DNA Data Bank of Japan
update report

National Institute of Genetics, JAPAN
Yasukazu “Yaz” NAKAMURA
中村保一

Dec. 9, ABC Symposium 2022

National Institute of Genetics

National Institute of Genetics: www.nig.ac.jp

Research Departments

Department of Informatics

- ARITA, Masanori — Biological Networks — Lab HP
- KAWAMOTO, Shoko — Genetic Informatics
- KUROKAWA, Ken — Genome Evolution
- MORI, Hiroshi — Genome Diversity
- NAKAMURA, Yasukazu — Genome Informatics — Lab HP
- OKUBO, Kousaku — Gene-Expression Analysis

Department of Genomics and Evolutionary Biology

- AKASHI, Hiroshi — Evolutionary Genetics — Lab HP
- IKEO, Kazuho — DNA Data Analysis — Lab HP
- INOUE, Ituro — Human Genetics — Lab HP
- KITANO, Jun — Ecological Genetics — Lab HP
- KURAKU, Shigehiro — Molecular Life History — Lab HP
- [illegible] Naruya — Population Genetics — Lab HP
- [illegible] TO, Yutaka — Plant Genetics
- [illegible] tsushi — Comparative Genomics

Department of Gene Function and Phenomics

- HIRATA, Tatsumi — Brain Function — Lab HP
- IWASATO, Takuji — Mammalian Neural Circuits — Lab HP
- KAWAKAMI, Koichi — Molecular and Developmental Biology — Lab HP
- KOIDE, Tsuyoshi — Mouse Genomics Resource — Lab HP
- MIYAGISHIMA, Shin-ya — Symbiosis and Cell Evolution — Lab HP
- NIKI, Hironori — Microbial Physiology — Lab HP
- NONOMURA, Ken-ichi — Plant Cytogenetics — Lab HP
- ODA, Yoshihisa — Cell Dynamics and Signaling — Lab HP
- SAGA, Yumiko — Mammalian Development — Lab HP
- SAKAI, Noriyoshi — Model Fish Genetics — Lab HP
- SAWA, Hitoshi — Multicellular Organization — Lab HP
- YONEHARA, Keisuke — Multiscale Sensory Structure — Lab HP

Department of Chromosome Science

- KANEMAKI, Masato — Molecular Cell Engineering — Lab HP
- KIMURA, Akatsuki — Cell Architecture — Lab HP
- MAESHIMA, Kazuhiro — Genome Dynamics — Lab HP
- SAITO, Kuniaki — Invertebrate Genetics — Lab HP
- SHIMAMOTO, Yuta — Physics and Cell Biology Laboratory — Lab HP

Center for Frontier Research

Center for Frontier Research

- KUBO, Fumi — Systems Neuroscience — Lab HP
- MURAYAMA, Yasuto — Chromosome Biochemistry

Intellectual Infrastructure Center

Bioinformation and DDBJ Center

- Division of High Performance Computing
- Division of Biological Databases
- Division of International Affairs

Advanced Genomics Center

- Sequencing Division
- Data Analysis Division

Genetic Resource Center

- Bioresource Management Division
- Plant Resource Development Division
- Division for development of genetic-engineered mouse resource
- Bioresource Database Division

Support Center

- IT Unit
- Radioisotope Unit
- Unit for Experimental Animal Care

NIG INNOVATION

- SUZUKI, Mutsuaki

Office for Research Development

- KURUSU, Mitsuhiko — HP

Office for Gender Equality

- HIRATA, Tatsumi — HP

Department of Administration

- General Affairs and Project Section
- Financial Affairs Section

Technical Section

- FURUUMI, Hiroyasu

Nakamura lab's genome works

Citrus unshiu
An orange

農研機構 NARO 国立研究開発法人 農業・食品産業技術総合研究機構

Gryllus bimaculatus
A cricket

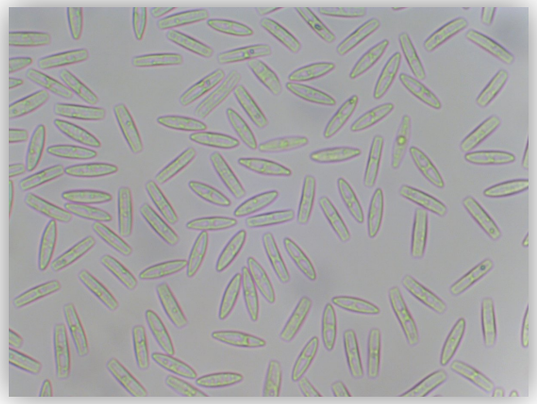
東京大学 THE UNIVERSITY OF TOKYO

京都大学 KYOTO UNIVERSITY
MONASH University
GMI GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY
KINDAI UNIVERSITY
NIBB *et al.*

Marchantia polymorpha
A liverwort

Felis catus

anicom medical アニコム先進医療研究所株式会社
公益財団法人 かずさDNA研究所
香港中文大學 The Chinese University of Hong Kong



Nitzschia spp.
A non-photosynthetic diatom

京都大学 KYOTO UNIVERSITY
筑波大学 University of Tsukuba
国立科学博物館 National Museum of Nature and Science

etc.

Tea tree (3,000)

Wasabi (100)

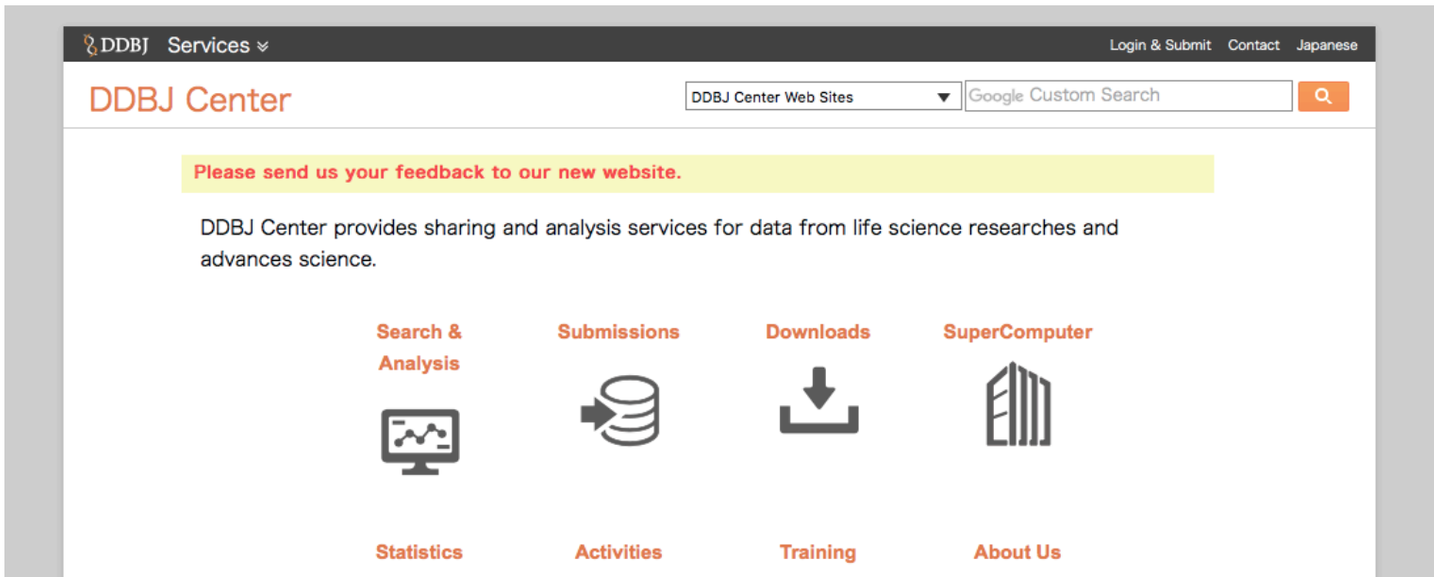
Strawberry (400)

DDBJ

DDBJ Database updates and computational infrastructure enhancement.

Osamu Ogasawara, Yuichi Kodama, Jun Mashima, Takehide Kosuge, Takatomo Fujisawa (2020)

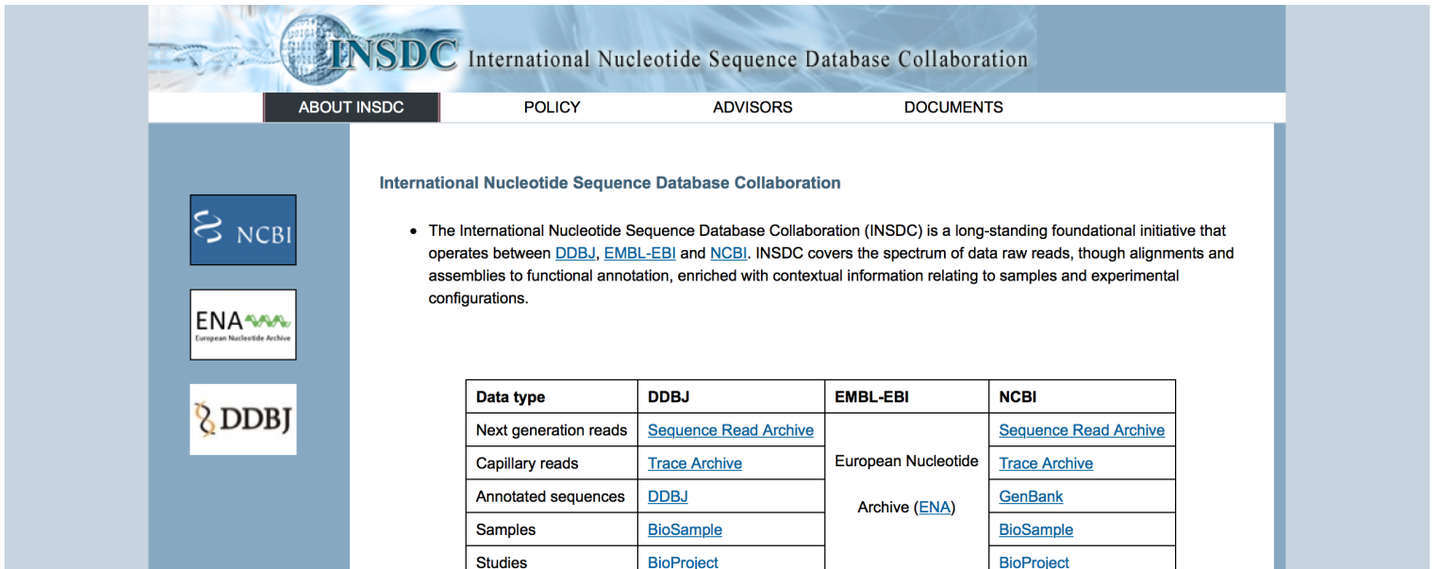
Nucleic Acids Research, **48,** Issue D1, Pages D45–D50, https://doi.org/10.1093/nar/gkz982.



INSDC

The international nucleotide sequence database collaboration.

Masanori Arita, Ilene Karsch-Mizrachi, Guy Cochrane
on behalf of the International Nucleotide Sequence Database Collaboration (2021)
Nucleic Acids Research, **49,** Issue D1, Pages D121–D124, https://doi.org/10.1093/nar/gkaa967



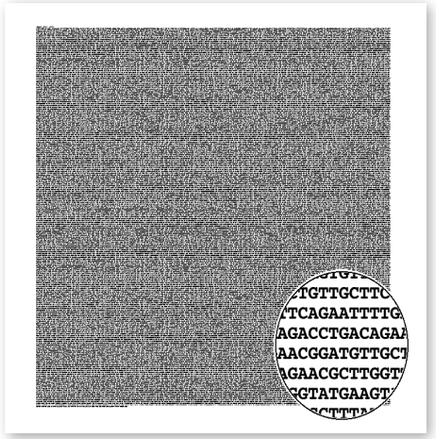
History

1995	EMBL data library was organized, and asked international cooperation for nucleotide sequence data bank to Japan.
1982	EMBL and GenBank started international cooperation, and invited Japan to participate their data bank.
1983	Aimed at contribution for international data bank to collect, to evaluate and to provide nucleotide sequence data, trial data loading was started.
1984	NIG; the National Institute of Genetics was reorganized as an Inter-University Research Institute. DDBJ began to work at NIG.
1986	DNA Database Advisory Committee organized.
1987	DDBJ release 1 was provided. By this release, we regard this year as official start of DDBJ operation.
1995.04	To operate DDBJ more efficiently, CIB; the Center for Information Biology was established in NIG.
2001.04	CIB was reorganized as CIB-DDBJ; the Center for Information Biology and DNA Data Bank of Japan
2004.04	NIG was reorganized as a member of ROIS; Research Organization of Information and Systems. DDBJ has also belonged to ROIS.
2005.05	DDBJ, EMBL-Bank and GenBank agreed to call their collaboration INSDC; International Nucleotide Sequence Database Collaboration; and to call the unified nucleotide sequence database INSD; the International Nucleotide Sequence Database.
2007.04	DBCLS; Database Center for Life Science was newly founded in ROIS
2009	DDBJ faculty staff have greatly been reshuffled. DDBJ collaborates with DBCLS more closely. INSDC added a collaborative meeting to deal with huge sequence data produced by the next generation sequencers (Sequence Read Archive) and traces produced by traditional sequencers (Trace Archive).
2012.04	DDBJ, expanding its DNA databank activities, was restructured as one of the Intellectual Infrastructure Project Centers of NIG, being separated from CIB.
2013.10	Collaborating with NBDC; National Bioscience Database Center, DDBJ Center started to operate the archive for all types of individual-level genetic and de-identified phenotypic data from human subjects, JGA; Japanese Genotype-phenotype Archive.

