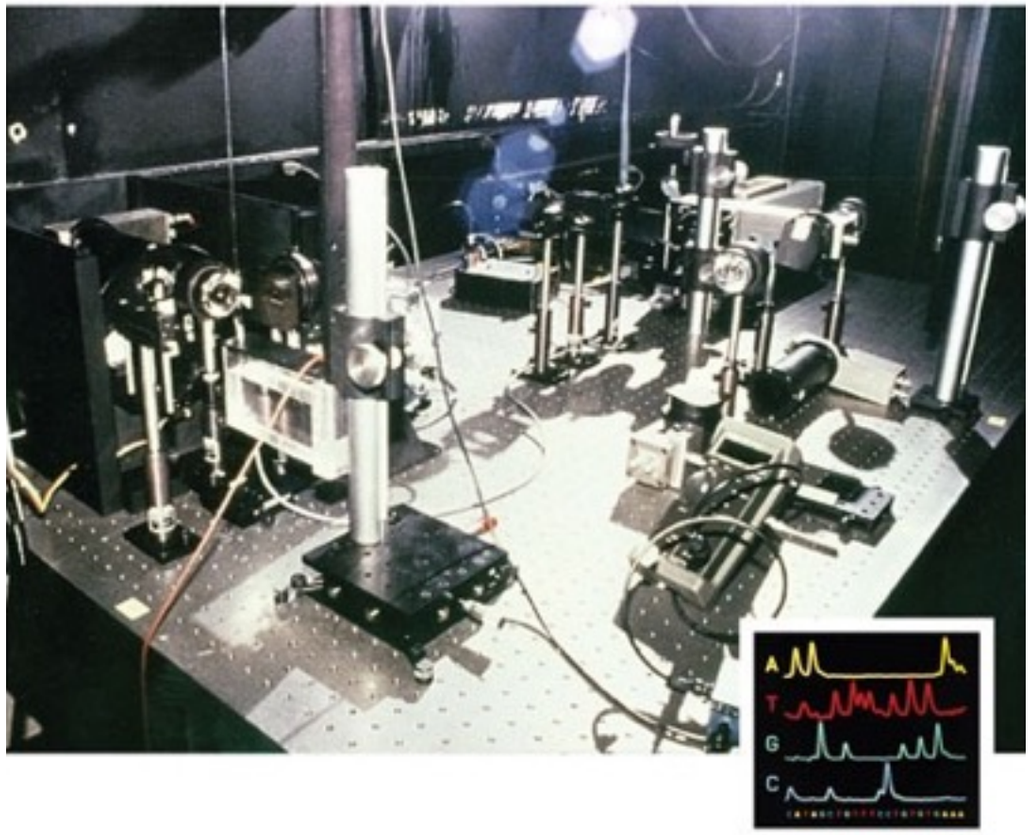
Digital Sequence Information: An Overview

Paul Oldham (PhD)

One World Analytics & United Nations University, Institute for the Advanced Study of Sustainability (UNU-IAS)

The first DNA sequencer 1985



A handheld USB sequencer 2015

DSI and the Nagoya Protocol

- What we now call Digital Sequence Information was raised repeatedly during the negotiations (in 2004 EU supported an independent research submission on this in UNEP/CBD/WG-ABS/3/INF/4). But, the discussion did not go anywhere. It has now come back.

- During the course of the negotiations the whole genomes of a growing number of organisms were sequenced. The technology for sequencing at scale has radically improved through Next Generation Sequencing (NGS). The cost of sequencing has dropped rapidly & handheld sequencers are now available.

- There are now over 2 trillion DNA bases in approx. 652 million sequences in GenBank. In 2003 when I started writing about this there were only 30 million sequences.

- This presentation will walk through some of the basics of DSI before turning to trends, costs, and geographic distribution. It will conclude with discussion of key issues arising and potential options.

Key Questions

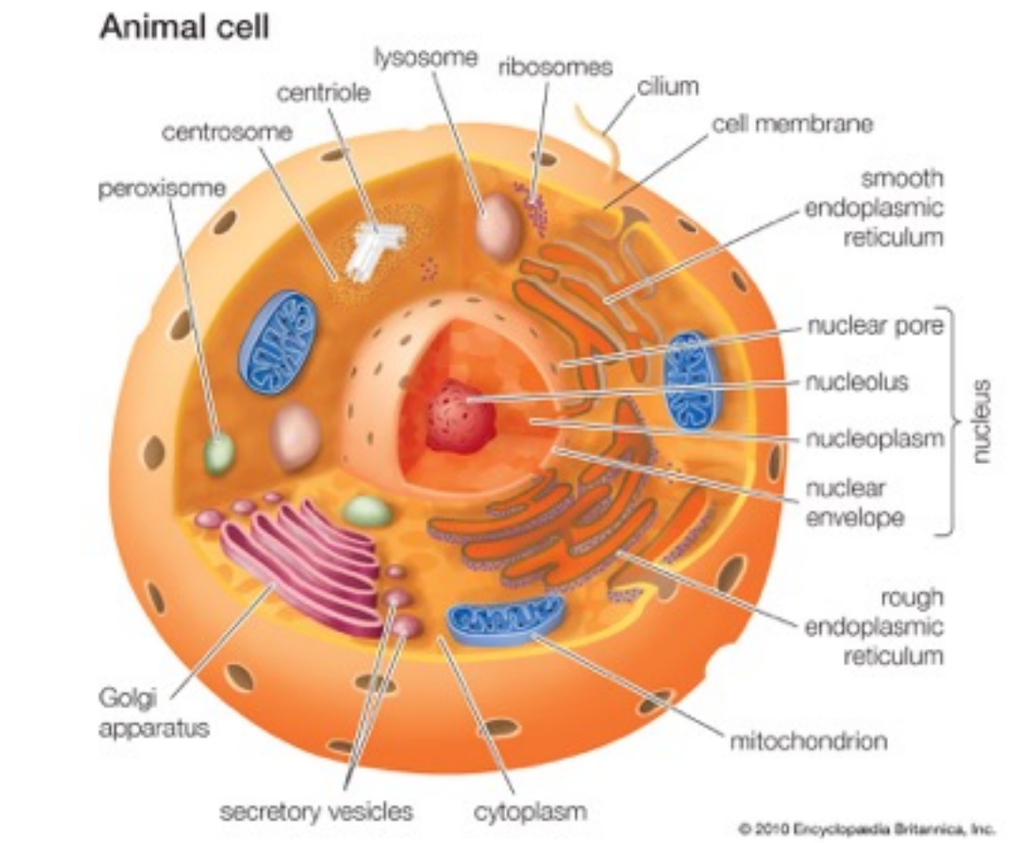
- What are the terms and conditions under which international electronic transfers are made?;
- Should electronic transfers be regulated?;
- What are the potential costs and benefits of the regulation of electronic transfers?;
- What forms of regulation of electronic transfers might be appropriate? (UNEP/CBD/WG-ABS/3/INF/4 at 15)

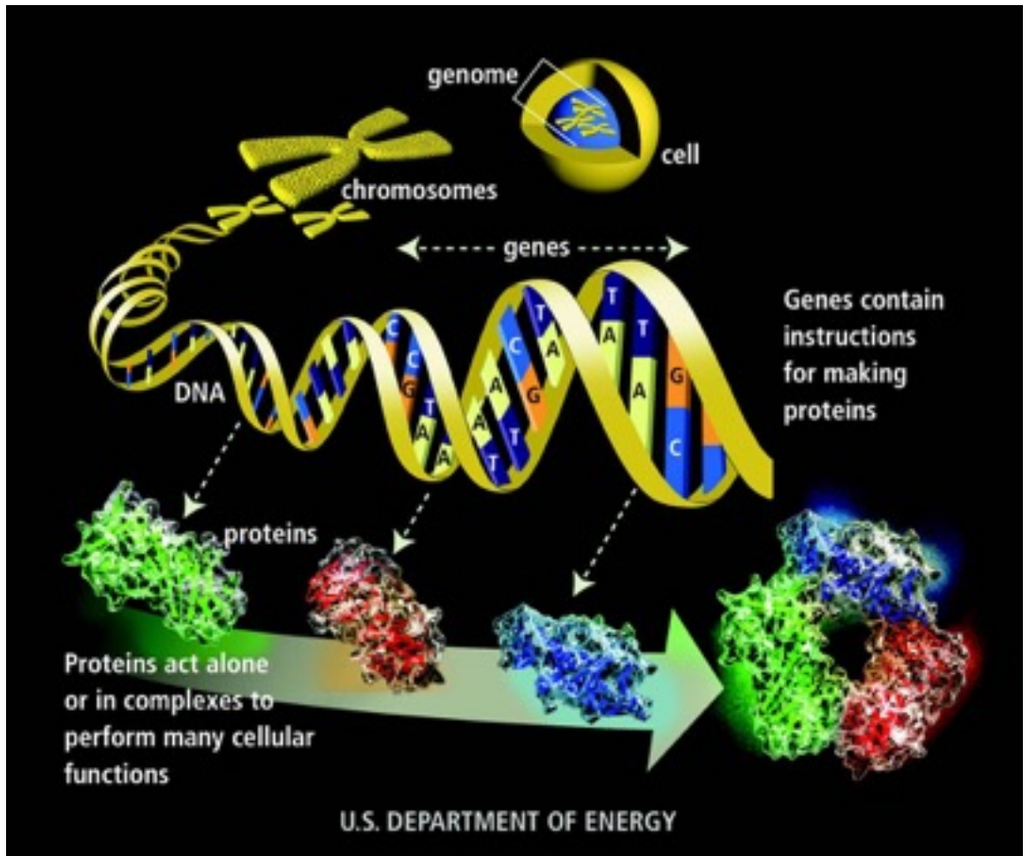
Key Issues

- Clarifying Trends
- Clarifying the who, what, where (and where from)
- Actual and Potential Uses
- Questions of Value
- Terms and Conditions of DNA Databases
- Options for Parties and Potential Consequences
- What else needs to be established?

The basics

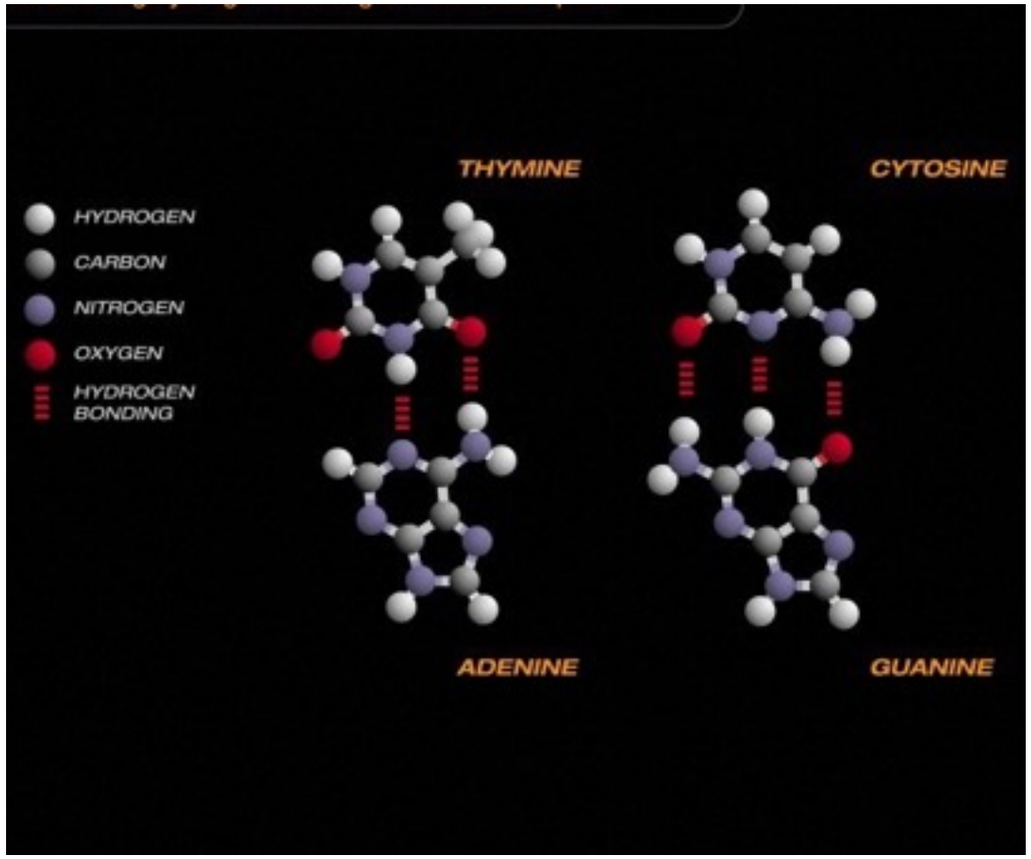
- The genome of an organism is contained in chromosomes situated in the cell nucleus.
- DNA within chromosomes are ordered into genes that lead to the expression of proteins that interact to perform cellular functions (which gets very complicated)
- All organisms (with the exception of some RNA viruses) are DNA based and may share genes that are highly conserved over evolutionary time. We can go into that later.

