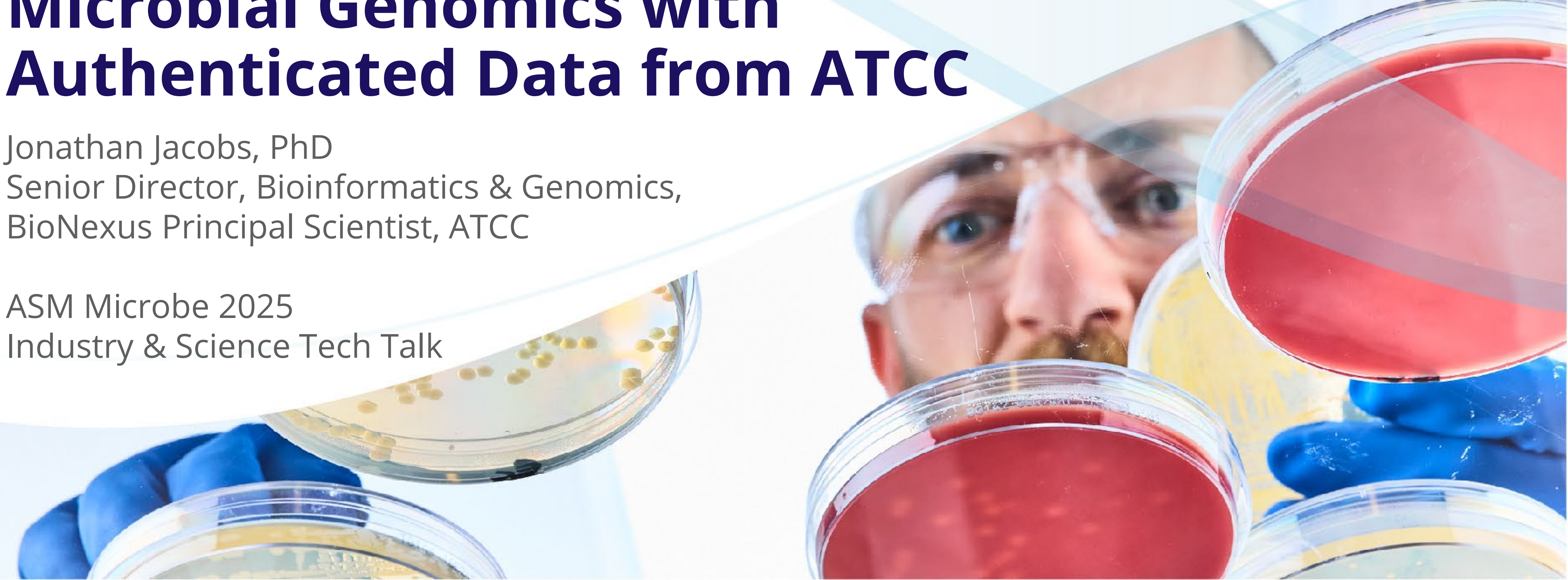
Boosting Reproducibility in Microbial Genomics with Authenticated Data from ATCC

Jonathan Jacobs, PhD
Senior Director, Bioinformatics & Genomics,
BioNexus Principal Scientist, ATCC

ASM Microbe 2025
Industry & Science Tech Talk



About ATCC

- Founded in 1925, ATCC® is a non-profit organization with HQ in Manassas, VA, and an R&D and Services center in Gaithersburg, MD
- World’s premier biological materials resource and standards development organization
- 5,000+ cell lines
- 80,000 microorganisms
- Genomic & synthetic nucleic acids
- Media/reagents

- ATCC® collaborates with and supports the scientific community with industry-standard biological products and innovative solutions
- Growing portfolio of products and services
- Sales and distribution in 150 countries, 20 international distributors
- Talented team of 600+ employees, over one-third with advanced degrees

bsi. ISO 9001 Quality Management

bsi. ISO 13485 Medical Devices Quality Management

ANAB ACCREDITED ISO/IEC 17025 TESTING LABORATORY

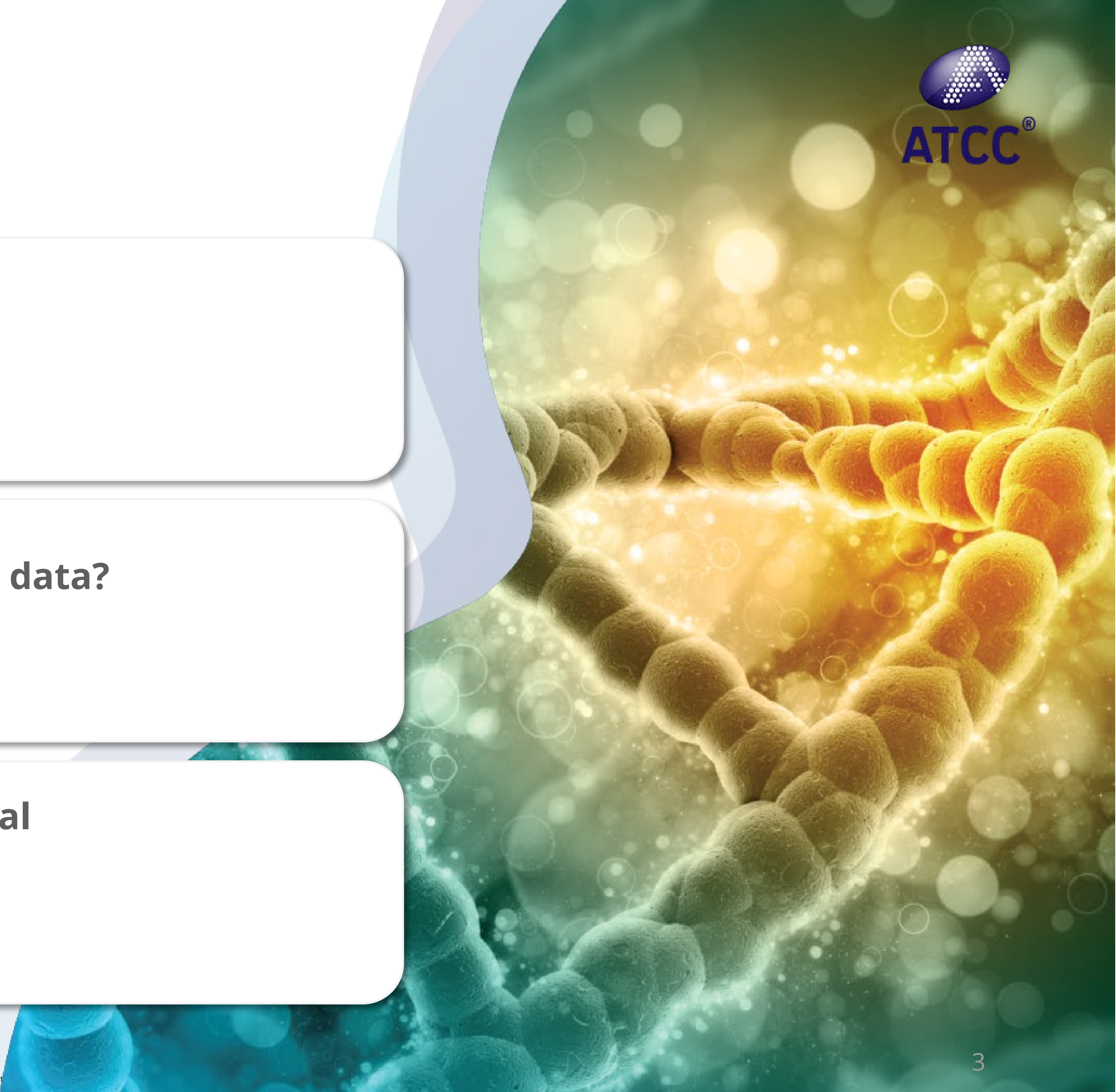
ANAB ACCREDITED ISO 17034 REFERENCE MATERIAL PRODUCER

Agenda

1 **The discovery loop**

2 **How trustworthy is that data?**
- Data provenance
- Examples of problems

3 **The ATCC® Genome Portal**
- Overview
- How to access data
- Data exploration tools



ATCC® Genome Portal

Download whole-genome sequences and annotations of ATCC® materials

Search for nucleotide sequences or genes within genomes

View genome assembly metadata and quality metrics

Learn more about the Genome Portal at www.atcc.org/genomeportal

Access the Genome Portal at genomes.atcc.org

5,500 Authenticated Microbial Reference Genomes
4,205 bacteria and archaea
439 viruses
352 fungi
4 protists

>250+ new genomes released every quarter!

Full data provenance

Illumina and Oxford Nanopore Technologies data used for most assemblies

Standardized
- Laboratory quality metrics
- Bioinformatics quality metrics
- De novo genome assembly pipelines
- Genome annotations

Powered by ONE CODEX

The discovery loop

How we generally do everything

The (improved) discovery loop

Have an idea

Plan experiments

Find data & materials

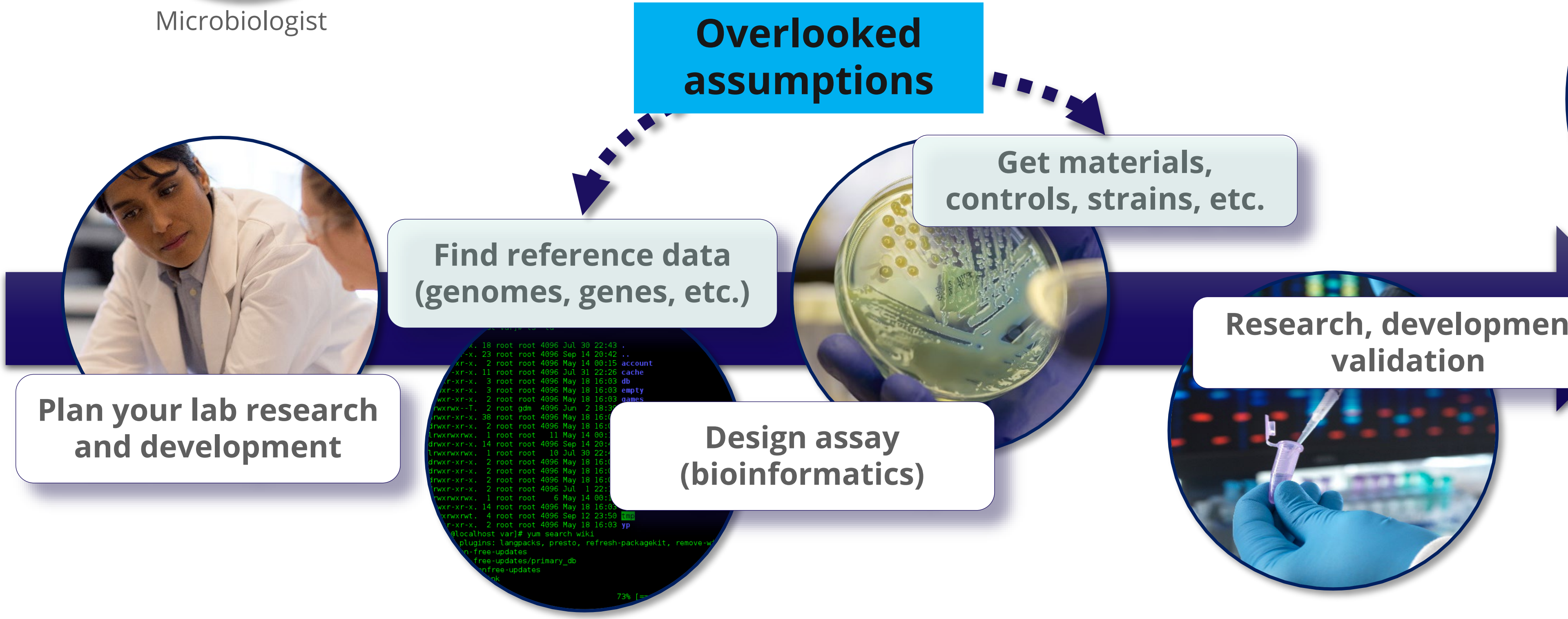
Do experiment

Analyze results

A common scenario…

Microbiologist

“I need to design a new assay for the detection of antibiotic-resistant bacteria.”



Overlooked assumptions

Plan your lab research and development

Find reference data (genomes, genes, etc.)

Design assay (bioinformatics)

Get materials, controls, strains, etc.

Research, development, validation

Unexpected results

Why?

The (improved) discovery loop

Have an idea

Use trusted data resources to discover new materials

Plan experiments

Use authenticated data to improve experimental planning

Find data & materials

- **Use authenticated materials & data**
- **Verify the source & history of external data**
- **Know the risks of using unverified data**

Do experiment

Analyze results

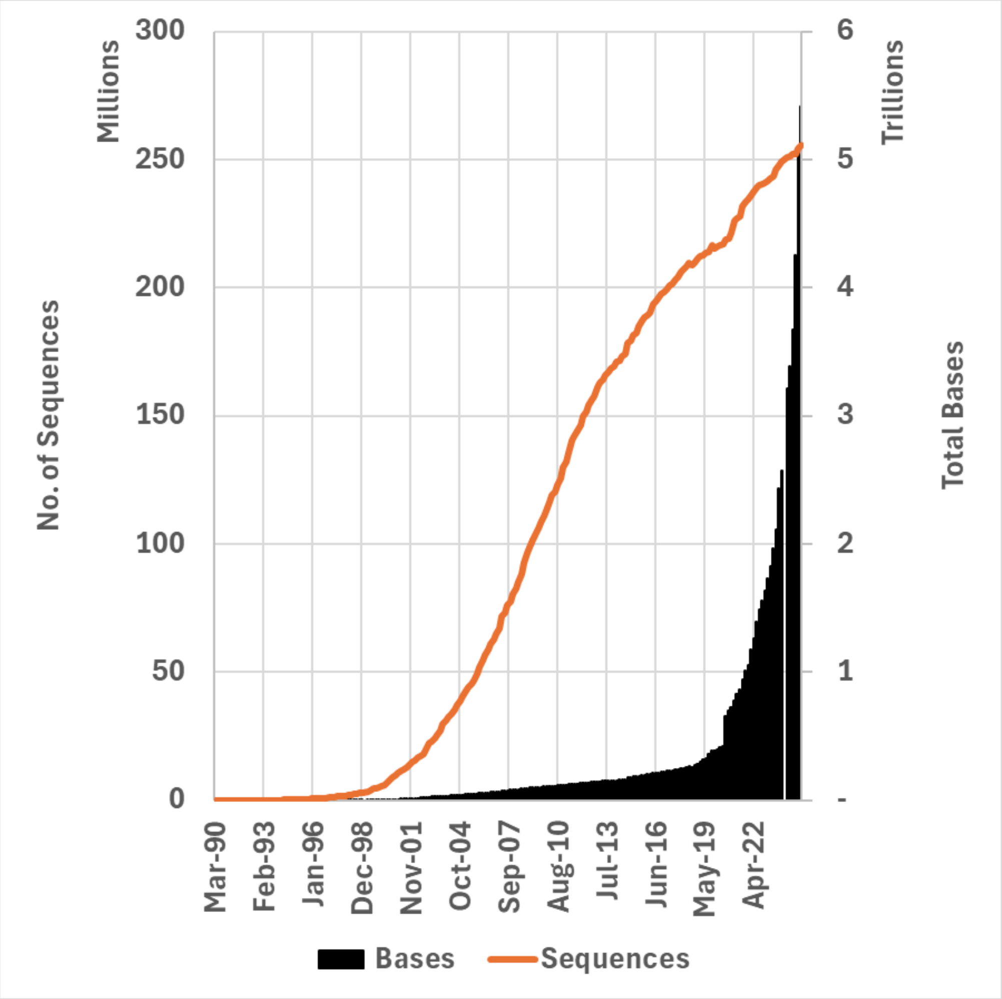
How trustworthy is that data?