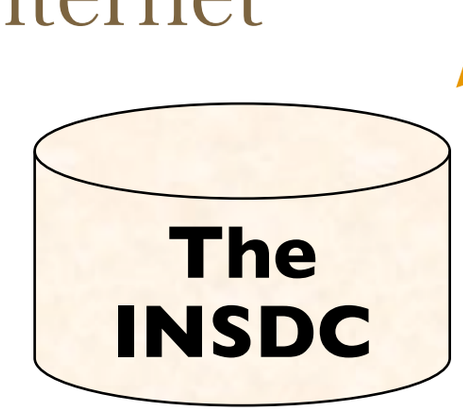
The business of DNA Databanks

DDBJ
DNA Data Bank of Japan

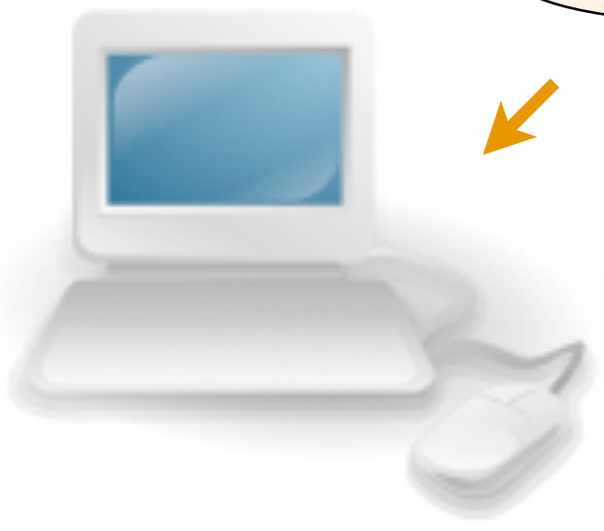
- Submitted Nucleotide Sequence
  - We check its data and metadata
  - We put it into the database
  - We make it public via the internet



Submission



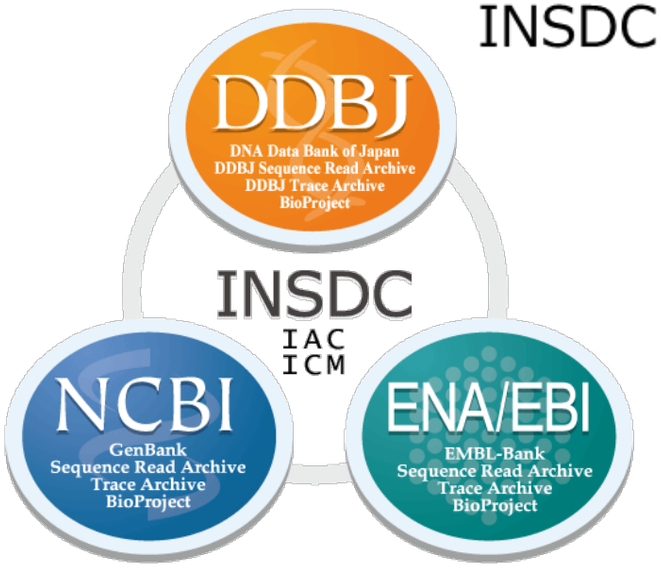
Open and Share



Databases at Bioinformation and DDBJ Center

	Annotated sequences	Capillary reads	NGS reads	Study	Sample	Assembly	Functional genomics	Variation	Genotype and phenotype	Metabolomics
NCBI	GenBank	Trace Archive	Sequence Read Archive	BioProject	BioSample	Assembly	GEO	dbSNP/dbVar	dbGaP	
EBI	European Nucleotide Archive (ENA)						ArrayExpress	EVA/DGVa	EGA	MetaboLights
DDBJ	DDBJ	Trace Archive	Sequence Read Archive	BioProject	BioSample	Assembly	GEA	JVar-SNP/SV	JGA	MetaboBank

INSDC: International Nucleotide Sequence Databank Collaboration



IAC: International Advisory Committee
ICM: International Collaborative Meeting

- GenBank (NCBI)
- ENA (EBI)
- DDBJ

DDBJ's Personnel (43 at this moment)

DDBJ
DNA Data Bank of Japan

NIG Faculties


Masanori Arita, PhD
Professor
Head of DDBJ Center


Yasukazu Nakamura, PhD
Professor
Head of International Affairs Division


Kosaku Okubo, MD, PhD
Professor


Osamu Ogasawara, PhD
Project Associate Professor
Head of HPC Division


Nozomu Sakurai, PhD
Project Associate Professor


Tomoya Tanjo, PhD
Assistant Professor
Head of HPC Division


Yasuhiro Tanizawa, PhD
Assistant Professor


Takatomo Fujisawa, PhD
Project Researcher
Head of Database Division

Database Division

Coordinators


Yuichi Kodama, PhD


Takehide Kosuge, PhD


Jun Mashima, PhD

Curators


Hideo Aono
Data Submission:MSS
Patent


Asami Fukuda
Data Submission: NGS


Andrea Ghelfi, PhD
Data Submission: MSS


Kazuhiro Hamaoka, PhD
Data Submission: NSSS


Tomoyo Iizuka, PhD
Data Submission: NSSS


Masahito Kawazoe
Data Submission: NSSS


Kyungbum Lee, PhD
Data Submission: MSS


Toshihisa Okido, PhD
Data Submission: MSS


Rie Sugita
Data Submission: NGS


Kimiko Suzuki
Data Submission: NGS


Kanae Takaki
Data Submission: NGS

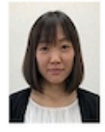

Toshiaki Tokimatsu, PhD
Data Submission:
MetaboBank, NGS


Tomoko Watanabe
Data Updates/Correction


Emi Yokoyama
Data Updates/Correction

Database System SE


Masahiro Fujimoto


Aimi Kawasaki


Hideki Mochizuki


Takahiro Suzuki


Tomoki Umeda


Koji Watanabe


Masahiro Yoshida

High Performance Computing Division(HPC)


Akiko Katsumata


Yoshihiro Okuda, PhD


Yaeko Takiguchi

Supercomputer System SE


Kazunori Aoki


Yuji Ashizawa


Tomohiro Hirai


Takehiro Kato


Yoshimasa Takahashi

Tadayoshi Watanabe

Secretaries

Mika Maki

Naoko Murakata

Alphabetical order.
Do not use the personal photos in this page without permission.

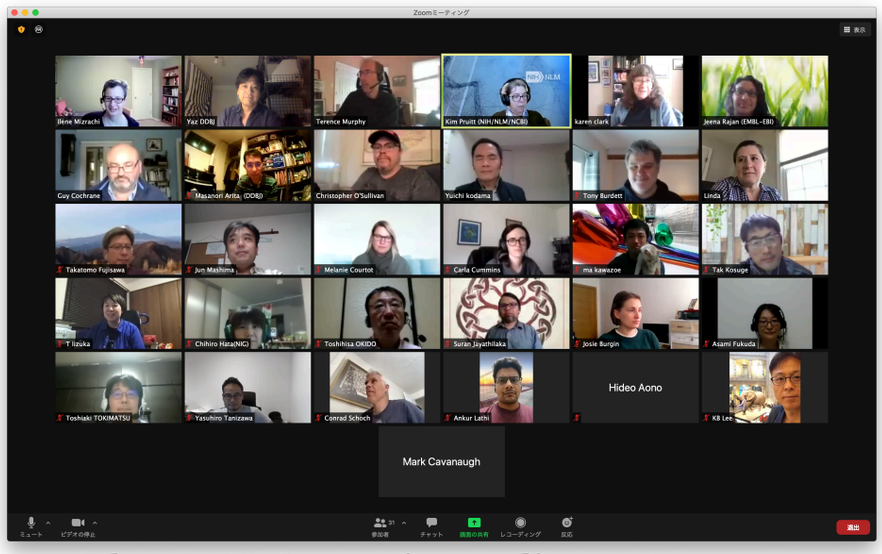
https://www.ddbj.nig.ac.jp/about/staff-e.html

DDBJ (from Release note 115) 43
Jun Mashima, Kazunori Aoki, Hideo Aono, Yuji Ashizawa, Yukino Dobashi, Mayumi Ejima, Masahiro Fujimoto, Asami Fukuda, Tomohiro Hirai, Michiaki Hiramatsu, Naofumi Ishikawa, Kenji Kato, Aimi Kawasaki, Yuichi Kodama, Junko Kohira, Takehide Kosuge, Kyungbum Lee, Mika Maki, Fujitaka Matsumori, Kimiko Mimura, Hideki Mochizuki, Naoko Murakata, Yoshiyuki Nogi, Toshihisa Okido, Yoshihiro Okuda, Maki Ono, Katsunaga Sakai, Yukie Sakon, Makoto Sato, Rie Sugita, Kimiko Suzuki, Takahiro Suzuki, Daisuke Takagi, Yaeko Takiguchi, Toshiaki Tokimatsu, Haru Tsutsui, Koji Watanabe, Tomohiko Yasuda, Emi Yokoyama, Masanori Arita, Takeshi Kawashima, Osamu Ogasawara, Kosaku Okubo, Nozomu Sakurai, Yasuhiro Tanizawa, Toshihisa Takagi, and Yasukazu Nakamura

ENA (from Release note 115) 27
Blaise Alako, Clara Amid, Lawrence Bower, Ana Cerdeno-Taraga, Iain Cleland, Richard Gibson, Neil Goodgame, Petra ten Hoopen, Mikyung Jang, Simon Kay, Rasko Leinonen, Xin Liu, Arnaud Oisel, Rodrigo Lopez, Hamish McWilliam, Nima Pakseresht, Sheila Plaister, Rajesh Radhakrishnan, Kethy Reddy, Stephane Riviere, Marc Rossello, Nicole Silvester, Dmitriy Smirnov, Ana Luisa Toribio, Daniel Vaughan, Vadim Zalunin and Guy Cochrane

GenBank (from Release note 227) 80
Mark Cavanaugh, Ilene Mizrachi, Michael Baxter, Shelby Bidwell, Lori Black, Larissa Brown, Vincent Calhoun, Larry Chlumsky, Karen Clark, Jianli Dai, Scott Durkin, Francescopaolo di Cello, Michel Eschenbrenner, Michael Fetchko, Linda Frisse, Andrea Gocke, Anjanette Johnston, Mark Landree, Jason Lowry, Richard McVeigh, DeAnne Olsen Cravaritis, Leigh Riley, Susan Schafer, Beverly Underwood, and Linda Yankie, Serge Bazhin, Evgueni Belyi, Colleen Bollin, Yoon Choi, Sergey Dikunov, Ilya Dondoshansky, Justin Foley, Viatcheslav Gorelenkov, Sergiy Gotvyanskyy, Eugene Gribov, Jonathan Kans, Leonid Khotomliansky, Michael Kimelman, Dmitri Kishchukov, Michael Kornbluh, Alex Kotliarov, Alexey Kuznetsov, Frank Ludwig, Anatoly Mnev, Jim Ostell, Vasuki Palanigobu, Anton Perkov, Andriy Petrow, Sergey Petrunin, Wenyao Shi, Denis Sinyakov, Thomas Smith, Vladimir Soussov, Elena Starchenko, Hanzhen Sun, Andrei Vereshchagin, Jewen Xiao, Eugene Yaschenko, Liwei Zhou, Slava Khotomliansky, Igor Lozitskiy, Craig Oakley, Eugene Semenov, Ben Slade, Constantin Vasilyev, Sherri Bailey, William Bocik, David Brodsky, Peter Cooper, Jada Lewis, Hanguan Liu, Bonnie Maidak, Wayne Matten, Scott McGinnis, Rana Morris, Monica Romiti, Eric Sayers, Tao Tao, Majda Valjavec-Gratian and Kim Pruitt

The INSDC meeting at EBI, May 2019



The INSDC meeting online, May 2021

Activities on the databases

Databases at Bioinformation and DDBJ Center

DDBJ DNA Data Bank of Japan

	Annotated sequences	Capillary reads	NGS reads	Study	Sample	Assembly	Functional genomics	Variation	Genotype and phenotype	Metabolomics
NCBI	GenBank	Trace Archive	Sequence Read Archive	BioProject	BioSample	Assembly	GEO	dbSNP/dbVar	dbGaP	
EBI		European Nucleotide Archive (ENA)					ArrayExpress	EVA/DGVa	EGA	MetaboLights
DDBJ	DDBJ	Trace Archive	Sequence Read Archive	BioProject	BioSample	Assembly	GEA	JVar-SNP/SV	JGA	MetaboBank

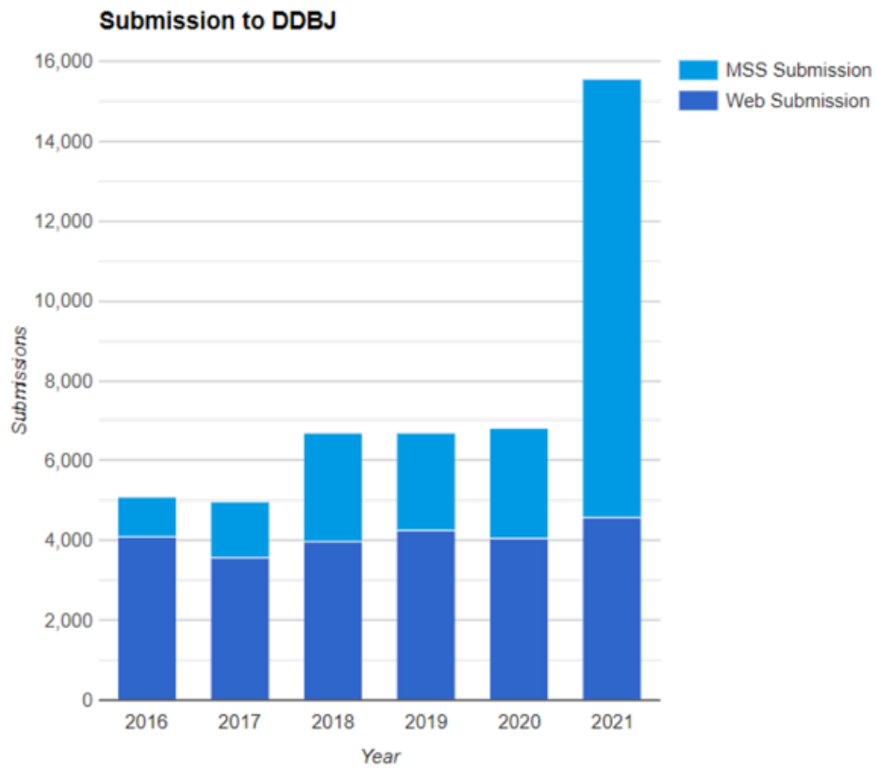
INSDC: International Nucleotide Sequence Databank Collaboration