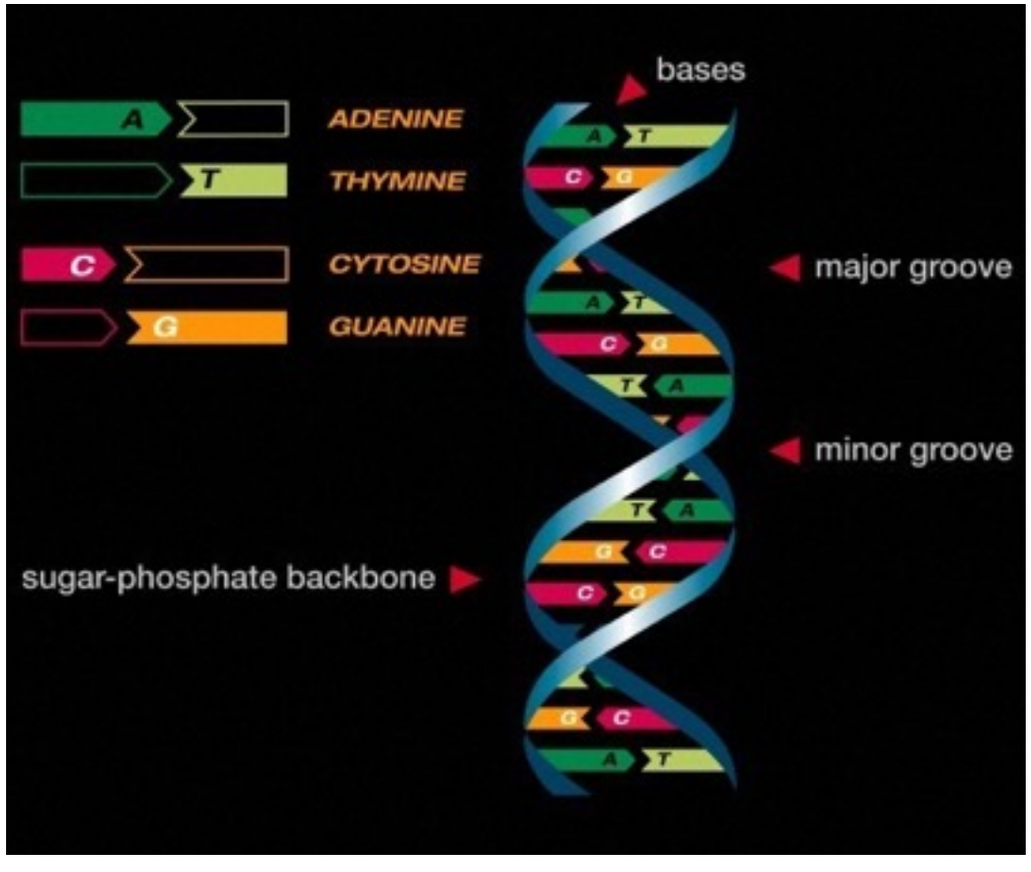




DNA & RNA

- Deoxyribonucleic acid (DNA) & Ribonucleic acid (RNA) molecules are the chemical foundations of cells and organisms.
- DNA molecules consist of four bases (A, C, T & G) that bind to each other in an ordered way (A & T, C & G) described as base pairs. RNA = Uracil instead of Thymine.
- There are different types of DNA and RNA molecules that are described in terms of sources such as mitochondrial DNA (mDNA) or functions (e.g. messenger or transfer = mRNA, tRNA)





Amino Acids

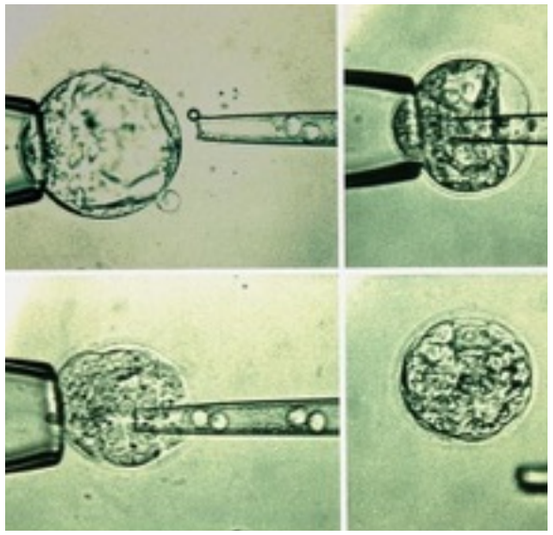
- The ordering of DNA codons (arrangements of four bases) are associated with the expression of amino acids that form the basis for building proteins
- There are 20 main amino acids that are expressed through codons.
- So the TTTC codon forms Leucine or Leu. While TTCG forms Serine or Ser.
- DNA is transcribed into amino acids and structured into proteins through bonding with RNA as the messenger that triggers gene expression in the cell.
- The important point here is that this is all expressible digitally as information.

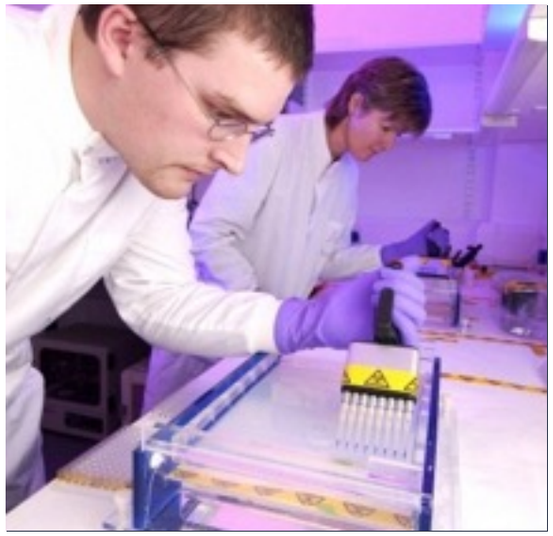
Sequence from the Rice Genome

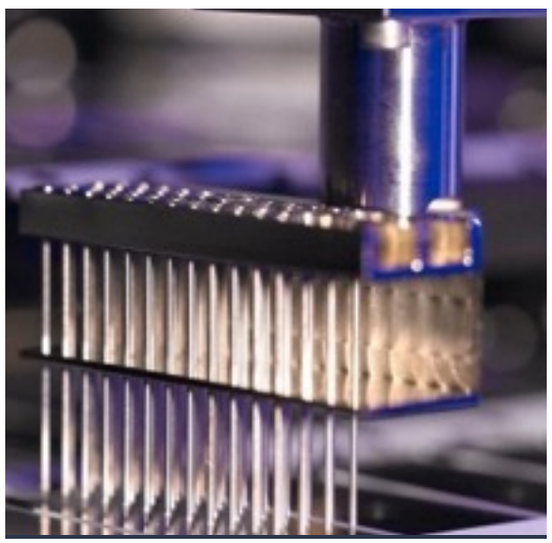
- atggggcgag ggaaagtaga gctgaaagcg gatcgagaac aagataagcc ggcaggtgac 60
- Met Gly Arg Gly Lys Val Glu Leu Lys Arg Ile Glu Asn Lys Ile Ser Arg Glu Val Thr 20

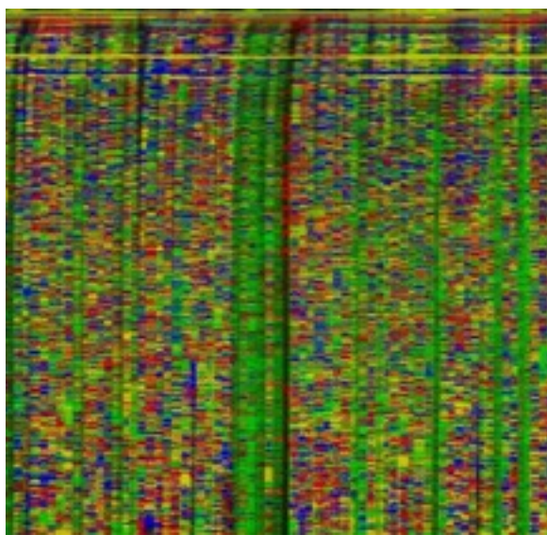
Sequencing

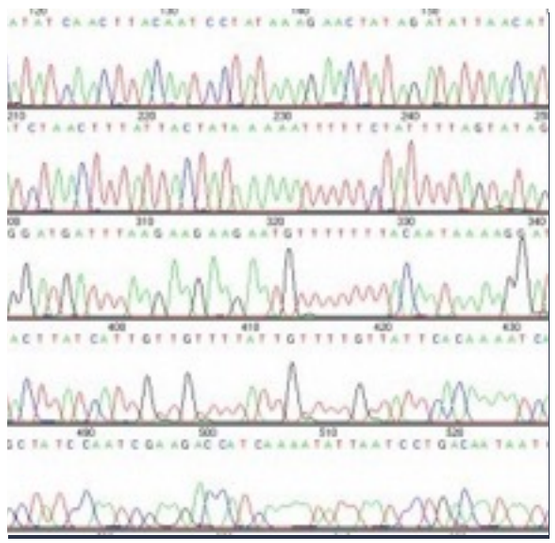
From Sanger Sequencing to Next Generation Sequencing (NGS)









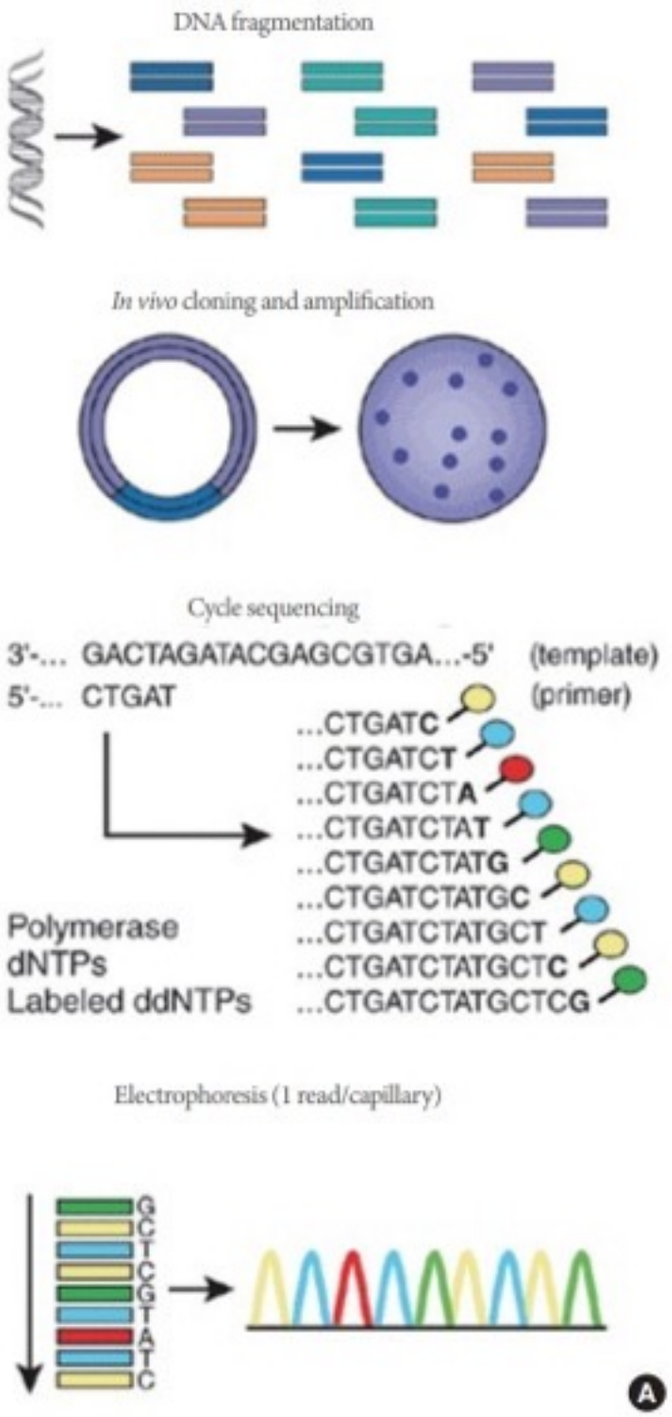


Sanger Sequencing (1977). extract into plasmids, culture colonies, extract & clean, sequence (gel), map

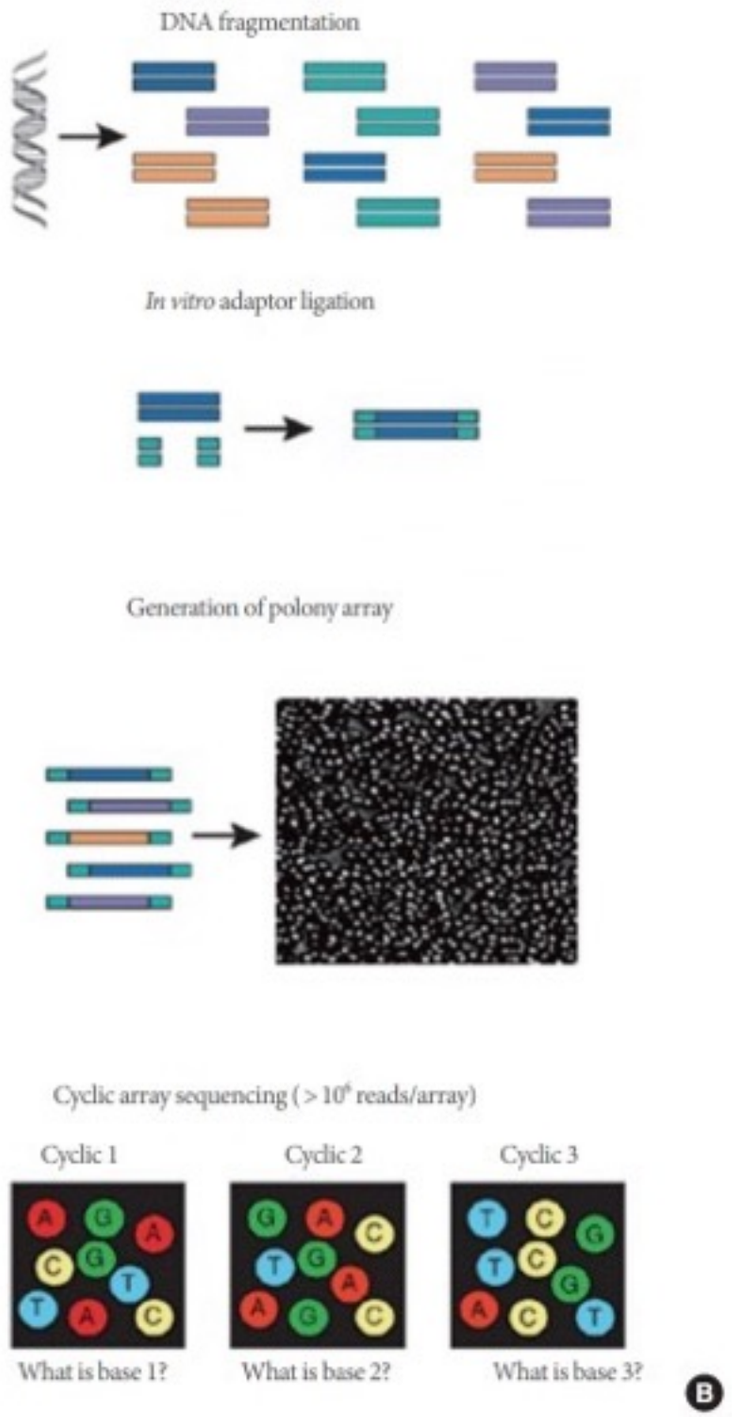
NGS

- Sanger Sequencing is accurate but slow. Next Generation Sequencing:
- Allows for the construction of libraries;
- No i*n vivo* cloning and colony picking. Done *in vitro;*
- Can be organised in arrays and highly parallel so can sequence faster and on a larger scale;
- Approaches include: pyrosequencing, sequencing by synthesis, ligation and phospholinked real time sequencing;
- Key companies include Roche, Illumina, Oxform Nanopore, Qiagen, Life Technologies, Complete Genomics, Helicos Biosciences, Pacific Biosciences.