A few stories to remember…

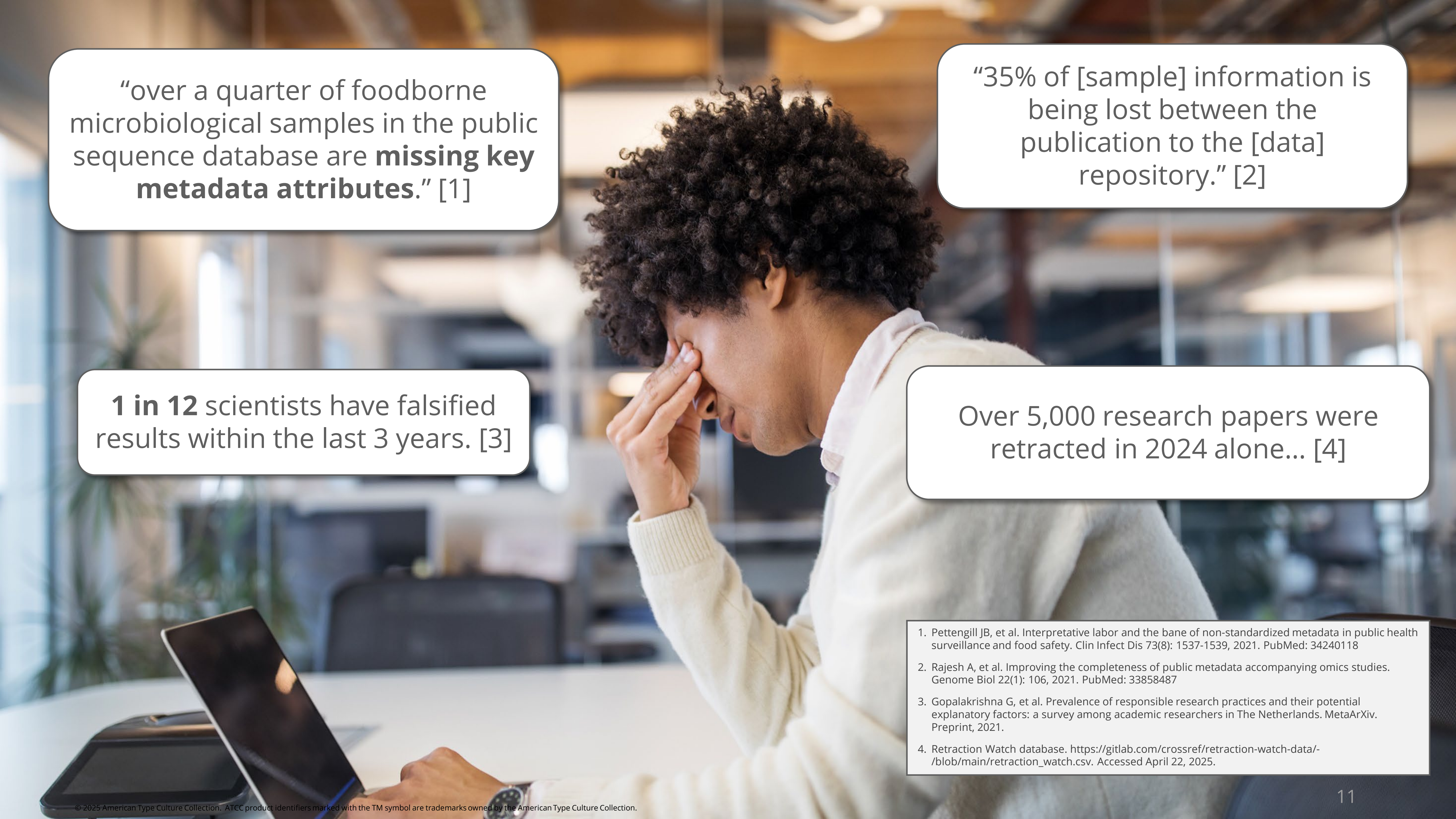
Authenticated, traceable, and reproducible?

Where do researchers turn to for “reference” genomes?

- NCBI - the *de facto* standard
    - From 1982 to the present, the number of bases in GenBank has **doubled approximately every 18 months.**
- Data submitted by thousands of labs, many with their own
    - laboratory protocols,
    - bioinformatics pipelines,
    - metadata curation preferences
- Very little human curation, mostly automated
- Highly variable quality
- Content is never retrospectively updated if methods or standards change
- ***NEVER*** authenticated by ATCC



National Center for Biotechnology Information. GenBank Statistics. NCBI. Published April 22, 2025. Accessed April 22, 2025. https://www.ncbi.nlm.nih.gov/genbank/statistics/



"over a quarter of foodborne microbiological samples in the public sequence database are **missing key metadata attributes**." [1]

"35% of [sample] information is being lost between the publication to the [data] repository." [2]

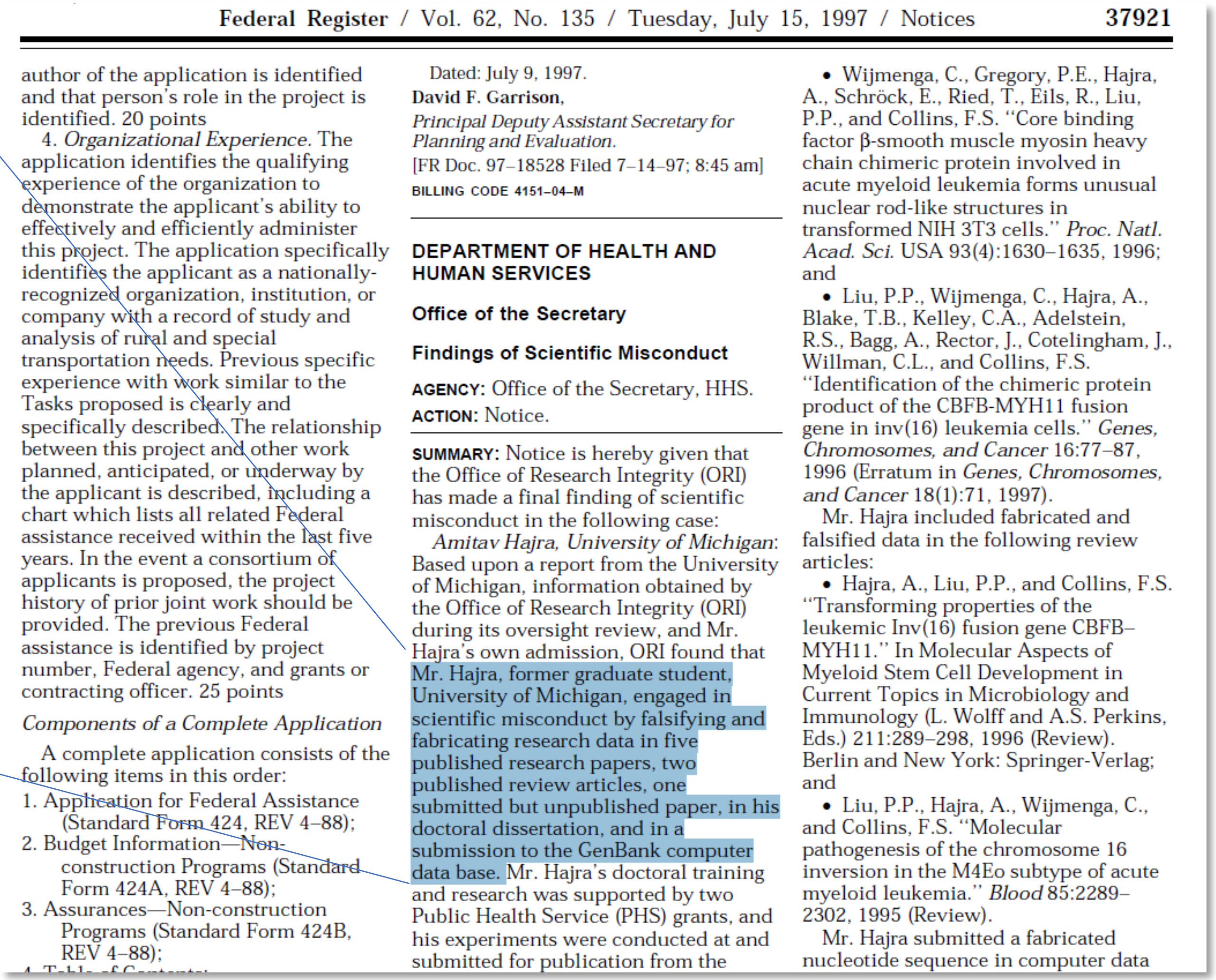
1 in 12 scientists have falsified results within the last 3 years. [3]

Over 5,000 research papers were retracted in 2024 alone… [4]

1. Pettengill JB, et al. Interpretative labor and the bane of non-standardized metadata in public health surveillance and food safety. Clin Infect Dis 73(8): 1537-1539, 2021. PubMed: 34240118
2. Rajesh A, et al. Improving the completeness of public metadata accompanying omics studies. Genome Biol 22(1): 106, 2021. PubMed: 33858487
3. Gopalakrishna G, et al. Prevalence of responsible research practices and their potential explanatory factors: a survey among academic researchers in The Netherlands. MetaArXiv. Preprint, 2021.
4. Retraction Watch database. https://gitlab.com/crossref/retraction-watch-data/-/blob/main/retraction_watch.csv. Accessed April 22, 2025.

Falsified data was deposited in GenBank as early as 1995

…graduate student "engaged in scientific misconduct by falsifying and fabricating research data in **five published research papers, two published review articles, one submitted but unpublished paper, in his doctoral dissertation**, and in a **submission to the GenBank computer data base.**" – The Federal Register, v62, n135 (1997)



Federal Register / Vol. 62, No. 135 / Tuesday, July 15, 1997 / Notices **37921**

author of the application is identified and that person's role in the project is identified. 20 points

4. *Organizational Experience.* The application identifies the qualifying experience of the organization to demonstrate the applicant's ability to effectively and efficiently administer this project. The application specifically identifies the applicant as a nationally-recognized organization, institution, or company with a record of study and analysis of rural and special transportation needs. Previous specific experience with work similar to the Tasks proposed is clearly and specifically described. The relationship between this project and other work planned, anticipated, or underway by the applicant is described, including a chart which lists all related Federal assistance received within the last five years. In the event a consortium of applicants is proposed, the project history of prior joint work should be provided. The previous Federal assistance is identified by project number, Federal agency, and grants or contracting officer. 25 points

Components of a Complete Application

A complete application consists of the following items in this order:

1. Application for Federal Assistance (Standard Form 424, REV 4–88);
2. Budget Information—Non-construction Programs (Standard Form 424A, REV 4–88);
3. Assurances—Non-construction Programs (Standard Form 424B, REV 4–88);

Dated: July 9, 1997.

David F. Garrison,

Principal Deputy Assistant Secretary for Planning and Evaluation.

[FR Doc. 97–18528 Filed 7–14–97; 8:45 am]

BILLING CODE 4151–04–M

DEPARTMENT OF HEALTH AND HUMAN SERVICES

Office of the Secretary

Findings of Scientific Misconduct

AGENCY: Office of the Secretary, HHS.

ACTION: Notice.

SUMMARY: Notice is hereby given that the Office of Research Integrity (ORI) has made a final finding of scientific misconduct in the following case:

Amitav Hajra, University of Michigan: Based upon a report from the University of Michigan, information obtained by the Office of Research Integrity (ORI) during its oversight review, and Mr. Hajra's own admission, ORI found that Mr. Hajra, former graduate student, University of Michigan, engaged in scientific misconduct by falsifying and fabricating research data in five published research papers, two published review articles, one submitted but unpublished paper, in his doctoral dissertation, and in a submission to the GenBank computer data base. Mr. Hajra's doctoral training and research was supported by two Public Health Service (PHS) grants, and his experiments were conducted at and submitted for publication from the

- Wijmenga, C., Gregory, P.E., Hajra, A., Schröck, E., Ried, T., Eils, R., Liu, P.P., and Collins, F.S. "Core binding factor β-smooth muscle myosin heavy chain chimeric protein involved in acute myeloid leukemia forms unusual nuclear rod-like structures in transformed NIH 3T3 cells." *Proc. Natl. Acad. Sci.* USA 93(4):1630–1635, 1996; and
- Liu, P.P., Wijmenga, C., Hajra, A., Blake, T.B., Kelley, C.A., Adelstein, R.S., Bagg, A., Rector, J., Cotelingham, J., Willman, C.L., and Collins, F.S. "Identification of the chimeric protein product of the CBFB-MYH11 fusion gene in inv(16) leukemia cells." *Genes, Chromosomes, and Cancer* 16:77–87, 1996 (Erratum in *Genes, Chromosomes, and Cancer* 18(1):71, 1997).

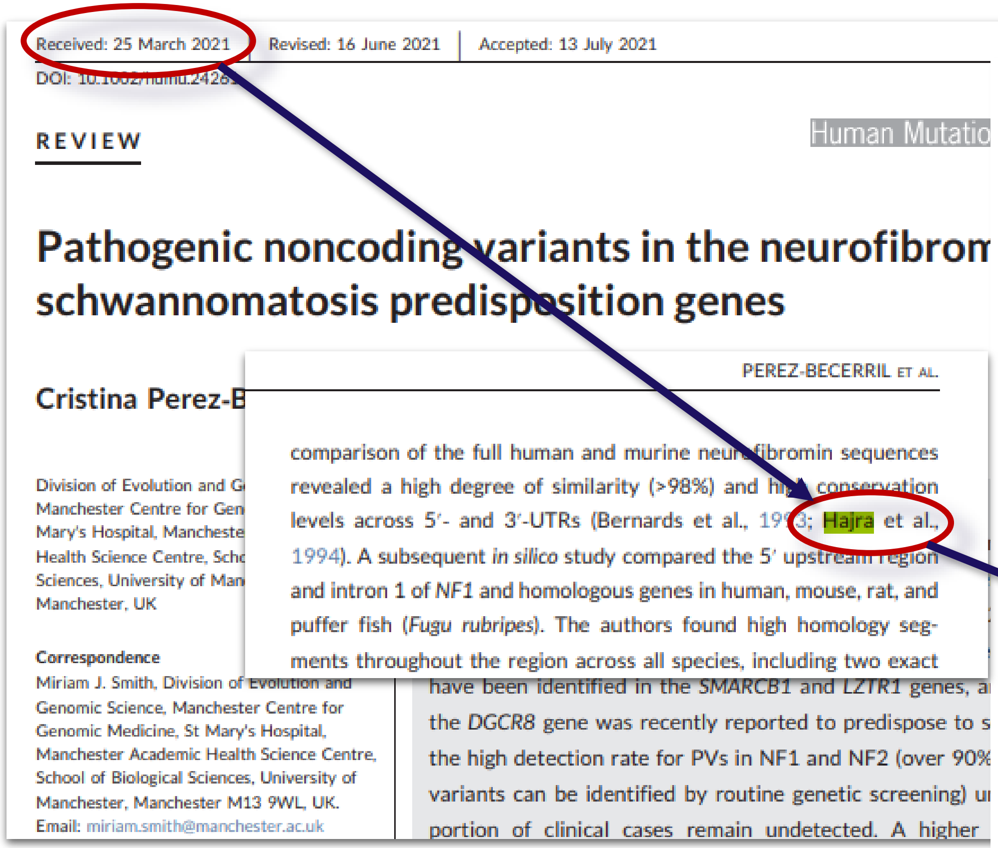
Mr. Hajra included fabricated and falsified data in the following review articles:

- Hajra, A., Liu, P.P., and Collins, F.S. "Transforming properties of the leukemic Inv(16) fusion gene CBFB–MYH11." In Molecular Aspects of Myeloid Stem Cell Development in Current Topics in Microbiology and Immunology (L. Wolff and A.S. Perkins, Eds.) 211:289–298, 1996 (Review). Berlin and New York: Springer-Verlag; and
- Liu, P.P., Hajra, A., Wijmenga, C., and Collins, F.S. "Molecular pathogenesis of the chromosome 16 inversion in the M4Eo subtype of acute myeloid leukemia." *Blood* 85:2289–2302, 1995 (Review).

Mr. Hajra submitted a fabricated nucleotide sequence in computer data

Office of the Secretary, Department of Health and Human Services. Findings of Scientific Misconduct. Federal Register 62(135): 37921, 1997.