An example for the “trad” database

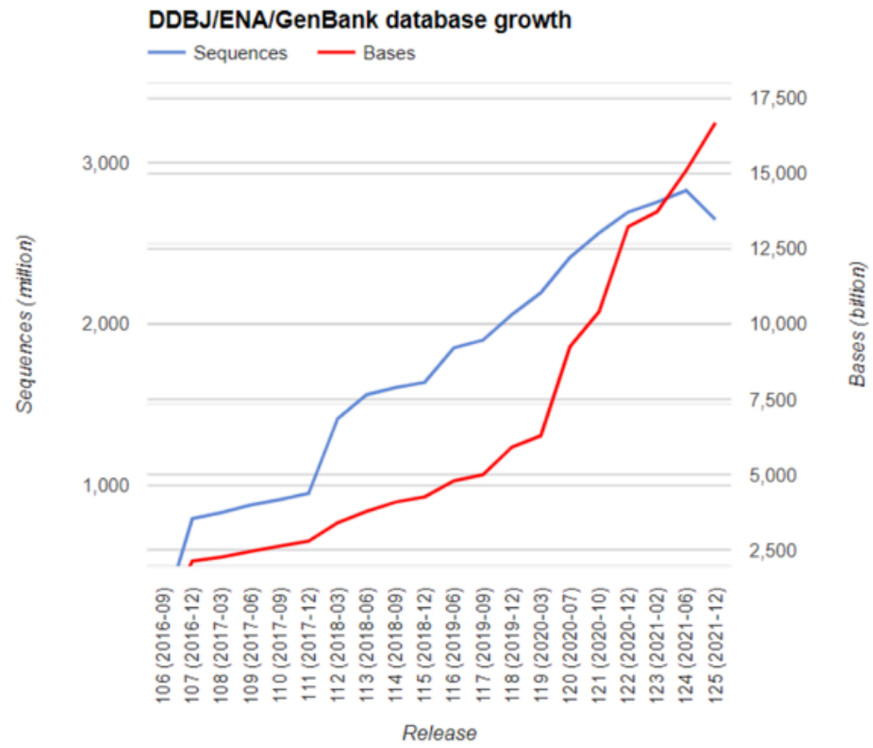
```
LOCUS       AB091058                2109 bp    DNA     linear   BCT 02-SEP-2003
DEFINITION  Gluconacetobacter xylinus cmcase, ccp genes for
            endo-beta-1,4-glucanase, cellulose complementing protein, complete
            cds.
ACCESSION   AB091058
VERSION     AB091058.1
KEYWORDS    .
SOURCE      Gluconacetobacter xylinus
  ORGANISM  Gluconacetobacter xylinus
            Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales;
            Acetobacteraceae; Gluconacetobacter.
REFERENCE   1  (bases 1 to 2109)
  AUTHORS   Kawano,S., Tajima,K., Uemori,Y., Yamashita,H., Erata,T.,
            Munekata,M. and Takai,M.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-AUG-2002) to the DDBJ/EMBL/GenBank databases.
            Contact:Kenji Tajima
            Hokkaido University, Graduate School of Engineering; N13W8,
            Kita-ku, Sapporo, Hokkaido 060-8628, Japan
REFERENCE   2
  AUTHORS   Kawano,S., Tajima,K., Uemori,Y., Yamashita,H., Erata,T.,
            Munekata,M. and Takai,M.
  TITLE     Cloning of Cellulose Synthesis Related Genes from Acetobacter
            xylinum ATCC23769 and ATCC53582: Comparison of Cellulose Synthetic
            Ability Between ATCC23769 and ATCC53582
  JOURNAL   Unpublished (2002)
COMMENT
FEATURES             Location/Qualifiers
     source          1..2109
                     /db_xref="taxon:28448"
                     /mol_type="genomic DNA"
                     /note="synonym:Acetobacter xylinum"
                     /organism="Gluconacetobacter xylinus"
                     /strain="ATCC 53582"
     CDS             10..1038
                     /codon_start=1
                     /gene="cmcase"
                     /product="endo-beta-1,4-glucanase"
                     /protein_id="BAC82540.1"
                     /transl_table=11
                     /translation="MSVMAAMGGAQVLSSTGAFADTAPDAVAQQWAIFRAKYLRPSGR
                     VVDTGNGGESHSEGQGYGMLFAASAGDLASFQSMWMWARTNLQHTNDKLFSWRFLKGH
                     QPPVPDKNNATDGDLLIALALGRAGKRFQRPDYIQDAMAIYGDVLNLMTMKAGPYVVL
                     MPGAVGFTKKDSVILNLSYYVMPSLLQAFDLTADPRWRQVMEDGIRLVSAGRFGQWRL
                     PPDWLAVNRATGALSIASGWPPRFSYDAIRVPLYFYWAHMLAPNVLADFTRFWNNFGA
                     NALPGWVDLTTGARSPYNAPPGYLAVAECTGLDSAGELPTLDHAPDYYSAALTLLVYI
                     ARAEETIK"
     CDS             1035..2096
                     /codon_start=1
                     /gene="ccp"
                     /product="cellulose complementing protein"
                     /protein_id="BAC82541.1"
                     /transl_table=11
                     /translation="MSASGSDEVAGGGQAGSPQDFQRVLRSFGVEGGQYSYRPFVDRS
                     FDVTGVPEAVERHFDQAEHDTAVEEQVTPAPQIAVAPPPPPVVPDPPAIVTETAPPPP
                     VVVSAPVTYEPPAAAVPAEPPVQEAPVQAAPVPPAPVPPIAEQAPPAAPDPASVPYAN
                     VAAAPVPPDPAPVTPAPQARVTGPNTRMVEPFSRPQVRTVQEGATPSRVPSRSMNAFP
                     RTSASSISERPVDRGVADEWSPVPKARLSPRERPRPGDLSFFFQGMRDTRDEKKFFPV
                     ASTRSVRSNVSRMTSMTKTDTNSSQASRPGSPVASPDGSPTMAEVFMTLGGRATELLS
                     PRPSLREALLRRRENEEES"
BASE COUNT          343 a          661 c          661 g          444 t
ORIGIN
        1 cgttccttta tgtcggtcat ggcggcgatg ggaggggcgc aggtgctttc atccaccggt
       61 gcgttcgcag acaccgcccc cgatgcggtc gcgcagcaat gggccatctt ccgcgccaag
      121 tatcttcgtc ccagcggacg tgtcgtggat acgggcaatg gtggcgaatc ccatagtgag
      181 gggcagggct atggcatgct ctttgccgcg tcggcggggg accttgcgtc gttccagtcg
      241 atgtggatgt gggcgcgcac caacctgcag cataccaatg acaagctgtt ttcctggcgg
      301 ttcctcaagg ggcatcagcc cccggtgccc gacaagaaca atgccacaga tggcgacctg
      361 ctgatcgcgc ttgcgcttgg tcgtgcgggc aagcgtttcc agcgccccga ttacattcag
      421 gacgccatgg ccatttatgg cgatgtgctg aacctgatga cgatgaaggc gggaccgtat
      481 gtcgtcctca tgcccggtgc tgtcggcttt accaagaagg acagcgtgat cctcaacctg
      541 tcctattacg tcatgccctc gctgctgcag gcgttcgacc ttacggccga cccgcgctgg
      601 cgtcaggtga tggaagacgg gattcgcctt gtttccgccg gccgtttcgg gcagtggcgc
      661 ctgccccccg actggctggc ggtgaatcgc gccaccggtg cgctgtcgat cgcatcggga
      721 tggccgccgc gcttttccta tgatgcgatt cgggtgccgc tttatttta ttgggcgcat
      781 atgctggcgc cgaacgtgtt ggctgatttc acccgattct ggaataattt cggggctaat
      841 gccctgccag gatgggttga tctgacaaca ggggcgcgtt cgccgtacaa cgccccgcct
      901 ggatatcttg ctgttgccga atgcacgggg cttgattctg ccggggaact cccgacactg
      961 gatcatgcgc ccgattatta ttccgcagcg ttgacgctgc tcgtttacat cgcgcgggcg
     1021 gaggagacta taaagtgagt gcttcagggt ctgatgaggt ggctgggga gggcaggctg
     1081 gaagtccgca ggattttcag cgggtcctgc gttcttttgg tgtcgaaggt gggcagtatt
     1141 cctaccggcc gtttgttgac cgttcctttg atgtgacagg cgtgcccgag gctgttgaaa
     1201 ggcacttcga tcaggcggag catgacacgg cggttgagga gcaggtcact cccgcgccac
     1261 aaatcgcggt cgcaccgcca ccgccgccag tcgttcctga cccgcccgcc atcgtgacgg
     1321 aaaccgcgcc cccgccgcct gtcgtggtca gcgctccggt cacgtatgaa cccccggctg
     1381 ccgccgtgcc ggcagagcct cccgttcagg aagcccccgt gcaggcggcg ccggttcccc
     1441 ccgcgcctgt gcccccgatt gcggagcagg ctcctcccgc ggcgccggac ccggcatccg
     1501 tgccgtatgc gaacgtcgcg gcagcacccg ttccacctga tcccgcaccg gttacgcctg
     1561 cgccgcaggc gcgcgtgacg gggccgaaca cccgtatggt ggagcccttt tcccgcccgc
     1621 aggtccgcac ggtgcaggag ggggcaaccc cgtcacgtgt accttcgcgt tcaatgaacg
     1681 ctttcccccg cacatcagca tcgtccataa gtgagcgtcc ggtggacagg ggtgttgccg
     1741 atgaatggag tcctgttccg aaggcacgcc tcagcccgcg ggagcgtccg cgtcccggcg
     1801 atctgagctt tttctttcag gggatgcgcg acacccgtga tgaaaagaag ttctttcccg
     1861 tggcgtccac gcgatcagtt cgttctaatg tttccaggat gaccagcatg accaagacag
     1921 acacgaattc ctctcaggct tctcgtcccg gcagccccgt cgcctcgcct gatgggtcgc
     1981 ccacaatggc cgaagtgttc atgacgctgg gtggtcgtgc gacggaactc ctcagccccc
     2041 gtccttcgct gcgggaggcg ctgttgcgtc gtcgtgaaaa cgaagaagaa tcctaaggcc
     2101 ctatattca

//
```

Trad in DDBJ



Annual number of registrations to Trad DDBJ.



Total number of sequences and bases released by DDBJ.

- Total number of registrations related to DDBJ entries in 2021 were 15,573 (4,574 web registrations and 10,999 MSS registrations). The rapid increase in the number of MSS submissions in 2021 was due to the large amount of MAG (Metagenome-Assembled Genome) data from TPA-WGS registered by The University of Tokyo. As of December 2021 DDBJ release 125, the total number of bases: 16,670,849,721,017 (about 17 trillion) and the total number of sequences: 2,650,249,718 (about 2.7 billion).

Databases at Bioinformation and DDBJ Center

DDBJ DNA Data Bank of Japan

	Annotated sequences	Capillary reads	NGS reads	Study	Sample	Assembly	Functional genomics	Variation	Genotype and phenotype	Metabolomics
NCBI	GenBank	Trace Archive	Sequence Read Archive	BioProject	BioSample	Assembly	GEO	dbSNP/dbVar	dbGaP	
EBI	European Nucleotide Archive (ENA)						ArrayExpress	EVA/DGVa	EGA	MetaboLights
DDBJ	DDBJ	Trace Archive	Sequence Read Archive	BioProject	BioSample	Assembly	GEA	JVar-SNP/SV	JGA	MetaboBank