Sanger Sequencing



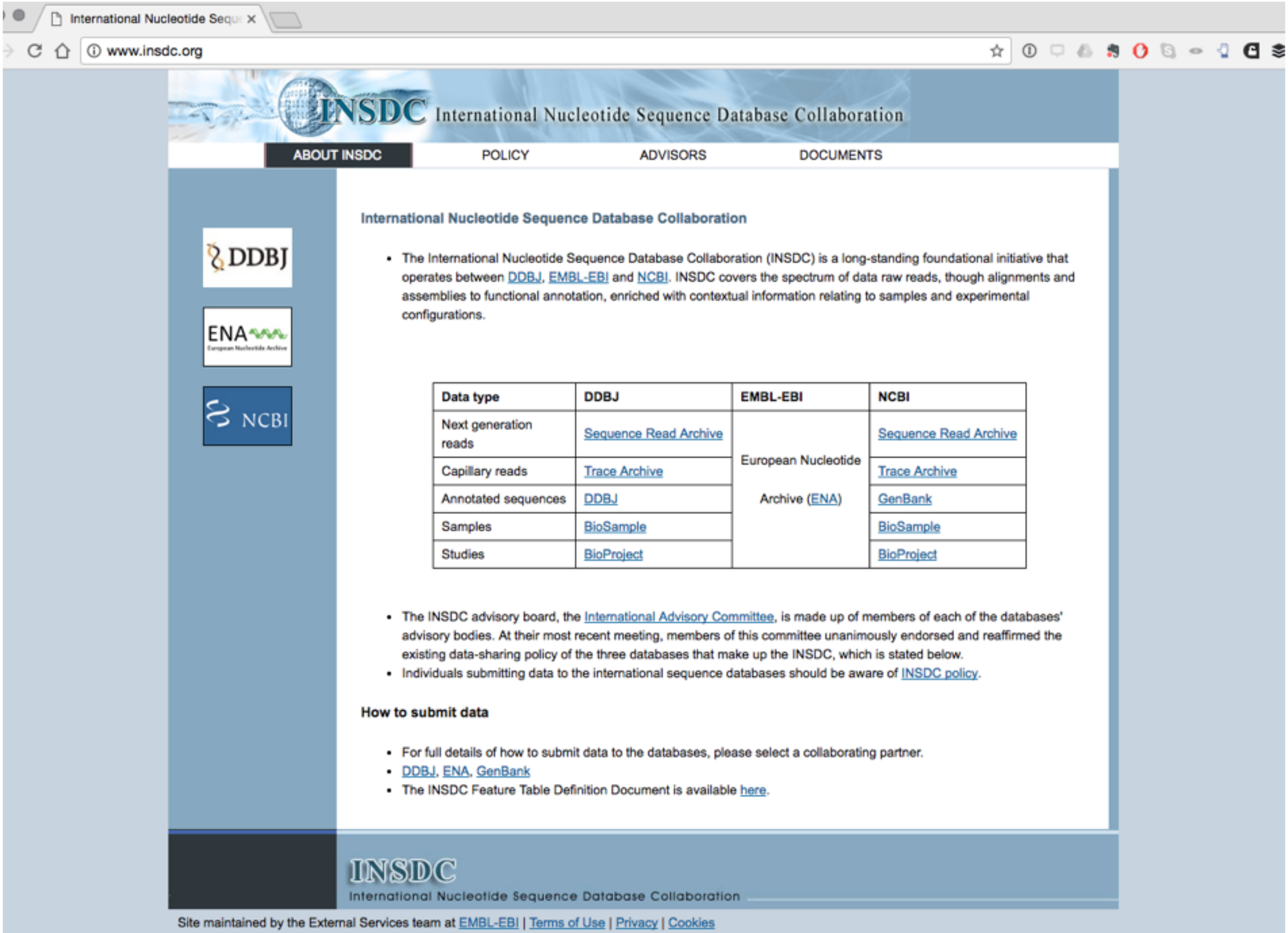
Next Generation Sequencing



Park & Kim (2016) Trends in Next Generation Sequencing and a New Era for Whole Genome Sequencing. 10.5213/inj.1632742.371

Trends

Deposits, Costs, Organisms and Actors



INSDC is made up of EMBL-EBI (EU), GenBank (US) & DDBJ (Japan)

INSDC International Nucleotide Sequence Database Collaboration

NCBI

ENA European Nucleotide Archive

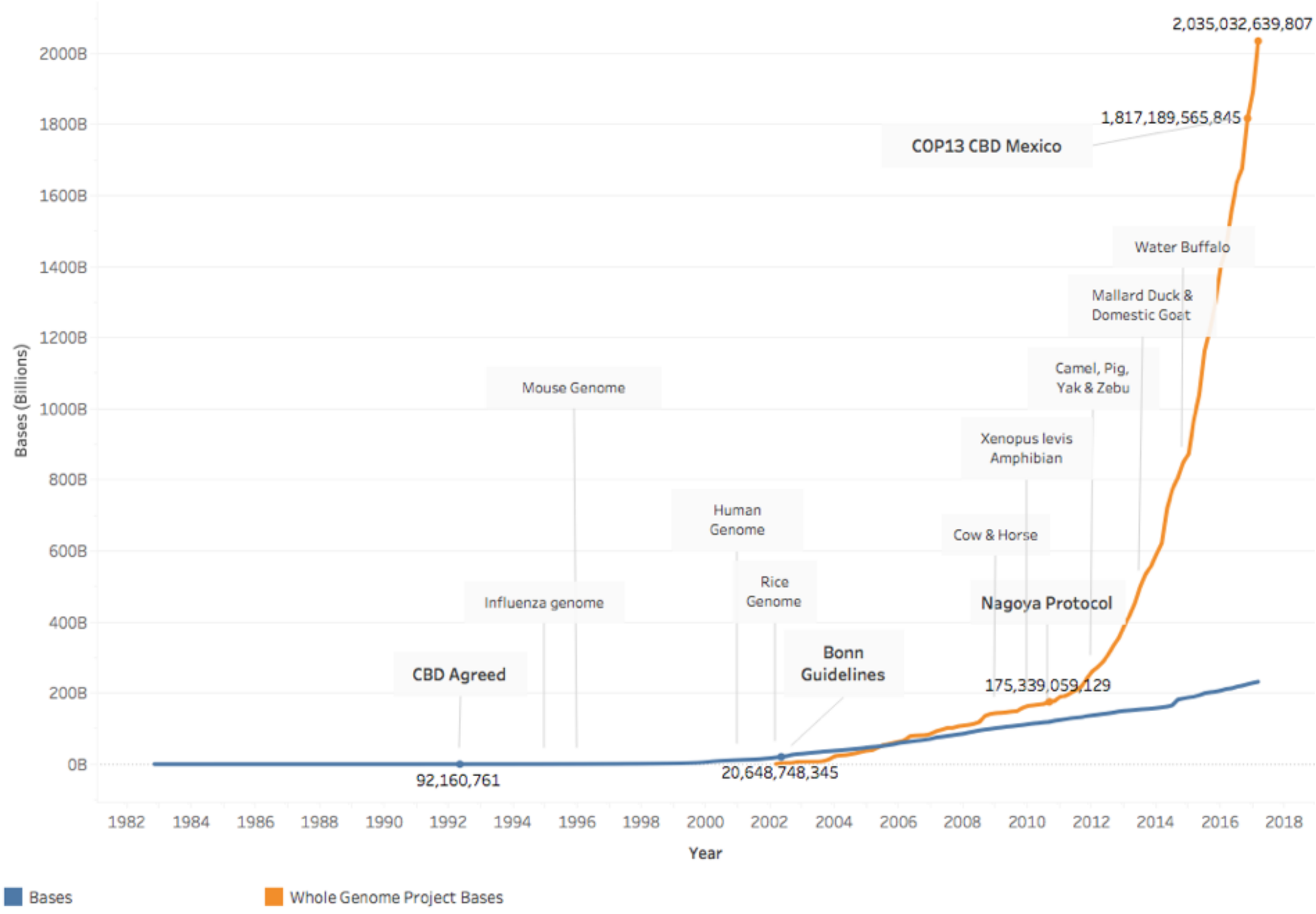
DDBJ

International Nucleotide Sequence Database Collaboration Policy

Soren Brunak, Antoine Danchin, Masahira Hattori, Haruki Nakamura, Kazuo Shinozaki, Tara Matise, Daphne Preuss (2002)
Nucleotide Sequence Database Policies
Science 298 (5597): 1333 15 Nov 2002

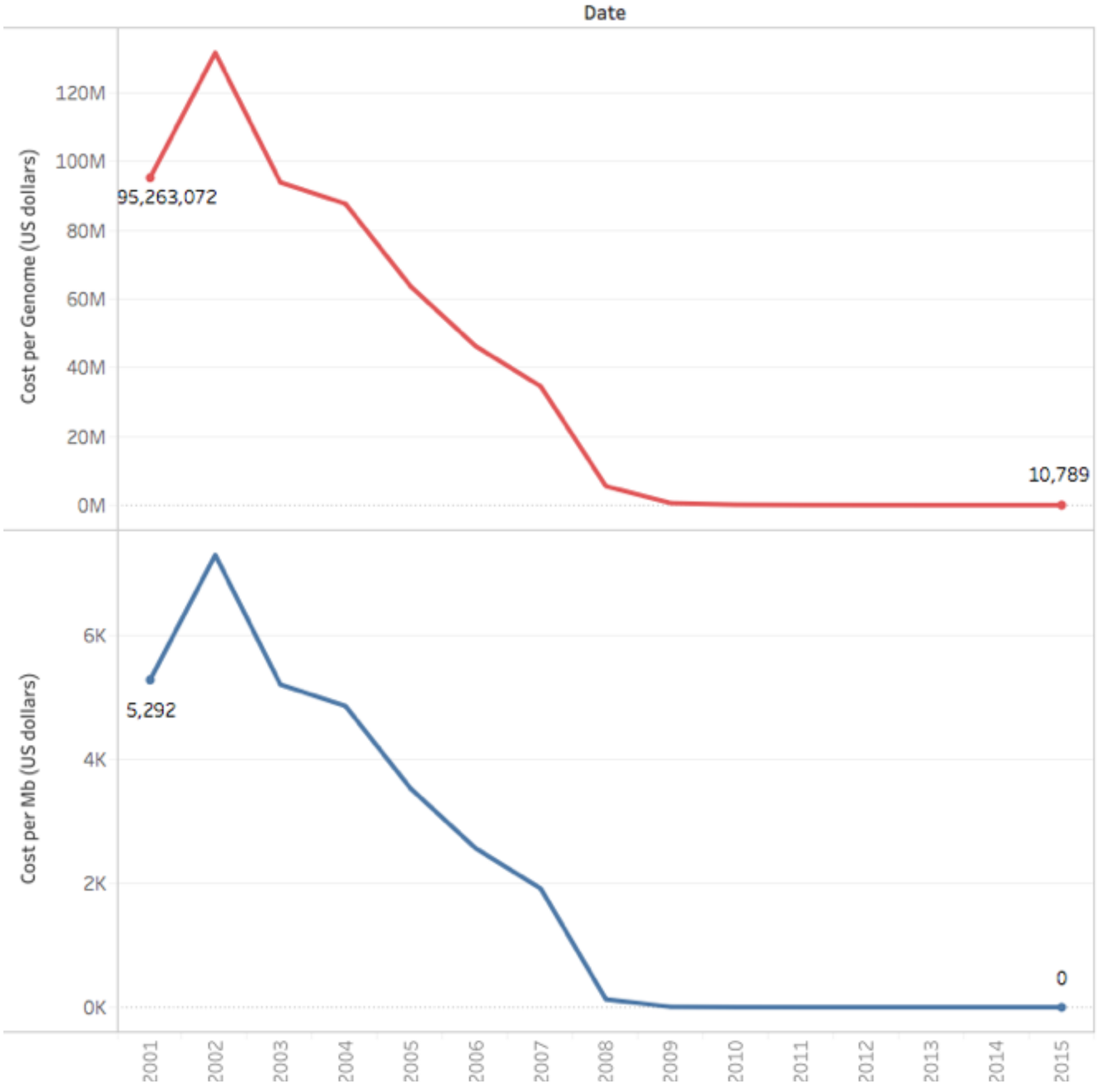
1. The INSD has a uniform policy of free and unrestricted access to all of the data records their databases contain. Scientists worldwide can access these records to plan experiments or publish any analysis or critique. Appropriate credit is given by citing the original submission, following the practices of scientists utilizing published scientific literature.
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, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party.
3. All database records submitted to the INSD will remain permanently accessible as part of the scientific record. Corrections of errors and update of the records by authors are welcome and erroneous records may be removed from the next database release, but all will remain permanently accessible by accession number.
4. Submitters are advised that the information displayed on the Web sites maintained by the INSD is fully disclosed to the public. It is the responsibility of the submitters to ascertain that they have the right to submit the data.
5. Beyond limited editorial control and some internal integrity checks (for example, proper use of INSD formats and translation of coding regions specified in CDS entries are verified), the quality and accuracy of the record are the responsibility of the submitting author, not of the database. The databases will work with submitters and users of the database to achieve the best quality resource possible.