30 years later, it's still being cited…



Received: 25 March 2021 | Revised: 16 June 2021 | Accepted: 13 July 2021

DOI: 10.1002/humu.2426

REVIEW

Human Mutatio

Pathogenic noncoding variants in the neurofibrom schwannomatosis predisposition genes

Cristina Perez-B

Division of Evolution and G
Manchester Centre for Gen
Mary's Hospital, Mancheste
Health Science Centre, Scho
Sciences, University of Man
Manchester, UK

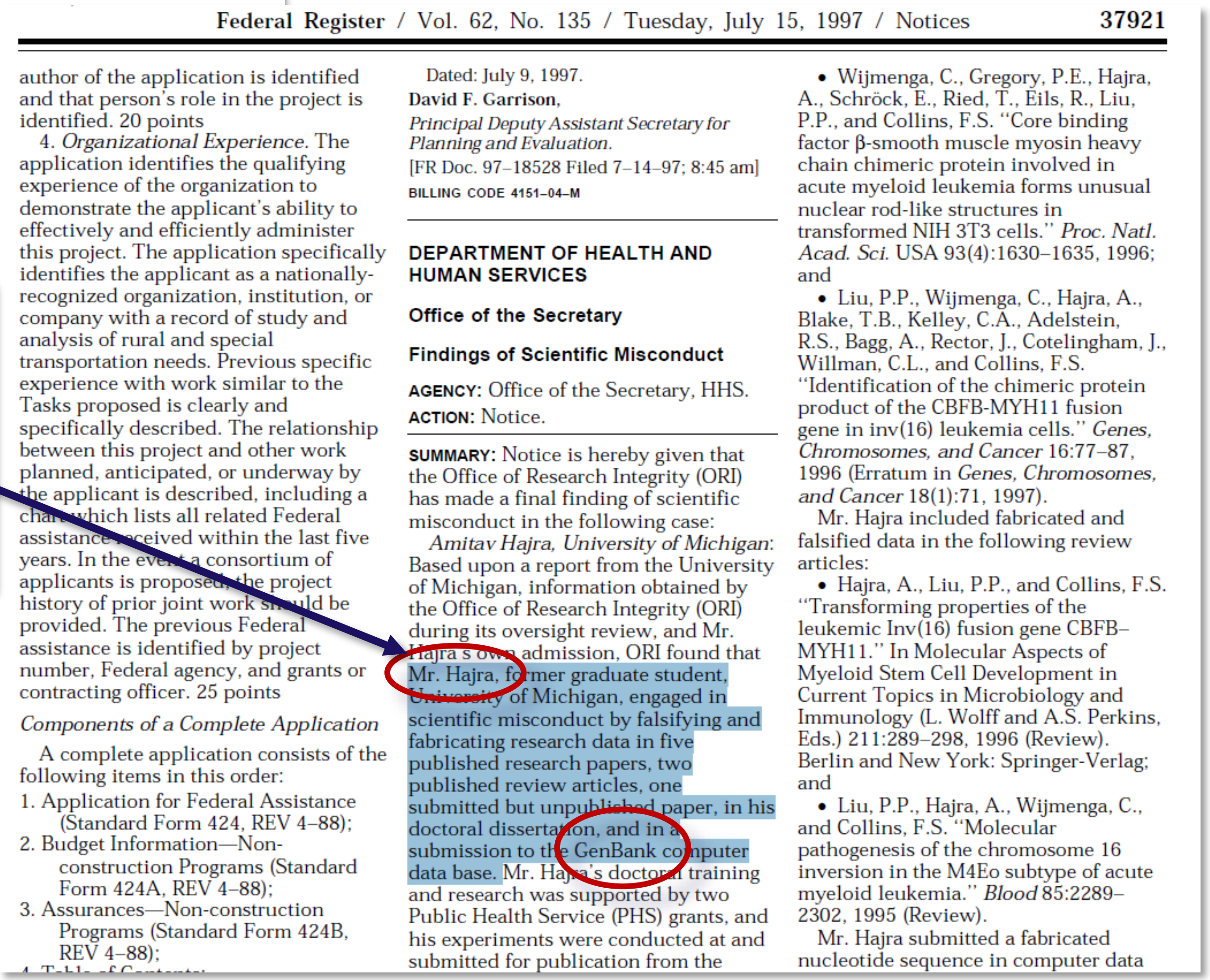
Correspondence
Miriam J. Smith, Division of Evolution and
Genomic Science, Manchester Centre for
Genomic Medicine, St Mary's Hospital,
Manchester Academic Health Science Centre,
School of Biological Sciences, University of
Manchester, Manchester M13 9WL, UK.
Email: miriam.smith@manchester.ac.uk

have been identified in the *SMARCB1* and *LZTR1* genes, a
the *DGCR8* gene was recently reported to predispose to s
the high detection rate for PVs in NF1 and NF2 (over 90%
variants can be identified by routine genetic screening) u
portion of clinical cases remain undetected. A higher

PEREZ-BECERRIL ET AL.

comparison of the full human and murine neurofibromin sequences revealed a high degree of similarity (>98%) and high conservation levels across 5′- and 3′-UTRs (Bernards et al., 1993; Hajra et al., 1994). A subsequent *in silico* study compared the 5′ upstream region and intron 1 of *NF1* and homologous genes in human, mouse, rat, and puffer fish (*Fugu rubripes*). The authors found high homology segments throughout the region across all species, including two exact

Perez-Becerril C, Evans DG, Smith MJ. Pathogenic noncoding variants in the neurofibromatosis and schwannomatosis predisposition genes. Hum Mutat 42(10):1187-1207, 2021. PubMed: 34273915



Federal Register / Vol. 62, No. 135 / Tuesday, July 15, 1997 / Notices 37921

author of the application is identified and that person's role in the project is identified. 20 points

4. *Organizational Experience.* The application identifies the qualifying experience of the organization to demonstrate the applicant's ability to effectively and efficiently administer this project. The application specifically identifies the applicant as a nationally-recognized organization, institution, or company with a record of study and analysis of rural and special transportation needs. Previous specific experience with work similar to the Tasks proposed is clearly and specifically described. The relationship between this project and other work planned, anticipated, or underway by the applicant is described, including a chart which lists all related Federal assistance received within the last five years. In the event a consortium of applicants is proposed, the project history of prior joint work should be provided. The previous Federal assistance is identified by project number, Federal agency, and grants or contracting officer. 25 points

Components of a Complete Application

A complete application consists of the following items in this order:

1. Application for Federal Assistance (Standard Form 424, REV 4–88);
2. Budget Information—Non-construction Programs (Standard Form 424A, REV 4–88);
3. Assurances—Non-construction Programs (Standard Form 424B, REV 4–88);

Dated: July 9, 1997.

David F. Garrison,
Principal Deputy Assistant Secretary for Planning and Evaluation.

[FR Doc. 97–18528 Filed 7–14–97; 8:45 am]

BILLING CODE 4151–04–M

DEPARTMENT OF HEALTH AND HUMAN SERVICES

Office of the Secretary

Findings of Scientific Misconduct

AGENCY: Office of the Secretary, HHS.

ACTION: Notice.

SUMMARY: Notice is hereby given that the Office of Research Integrity (ORI) has made a final finding of scientific misconduct in the following case:

Amitav Hajra, University of Michigan: Based upon a report from the University of Michigan, information obtained by the Office of Research Integrity (ORI) during its oversight review, and Mr. Hajra's own admission, ORI found that Mr. Hajra, former graduate student, University of Michigan, engaged in scientific misconduct by falsifying and fabricating research data in five published research papers, two published review articles, one submitted but unpublished paper, in his doctoral dissertation, and in a submission to the GenBank computer data base. Mr. Hajra's doctoral training and research was supported by two Public Health Service (PHS) grants, and his experiments were conducted at and submitted for publication from the

- Wijmenga, C., Gregory, P.E., Hajra, A., Schröck, E., Ried, T., Eils, R., Liu, P.P., and Collins, F.S. "Core binding factor β-smooth muscle myosin heavy chain chimeric protein involved in acute myeloid leukemia forms unusual nuclear rod-like structures in transformed NIH 3T3 cells." *Proc. Natl. Acad. Sci.* USA 93(4):1630–1635, 1996; and
- Liu, P.P., Wijmenga, C., Hajra, A., Blake, T.B., Kelley, C.A., Adelstein, R.S., Bagg, A., Rector, J., Cotelingham, J., Willman, C.L., and Collins, F.S. "Identification of the chimeric protein product of the CBFB-MYH11 fusion gene in inv(16) leukemia cells." *Genes, Chromosomes, and Cancer* 16:77–87, 1996 (Erratum in *Genes, Chromosomes, and Cancer* 18(1):71, 1997).

Mr. Hajra included fabricated and falsified data in the following review articles:

- Hajra, A., Liu, P.P., and Collins, F.S. "Transforming properties of the leukemic Inv(16) fusion gene CBFB–MYH11." In Molecular Aspects of Myeloid Stem Cell Development in Current Topics in Microbiology and Immunology (L. Wolff and A.S. Perkins, Eds.) 211:289–298, 1996 (Review). Berlin and New York: Springer-Verlag; and
- Liu, P.P., Hajra, A., Wijmenga, C., and Collins, F.S. "Molecular pathogenesis of the chromosome 16 inversion in the M4Eo subtype of acute myeloid leukemia." *Blood* 85:2289–2302, 1995 (Review).

Mr. Hajra submitted a fabricated nucleotide sequence in computer data

Office of the Secretary, Department of Health and Human Services. Findings of Scientific Misconduct. Federal Register 62(135): 37921, 1997.

Falsified sequencing to support a false phylogeny

Biochemical Systematics and Ecology
Volume 96, June 2021, 104263

Scientific data laundering: Chimeric mitogenomes of a sparrowhawk and a nightjar covered-up by forged phylogenies

George Sangster [a], Jolanda A. Luksenburg [b] [c]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.bse.2021.104263

Get rights and content

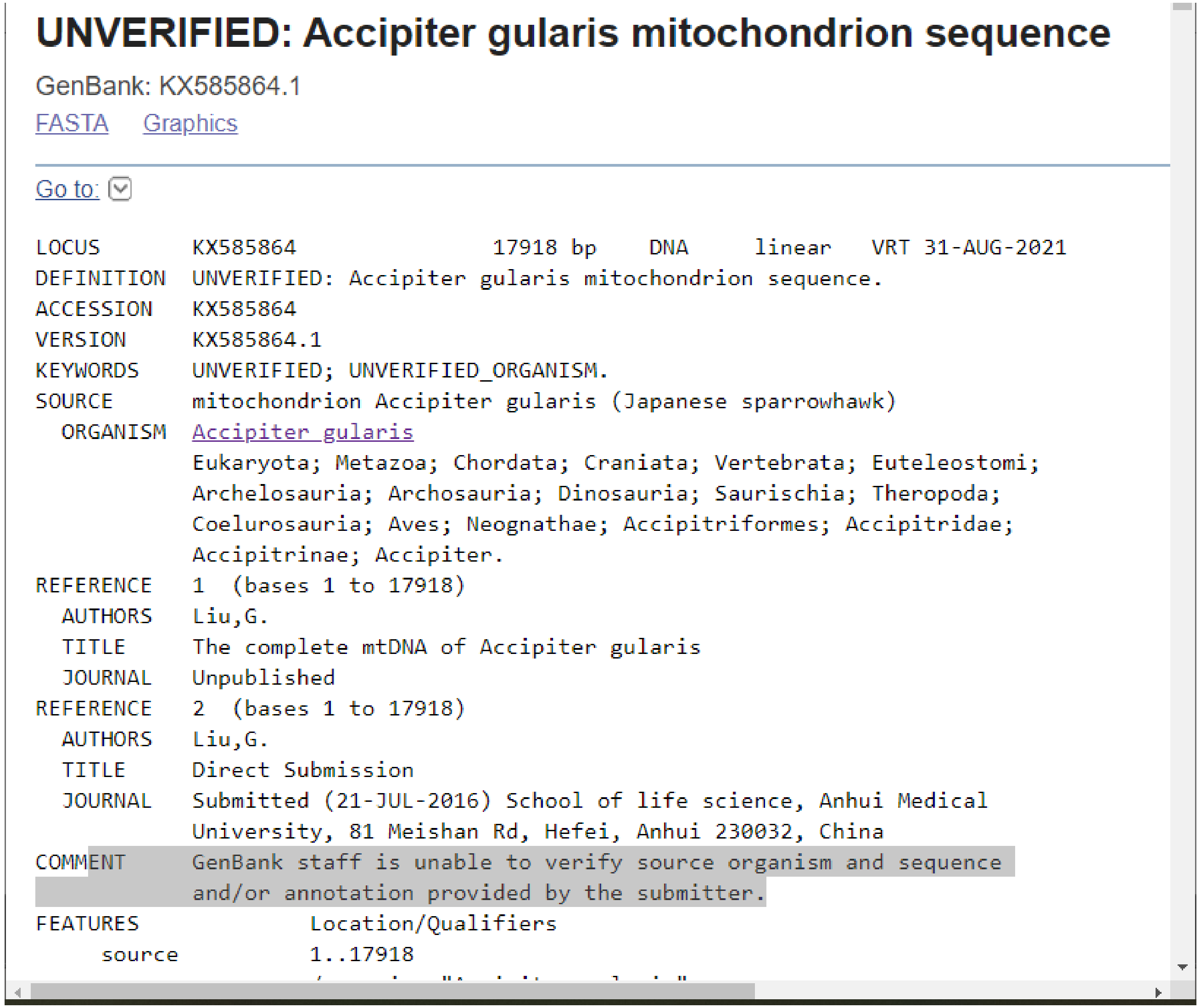
Highlights

- This manuscript presents evidence that a complete mitochondrial genome of a sparrowhawk published by Gang Liu and colleagues in a paper in *Biochemical Systematics and Ecology* in 2017 is not an authentic sequence of this species but represents a chimera of three different species (a sparrowhawk, a buzzard and a dove).
- The manuscript also presents evidence that the authors of the aforementioned paper have fabricated false phylogenies to cover-up this problematic genome, and that of a nightjar previously published by another team, which is also a chimera (of two owls). To our knowledge this is the first known case of scientific fraud in phylogenetics.

"The evidence indicates that Liu et al. (2017) published phylogenies that were not based on existing data **but were fabricated to reflect preconceived ideas** about phylogenetic relationships." – Sangster & Luksenburg (2021)

Sangster G, Luksenburg JA. Scientific data laundering: Chimeric mitogenomes of a sparrowhawk and a nightjar covered-up by forged phylogenies. Biochem Syst Ecol 96: 104263, 2021.

Unfortunately, the data is still in GenBank…



UNVERIFIED: Accipiter gularis mitochondrion sequence

GenBank: KX585864.1

FASTA Graphics

Go to:

```
LOCUS       KX585864               17918 bp    DNA     linear   VRT 31-AUG-2021
DEFINITION  UNVERIFIED: Accipiter gularis mitochondrion sequence.
ACCESSION   KX585864
VERSION     KX585864.1
KEYWORDS    UNVERIFIED; UNVERIFIED_ORGANISM.
SOURCE      mitochondrion Accipiter gularis (Japanese sparrowhawk)
  ORGANISM  Accipiter gularis
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda;
            Coelurosauria; Aves; Neognathae; Accipitriformes; Accipitridae;
            Accipitrinae; Accipiter.
REFERENCE   1  (bases 1 to 17918)
  AUTHORS   Liu,G.
  TITLE     The complete mtDNA of Accipiter gularis
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 17918)
  AUTHORS   Liu,G.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-JUL-2016) School of life science, Anhui Medical
            University, 81 Meishan Rd, Hefei, Anhui 230032, China
COMMENT     GenBank staff is unable to verify source organism and sequence
            and/or annotation provided by the submitter.
FEATURES             Location/Qualifiers
     source          1..17918
```

National Center for Biotechnology Information. Sequence: KX585864.1. NCBI. Published April 22, 2025. Accessed April 22, 2025. https://www.ncbi.nlm.nih.gov/nuccore/KX585864.1

- Labeled as “UNVERIFIED,” but the sequence remains in GenBank
- And can be returned with a BLAST search
- GenBank record “comments” aren’t visible directly in BLAST results

Intentional falsification is rare… but… what about accidents?

Over 2 million "accidents" might be something else…

Mukherjee et al. Standards in Genomic Sciences 2015, 10:18
http://www.standardsingenomics.com/content/10/1/18

Standards in Genomic Sciences

COMMENTARY Open Access

Large-scale contamination of microbial isolate genomes by Illumina PhiX control

Supratim Mukherjee[1*], Marcel Huntemann[1], Natalia Ivanova[1], Nikos C Kyrpides[1

Abstract

With the rapid growth and development of sequencing technologies, genomes
exploring solutions to some of the world's biggest challenges such as searching
exploration of genomic dark matter. However, progress in sequencing has been
that can occur during template or library preparation, sequencing, imaging or d
screened over 18,000 publicly available microbial isolate genome sequences in t
database and identified more than 1000 genomes that are contaminated with P
during Illumina sequencing runs. Approximately 10% of these genomes have be
contaminated genomes were sequenced under the Human Microbiome Project
contamination from various sources and are usually eliminated during downstre
of PhiX contaminated genomes indicates a lapse in either the application or eff
measures. The presence of PhiX contamination in several publicly available isola
errors when such data are used in comparative genomics analyses. Such contam
far-reaching consequences in the form of erroneous data interpretation and ana
measures to proofread raw sequences before releasing them to the broader scie

Keywords: Next-generation sequencing, PhiX, Contamination, Comparative gen

Background

The ability to produce large numbers of high-quality, low-cost reads has revolutionized the field of microbiology [1-3]. Starting from a meager 1575 registered projects in September 2005, there has been a steady increase in the number of sequencing projects according to the Genomes OnLine Database [4]. As of November 17th 2014, there were 41,553 bacterial and archaeal isolate genome sequencing projects reported in GOLD [4,5]. This explosion of genome sequencing projects especially during the last 5 years has been largely catalyzed by the development of several next-generation sequencing platforms offering rapid and accurate genome information at a low cost. Among the different NGS technologies available commercially, the sequencing by

Despite its high
platform does come
need to be address
One such challeng
used as a quality ar
runs. PhiX is an i
with a single-strand
5386 nucleotides ar
sequenced by Fred
defined genome se
used as a control f
majority of its librar
using PhiX at a low
raised up to 40% f
on the concentratio

Steinegger and Salzberg Genome Biology (2020) 21:115
https://doi.org/10.1186/s13059-020-02023-1

Genome Biology

METHOD Open Access

Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank

Martin Steinegger[1,2,3*] and Steven L. Salzberg[2,4,5]

*Correspondence:
martin.steinegger@snu.ac.kr
[1] School of Biological Sciences, Seoul National University, Seoul, 08826, South Korea
[2] Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, 21218 Baltimore, Maryland, USA
Full list of author information is available at the end of the article

Abstract

Genomic analyses are sensitive to contamination in public databases caused by incorrectly labeled reference sequences. Here, we describe Conterminator, an efficient method to detect and remove incorrectly labeled sequences by an exhaustive all-against-all sequence comparison. Our analysis reports contamination of 2,161,746, 114,035, and 14,148 sequences in the RefSeq, GenBank, and NR databases, respectively, spanning the whole range from draft to "complete" model organism genomes. Our method scales linearly with input size and can process 3.3 TB in 12 days on a 32-core computer. Conterminator can help ensure the quality of reference databases. Source code (GPLv3): https://github.com/martin-steinegger/conterminator

Keywords: Genomes, Contamination, Software, RefSeq, GenBank

Downloaded from genome.cshlp.org on October 20, 2021 - Published by Cold Spring Harbor Laboratory Press

Research

Human contamination in bacterial genomes has created thousands of spurious proteins

Florian P. Breitwieser,[1] Mihaela Pertea,[1,2] Aleksey V. Zimin,[1,3]

thans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore,
er Science, Whiting School of Engineering, Johns Hopkins University, Baltimore,
cal Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA;
f Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA

lished genomes can cause numerous problems for downstream analyses, partic-
omics projects. Our large-scale scan of complete and draft bacterial and archaeal
als that 2250 genomes are contaminated by human sequence. The contaminant
human repeat regions, which themselves are not adequately represented in the
8. The absence of the sequences from the human assembly offers a likely expla-
blies. In some cases, the contaminating contigs have been erroneously annotated
ich over time have propagated to create spurious protein "families" across mul-
As a result, 3437 spurious protein entries are currently present in the widely
report here an extensive list of contaminant sequences in bacterial genome as-
hem. We found that nearly all contaminants occurred in small contigs in draft
small contigs from draft genome assemblies may mitigate the issue of contami-
genuine genomic sequences.

article.]

assemblies in the NCBI and UCSC Genome Browser databases were contaminated with the primate-specific *AluY* repeats (Longo et al. 2011). Although validation pipelines have improved substantially since then (Tatusova et al. 2016; Haft et al. 2018), some contaminants still remain, as we describe below. Furthermore, when open reading frames (ORFs) in the contaminated contigs get annotated as protein-coding genes, their protein sequence may be added to other databases. Once in those databases, these spurious proteins may in turn be used in future annotation, leading to the so-called "transitive catastrophe" problem where errors are propagated widely (Karp 1998; Salzberg 2007; Danchin et al. 2018). Indeed, one study found that the percentage of misannotated entries in the NCBI nonredundant (nr) protein collection, which is used for thousands of BLAST searches every day, has been increasing over time (Schnoes et al. 2009).