INSDC: International Nucleotide Sequence Databank Collaboration

NGS's

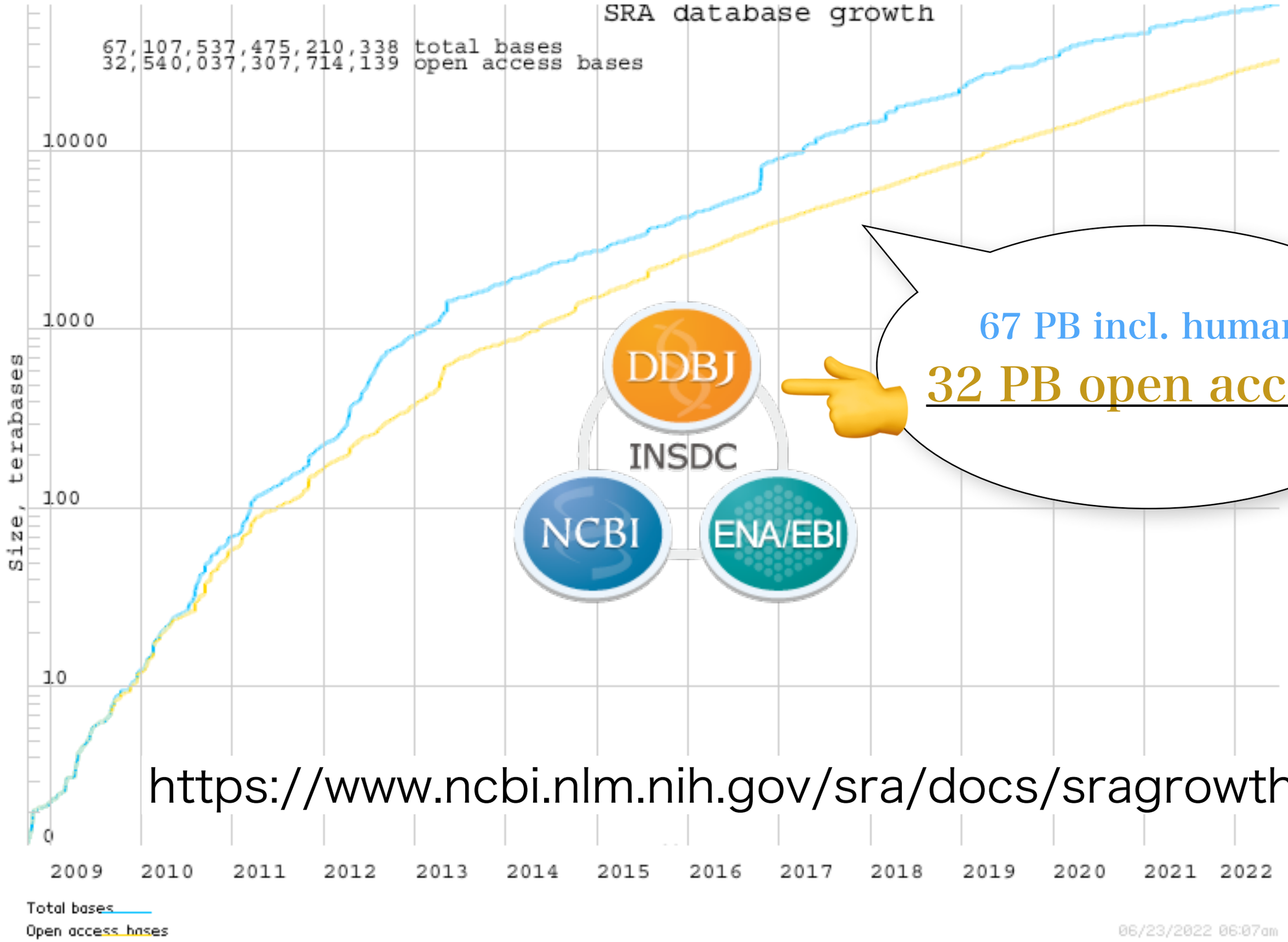
PacBio: Sequel II System

illumina: NovaSeq System

Oxford NANOPORE MinION / GridION

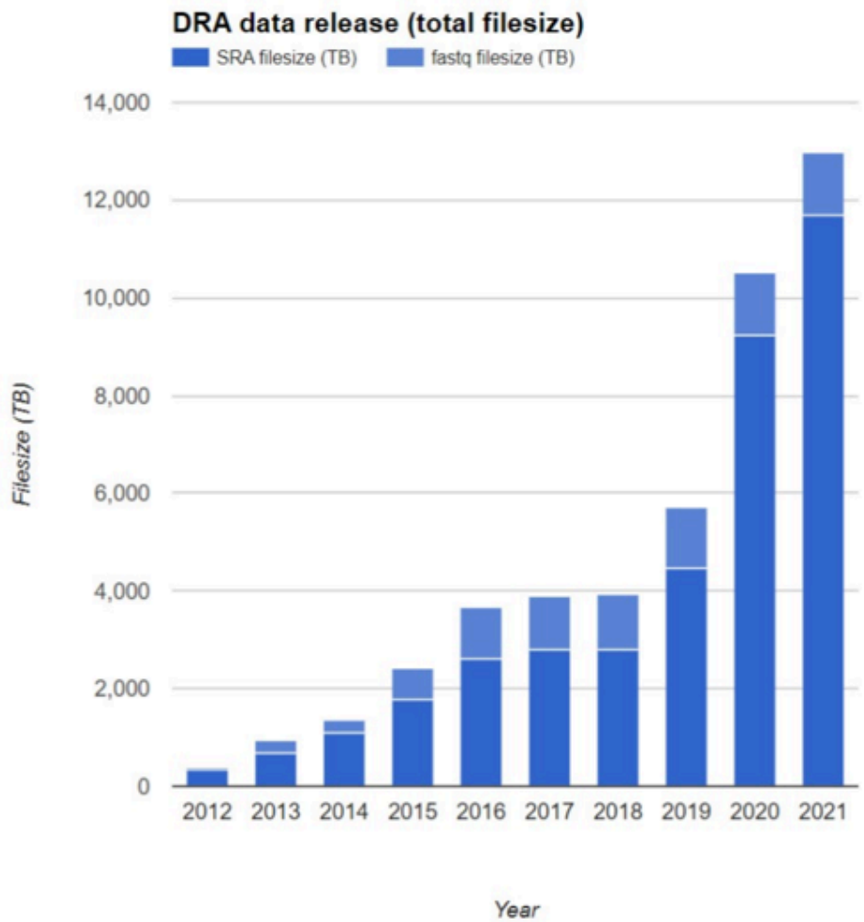
MGI DNBSEQ

SRA (NGS-archive) growth

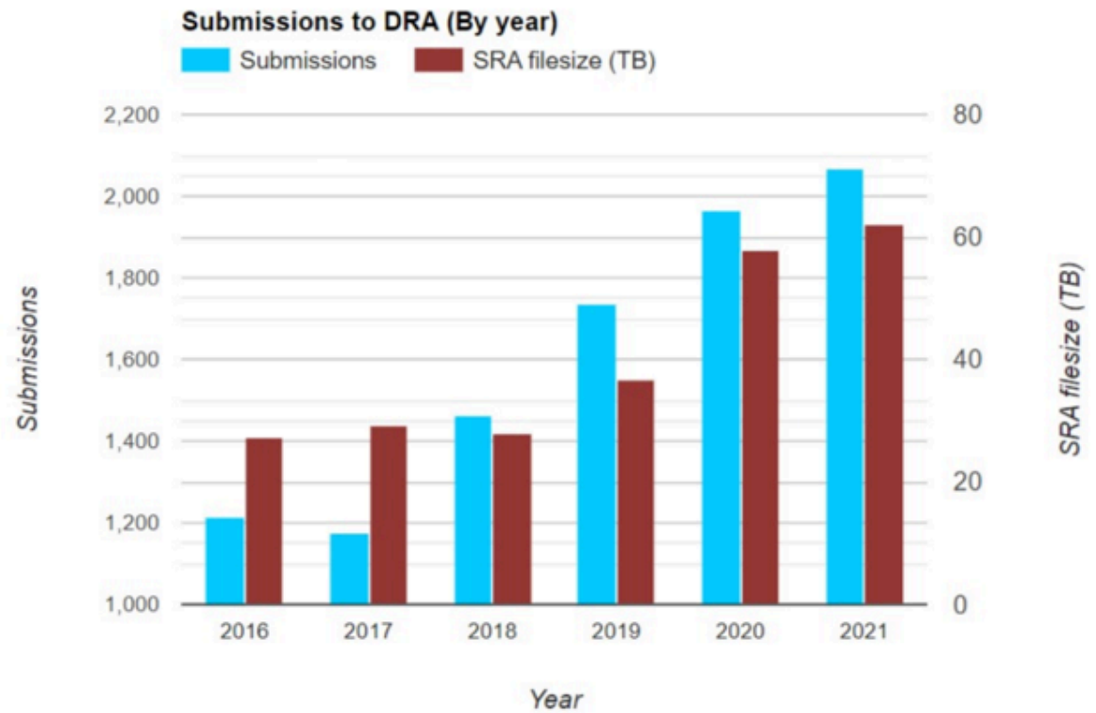


https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

SRA in DDBJ (DRA)



Amount of SRA and fastq files in DRA



Number of registrations and amount of files per year

- The number of DDBJ SRA (DRA) registrations in 2021 was 2,066 (62 TB). Total SRA and fastq file size published were 11.7PB and 1.3PB, respectively. In December 2021.

Databases at Bioinformation and DDBJ Center

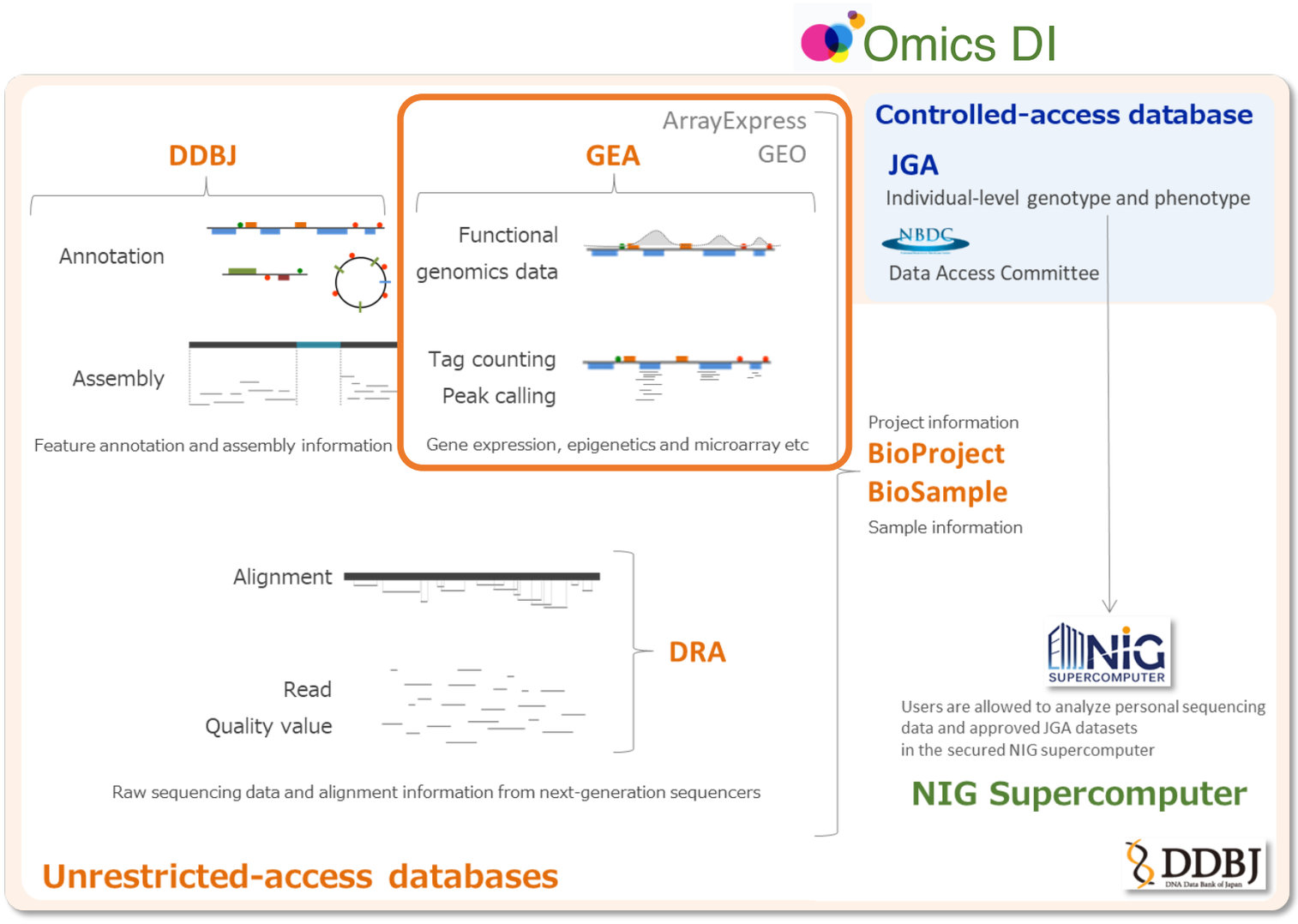
DDBJ DNA Data Bank of Japan

	Annotated sequences	Capillary reads	NGS reads	Study	Sample	Assembly	Functional genomics	Variation	Genotype and phenotype	Metabolomics
NCBI	GenBank	Trace Archive	Sequence Read Archive	BioProject	BioSample	Assembly	GEO	dbSNP/dbVar	dbGaP	
EBI	European Nucleotide Archive (ENA)						ArrayExpress	EVA/DGVa	EGA	MetaboLights
DDBJ	DDBJ	Trace Archive	Sequence Read Archive	BioProject	BioSample	Assembly	GEA	JVar-SNP/SV	JGA	MetaboBank

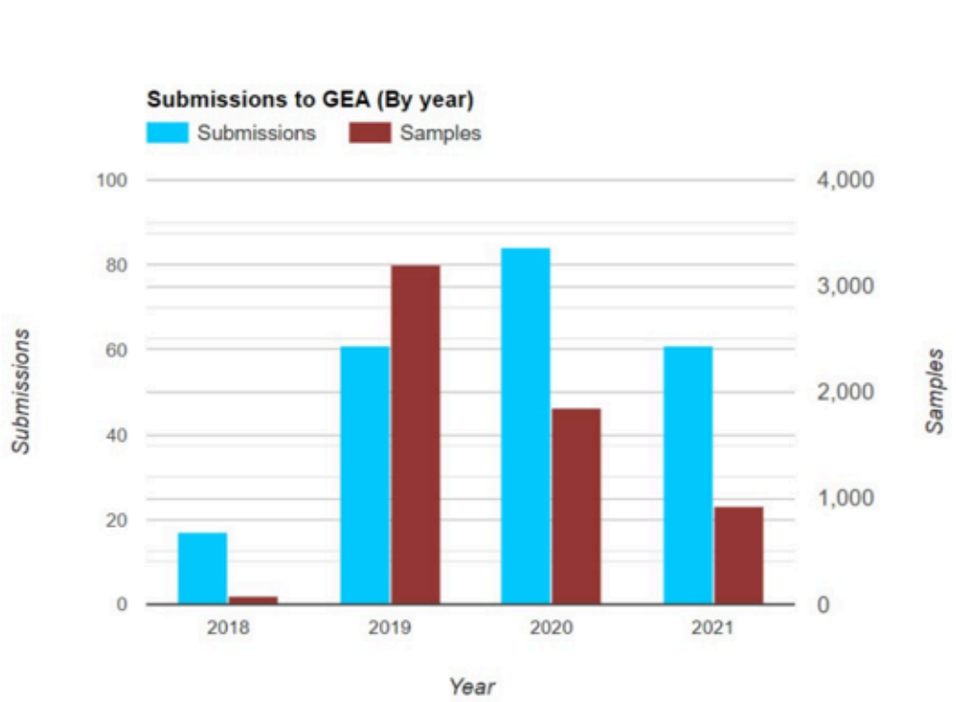
INSDC: International Nucleotide Sequence Databank Collaboration