Genbank: Trends in Bases & Whole Genome Project Bases (cumulative) with landmarks

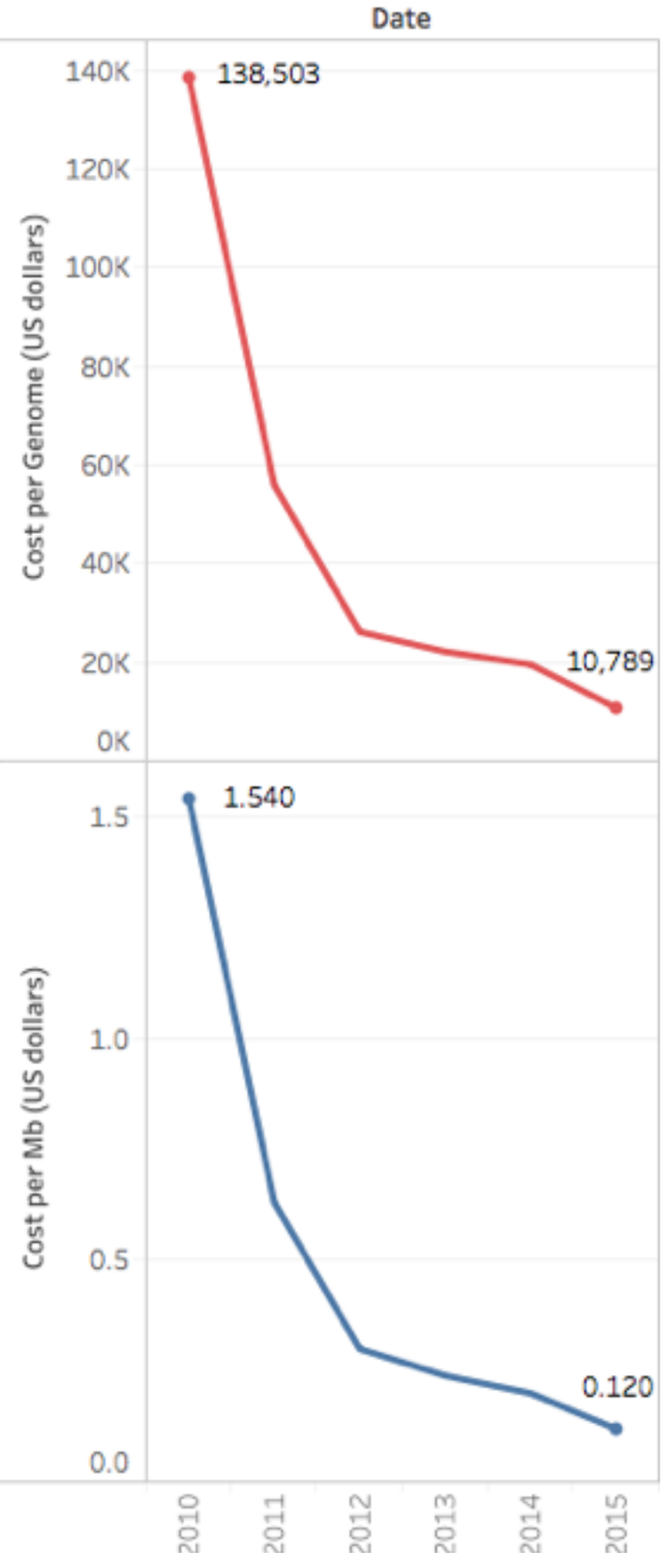


Source: GenBank and WGS Statistics. Note, the version of the data above uses a uniform scale.

Trends in Sequencing Costs (US dollars)



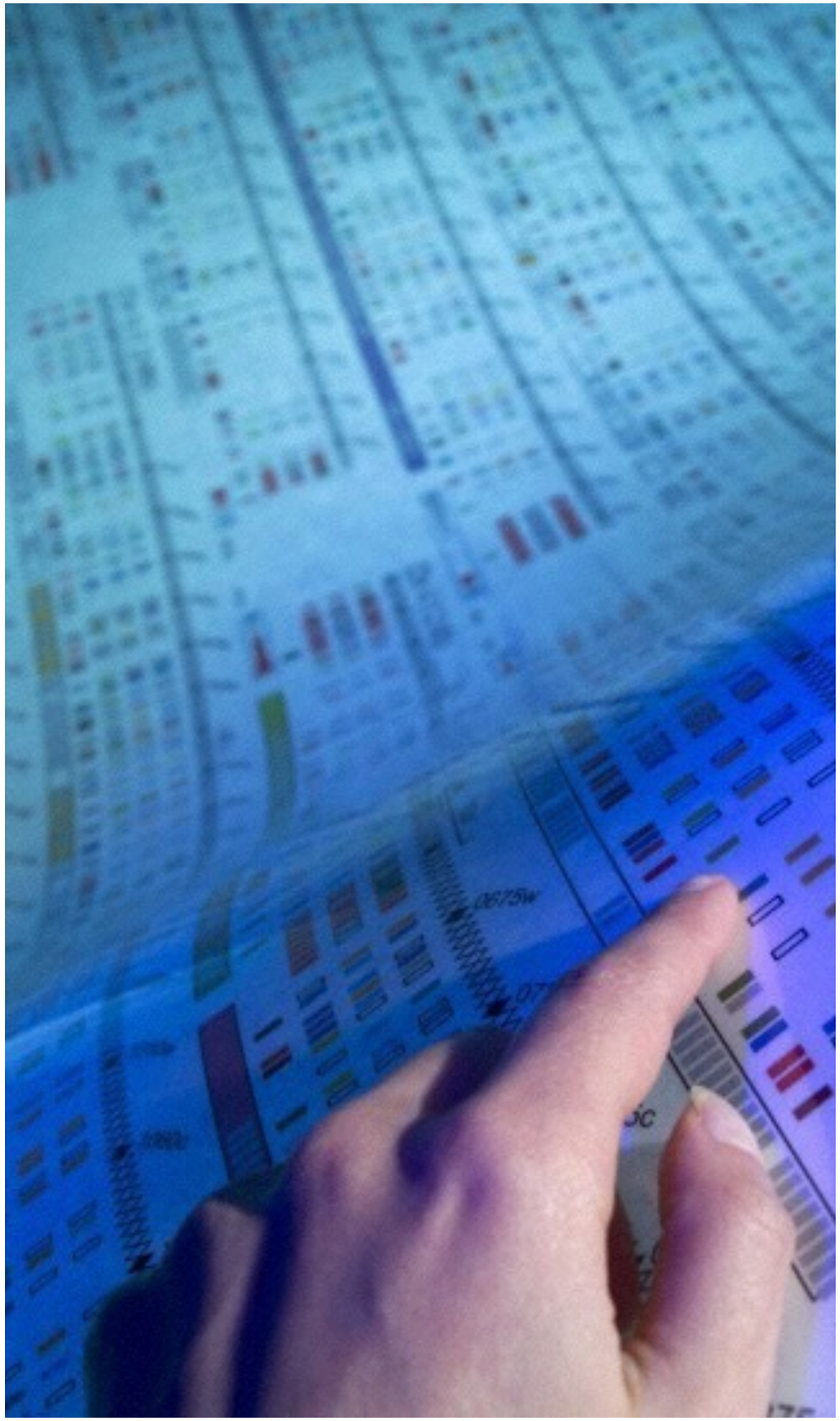
Trends 2010 to 2015



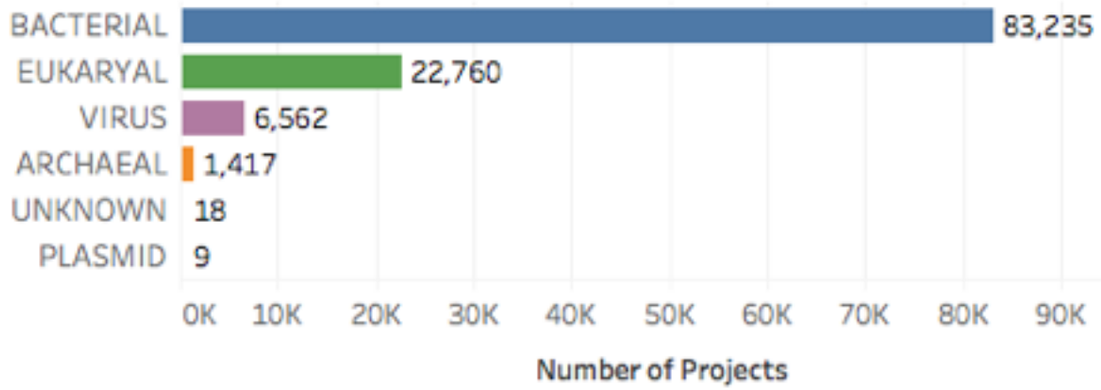
Mb = Megabase or 1,000,000 base pairs. Source: National Human Genome Research Institute. DNA Sequencing Costs: Data. Raw data: Sequencing Cost Table.

Whole Genome Sequencing Projects

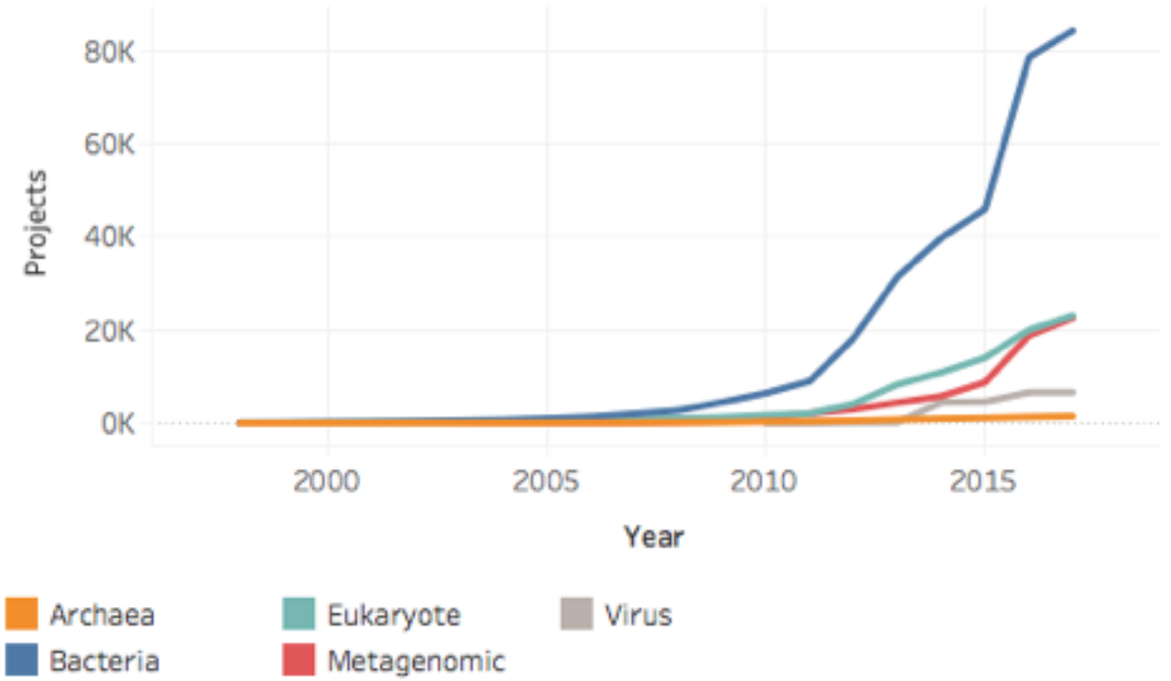
- A growing trend towards whole genome sequencing as sequencing costs fall;
- The Genomes Online Database (GOLD) is the main source of data on global projects;
- Data on projects is submitted voluntarily and may be incomplete. Data fields on funding are incomplete and others require further clean up… with those caveats…