Contamination of genomic sequences can be particularly problematic for metagenomic studies. For example, if a genome labeled as species X contains fragments of the human genome, then any sample containing human DNA might erroneously be identified as also containing species X. Since human DNA is virtually always present in the environment of sequencing laboratories, human contamination is very common in sequencing experiments of all types. Contamination of laboratory reagents with

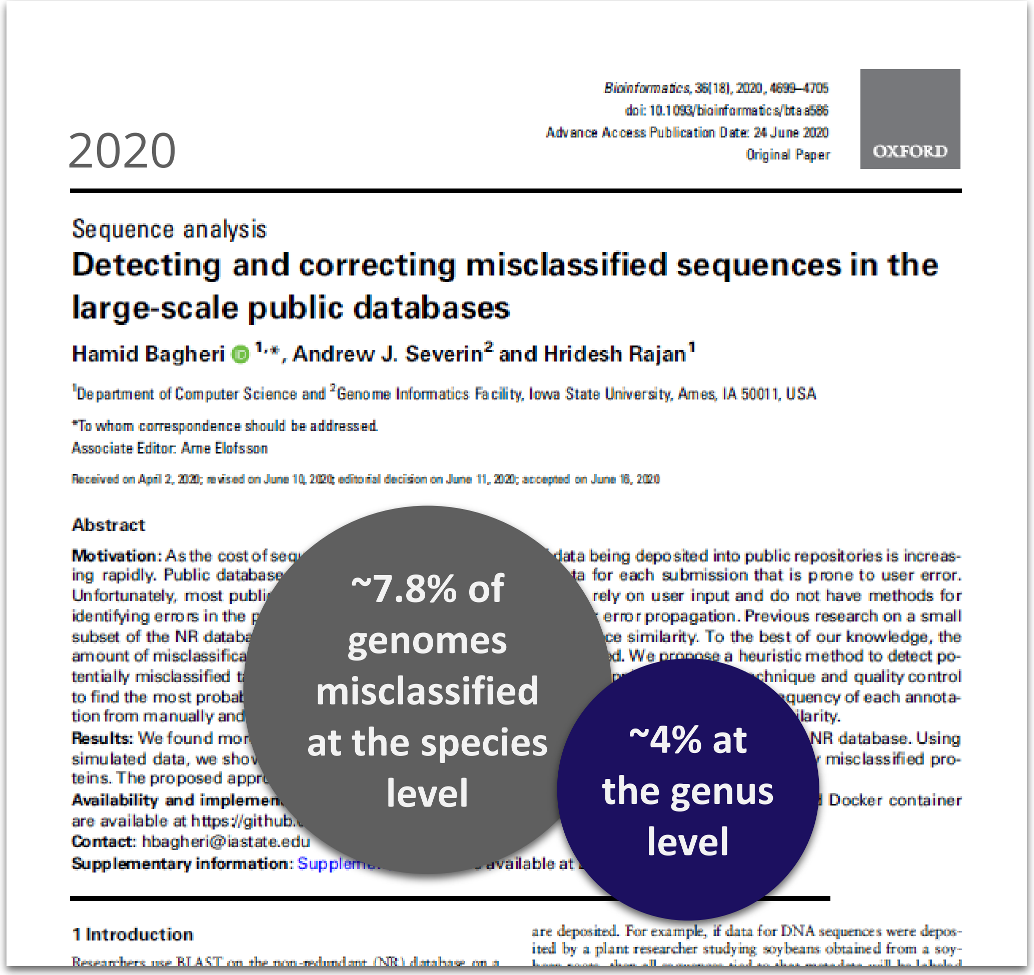
Mukherjee S, Huntemann M, Ivanova N, et al. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. Stand Genomic Sci 10: 18, 2015. PubMed: 26203331

Steinegger M, Salzberg SL. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biol 21(1): 115, 2020. PubMed: 32398145

Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. Human contamination in bacterial genomes has created thousands of spurious proteins. Genome Res 29(6): 954-960, 2019. PubMed: 31064768

Poor quality genomes can result in misclassification

Multiple papers have found widespread misclassification based on sequencing data (examples below)



2020

Bioinformatics, 36(18), 2020, 4699–4705
doi: 10.1093/bioinformatics/btaa586
Advance Access Publication Date: 24 June 2020
Original Paper

OXFORD

Sequence analysis

Detecting and correcting misclassified sequences in the large-scale public databases

Hamid Bagheri [1,*], **Andrew J. Severin** [2] **and Hridesh Rajan** [1]

[1]Department of Computer Science and [2]Genome Informatics Facility, Iowa State University, Ames, IA 50011, USA

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on April 2, 2020; revised on June 10, 2020; editorial decision on June 11, 2020; accepted on June 16, 2020

Abstract

Motivation: As the cost of seq … data being deposited into public repositories is increas-
ing rapidly. Public database … ta for each submission that is prone to user error.
Unfortunately, most publi … rely on user input and do not have methods for
identifying errors in the p … error propagation. Previous research on a small
subset of the NR databa … ce similarity. To the best of our knowledge, the
amount of misclassifica … d. We propose a heuristic method to detect po-
tentially misclassified t … chnique and quality control
to find the most probab … quency of each annota-
tion from manually and … larity.
Results: We found mor … NR database. Using
simulated data, we show … misclassified pro-
teins. The proposed appr …
Availability and implemen … Docker container
are available at https://github. …
Contact: hbagheri@iastate.edu
Supplementary information: Suppleme … available at

1 Introduction

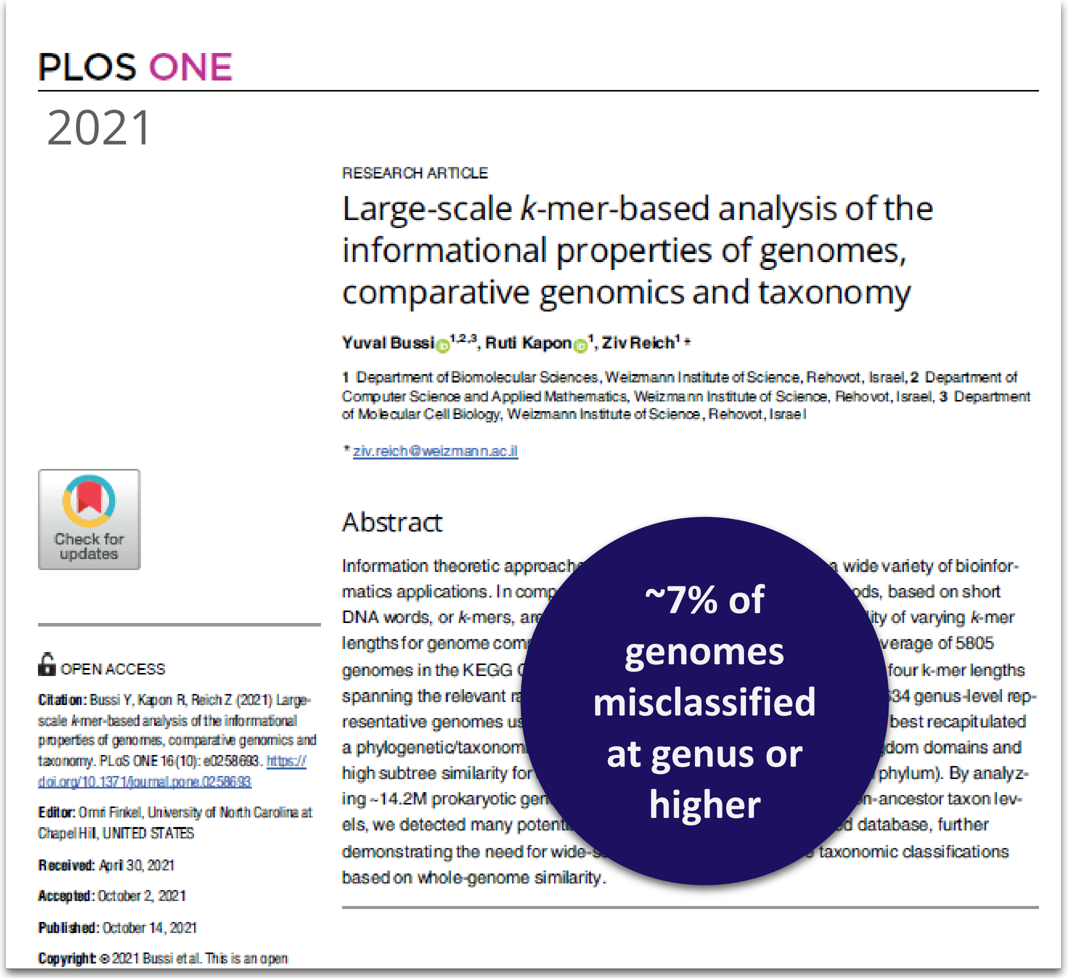
Researchers use BLAST on the non-redundant (NR) database on a

are deposited. For example, if data for DNA sequences were deposited by a plant researcher studying soybeans obtained from a soy-

~7.8% of genomes misclassified at the species level

~4% at the genus level

Bagheri H, Severin AJ, Rajan H. Detecting and correcting misclassified sequences in the large-scale public databases. Bioinformatics 36(18): 4699-4705, 2020. PubMed: 32579213



PLOS ONE

2021

RESEARCH ARTICLE

Large-scale *k*-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy

Yuval Bussi[1,2,3], **Ruti Kapon**[1], **Ziv Reich**[1]*

1 Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot, Israel, **2** Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, **3** Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel

* ziv.reich@weizmann.ac.il

Check for updates

OPEN ACCESS

Citation: Bussi Y, Kapon R, Reich Z (2021) Large-scale *k*-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. PLoS ONE 16(10): e0258693. https://doi.org/10.1371/journal.pone.0258693

Editor: Omri Finkel, University of North Carolina at Chapel Hill, UNITED STATES

Received: April 30, 2021

Accepted: October 2, 2021

Published: October 14, 2021

Abstract

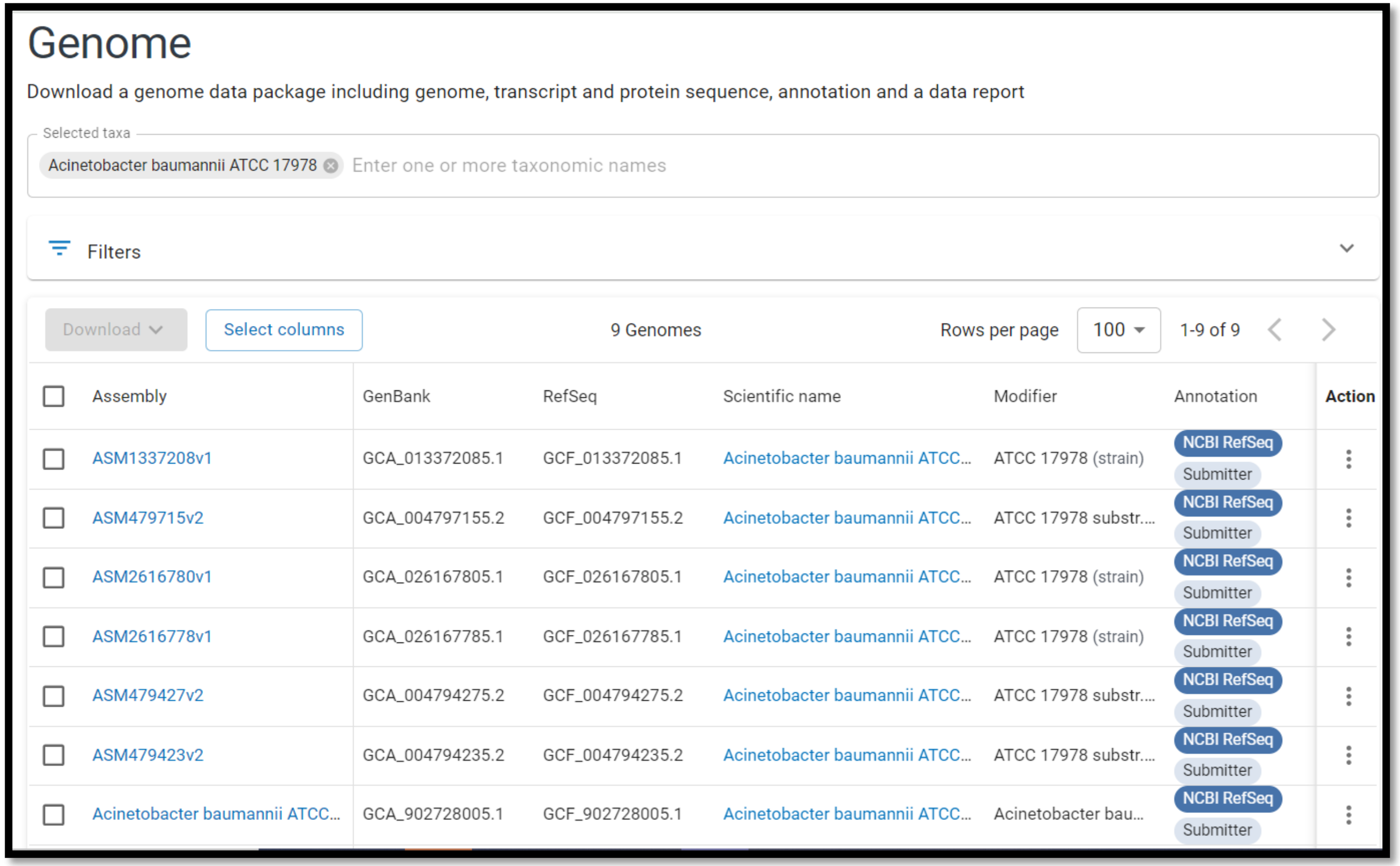
Information theoretic approach … a wide variety of bioinfor-
matics applications. In comp … ods, based on short
DNA words, or *k*-mers, ar … lity of varying *k*-mer
lengths for genome comp … verage of 5805
genomes in the KEGG … four k-mer lengths
spanning the relevant ra … 34 genus-level rep-
resentative genomes us … best recapitulated
a phylogenetic/taxonom … dom domains and
high subtree similarity for … phylum). By analyz-
ing ~14.2M prokaryotic gen … n-ancestor taxon lev-
els, we detected many potent … d database, further
demonstrating the need for wide- … taxonomic classifications
based on whole-genome similarity.