Genomic Expression Archive

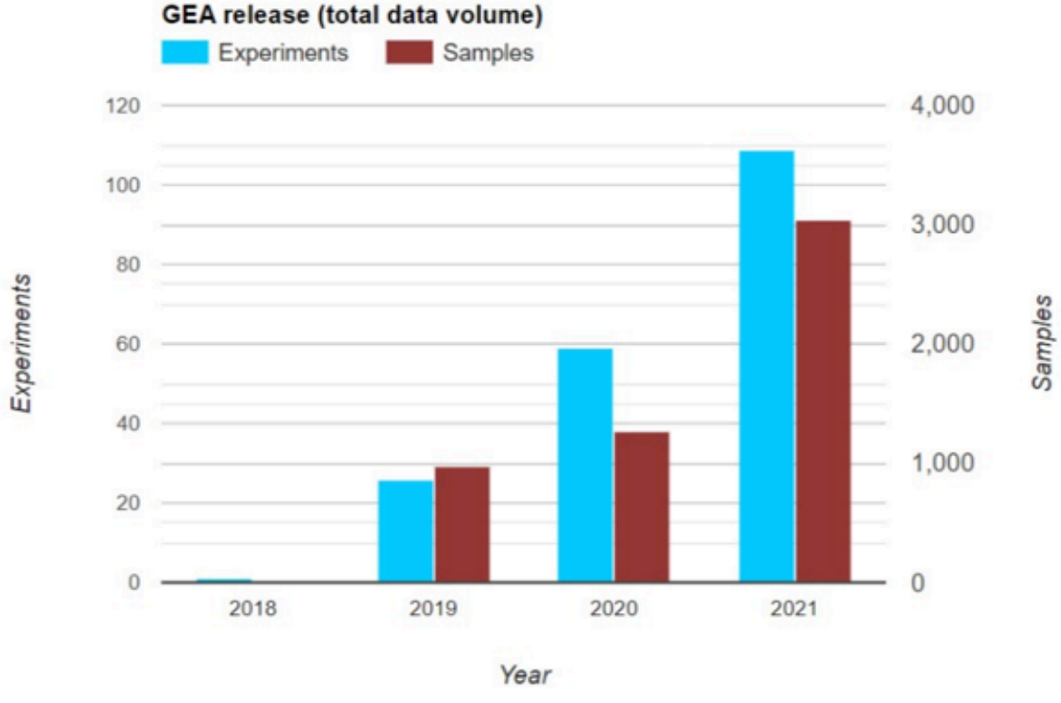
We have launched a long-awaiting database ‘Genomic Expression Archive (GEA)’ for functional genomics data (a counterpart of GEO and ArrayExpress). The GEA archives data in the standard MAGE-TAB format and metadata will be indexed by EBI’s Omics DI.



GEA: Genomic Expression Archive



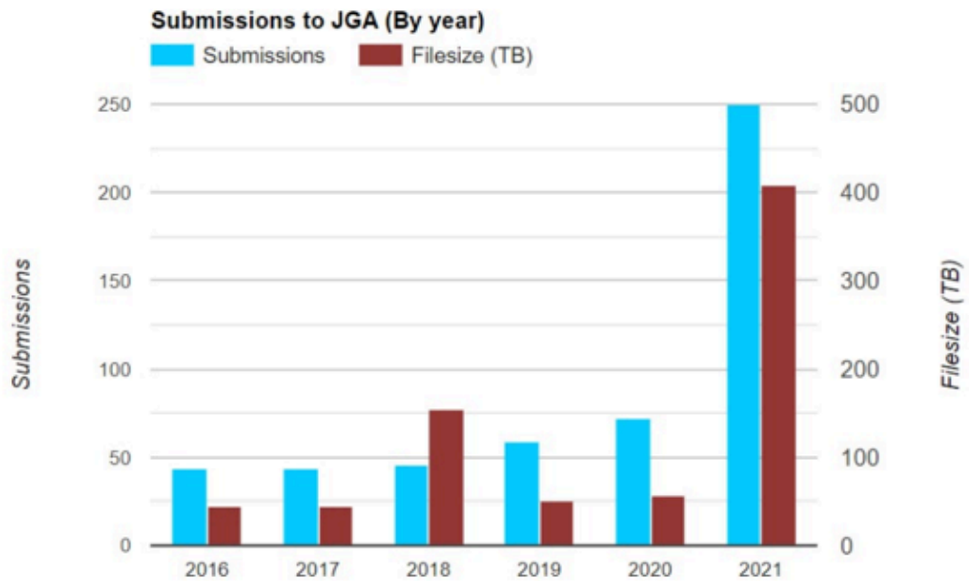
GEA's submissions and samples numbers by year.



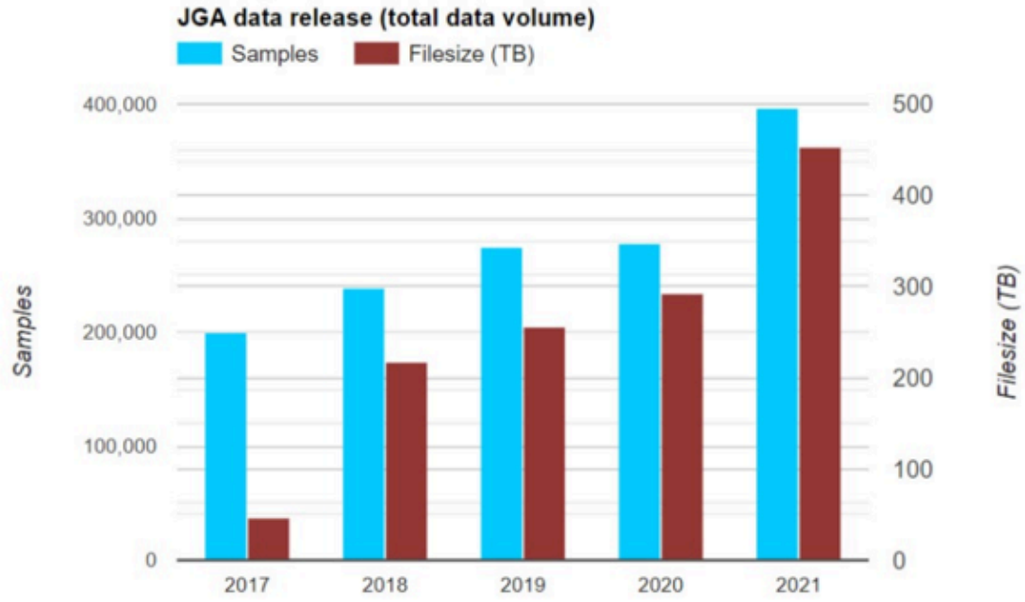
Number of relased experiments and samples from GEA.

- GEA is a database of gene expression data from microarrays and NGS in DDBJ. In 2021, we had 61 data submissions, 50 (ca. 110 experiments / 3,000 samples) were publicly available.

Japanese Genotype-phenotype Archive (JGA)



JGA's submissions and samples numbers by year.

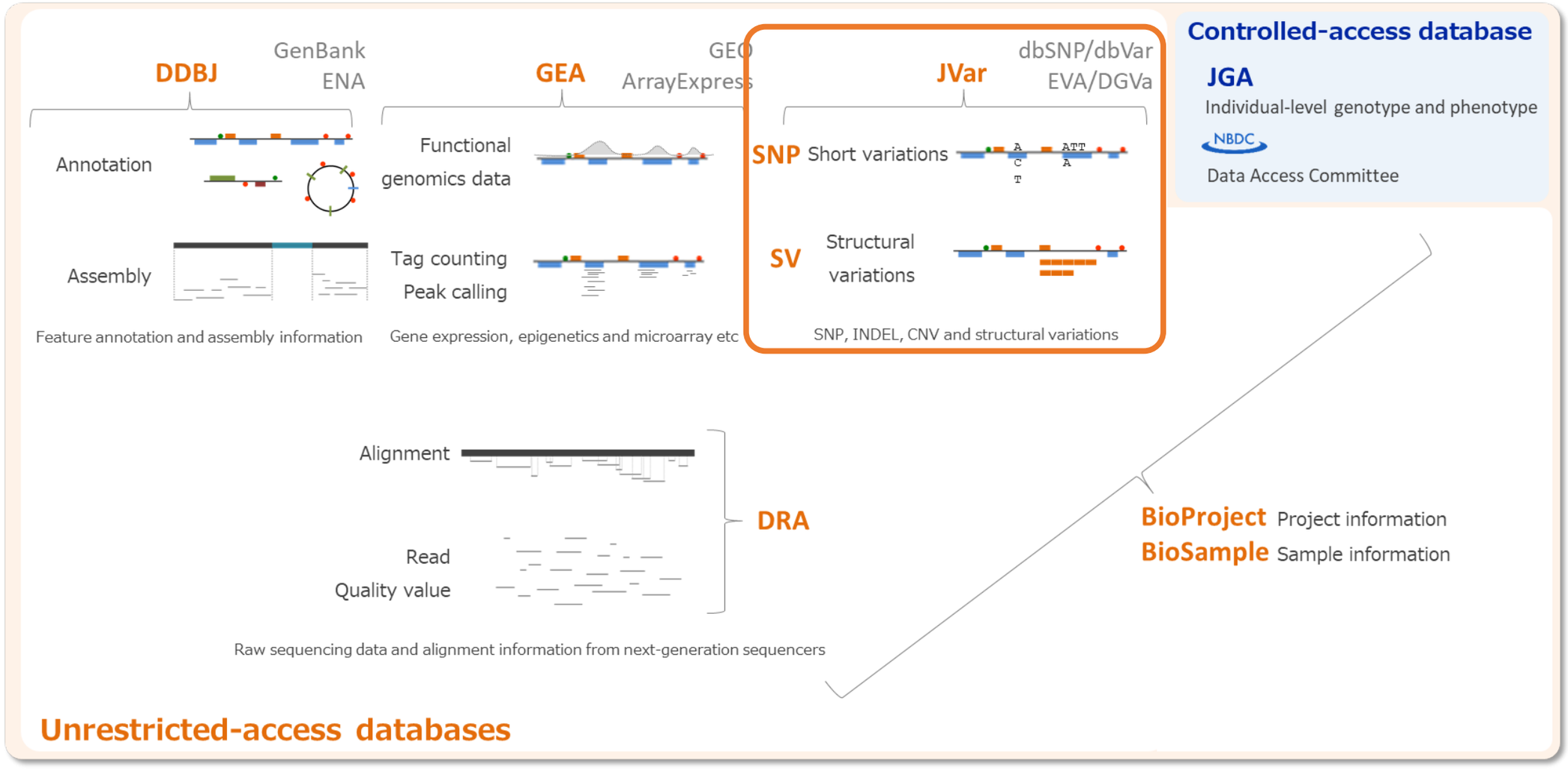


Number of relased experiments and samples from JGA.

- A control-access database for personal genome phenotype/genotype information operated jointly with NBDC. Registration and use of the database are subject to NBDC review. In 2021, 250 data and 409 TB were submitted, more than a 3-fold increase compared to the previous year. As of the end of 2021, 240 studies, 396,471 samples, and 453 TB of data were released. The number of applications approved by NBDC in 2021 totaled 43 (domestic: 31, overseas: 12). A cumulative total number is 178.

Japan Variation Archive

We have launched two new databases ‘Japan Variation Archive (JVar)’ for human variation data. JVar consists of two parts ‘JVar-SNP’ (counterpart of dbSNP) and ‘JVar-SV’ (a counterpart of dbVar). **JVar accepts human-data only.**