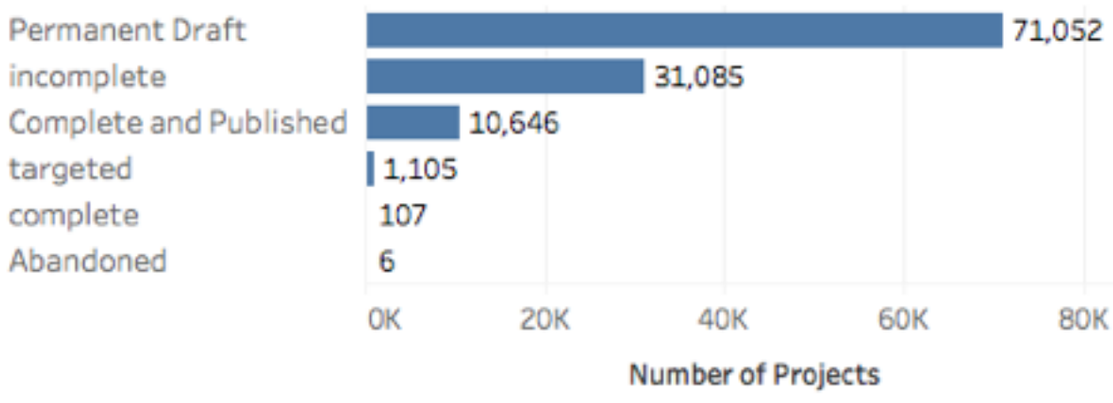
Whole Genome Projects by Domain



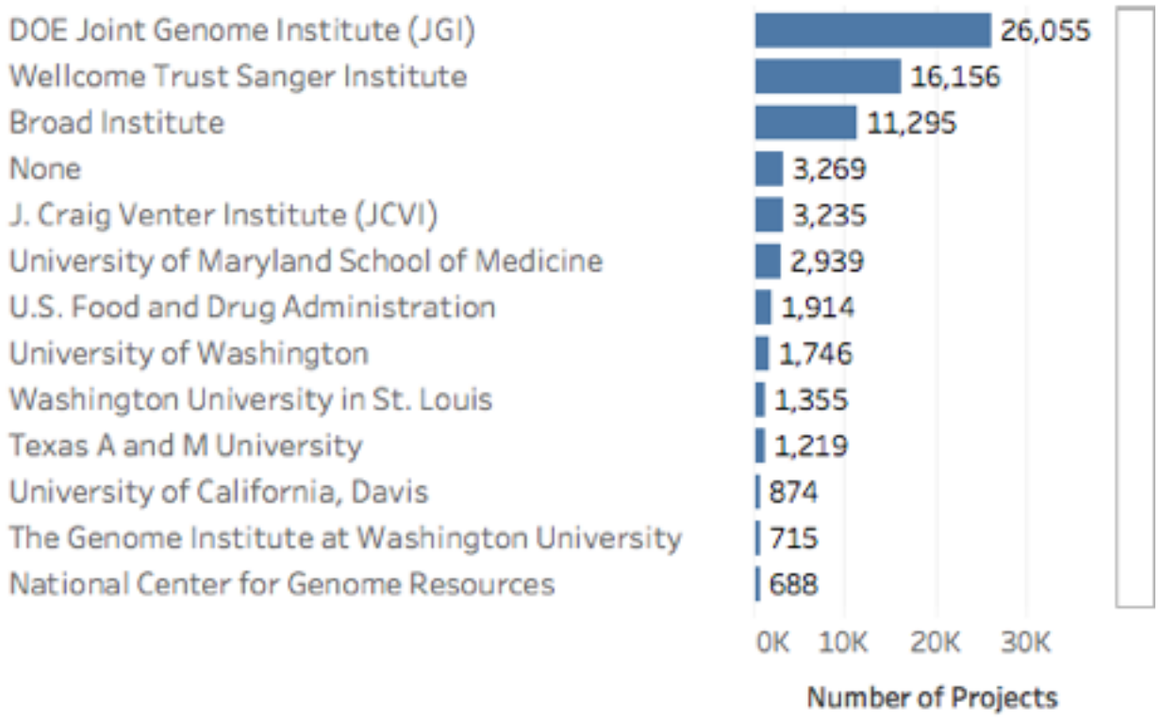
Genome Project Trends by Domain



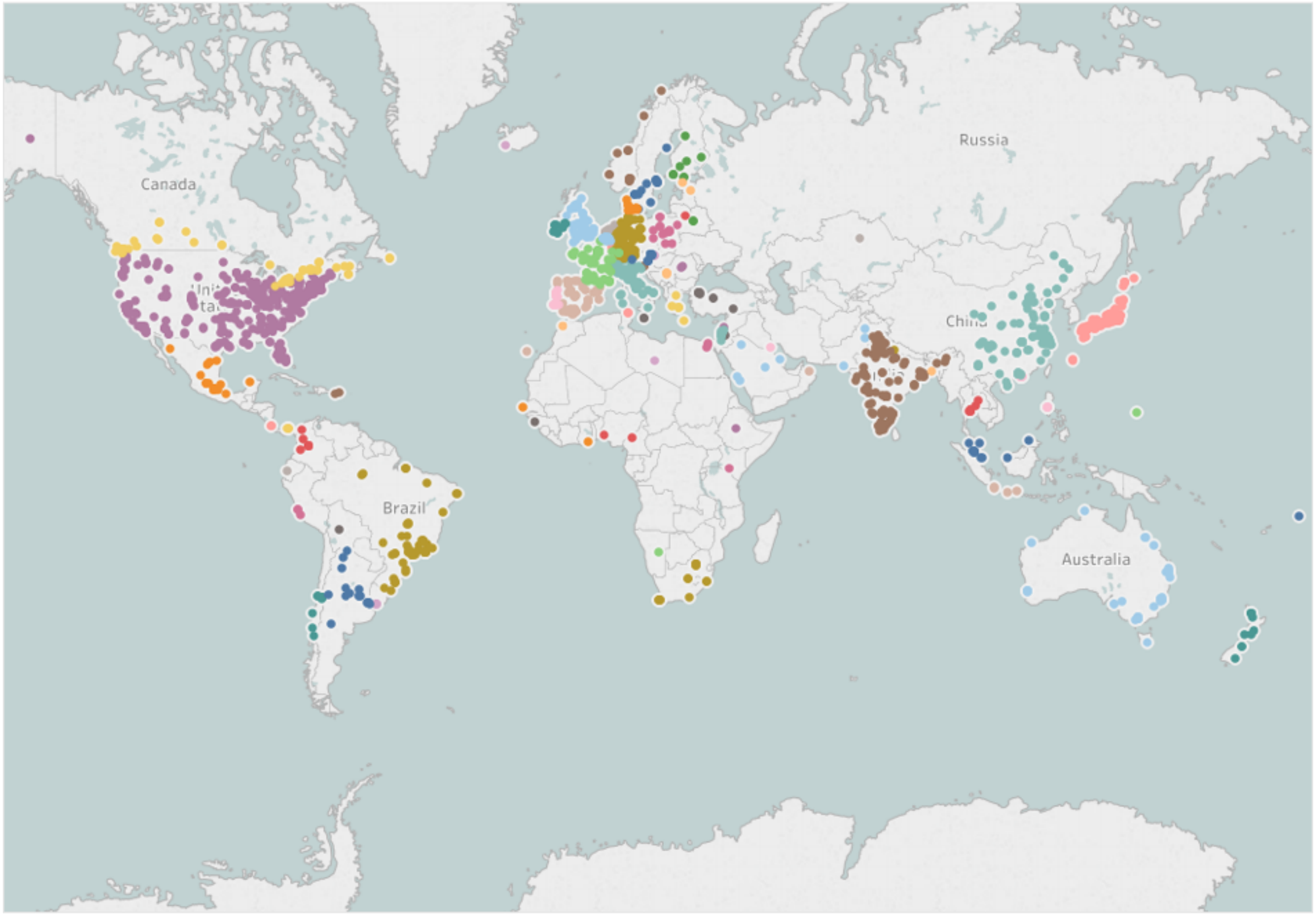
Project Status



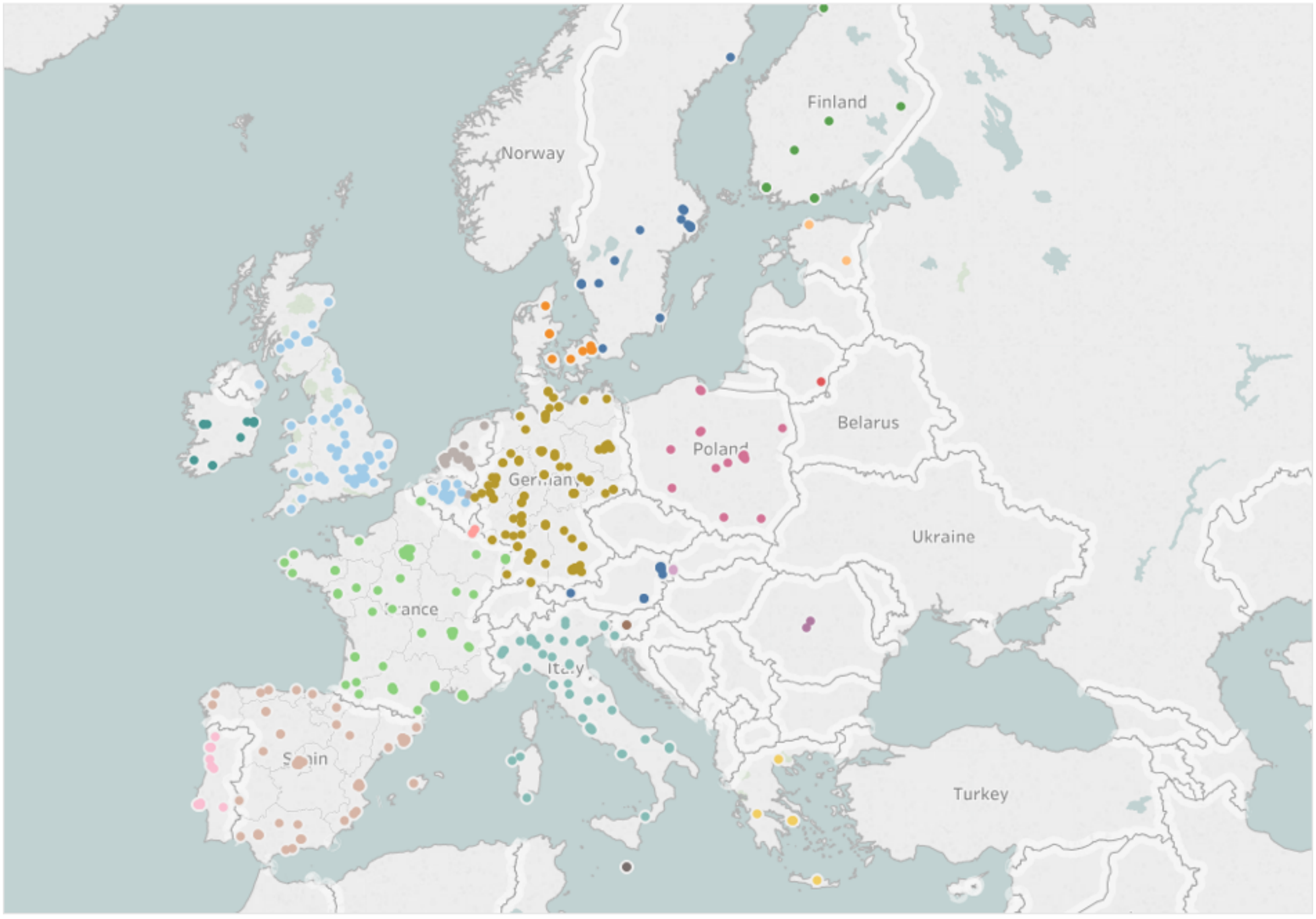
Genome Sequencing Centres (raw)



Source: Genomes Online Database (GOLD) data file 27/05/2017. https://gold.jgi.doe.gov/downloadexceldata

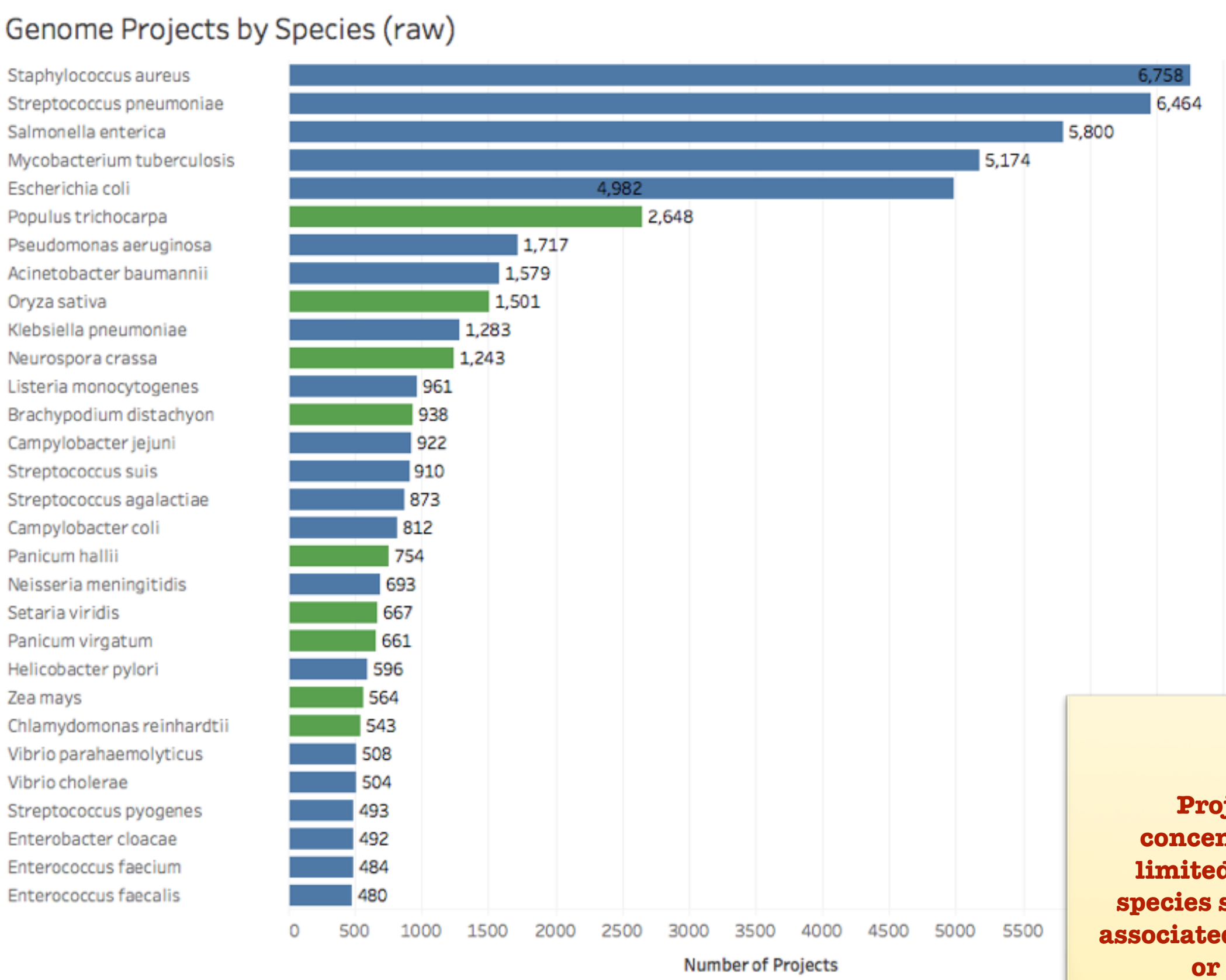


Genome Sequencing Organisations (raw, not validated, +3,000)



Organisations with Whole Genome Projects in the EU

Genome Projects by Species (raw)



Projects are concentrated in a limited number of species such as those associated with disease or models