~7% of genomes misclassified at genus or higher

Bussi Y, Kapon R, Reich Z. Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. PLoS One 16(10): e0258693, 2021. PubMed: 34648558

Which reference? 9 and growing…

Acinetobacter baumannii (ATCC® 17978™)



Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa: Acinetobacter baumannii ATCC 17978 — Enter one or more taxonomic names

9 Genomes — Rows per page 100 — 1-9 of 9

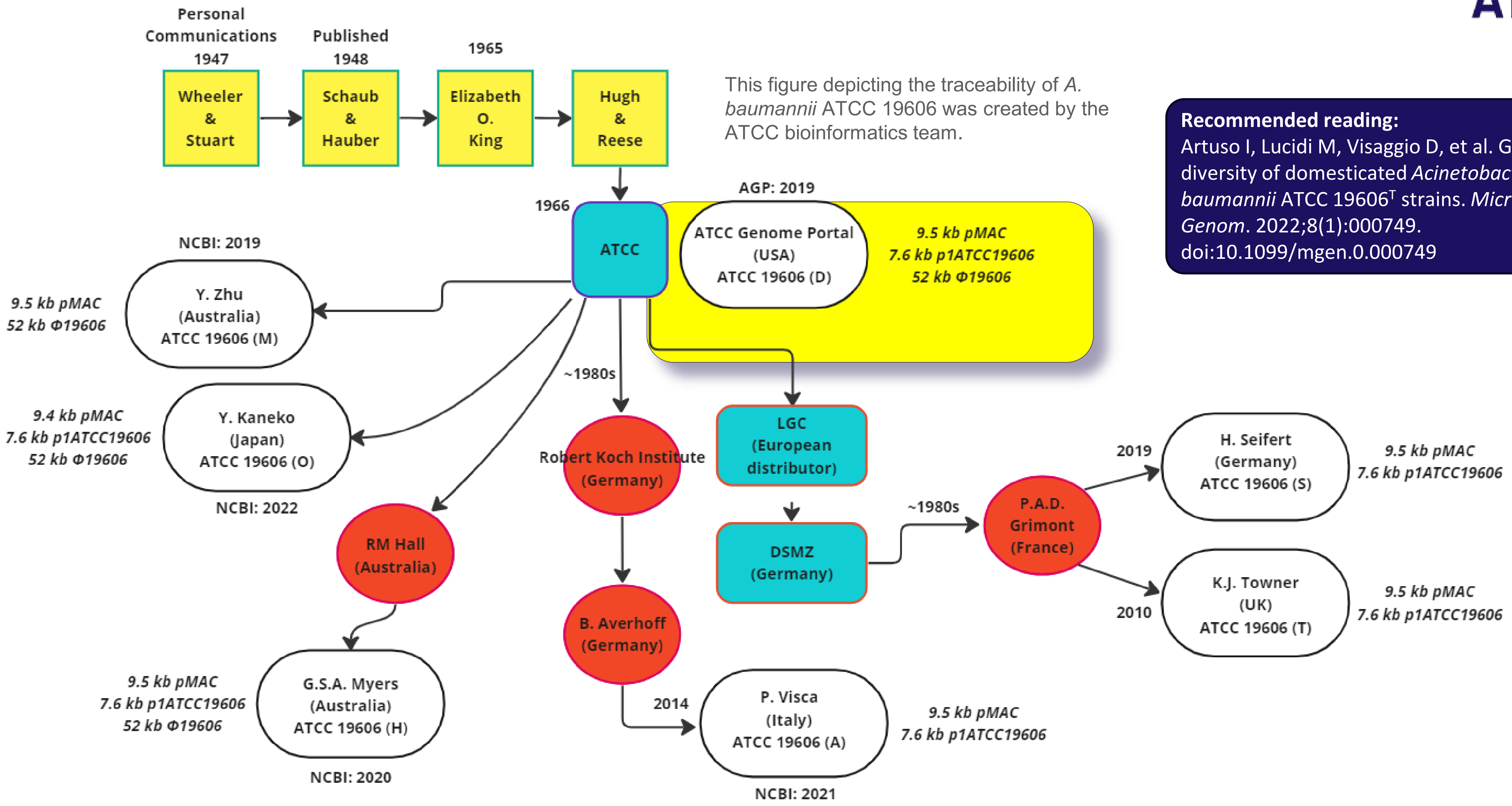
☐	Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action
☐	ASM1337208v1	GCA_013372085.1	GCF_013372085.1	Acinetobacter baumannii ATCC...	ATCC 17978 (strain)	NCBI RefSeq Submitter	⋮
☐	ASM479715v2	GCA_004797155.2	GCF_004797155.2	Acinetobacter baumannii ATCC...	ATCC 17978 substr....	NCBI RefSeq Submitter	⋮
☐	ASM2616780v1	GCA_026167805.1	GCF_026167805.1	Acinetobacter baumannii ATCC...	ATCC 17978 (strain)	NCBI RefSeq Submitter	⋮
☐	ASM2616778v1	GCA_026167785.1	GCF_026167785.1	Acinetobacter baumannii ATCC...	ATCC 17978 (strain)	NCBI RefSeq Submitter	⋮
☐	ASM479427v2	GCA_004794275.2	GCF_004794275.2	Acinetobacter baumannii ATCC...	ATCC 17978 substr....	NCBI RefSeq Submitter	⋮
☐	ASM479423v2	GCA_004794235.2	GCF_004794235.2	Acinetobacter baumannii ATCC...	ATCC 17978 substr....	NCBI RefSeq Submitter	⋮
☐	Acinetobacter baumannii ATCC...	GCA_902728005.1	GCF_902728005.1	Acinetobacter baumannii ATCC...	Acinetobacter bau...	NCBI RefSeq Submitter	⋮

National Center for Biotechnology Information (NCBI). Genome Data Viewer [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [cited 2025 Apr 22]. Available from: https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=400667

- Unverified chain of custody.
- Growth conditions?
- DNA extraction methods?
- DNA sequencing platforms?
- *de novo* assembly methods?

*How do researchers *know* which data set to use for their research?*

Which reference is the “right” reference?



Personal Communications 1947
Wheeler & Stuart

Published 1948
Schaub & Hauber

1965
Elizabeth O. King

Hugh & Reese

1966
ATCC

AGP: 2019
ATCC Genome Portal (USA) ATCC 19606 (D)
9.5 kb pMAC
7.6 kb p1ATCC19606
52 kb Φ19606

NCBI: 2019
Y. Zhu (Australia) ATCC 19606 (M)
9.5 kb pMAC
52 kb Φ19606

Y. Kaneko (Japan) ATCC 19606 (O)
NCBI: 2022
9.4 kb pMAC
7.6 kb p1ATCC19606
52 kb Φ19606

RM Hall (Australia)

G.S.A. Myers (Australia) ATCC 19606 (H)
NCBI: 2020
9.5 kb pMAC
7.6 kb p1ATCC19606
52 kb Φ19606

~1980s
Robert Koch Institute (Germany)

B. Averhoff (Germany)

2014
P. Visca (Italy) ATCC 19606 (A)
NCBI: 2021
9.5 kb pMAC
7.6 kb p1ATCC19606

LGC (European distributor)

DSMZ (Germany)

~1980s
P.A.D. Grimont (France)

2019
H. Seifert (Germany) ATCC 19606 (S)
9.5 kb pMAC
7.6 kb p1ATCC19606

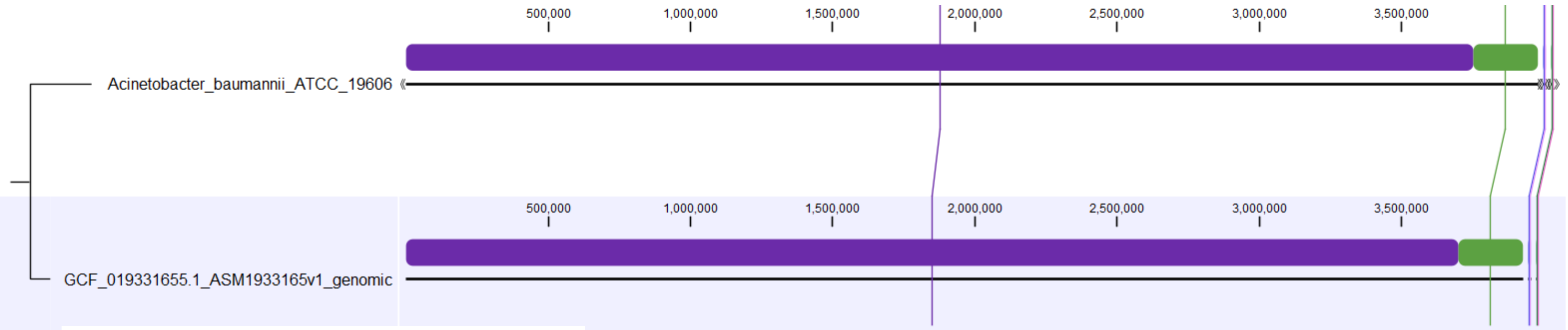
2010
K.J. Towner (UK) ATCC 19606 (T)
9.5 kb pMAC
7.6 kb p1ATCC19606

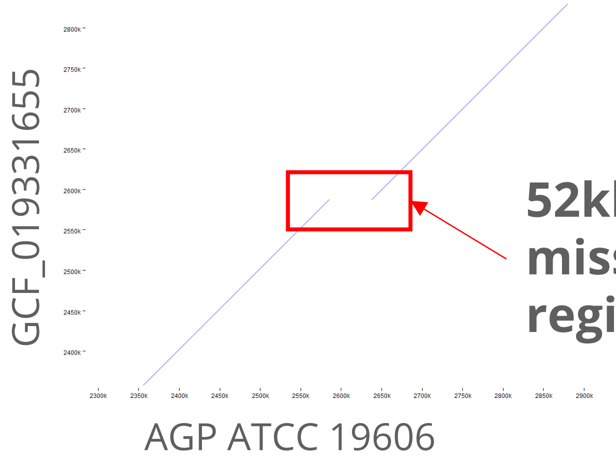
This figure depicting the traceability of *A. baumannii* ATCC 19606 was created by the ATCC bioinformatics team.

Recommended reading:
Artuso I, Lucidi M, Visaggio D, et al. Genome diversity of domesticated *Acinetobacter baumannii* ATCC 19606^T strains. *Microb Genom*. 2022;8(1):000749. doi:10.1099/mgen.0.000749

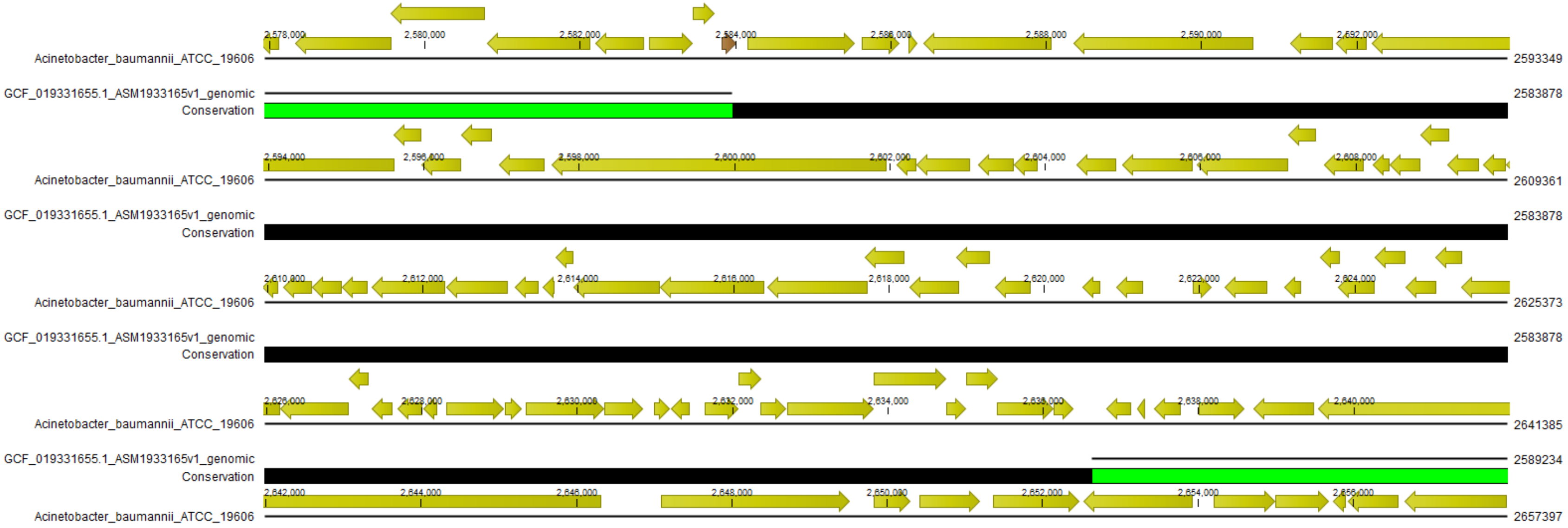
The NCBI “reference genome”

Comparison of ATCC® Genome Portal vs. RefSeq Assemblies





52kb missing region



refseq assembly

- 52kb region missing, which includes 74 genes
- 51 are not found anywhere else in the refseq assembly

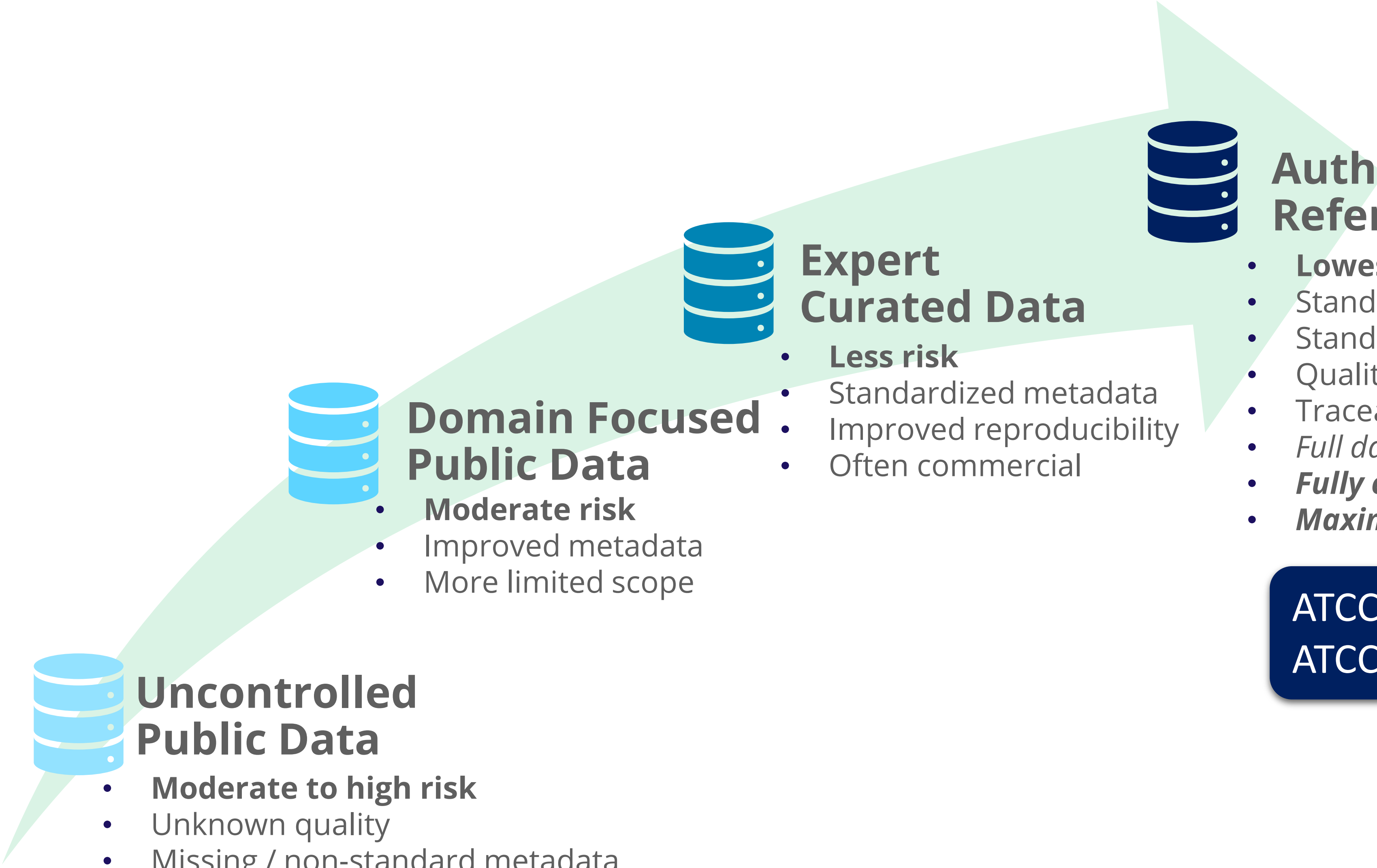
Examples of other reliability issues

- Misclassification of type strains used in FDA-approved probiotic foods.
- Misclassification of control strains used by clinical microbiology labs for a widely used AMR testing platform.
- Different phenotypes for the “same” strains.
- Unknown history or chain-of-custody of materials or data.
- There’s no “track changes” with genome assemblies.
- Accidental mislabeling of files or rows in a table can lead to incorrect links between NGS data and metadata.