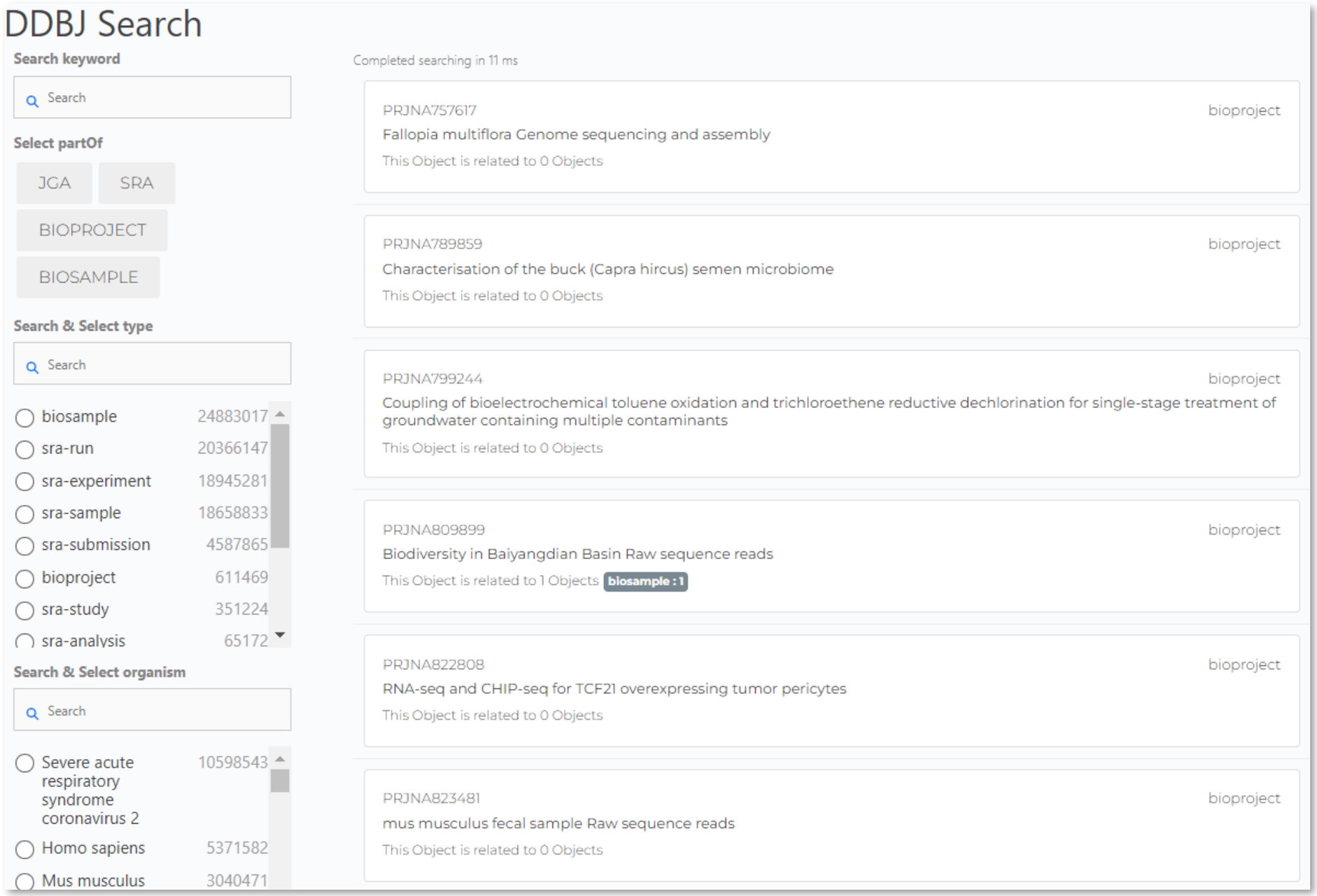


Developments

DDBJ Search covers BioProject/BioSample/SRA

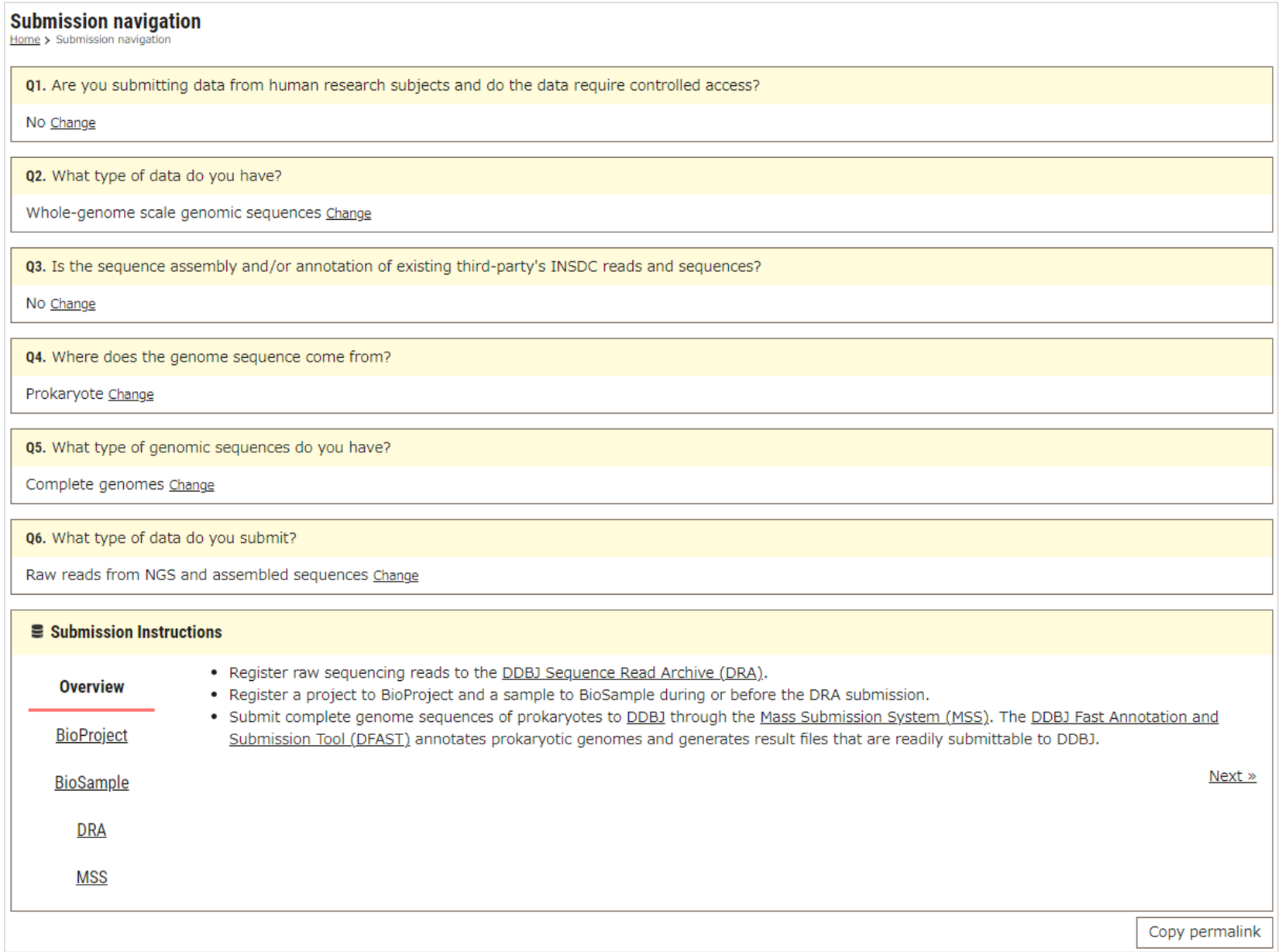
The DDBJ Search has covered BioProject/BioSample/SRA metadata in addition to JGA since October 2021.

Takatomo Fujisawa, PhD
Project Researcher
Head of Database Division



https://ddbj.nig.ac.jp/search

Submission navigation



https://www.ddbj.nig.ac.jp/submission-navigation-e.html?state=3Ef9JWrZHUjyn4GKTxeEMqukQe4o

In the submission navigation site released in April 2022, submitters can know how to submit their data by simply answering several questions.

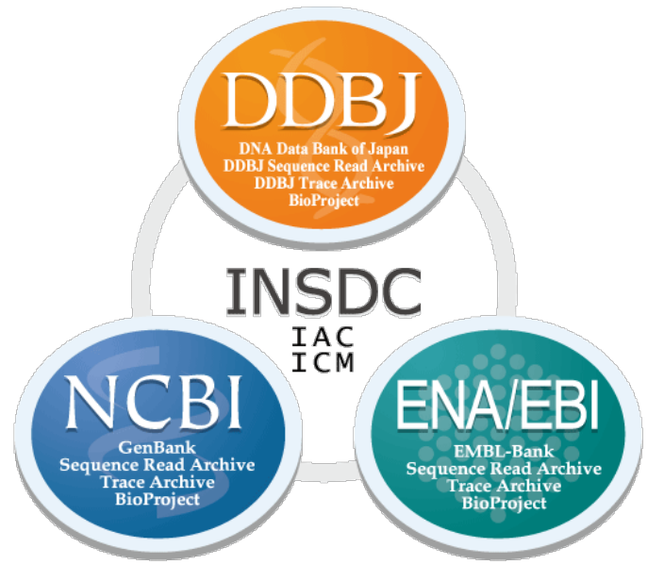
Pipeline for bacterial genome -> Official Submission Tool

TANIZAWA Yasuhiro, PhD
Assistant Professor

https://dfast.nig.ac.jp

Taxonomic assignment of genome is checked by using average nucleotide identity (ANI).
In 2020, 92% of bacterial genomes were annotated and submitted through DFAST.

Release of part of 10K microbe type strain



DDBJ released genome sequence data of 689 type strains, which had been submitted by the WFCC-MIRCEN World Data Centre for Microorganisms (WDCM) and China National Microbial Data Center (NMDC) at IMCAS. This is part of the Global Catalogue of Microorganisms (GCM) 10K type strain sequencing project.

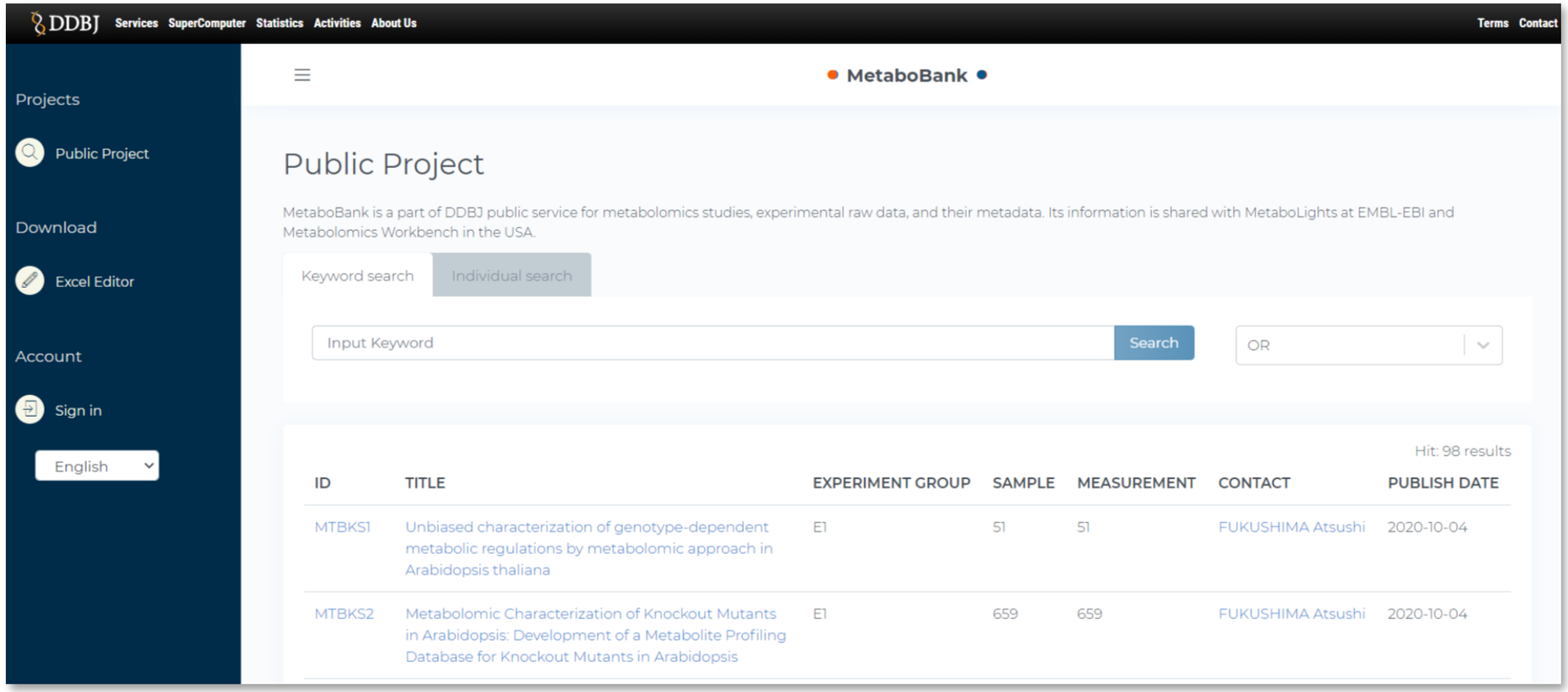
Reference

https://doi.org/10.1099/ijsem.0.003276

MetaboBank

In October 2020, we have released ‘MetaboBank’ for metabolomics data as a member of MetabolomeXchange

SAKURAI Nozomu, PhD
Project Associate Professor

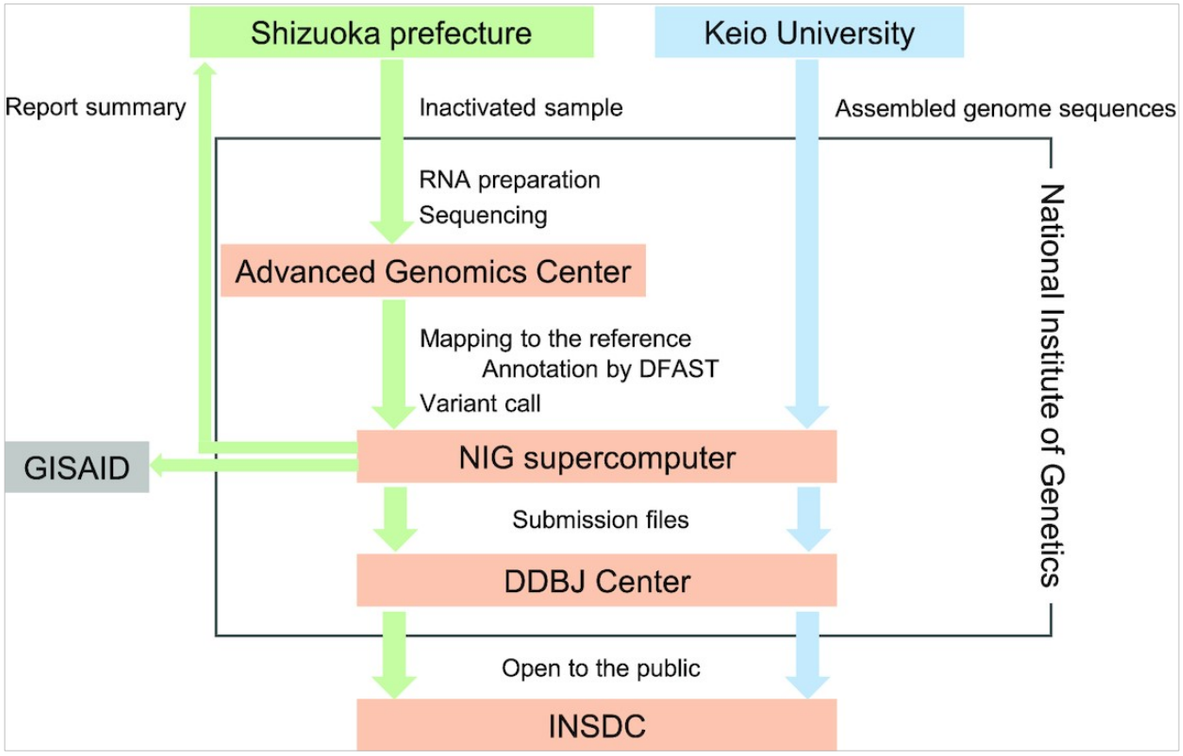


MetabolomeXchange

An international data aggregation and notification service for metabolomics.