JGI GOLD
GENOMES ONLINE DATABASE

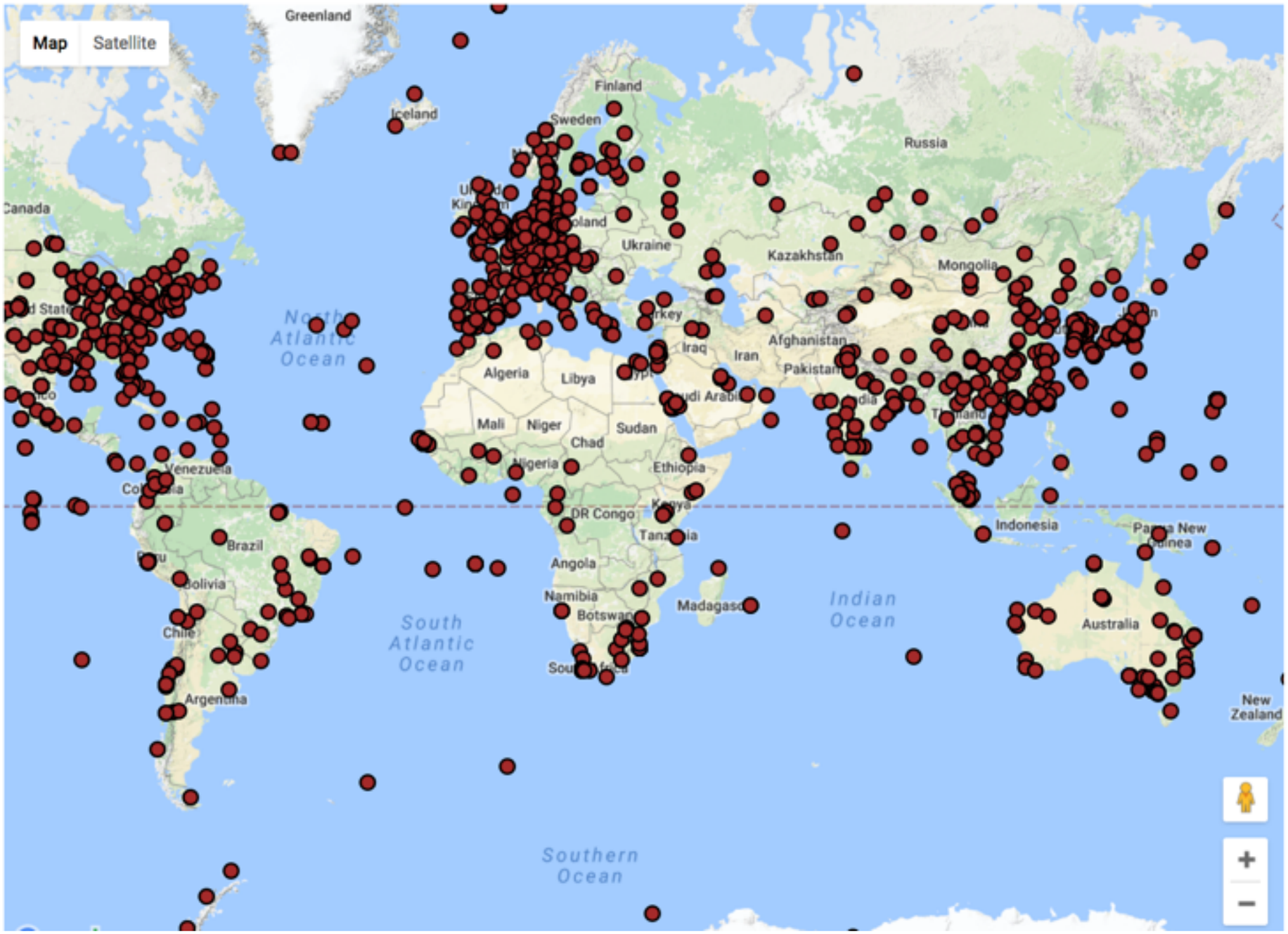
JGI HOME LOG IN

Home Search Distribution Graphs Biogeographical Metadata Statistics GOLD Usage Policy Team Help News

Studies	28,794
Biosamples	22,284
Sequencing Projects	136,099
Analysis Projects	112,912
Organisms	272,928

Organism Distribution Map

Data represents a sample of Organisms in GOLD



Origin data not readily accessible at present for further analysis

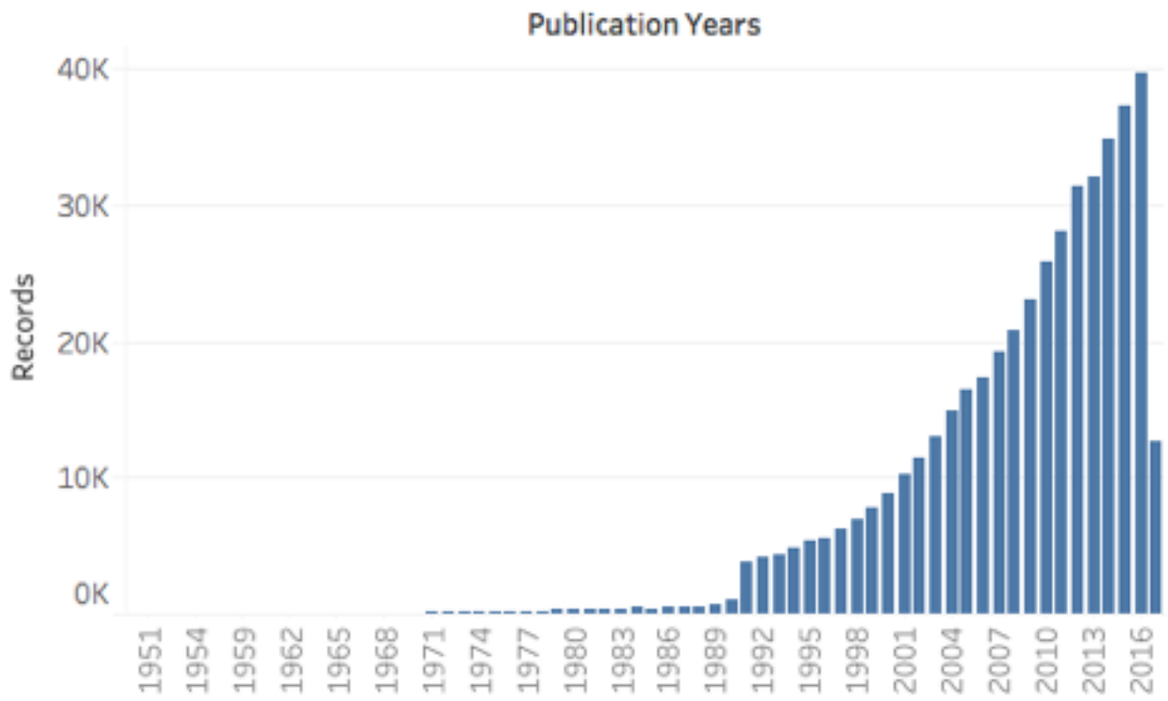
Scientific Publications in Genomics

Scientific outputs from DSI research can be illuminated using publication data for subjects such as genomics. This can also assist with identifying key areas of research and key actors.

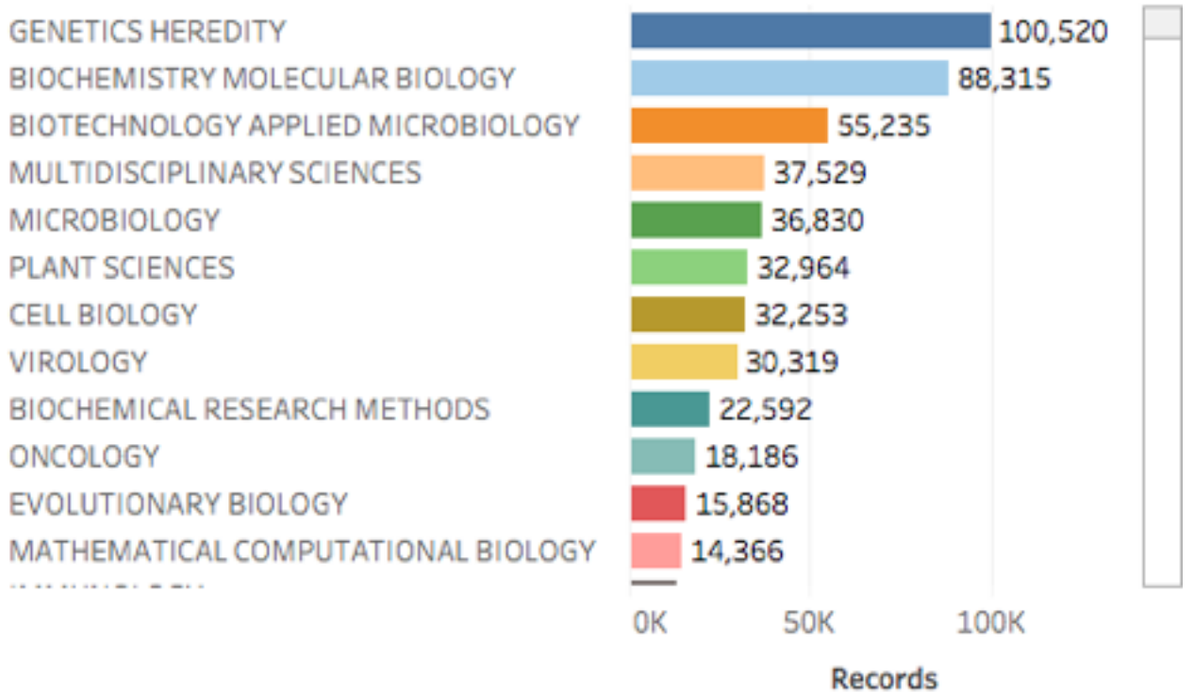
The Omics

- Functional Genomics (transcription, translation, protein - protein interactions)
- Structural Genomics (description of all proteins encoded by a genome)
- Epigenomics (factors influencing phenotypes)
- Metagenomics (sequencing environmental samples for taxonomy etc.)
- Synthetic Genomics (synthetic biology, engineering new genetic components and organisms from scratch)
- Conservation genomics (informing conservation decision making)
- Proteomics (understanding the protein complement of a cell or organism)
- Molecular Taxonomy, Cladistics and DNA barcoding
- … yet more omics

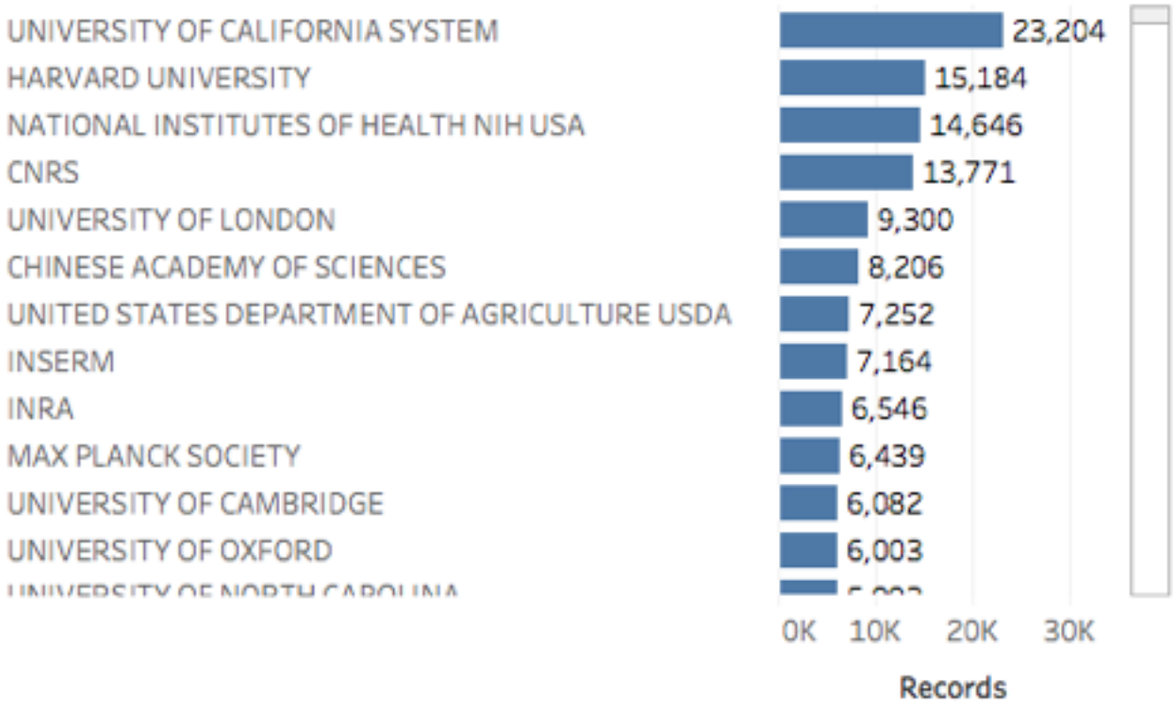
Genomics Scientific Publications



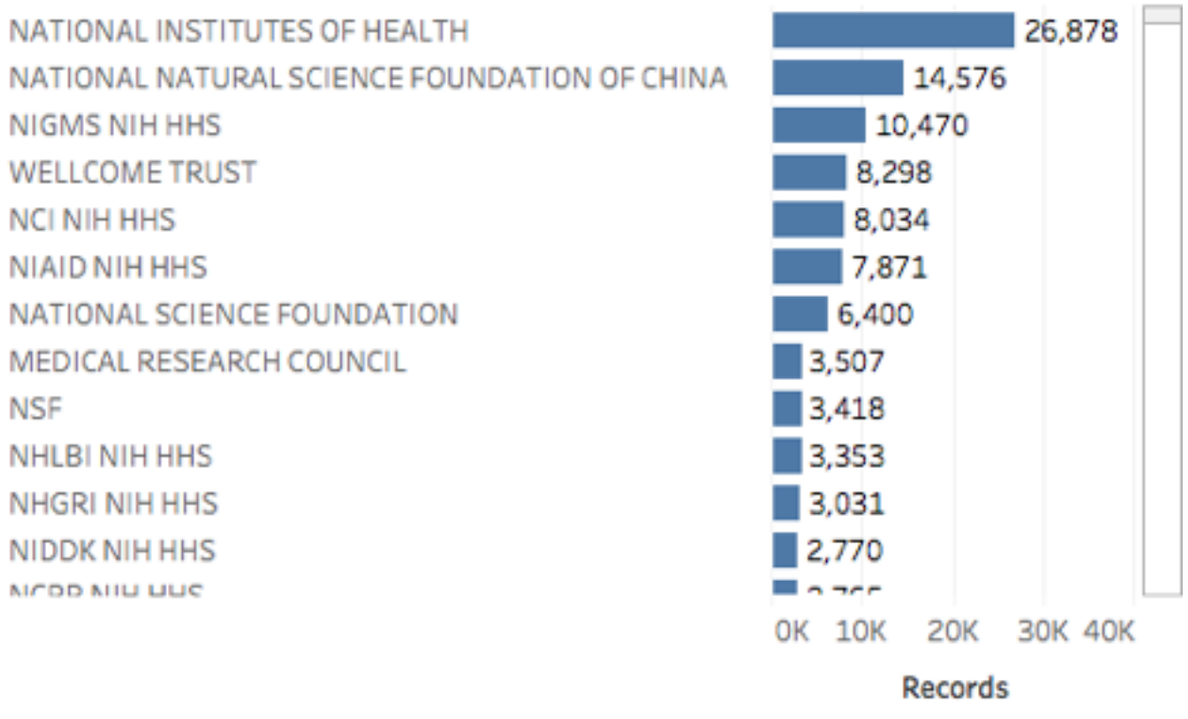
Gernomics by Subject Area



Genomics Top Organisations (Raw)