The ATCC® Genome Portal

The *only* source for authenticated genome assemblies for ATCC® materials

The arc of database quality in genomics



Uncontrolled Public Data

- **Moderate to high risk**
- Unknown quality
- Missing / non-standard metadata
- Often no data provenance

Domain Focused Public Data

- **Moderate risk**
- Improved metadata
- More limited scope

Expert Curated Data

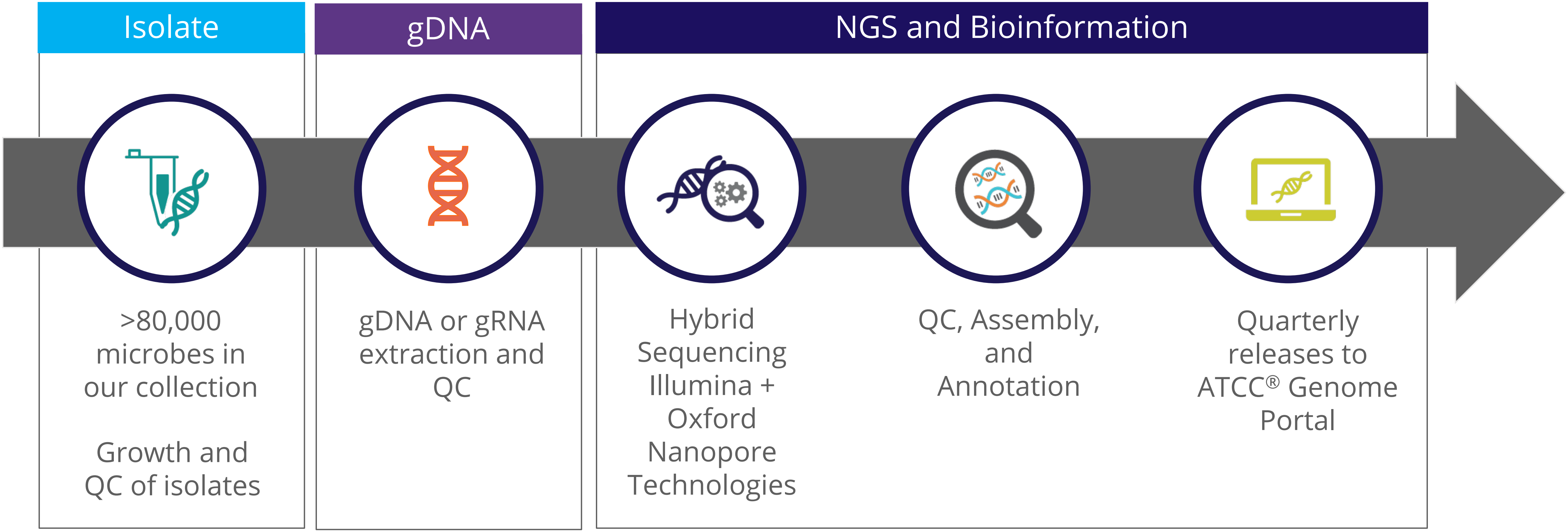
- **Less risk**
- Standardized metadata
- Improved reproducibility
- Often commercial

Authenticated Reference Data

- **Lowest risk**
- Standardized laboratory methods
- Standardized bioinformatics methods
- Quality assurance (ISO 9001)
- Traceable to materials in a biorepository
- *Full data provenance*
- ***Fully curated and authenticated***
- ***Maximum reproducibility***

ATCC® Genome Portal
ATCC® Cell Line Land

Authenticated physical material coupled with reference-quality genome sequences



Isolate

>80,000 microbes in our collection

Growth and QC of isolates

gDNA

gDNA or gRNA extraction and QC

NGS and Bioinformation

Hybrid Sequencing Illumina + Oxford Nanopore Technologies

QC, Assembly, and Annotation

Quarterly releases to ATCC® Genome Portal

- Fully traceable and authenticated to ATCC® materials
- All genome assemblies produced in-house at ATCC® in an ISO-certified laboratory

“Authenticated Genome”

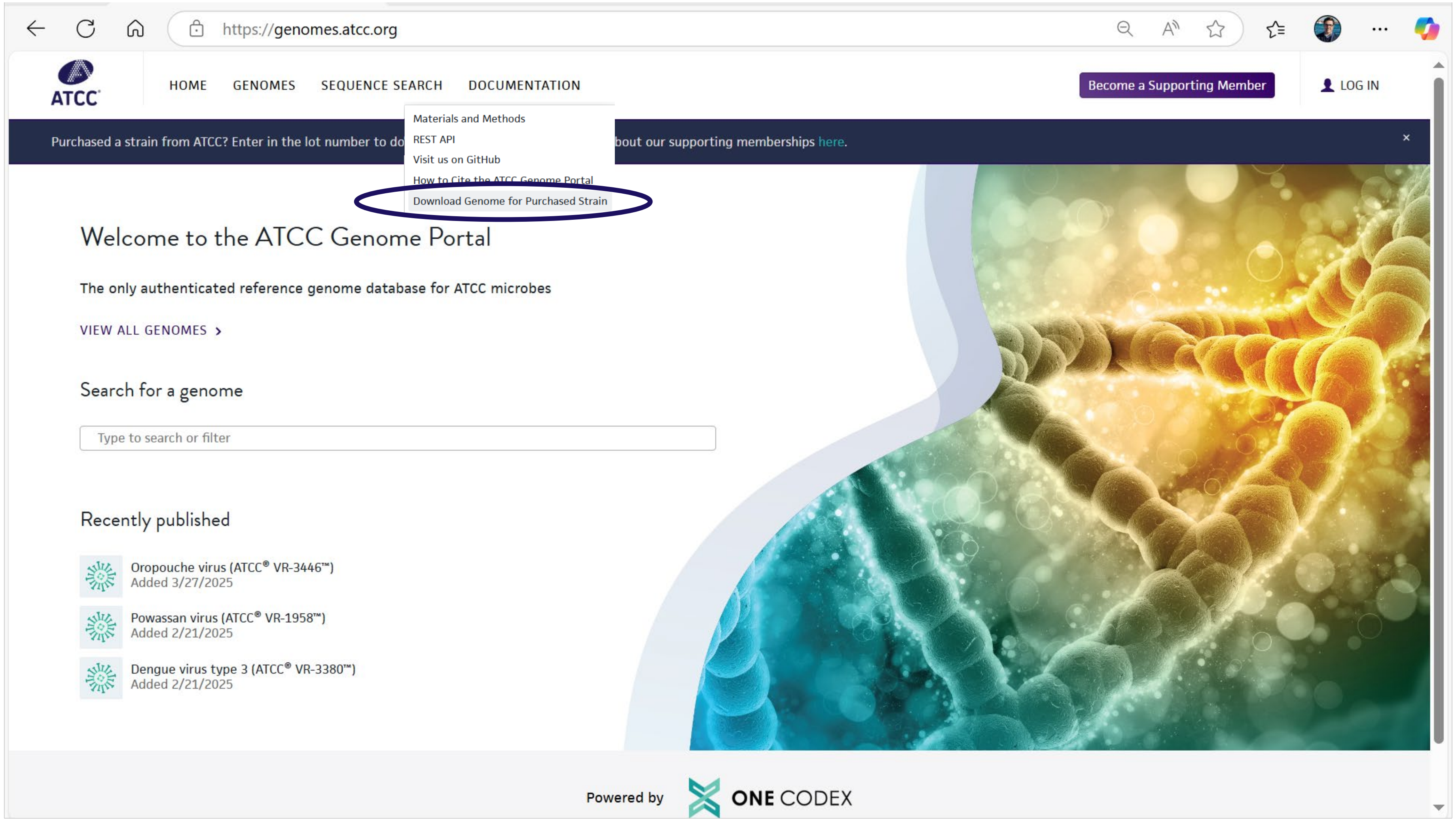
1. ***Traceable** authenticated materials*
2. ***Documented** methods (i.e., ISO)*
3. *Exceeds **standardized** quality metrics*
4. *Full end-to-end **data provenance** (from materials to data)*

The ATCC® Genome Portal

Accessing our data

If customers have purchased the physical product…

Download genome(s) with the lot number



https://genomes.atcc.org

Purchased a strain from ATCC? Enter in the lot number to do … bout our supporting memberships here.

Welcome to the ATCC Genome Portal

The only authenticated reference genome database for ATCC microbes

VIEW ALL GENOMES ›

Search for a genome

Type to search or filter

Recently published

Oropouche virus (ATCC® VR-3446™)
Added 3/27/2025

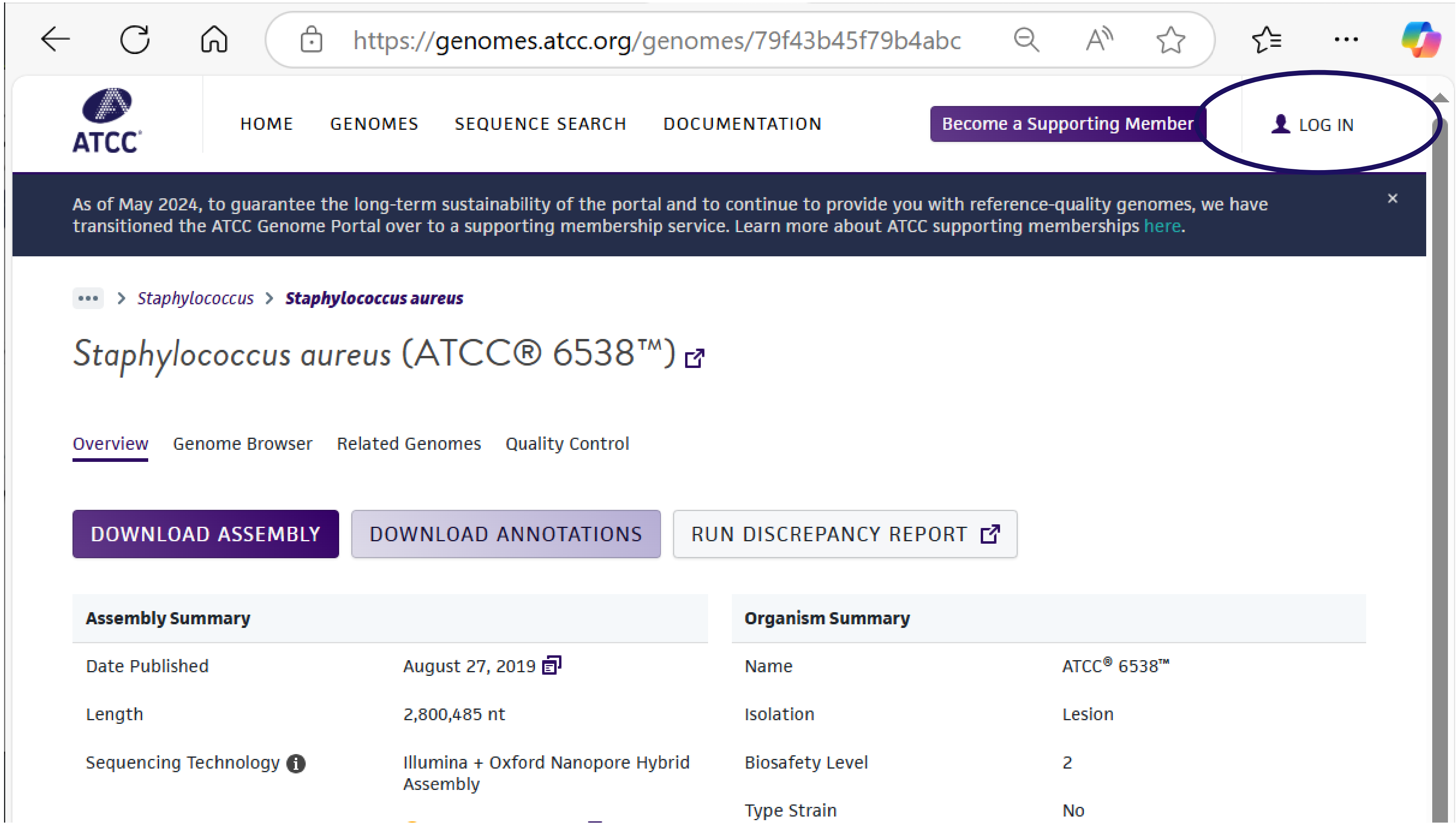
Powassan virus (ATCC® VR-1958™)
Added 2/21/2025

Dengue virus type 3 (ATCC® VR-3380™)
Added 2/21/2025

Powered by ONE CODEX

If customers have purchased the physical product…

Download genome(s) with the lot number



As of May 2024, to guarantee the long-term sustainability of the portal and to continue to provide you with reference-quality genomes, we have transitioned the ATCC Genome Portal over to a supporting membership service. Learn more about ATCC supporting memberships here.

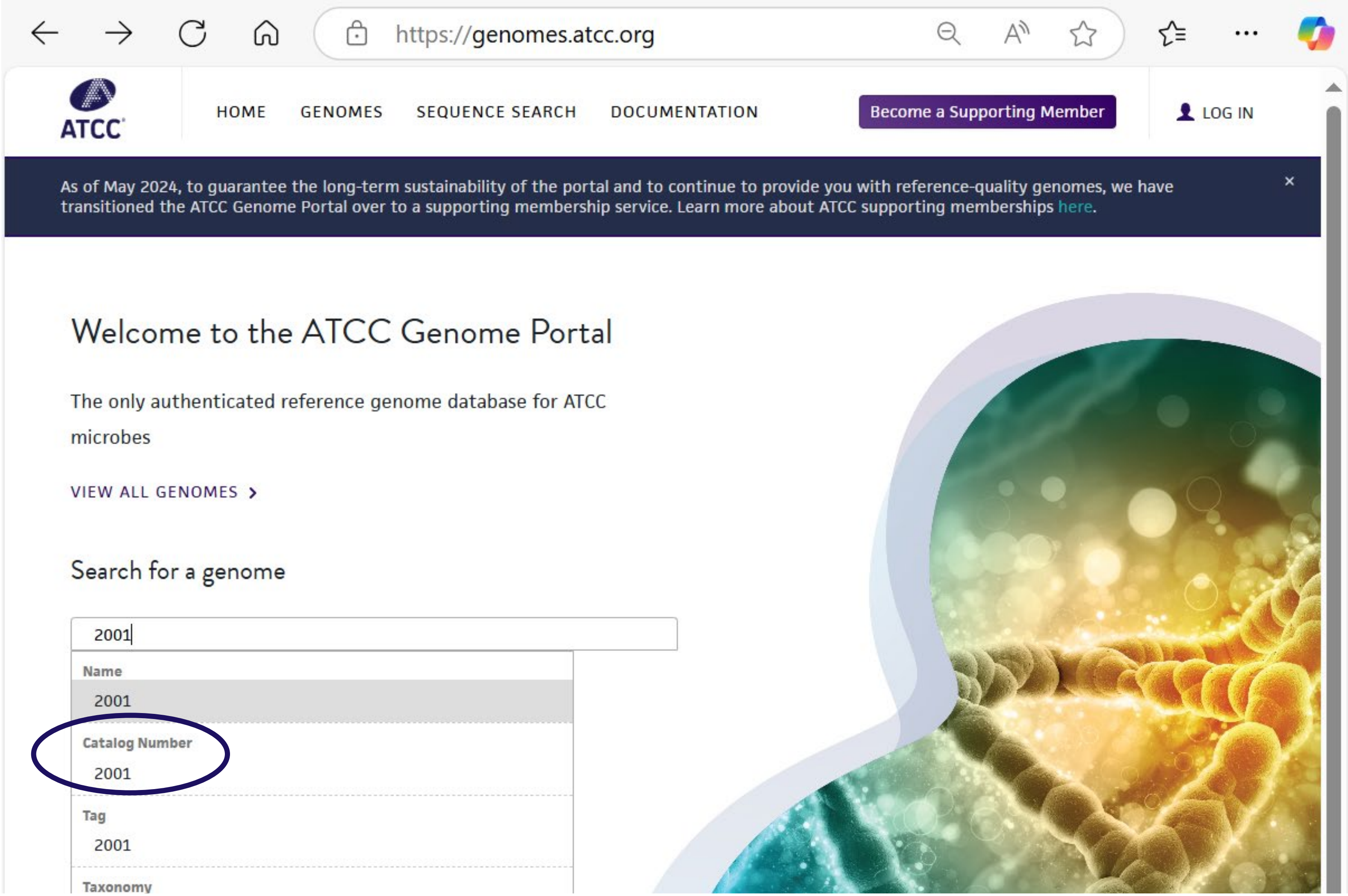
Staphylococcus aureus (ATCC® 6538™)

Assembly Summary	
Date Published	August 27, 2019
Length	2,800,485 nt
Sequencing Technology	Illumina + Oxford Nanopore Hybrid Assembly

Organism Summary	
Name	ATCC® 6538™
Isolation	Lesion
Biosafety Level	2
Type Strain	No

If customers have purchased the physical product…

Download genome(s) with the lot number



As of May 2024, to guarantee the long-term sustainability of the portal and to continue to provide you with reference-quality genomes, we have transitioned the ATCC Genome Portal over to a supporting membership service. Learn more about ATCC supporting memberships here.