Responses to COVID-19 update

Left, Heita Kawakatsu (Governor, Shizuoka Prefecture)
Right, Fumio Hanaoka (Director-General, National Institute of Genetics)



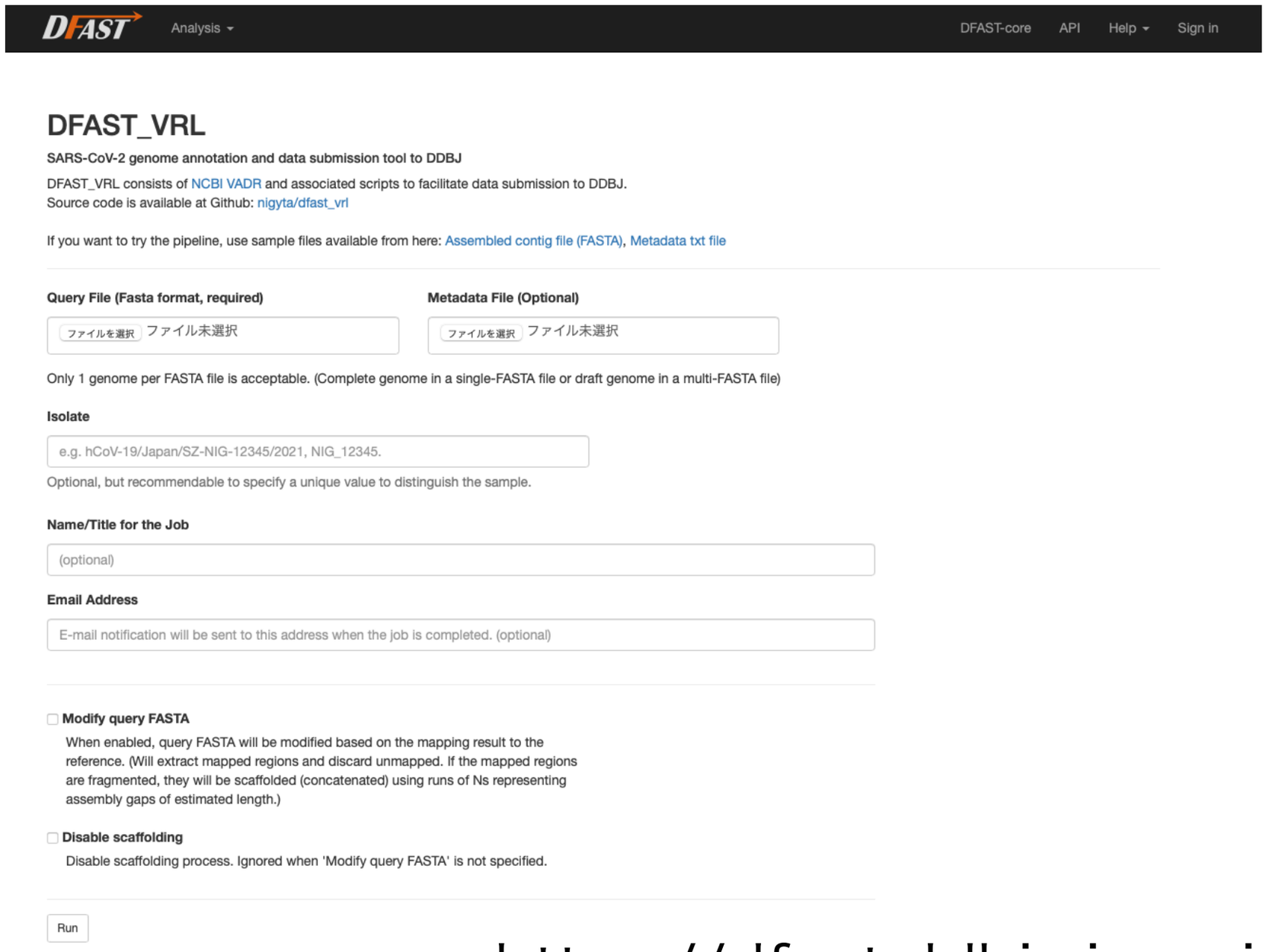
Genome sequencing, analysis and registration to INSDC for SARS-CoV-2 samples.

The signing of a memorandum of cooperation with Shizuoka prefecture to implement whole-genome analysis of novel coronavirus in April 2021. SARS-CoV-2 genome sequences were shared among INSDC and GISAID.

⇒ Toward the open sharing of SARS-CoV-2 genomic data in Japan, we have started the “Japan COVID-19 Open Data Consortium”.

DFAST_VRL: SARS-CoV-2 genome annotation and data submission tool to DDBJ

TANIZAWA Yasuhiro, PhD
Assistant Professor



https://dfast.ddbj.nig.ac.jp/dfv/

(DFAST, an automated annotation pipeline for bacterial genome submission, was used for 75.6% of the microbial genomes registered to DDBJ in 2021)

NIG Supercomputer

NIG SuperComputer replacement and migration

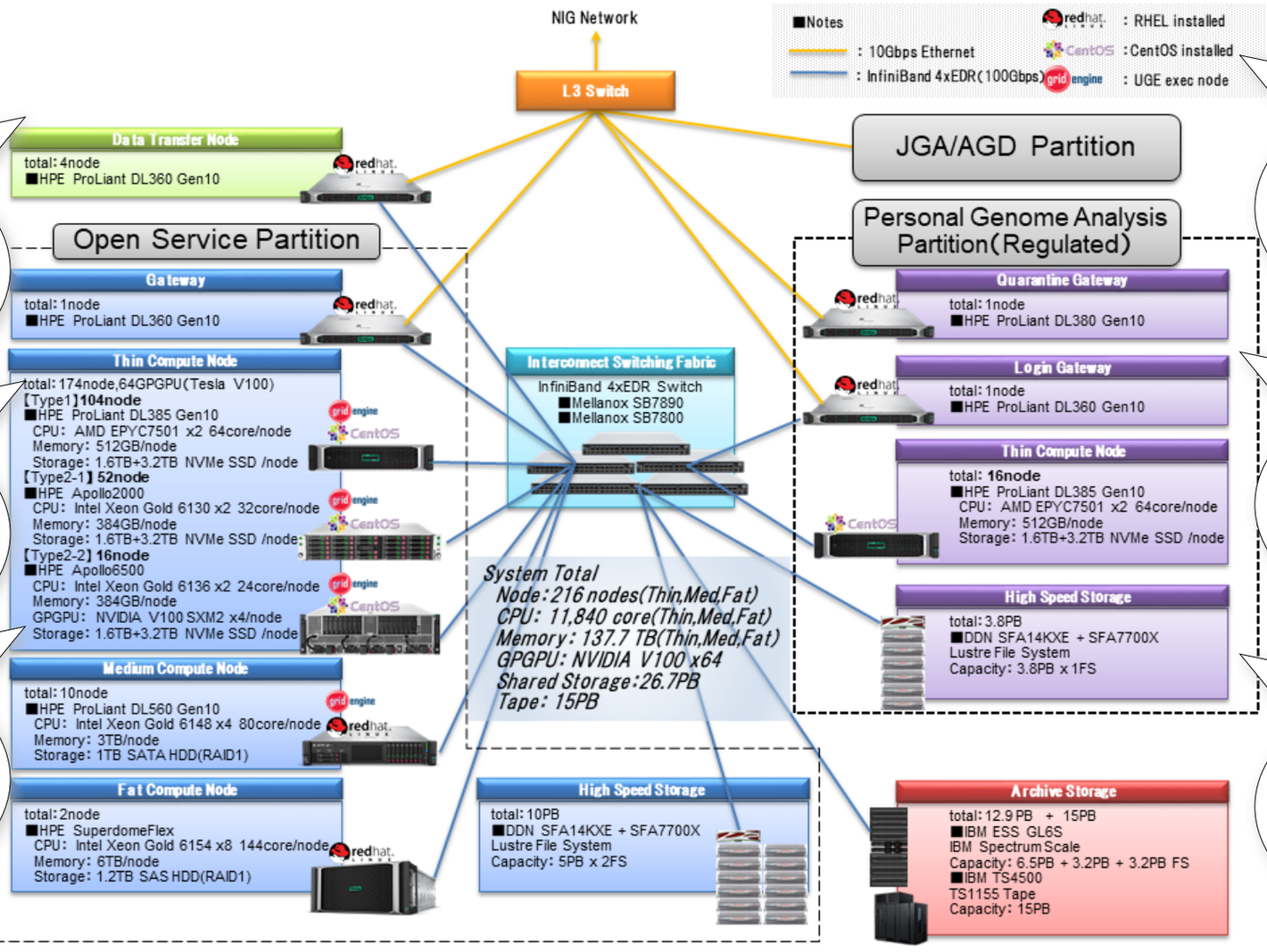
We have renewed the NIG SuperComputer in 2019 (7 years since the last replacement).
Singularity/Docker containers are available for users.

- CPU cores: 11,000 cores
- 43.6 PB HSM strage
- **Singularity/Docker container images**

13.6 PB
Lustre

15 PB
GPFS

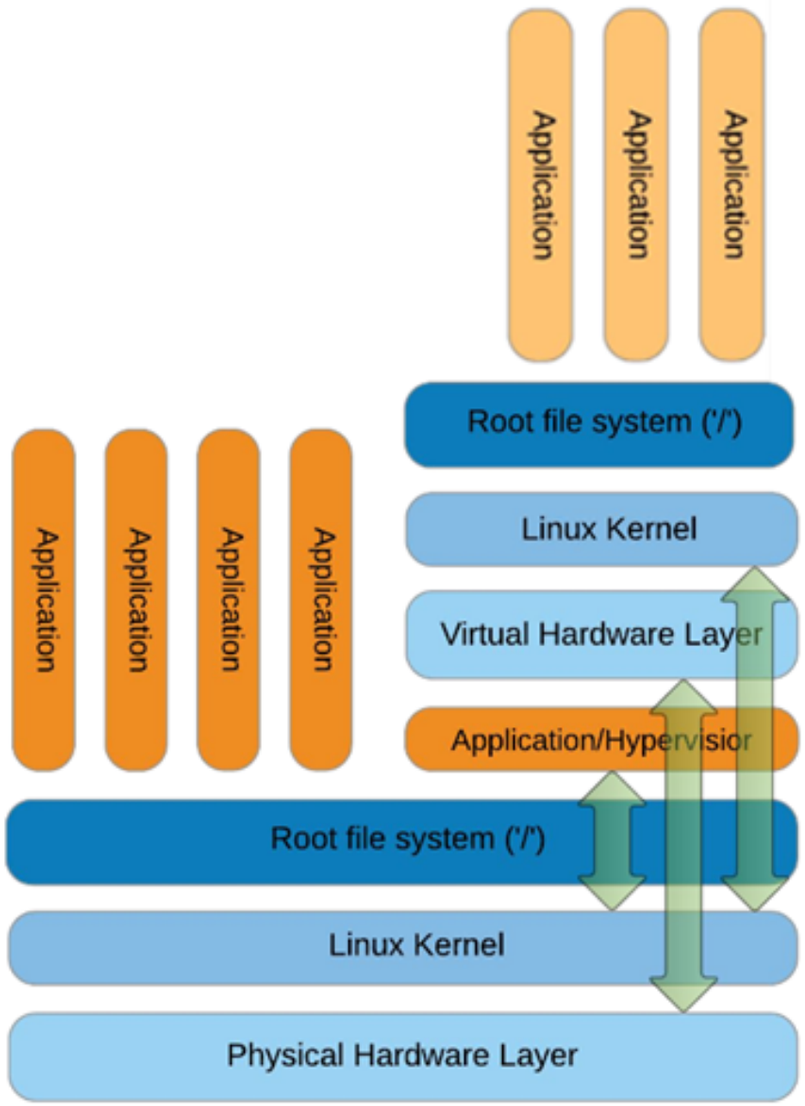
15 PB
Tape strage



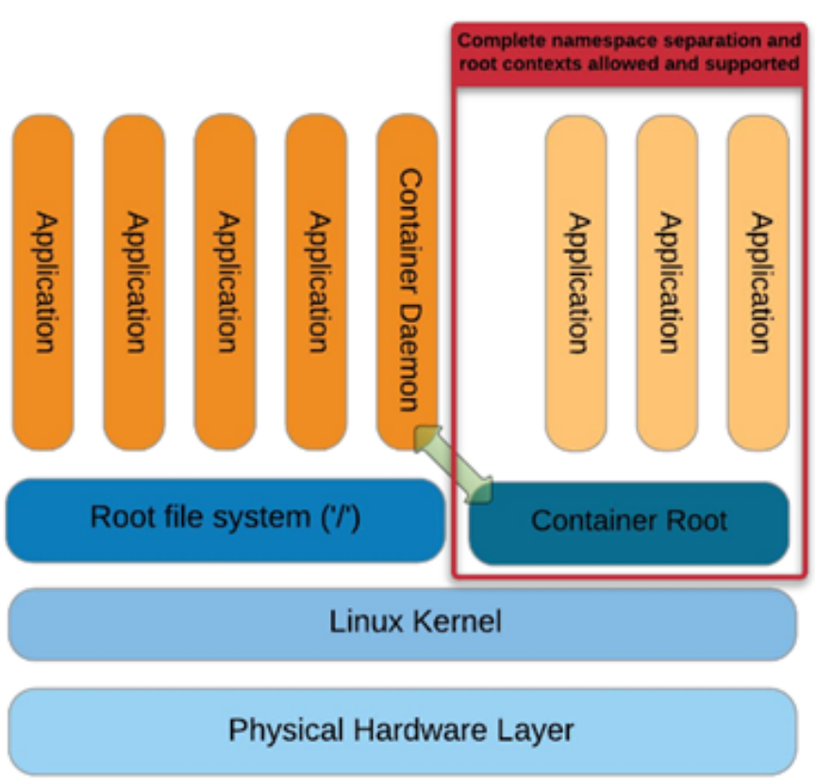
"thin"
512 / 384 GB
64 / 32 cores
x 172 nodes