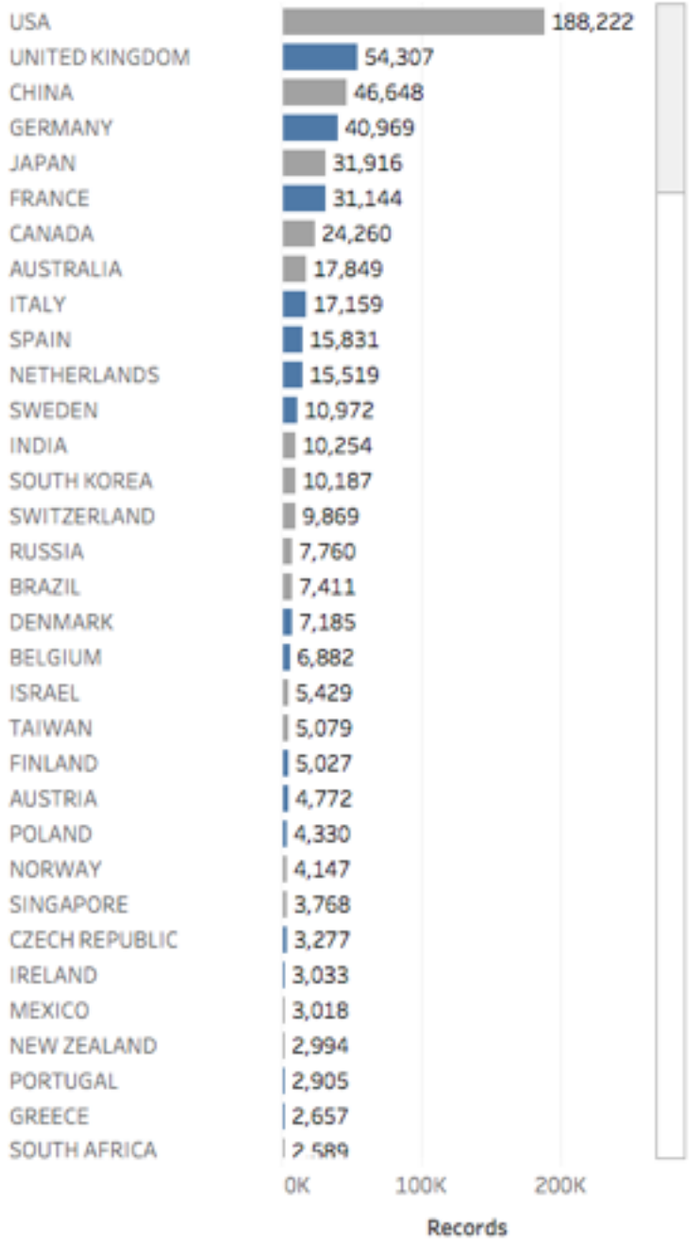


Genomics Funding (Raw)



Source: Search of Web of Science Core Collection topic field for genome, genomes and genomics 29/05/2017. Organisation and Funding agency data has not been cleaned and is classified as raw. This will affect rankings.

Genomics Publications Rank Country



Genomics Publications

"…it is conceivable that technological innovation may one-day permit the *in situ* extraction of genetic material and transfer of data to electronic form without the necessity of the collection, taxonomic identification and storage of field samples."

- Oldham 2004: UNEP/CBD/WG-ABS/3/INF/4 at 17

Pint-sized DNA sequencer impresses first users

Portable device offers on-the-spot data to fight disease, catalogue species and more.

Erika Check Hayden

05 May 2015 | Corrected: 11 May 2015

PDF Rights & Permissions

MUSE/Science Museum of Trento

The MinION device can sequence small genomes, such as those of bacteria and viruses, displaying the results as they are generated.

In April, Joshua Quick boarded a plane to Guinea with three genetic sequencers packed in his luggage. That fact alone is astonishing: most sequencing machines are much too heavy and delicate to travel as checked baggage in the hold of a commercial airliner. What came next was even more impressive. For 12 days, Quick used these sequencers — called MinIONs — to read the genomes of Ebola viruses from 14 patients in as little as 48 hours after samples were collected.

Pssst...

The best-kept secrets to winning grants

With competition for research funding approaching an all-time high, experts reveal their top tips and tricks.

Like Share 351k people like this. Be the first of your friends.

Recent | Read | Commented

1. **Crowd-based peer review can be good and fast**
   Nature | 30 May 2017
2. **Govern land as a global commons**
   Nature | 30 May 2017
3. **Monument to peer review unveiled in Moscow**
   Nature | 26 May 2017

nature briefing

The best science news from *Nature* and beyond, direct to your inbox every day.

Email*

Subscribe

Nature 2015

VolTRAX | List of open-source bioinform | New Tab

Secure https://nanoporetech.com/products/voltrax

Rapid, programmable, portable, disposable sample processor

VolTRAX | VolTRAX Introduction Programme

About VolTRAX

Oxford Nanopore offers a range of options for converting your original biological sample to a form ready for application into a nanopore sensing device.

Oxford Nanopore has developed VolTRAX – a small device designed to perform library preparation automatically, so that a user can get a biological sample ready for analysis, hands free.

VolTRAX Introduction Programme >

Do you need any help?

Portable Sample Preparation US$2300

About MinION

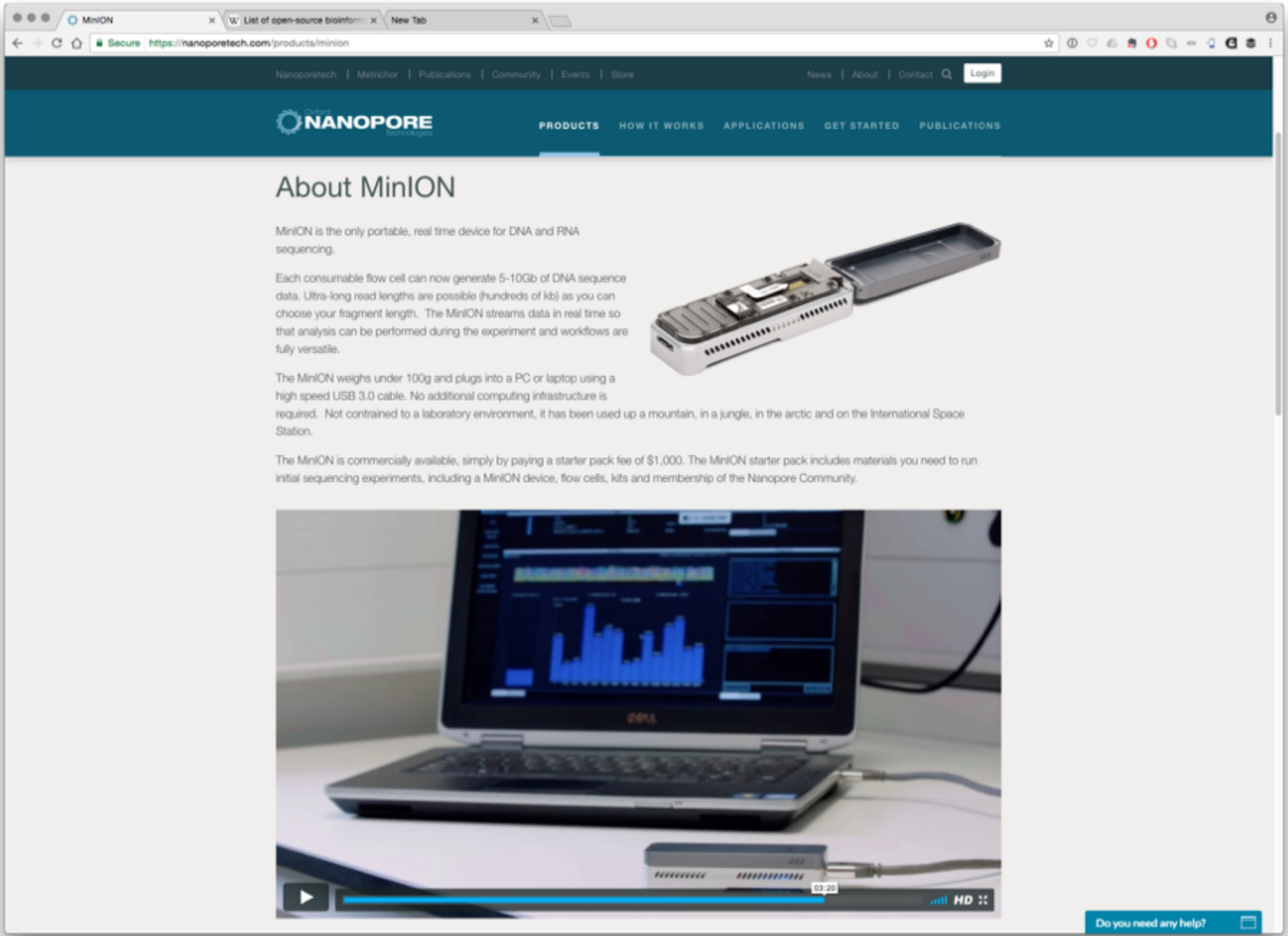
MinION is the only portable, real time device for DNA and RNA sequencing.

Each consumable flow cell can now generate 5-10Gb of DNA sequence data. Ultra-long read lengths are possible (hundreds of kb) as you can choose your fragment length. The MinION streams data in real time so that analysis can be performed during the experiment and workflows are fully versatile.

The MinION weighs under 100g and plugs into a PC or laptop using a high speed USB 3.0 cable. No additional computing infrastructure is required. Not contrained to a laboratory environment, it has been used up a mountain, in a jungle, in the arctic and on the International Space Station.

The MinION is commercially available, simply by paying a starter pack fee of $1,000. The MinION starter pack includes materials you need to run initial sequencing experiments, including a MinION device, flow cells, kits and membership of the Nanopore Community.

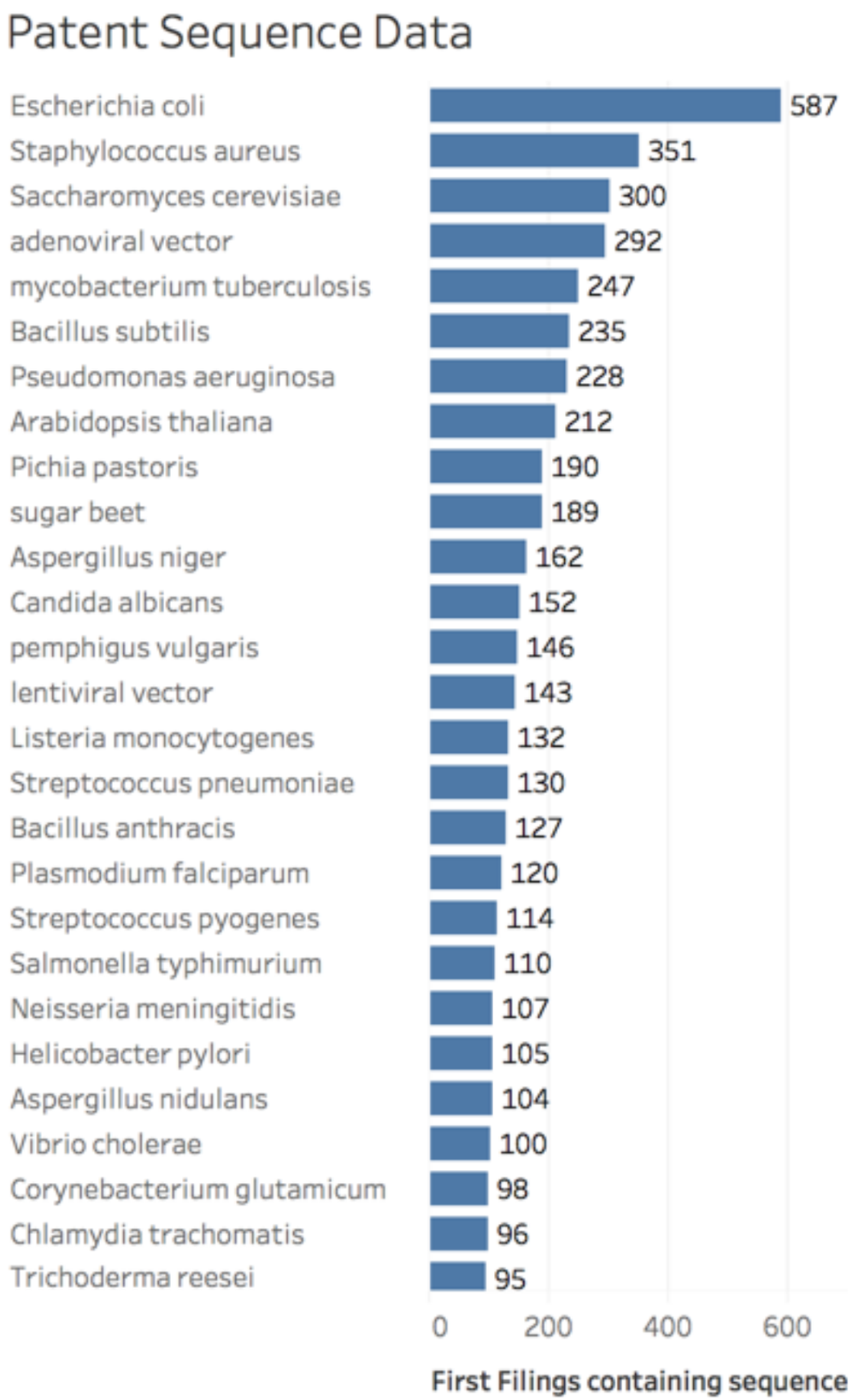
Real Time Portable DNA sequencing. US $100 basic pack, US$4999



Sequencing links to cloud based analytics (note may be outside jurisdiction of provider)

Intellectual Property Issues

- Issues around the implications of patent rights have been widely debated (and patent regulations have increasingly been restricted). However, it is important to bear in mind that in addition to patents DSI as data involves
- Copyright (in sequences)
- Database Rights (in applicable jurisdictions)
- Prepublication and data access agreements appear to have played an important role in large sequencing projects (e.g. human and wheat genome)



Source: WIPO PCT Filings 1997-2013

Implications for ABS

- Developing countries are likely to incorporate articles into national legislation and ABS contracts on sequence data. This is logical but the question is the impact relative to the gain.

- The International Nucleotide Sequence Database Collaboration effectively asserts that DNA, RNA and amino acid sequence data belongs in the public domain (unrestricted use). That could be a good thing…but…

- Provider countries are likely to question the legitimacy of this assertion and may turn to countries that are not part of the INSDC that will meet requirements such as renewed PIC & MAT for the use of sequence data.