Welcome to the ATCC Genome Portal

The only authenticated reference genome database for ATCC microbes

VIEW ALL GENOMES ›

Search for a genome

2001

Name
2001

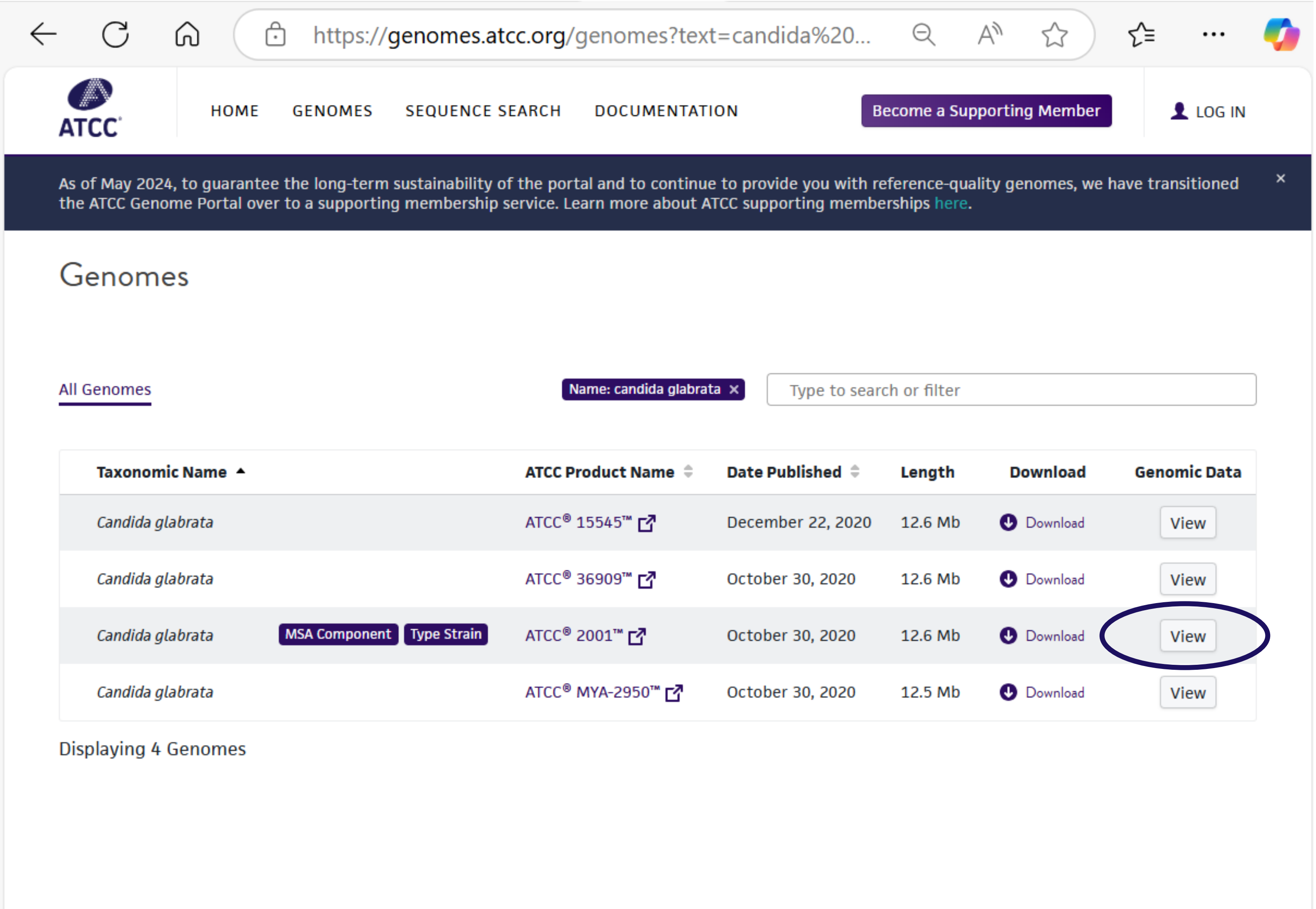
Catalog Number
2001

Tag
2001

Taxonomy

If customers have purchased the physical product...

Download genome(s) with the lot number



As of May 2024, to guarantee the long-term sustainability of the portal and to continue to provide you with reference-quality genomes, we have transitioned the ATCC Genome Portal over to a supporting membership service. Learn more about ATCC supporting memberships here.

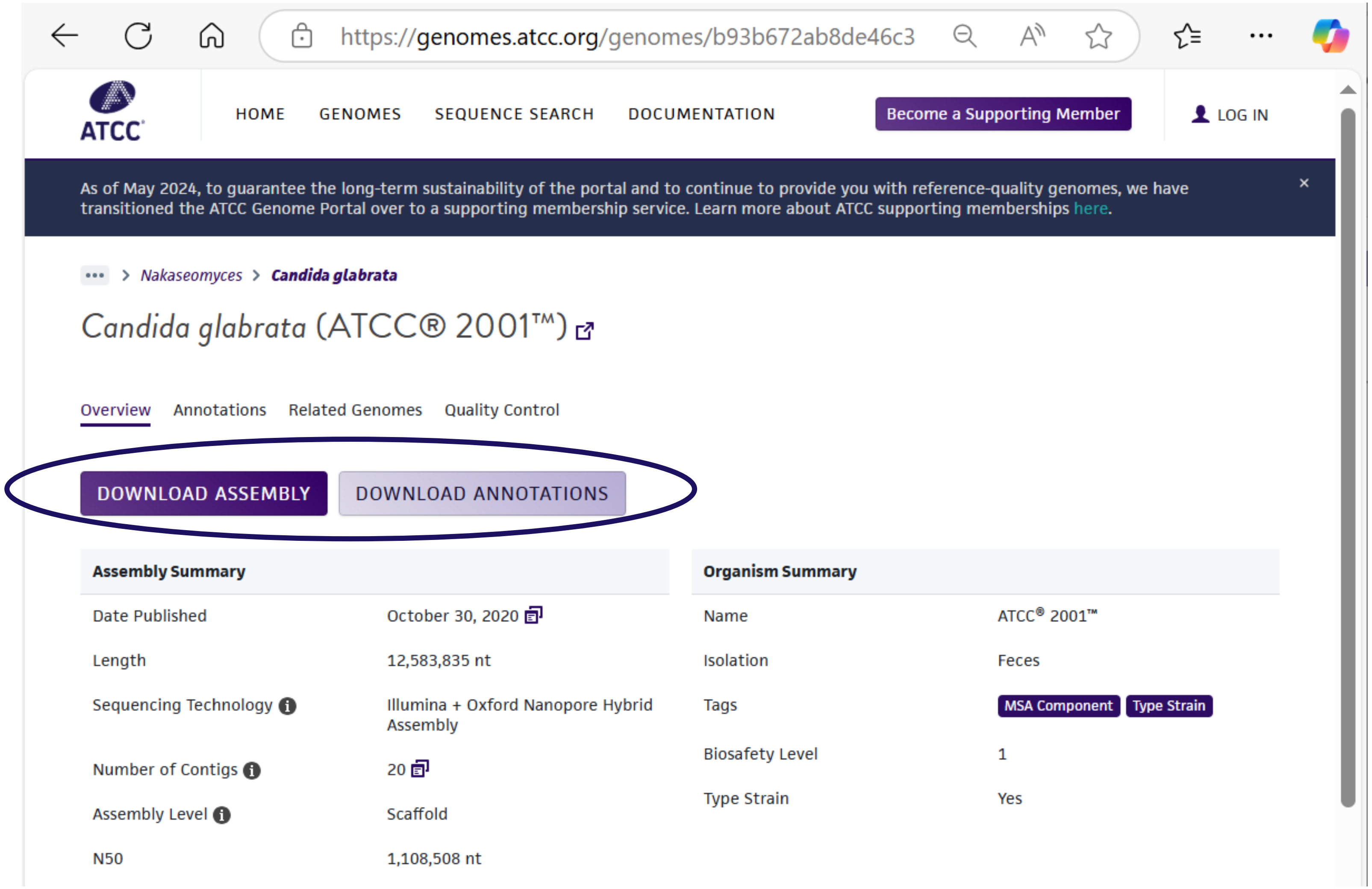
Genomes

Taxonomic Name ▲		ATCC Product Name	Date Published	Length	Download	Genomic Data
Candida glabrata		ATCC® 15545™	December 22, 2020	12.6 Mb	Download	View
Candida glabrata		ATCC® 36909™	October 30, 2020	12.6 Mb	Download	View
Candida glabrata	MSA Component Type Strain	ATCC® 2001™	October 30, 2020	12.6 Mb	Download	View
Candida glabrata		ATCC® MYA-2950™	October 30, 2020	12.5 Mb	Download	View

Displaying 4 Genomes

If customers have purchased the physical product…

Download genome(s) with the lot number



https://genomes.atcc.org/genomes/b93b672ab8de46c3

As of May 2024, to guarantee the long-term sustainability of the portal and to continue to provide you with reference-quality genomes, we have transitioned the ATCC Genome Portal over to a supporting membership service. Learn more about ATCC supporting memberships here.

••• > *Nakaseomyces* > ***Candida glabrata***

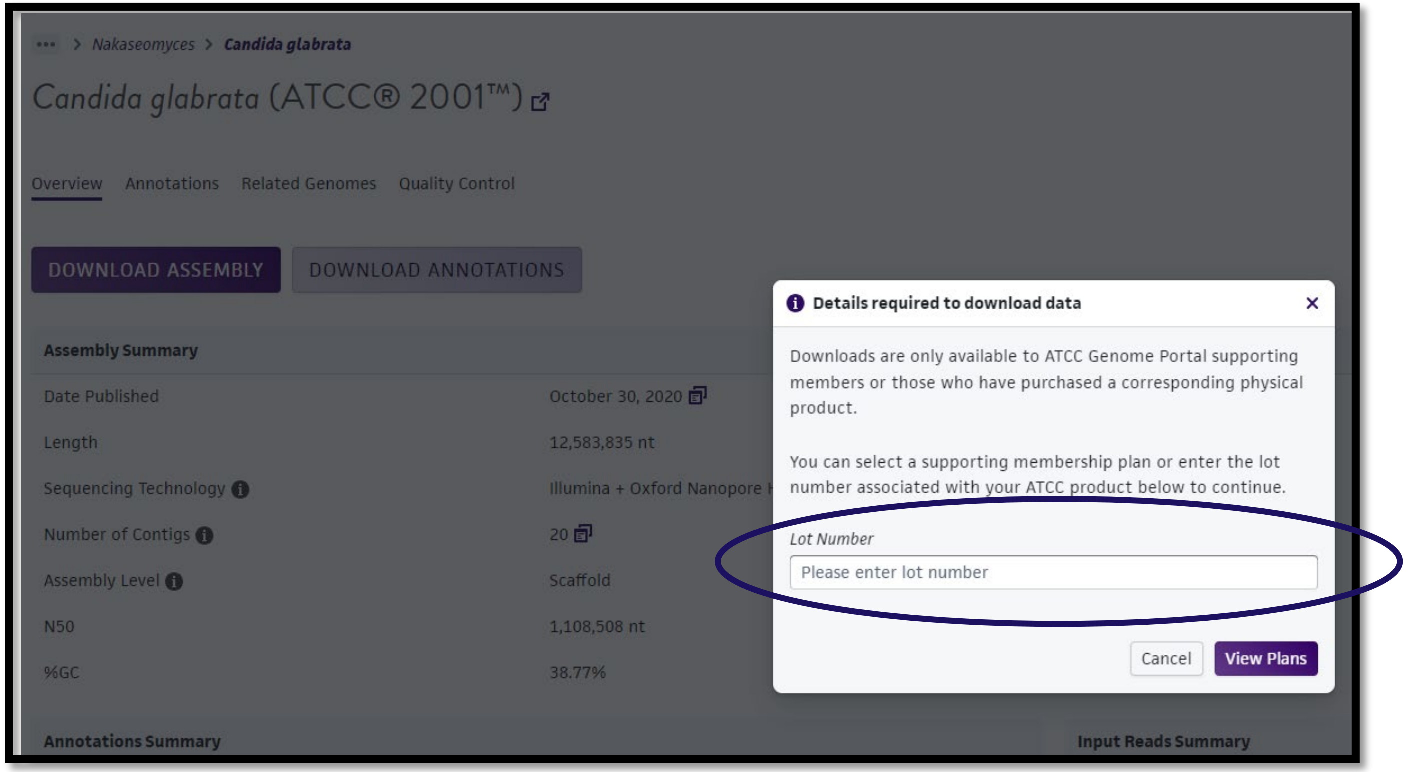
Candida glabrata (ATCC® 2001™)

Assembly Summary	
Date Published	October 30, 2020
Length	12,583,835 nt
Sequencing Technology	Illumina + Oxford Nanopore Hybrid Assembly
Number of Contigs	20
Assembly Level	Scaffold
N50	1,108,508 nt

Organism Summary	
Name	ATCC® 2001™
Isolation	Feces
Tags	MSA Component Type Strain
Biosafety Level	1
Type Strain	Yes

If customers have purchased the physical product…

Download genome(s) with the lot number



Candida glabrata (ATCC® 2001™)

Assembly Summary

Date Published	October 30, 2020
Length	12,583,835 nt
Sequencing Technology	Illumina + Oxford Nanopore
Number of Contigs	20
Assembly Level	Scaffold
N50	1,108,508 nt
%GC	38.77%

Annotations Summary

Input Reads Summary

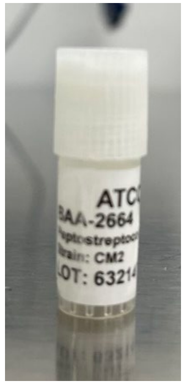
Details required to download data

Downloads are only available to ATCC Genome Portal supporting members or those who have purchased a corresponding physical product.

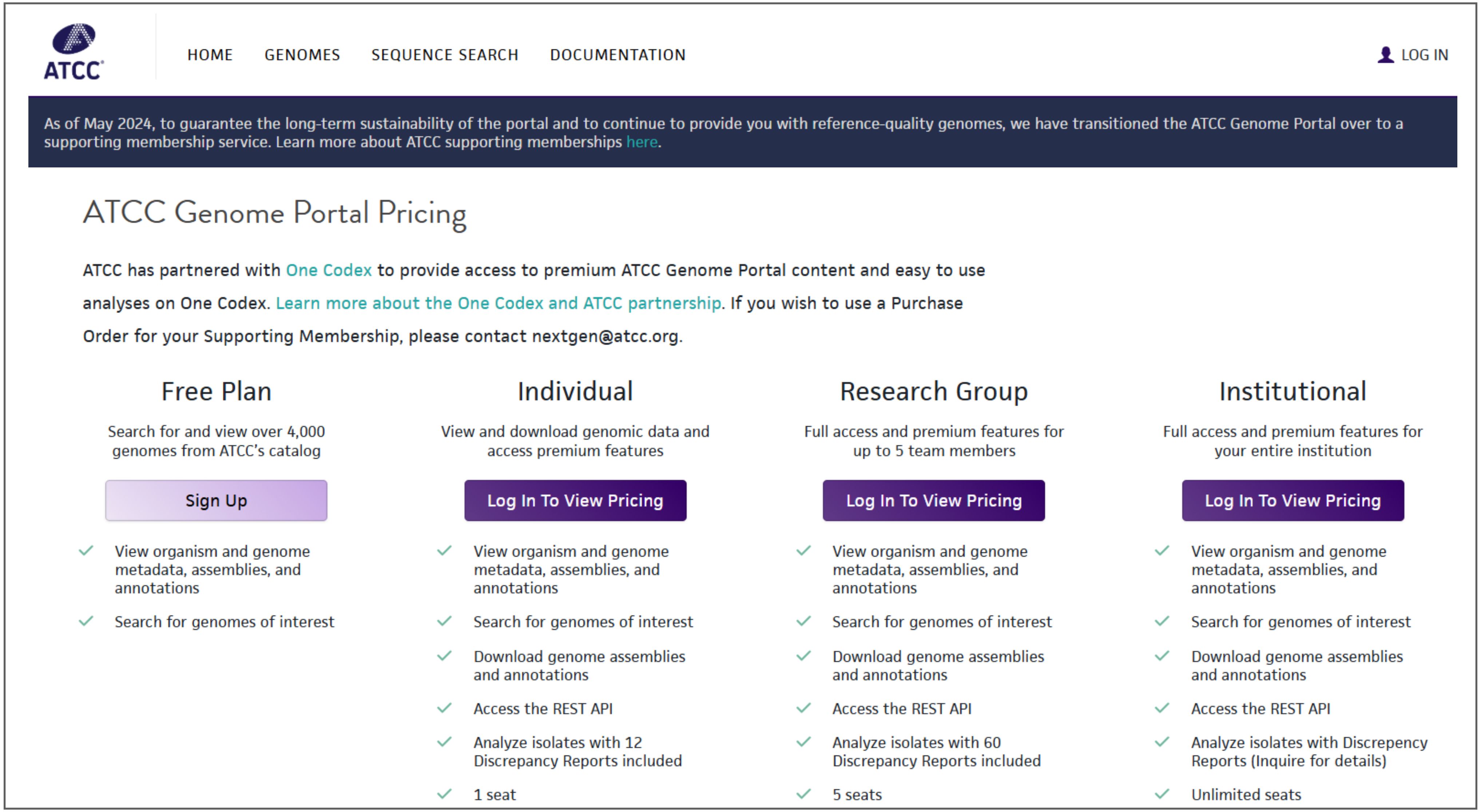
You can select a supporting membership plan or enter the lot number associated with your ATCC product below to continue.