"medium"
3 TB
memory
x 10

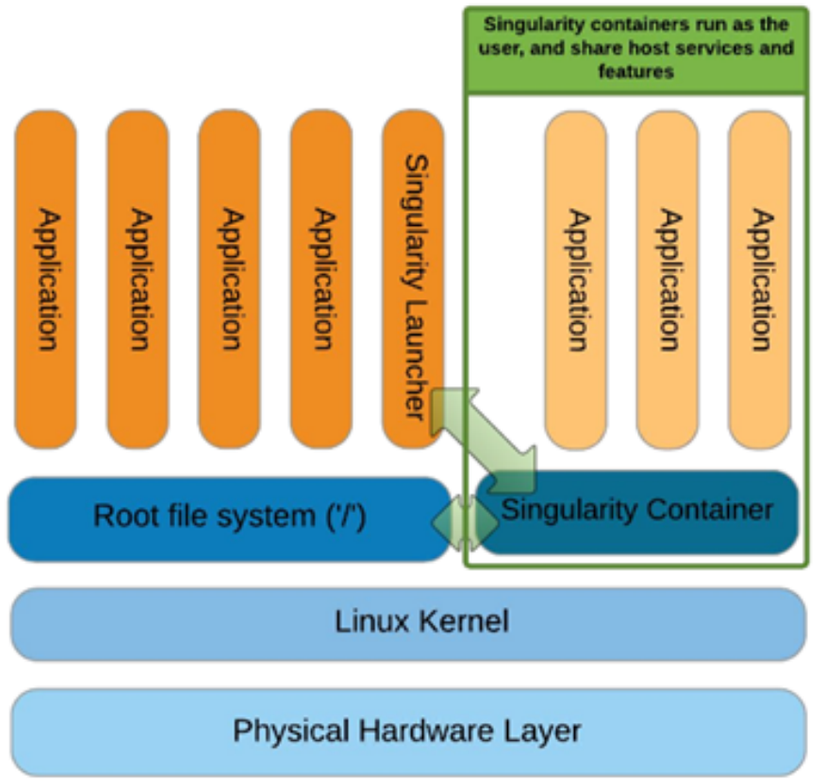
"fat"
12 TB
memory



General VM
eg ESXi



General Container
eg Docker



HPC Container
Singularity

Source: [Greg Kurtzer keynote at HPC Advisory Council 2017 @ Stanford]

BioContainers

Registry | Documentation | GitHub | RestFul API | Twitter | Scholar | Resources

Search

> 2,000 Singularity (Docker) container images at Biocontainers

Search

Sorts by: Default | Sorts Order: Asc | Refine Results: license | Select | Search

2pg_cartesian 22.0K

2pg cartesian is a framework of optimization algorithms for protein structure prediction

license Apache-2.0

a5-miseq 12.4K

A5-miseq is a pipeline for assembling dna sequence data generated on the illumina sequencing platform. this readme will ...

license GPL-3.0

abacas 20.5K

Close gaps in genomic alignments from short reads

license GPL-2+

abacat 636

abacat - A BActerial genome Curation and Annotation Toolkit

license MIT

abawaca 12.0K

abawaca (A Binning Algorithm Without A Cool Acronym) is a binning program that can take advantage of different types of ...

license MIT

abeona 562.0K

A simple transcriptome assembler based on kallisto and cortex graphs.

license Apache-2.0

abismal 1.7K

Tool description is coming

abnumber 12.6K

Abnumber antibody numbering using anarci

license Not available

abpoa 101.9K

an SIMD-based C library for fast partial order alignment using adaptive band

license MIT

abra2 24.2K

ABRA is a realigner for next generation sequencing data. It uses localized assembly and global realignment to align read...

license MIT

abricate 932.9K

Mass screening of contigs for antibiotic resistance genes

license GPL-2.0

abritamr 5.3K

Running amrfinderplus for mdu

license Apache-2.0

abruijn 3.4K

Assembly of Long Error-Prone Reads Using De Bruijn Graphs

license BSD

abundancebin 253.6K

Abundance-based tool for binning metagenomic sequences

license Custom License

abyss 3.9M

De novo genome sequence assembler.

license GPL-3.0+

abyss-k128 28.1K

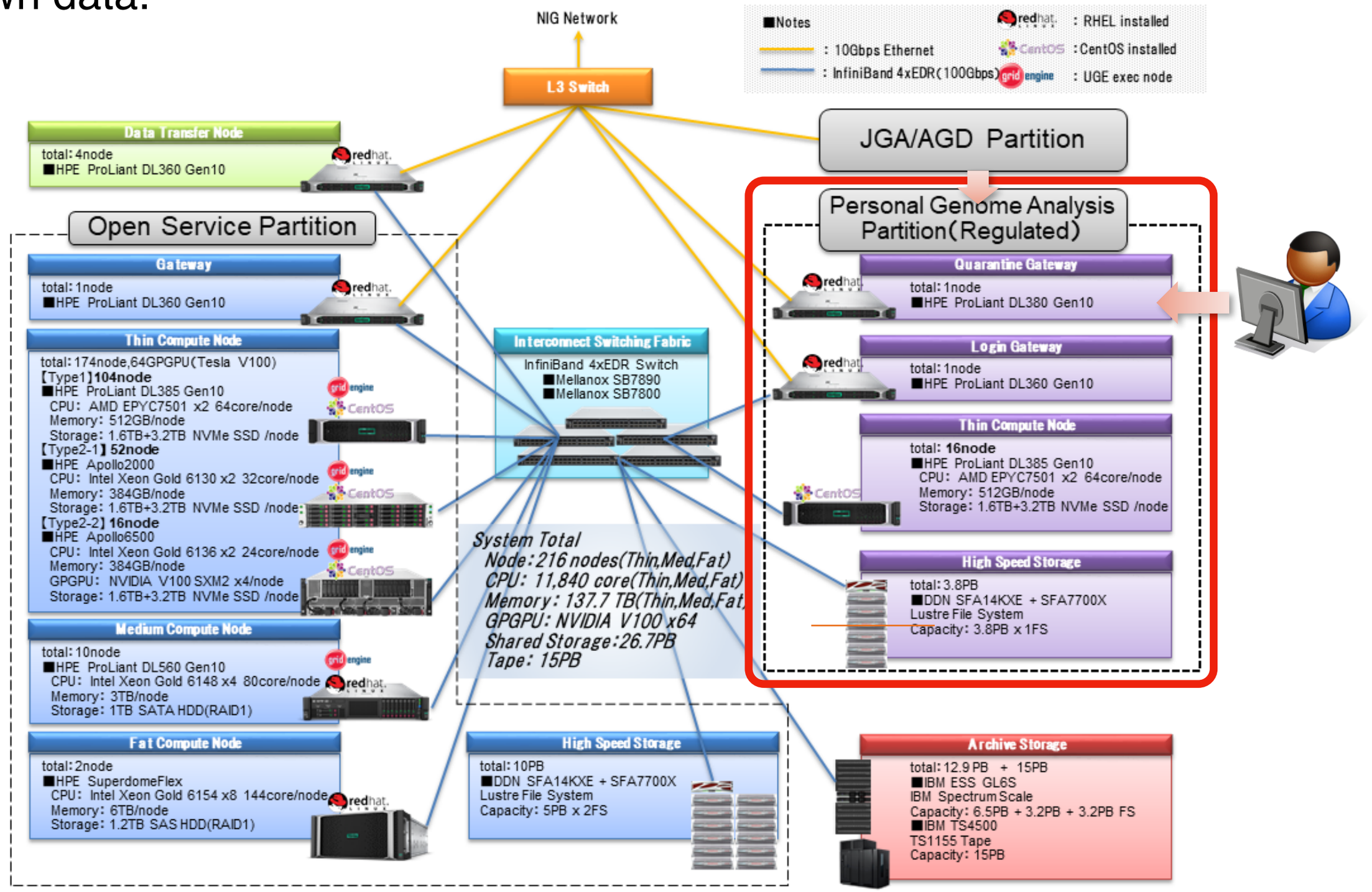
Assembly By Short Sequences - a de novo, parallel, paired-end sequence assembler

license GPL-3.0

Secured partition for personal genome analysis

DDBJ DNA Data Bank of Japan

We have operated the secured supercomputer environment since 2018.
Users can download and analyze JGA datasets in combination with their own data.



CORRESPONDENCE | 04 March 2021

nature

Sequence data: expand comprehensive access

Richard J. Roberts

Scientific data must not be 'balkanized' into multiple databases, each with its own rules and restrictions.

Almost 40 years ago, GenBank and the EMBL databank started independently. They soon joined forces and, with the DNA Database of Japan, formed a repository now called the International Nucleotide Sequence Data Collaboration (INSDC). China is now set to join. The INSDC has been one of the world's most successful initiatives to collect and share scientific data (see *Nature* **590**, 183–184; 2021). As DNA sequence data accumulate at ever-greater rates, the need for INSDC to continue and expand has never been more urgent.

The COVID-19 pandemic is an excellent example of data sharing leading to effective science (see *Nature* **590**, 195–196; 2021). The first sequence of the SARS-CoV-2 virus was released by Yong-Zhen Zhang on 11 January 2020 and was released completely openly that same day in the INSDC databases (accession #MN908947). This enabled the development of rapid PCR-based tests for the viral RNA and jump-started vaccine development.

As international advisers to the INSDC, we call on the scientific community to help ensure that this openness and sharing grows to include many more types of data, so that scientists can use the INSDC to catalyse ever more biological discoveries.

Nature **591**, 202 (2021)

INSDC 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.

https://www.insdc.org/policy.html

Documents for the INSDC (in prep)

- **Founders' Agreement**
  - NCBI, EBI, DDBJ
  - Committees (Advisor, Executive etc.)

- **Membership Arrangement**
  - Membership requirements / metrics
  - Types for the membership (Full, Associate, Submission etc.)

- It will be finished and open this year.

GSA News: GSA has collected all of the metadata of the genome sequence data in INSDC and provides global search and hot data download services.

Genome Sequence Archive

The Genome Sequence Archive (GSA) is a data repository for collecting, archiving, managing and sharing raw sequence data, which is the first repository of the genome sequence data with international journal recognition in China.

Submit	Download	Browse	Document
Submit data to GSA	Download data to your computer	Browse publicly available records	Find help information and documents

Collected and Archived Data

623,370	29,577,117	22,027,652	23,557,898
PROJECTS	SAMPLES	EXPERIMENTS	RUNS

Data of Concern

- Monkeypox Virus Sequence Data
- Omicron Sequence Data in South Africa
- SARS-CoV-2 Sequence Data
- Human Genetic Resources Data
- Open Archive for Miscellaneous Data

Help & Support

If you have any question or would like to give us any suggestion/comment or report a bug, please feel free to contact us.

Email: gsa@big.ac.cn

QQ group: 548170081

We highly appreciate your comments and suggestions for further improvements.

GSA-supported Deposition

Data submissions to GSA have been reported by multiple high-profile journals. GSA has been

Submit BioProject

BioProject ID : PRJKA220470 | Project title : Development of industrial bio-materials research based on biodiversity

Submitter

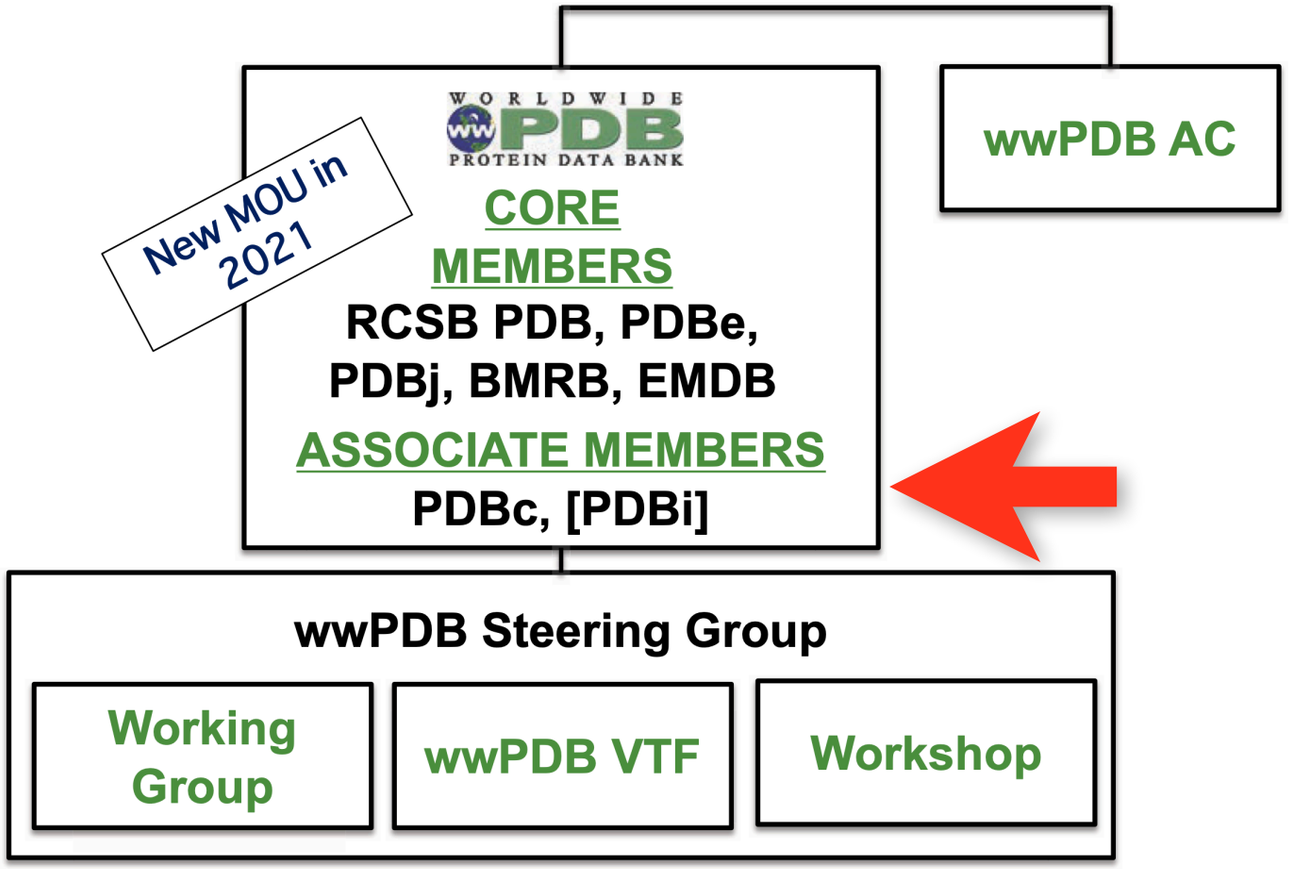
Name	KIM JONG-HOON
Organization	KRIBB
Department	Microbiome convergence research center

Date

Registration Date	2022-10-05

Project Design

NTIS Number	-
Project Title	Development of industrial bio-materials research based on biodiversity
Relevance	industrial
Description	Development of industrial bio-materials research based on biodiversity



PDB expand: 2021

PDBj is responsible for training and auditing the quality of PDBc.