- If providers do introduce regulations on DSI (which looks fairly inevitable at present) the question would be how to operationalise that?

Minimum information about species barcode nucleotide sequence

The Species BARCODE Data Standard is a biodiversity standard formulated by the Consortium for the Barcode of Life (CBOL) for reporting minimum information about species barcode nucleotide sequences. The CBOL specifies requirements on reporting sample provenance information and on sequence quality with the aim to create a reference library of barcode DNA sequences integrated with related biodiversity information, such as taxonomy, specimen vouchers or geo-reference. Ultimately, DNA barcoding shall serve as a global standard for species identification.

The International Barcode of Life project (iBOL) develops a DNA barcode reference library that will serve as DNA-based identification system for multi-cellular life.

The Barcode of Life Data Systems (BOLD) is the central informatics platform for DNA barcoding providing acquisition, storage, analysis and publication of DNA barcode records.

A suitable species barcode marker has to meet several criteria. Ideally, the barcode marker (1) can be easily amplified in one read following a standardised protocol, (2) is on both sides flanked by a highly conserved region for reliable primers annealing, (3) is capable of organism identification on a species level.

Currently, the CBOL approves as effective barcodes the following loci:

- for metazoa, the cytochrome c oxidase 1 (cox1) gene region
- for land plants, a two-locus barcode, the ribulose-bisphosphate carboxylase (rbcL) and maturaseK (matK) gene regions (with recommendation to collect also non-coding regions, such as the chloroplast trnH-psbA spacer region)
- for fungi, the ribosomal internal transcribed spacer (ITS) region

INSDC records that meet the criteria of Species BARCODE Data Standard have the keyword 'BARCODE'.

The MIMARKS includes the Species BARCODE Data Standard, which means that a MIMARKS-compliant dataset is also Species BARCODE compliant.

Species BARCODE data submission

The Species BARCODE reporting requirements are devided into mandatory (available here), highly recommended (available here) and optional (available here) irrespective of the sequenced marker locus.

Submit & Update

- Data formats
- Uploading data files
- Reads
- Sequences
- Genome assemblies
- Taxonomy
- Sample checklists
- Environmental
- Epigenomic
- Species BARCODE
- Metadata model
- Register submission account
- Programmatic XML submissions

Popular

- Submit and update
- Sequence submissions
- Genome assembly submissions
- Submitting environmental sequences
- Citing ENA data
- Rest URLs for data retrieval
- Rest URLs to search ENA

Latest ENA news

27 Apr 2017: New ENA discov

ENA has launched a new API t
search across all data
types: https://www.ebi.ac.uk/e

ENA has a range of submission forms with requirements. This is for DNA barcodes

sequences now available

Mandatory Species BARCODE checklist

Field	Description	Example
Organism name;	Formal taxonomic name of this metozoan organism or informal name if unpublished/unidentified.	Arabidopsis thaliana
Bio-repository data	Reference to physical specimen from which the sequence was obtained (e.g. curated museum collection, living specimen), can be structured or unstructured.	structured YMUK:12345 unstructured ABCD-12345
Country	Political name of country or ocean in which a sequenced sample or isolate was collected.	France, Mediterranean Sea
Translation table	Mitochondrial translation table for this organism. Choose between vertebrate (table 2) and invertebrate (table 5) codes.	2
Codon Start (required to determine reading frame)	The codon start for the reading frame which should be translated is the coordinate of the base for the fisrt complete codon.	3
Forward Primer Name	Name of the forward direction PCR primer.	ArthFW1
Forward Primer Sequence	Sequences should be given in the IUPAC degenerate-base alphabet, except for the modified bases; those must be included within angle brackets.	GACATTGKG<I>T
Reverse Primer Name	Name of the reverse direction PCR primer.	ArthRV1
Reverse Primer Sequence	Sequences should be given in the IUPAC degenerate-base alphabet, except for the modified bases; those must be included within angle brackets.	CATGRTTAGAC

Highly recommended Species BARCODE checklist

Field	Description	Example
Latitude/Longitude	Geographical coordinates of the location where the specimen was collected, in decimal degrees (to 2 places).	47.94, -12.45
Identified by	The person that identified the organism/sample.	John White
Collector	Name of the person that originally collected the sample/organism	John White
Collection Date	Date of collection of the original sample/organism	12-Apr-2013

Mandatory disclosure of Country and additional voluntary options available

Previous work explored the use of creative commons style licensing models for ABS

UNEP

Convention on Biological Diversity

Distr.
GENERAL

UNEP/CBD/WG-ABS/8/INF/3
30 July 2009

ENGLISH ONLY

AD HOC OPEN-ENDED WORKING GROUP
ON ACCESS AND BENEFIT-SHARING
Eighth meeting
Montreal, 9-15 November 2009
Item 3 of the provisional agenda*

THE ROLE OF COMMONS/OPEN SOURCE LICENCES IN THE INTERNATIONAL REGIME ON ACCESS TO GENETIC RESOURCES AND BENEFIT-SHARING

Note by the Executive Secretary

1. The Executive Secretary is pleased to circulate herewith, for the information of participants in the eighth meeting of the Ad Hoc Open-ended Working Group on Access and Benefit-sharing, a discussion paper on the role of commons/open source licences in the international regime on access to genetic resources and benefit-sharing, ESRC Centre for Economic and Social Aspects of Genomics (Cesagen), University of Lancaster and the Peruvian Society for Environmental Law (SPDA). This paper is referred to in the first paragraph of the suggestions on operational text submitted by Cesagen, which is also available at https://www.cbd.int/abs/submissions/abswg-08-cesagen-en.pdf.

2. The paper is being circulated in the form and language in which it was received by the Secretariat.

https://www.cbd.int/kb/record/meetingDocument/62145?RecordType=meetingDocument

Share your work - Creative Co ×

Secure https://creativecommons.org/share-your-work/

Dictionaries | Travel | Synthetic Biology… | PLOS Biology: Op… | SelectorGadget | pauloldham.net/us… | Login | Thomson In… | SelectorGadget | Other Bookmarks

Choose a license

This chooser helps you determine which Creative Commons License is right for you in a few easy steps. If you are new to Creative Commons, you may also want to read Licensing Considerations before you get started.

Get Started

- Open science
- Policy / advocacy / copyright reform
- Press
- Share your work
- Technology
- Uncategorized
- Weblog

In this section

- Licensing considerations
- Licensing types
- Public domain
- Places to share
- For developers

CONNECT WITH CREATIVE COMMONS

Your email

Sign up to Our Newsletter

Choose a License

Secure https://creativecommons.org/choose/

License Features

Your choices on this panel will update the other panels on this page.

Allow adaptations of your work to be shared?

Yes No Yes, as long as others share alike

Allow commercial uses of your work?

Yes No

Selected License

Attribution-ShareAlike 4.0 International

This is a Free Culture License!

Choose a License | Creative Commons — Attribut

Secure https://creativecommons.org/licenses/by-sa/4.0/

Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Choose a License | Creative Commons — Attribut

Secure https://creativecommons.org/licenses/by-sa/4.0/legalcode

creative commons

Attribution-ShareAlike 4.0 International

Official translations of this license are available in other languages.

Creative Commons Corporation ("Creative Commons") is not a law firm and does not provide legal services or legal advice. Distribution of Creative Commons public licenses does not create a lawyer-client or other relationship. Creative Commons makes its licenses and related information available on an "as-is" basis. Creative Commons gives no warranties regarding its licenses, any material licensed under their terms and conditions, or any related information. Creative Commons disclaims all liability for damages resulting from their use to the fullest extent possible.

Using Creative Commons Public Licenses

Creative Commons public licenses provide a standard set of terms and conditions that creators and other rights holders may use to share original works of authorship and other material subject to copyright and certain other rights specified in the public license below. The following considerations are for informational purposes only, are not exhaustive, and do not form part of our licenses.

Considerations for licensors: Our public licenses are intended for use by those authorized to give the public permission to use material in ways otherwise restricted by copyright and certain other rights. Our licenses are irrevocable. Licensors should read and understand the terms and conditions of the license they choose before applying it. Licensors should also secure all rights necessary before applying our licenses so that the public can reuse the material as expected. Licensors should clearly mark any material not subject to the license. This includes other CC-licensed material, or material used under an exception or limitation to copyright. More considerations for licensors.

Considerations for the public: By using one of our public licenses, a licensor grants the public permission to use the licensed material under specified terms and conditions. If the licensor's permission is not necessary for any reason–for example, because of any applicable exception or limitation to copyright–then that use is not regulated by the license. Our licenses grant only permissions under copyright and certain other rights that a licensor has authority to grant. Use of the licensed material may still be restricted for other reasons, including because others have copyright or other rights in the material. A licensor may make special requests, such as asking that all changes be marked or described. Although not required by our licenses, you are encouraged to respect those requests where reasonable. More considerations for the public.

The regulatory challenge

- We intend to test this kind of approach in a project with partners in Kenya… with the aim of finding a way forward.

- The use of Creative Commons style licences for Sequence Data would run straight into the no restriction requirements of the INSDC. My understanding is that this policy originated from the efforts by companies to use restrictive licensing on genome sequence data. Science would have been the loser.

- However. providers will be confronted with the challenge of how to protect their interest and at the same time promote scientific research and cooperations and innovation for genetic resources. A pure public domain argument is unlikely to gain traction…

Conclusions

- If provider countries go down the route of introducing requirements on digital sequence data into their legislation and MAT, this is likely to have significant consequences for scientific research (notably taxonomy) and over the longer term for innovation.

- At the same time pursuing a pure public domain argument is unlikely (in my view) to succeed because it will not address provider concerns.

- A middle ground may be possible but it would need to be simple (in terms of options) in order to address scale measured in terms of billions and trillions of bases from organisms distributed around the world. Models for this already exist but would need to be adjusted for ABS needs.

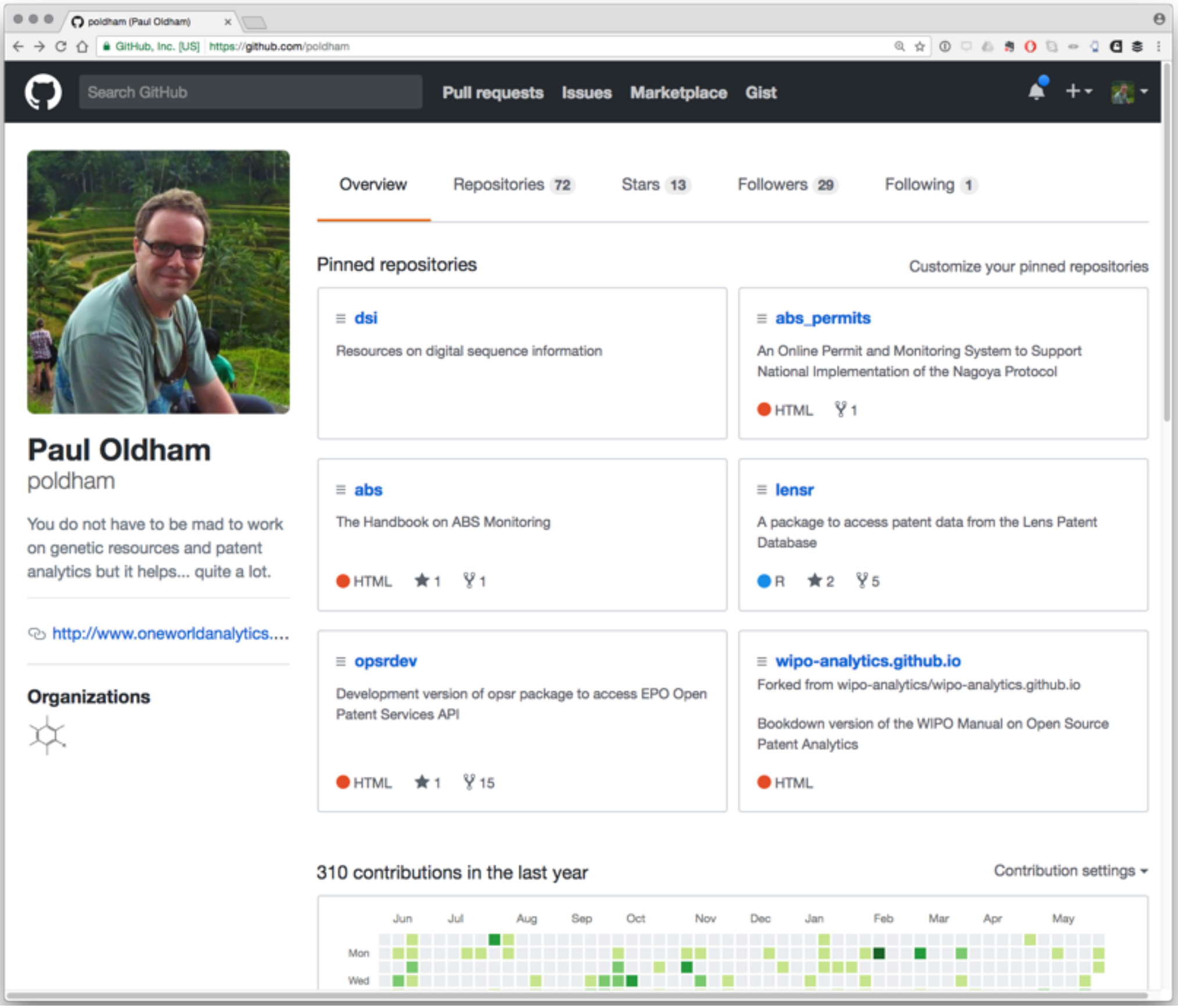
- The alternative may possibly be fragmentation of DSI into multiple silos depending on the willingness of database providers to accept provider country conditions.

The global public goods dimension

- Genomes and genomic databases have been treated as a global public good. In the context of the Nagoya Protocol it is important to emphasise the opportunities that may exist for international cooperation in taxonomy, conservation genomics, to address human health issues (e.g. neglected diseases) or identify strategies for adaptation to climate change that are enabled by genomics and DSI.

- The investments and international collaboration that exists in genome sequencing and genomic research are valuable in themselves in terms of knowledge and technology transfer and capacity-building. Above all perhaps they have value in advancing knowledge and understanding of biodiversity and genetic resources. More could be made to highlight this at the expense of the perils of the pursuit of hyperownership in the context of the promises of biotechnology.

Access the presentation and other materials at dsi



https://github.com/poldham