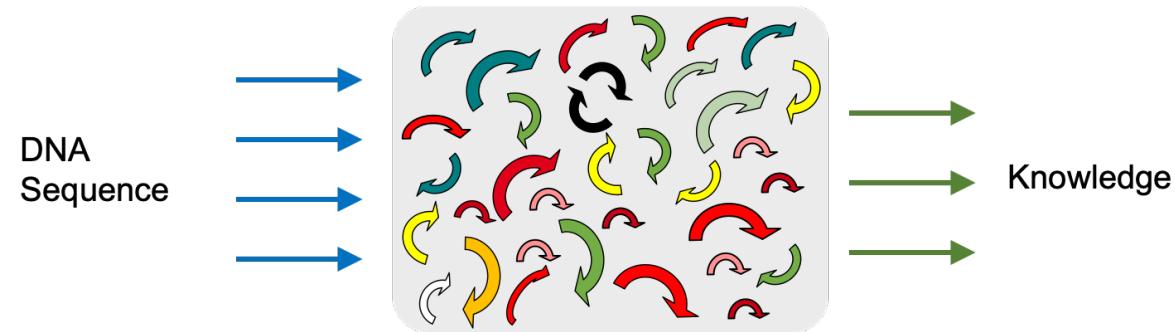
Opportunity:
Location information is scientifically informative

What if.. collaboration networks were tracked?

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

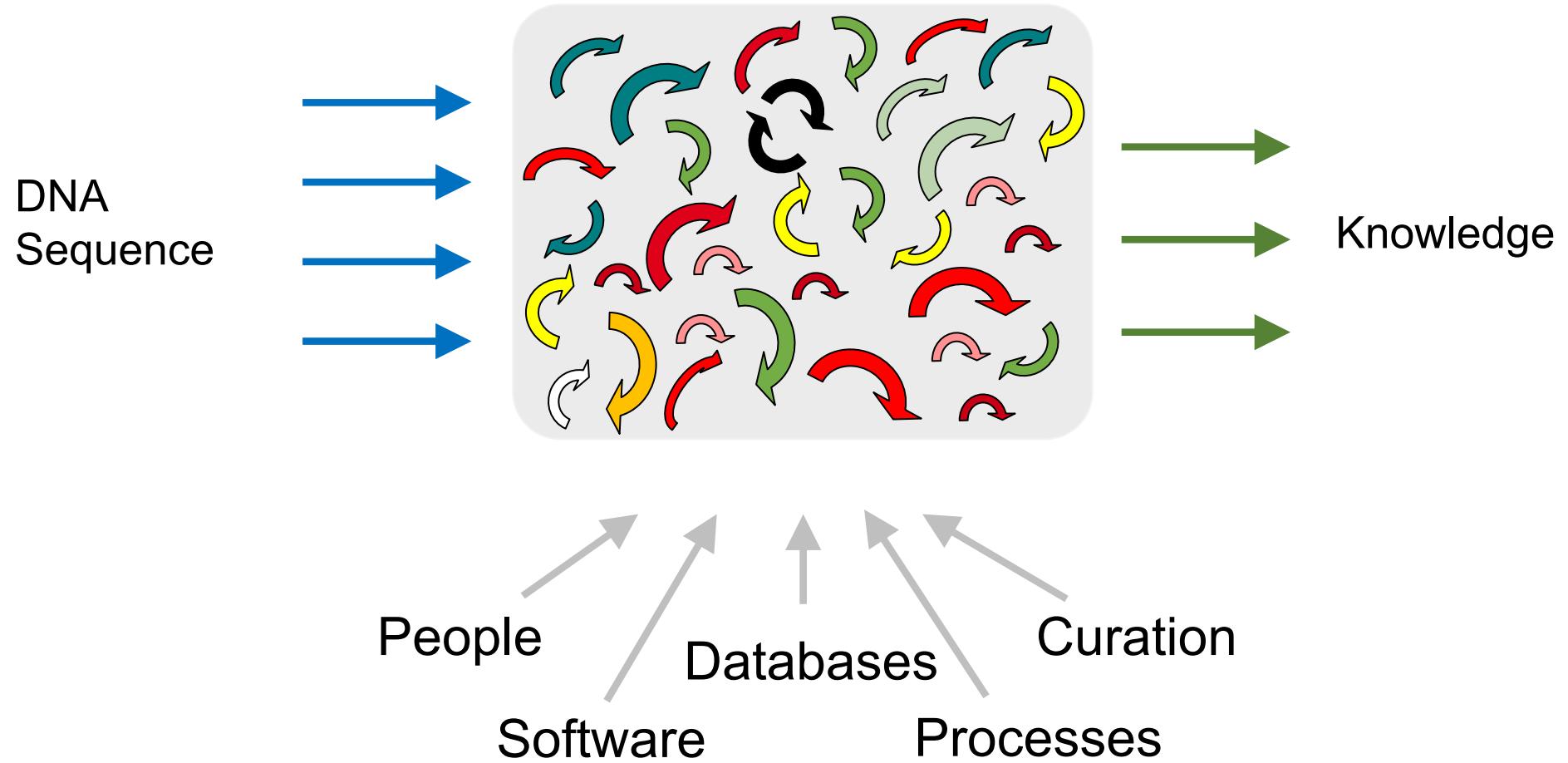


- Scientists would be able to identify potential collaborators
- Co-owned data or co-publishing authors may reveal data links



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>