Lot Number

Please enter lot number



Access to the entire database…



As of May 2024, to guarantee the long-term sustainability of the portal and to continue to provide you with reference-quality genomes, we have transitioned the ATCC Genome Portal over to a supporting membership service. Learn more about ATCC supporting memberships here.

ATCC Genome Portal Pricing

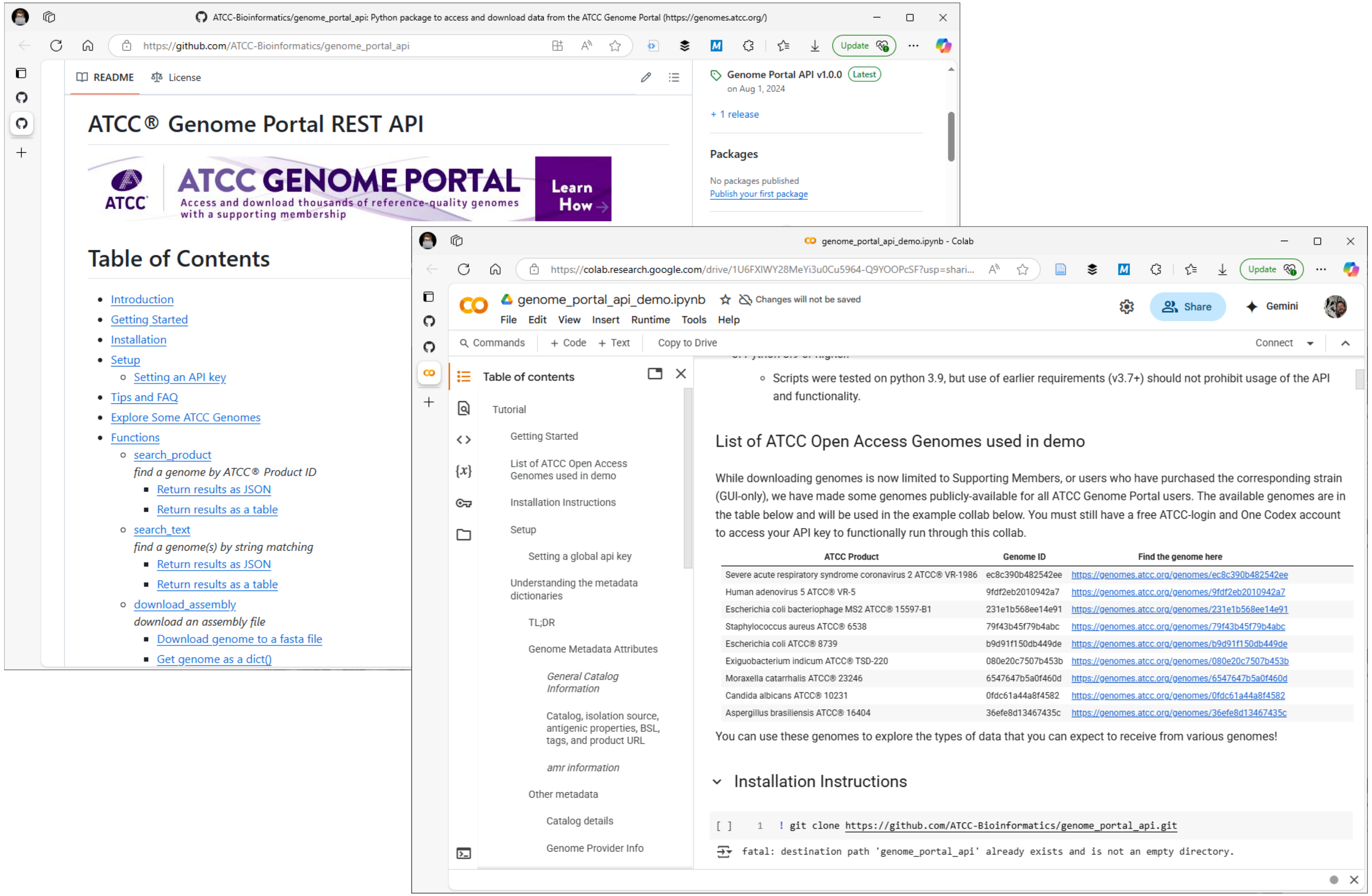
ATCC has partnered with One Codex to provide access to premium ATCC Genome Portal content and easy to use analyses on One Codex. Learn more about the One Codex and ATCC partnership. If you wish to use a Purchase Order for your Supporting Membership, please contact nextgen@atcc.org.

Free Plan	Individual	Research Group	Institutional
Search for and view over 4,000 genomes from ATCC's catalog	View and download genomic data and access premium features	Full access and premium features for up to 5 team members	Full access and premium features for your entire institution
✓ View organism and genome metadata, assemblies, and annotations	✓ View organism and genome metadata, assemblies, and annotations	✓ View organism and genome metadata, assemblies, and annotations	✓ View organism and genome metadata, assemblies, and annotations
✓ Search for genomes of interest	✓ Search for genomes of interest	✓ Search for genomes of interest	✓ Search for genomes of interest
	✓ Download genome assemblies and annotations	✓ Download genome assemblies and annotations	✓ Download genome assemblies and annotations
	✓ Access the REST API	✓ Access the REST API	✓ Access the REST API
	✓ Analyze isolates with 12 Discrepancy Reports included	✓ Analyze isolates with 60 Discrepancy Reports included	✓ Analyze isolates with Discrepency Reports (Inquire for details)
	✓ 1 seat	✓ 5 seats	✓ Unlimited seats

Programmatic access via our REST-API

https://github.com/ATCC-Bioinformatics/genome_portal_api

- **Documentation** available on GitHub.
- **Tutorials** available on Google Colab. Open-access assemblies available for testing. No Supporting Membership required.



ATCC® Genome Portal REST API

Table of Contents

- Introduction
- Getting Started
- Installation
- Setup
  - Setting an API key
- Tips and FAQ
- Explore Some ATCC Genomes
- Functions
  - search_product
    find a genome by ATCC® Product ID
    - Return results as JSON
    - Return results as a table
  - search_text
    find a genome(s) by string matching
    - Return results as JSON
    - Return results as a table
  - download_assembly
    download an assembly file
    - Download genome to a fasta file
    - Get genome as a dict()

- Scripts were tested on python 3.9, but use of earlier requirements (v3.7+) should not prohibit usage of the API and functionality.

List of ATCC Open Access Genomes used in demo

While downloading genomes is now limited to Supporting Members, or users who have purchased the corresponding strain (GUI-only), we have made some genomes publicly-available for all ATCC Genome Portal users. The available genomes are in the table below and will be used in the example collab below. You must still have a free ATCC-login and One Codex account to access your API key to functionally run through this collab.

ATCC Product	Genome ID	Find the genome here
Severe acute respiratory syndrome coronavirus 2 ATCC® VR-1986	ec8c390b482542ee	https://genomes.atcc.org/genomes/ec8c390b482542ee
Human adenovirus 5 ATCC® VR-5	9fdf2eb2010942a7	https://genomes.atcc.org/genomes/9fdf2eb2010942a7
Escherichia coli bacteriophage MS2 ATCC® 15597-B1	231e1b568ee14e91	https://genomes.atcc.org/genomes/231e1b568ee14e91
Staphylococcus aureus ATCC® 6538	79f43b45f79b4abc	https://genomes.atcc.org/genomes/79f43b45f79b4abc
Escherichia coli ATCC® 8739	b9d91f150db449de	https://genomes.atcc.org/genomes/b9d91f150db449de
Exiguobacterium indicum ATCC® TSD-220	080e20c7507b453b	https://genomes.atcc.org/genomes/080e20c7507b453b
Moraxella catarrhalis ATCC® 23246	6547647b5a0f460d	https://genomes.atcc.org/genomes/6547647b5a0f460d
Candida albicans ATCC® 10231	0fdc61a44a8f4582	https://genomes.atcc.org/genomes/0fdc61a44a8f4582
Aspergillus brasiliensis ATCC® 16404	36efe8d13467435c	https://genomes.atcc.org/genomes/36efe8d13467435c

You can use these genomes to explore the types of data that you can expect to receive from various genomes!

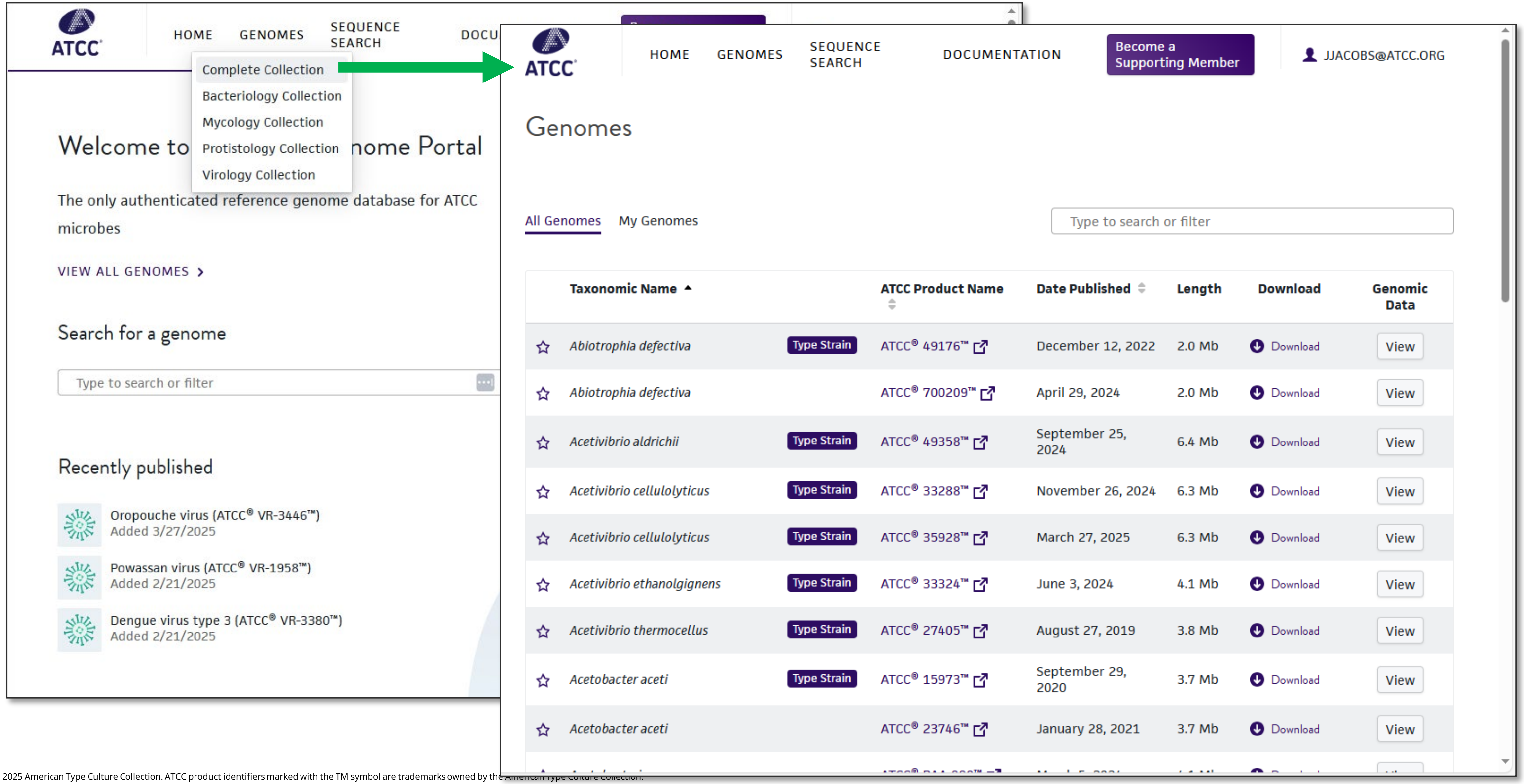
Installation Instructions

```
[ ]  ! git clone https://github.com/ATCC-Bioinformatics/genome_portal_api.git
fatal: destination path 'genome_portal_api' already exists and is not an empty directory.
```

The ATCC® Genome Portal

Search, exploration, and analysis tools

Browse for your data



Welcome to … nome Portal

The only authenticated reference genome database for ATCC microbes

Search for a genome

Recently published

Oropouche virus (ATCC® VR-3446™)
Added 3/27/2025

Powassan virus (ATCC® VR-1958™)
Added 2/21/2025

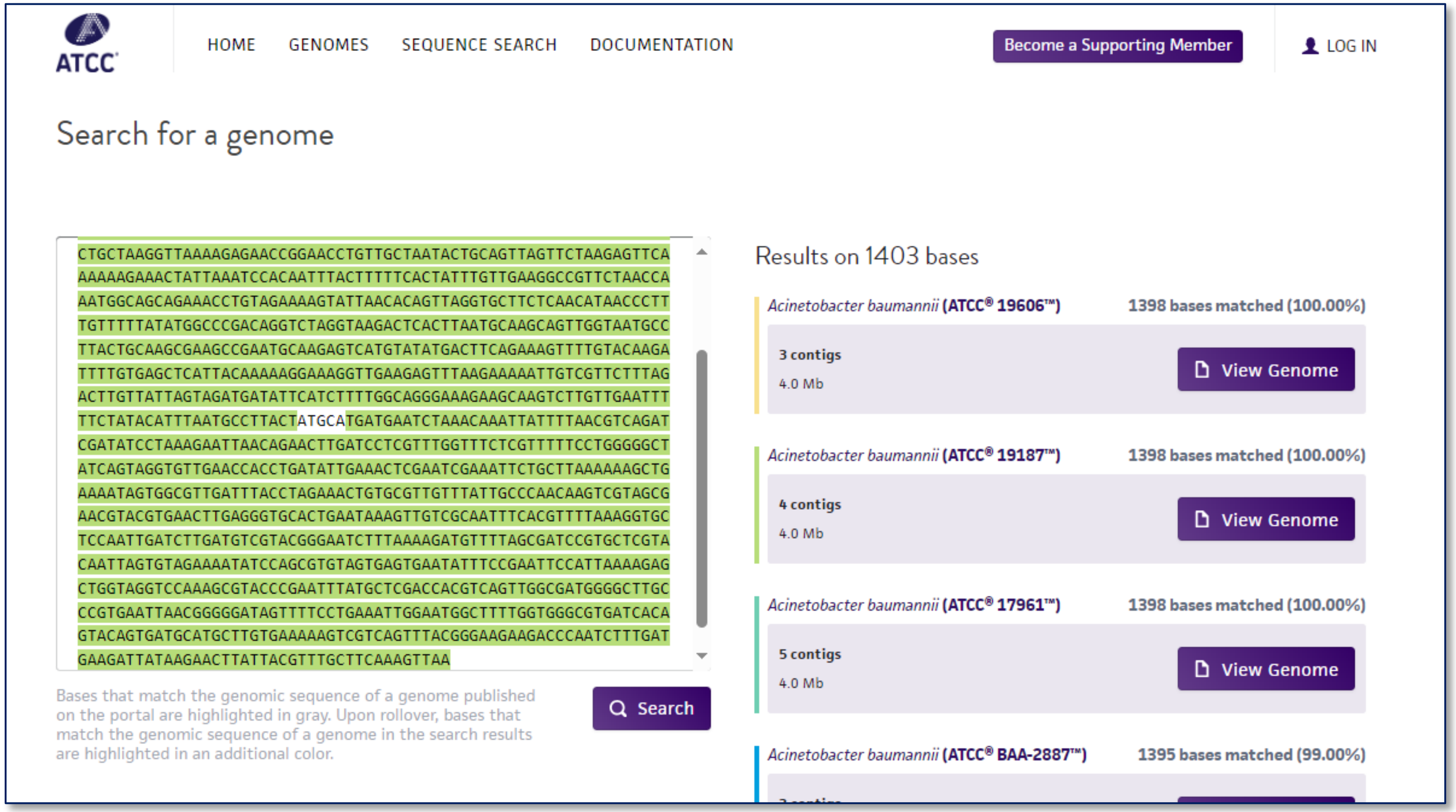
Dengue virus type 3 (ATCC® VR-3380™)
Added 2/21/2025

Genomes

	Taxonomic Name ▲		ATCC Product Name	Date Published	Length	Download	Genomic Data
☆	*Abiotrophia defectiva*	Type Strain	ATCC® 49176™	December 12, 2022	2.0 Mb	Download	View
☆	*Abiotrophia defectiva*		ATCC® 700209™	April 29, 2024	2.0 Mb	Download	View
☆	*Acetivibrio aldrichii*	Type Strain	ATCC® 49358™	September 25, 2024	6.4 Mb	Download	View
☆	*Acetivibrio cellulolyticus*	Type Strain	ATCC® 33288™	November 26, 2024	6.3 Mb	Download	View
☆	*Acetivibrio cellulolyticus*	Type Strain	ATCC® 35928™	March 27, 2025	6.3 Mb	Download	View
☆	*Acetivibrio ethanolgignens*	Type Strain	ATCC® 33324™	June 3, 2024	4.1 Mb	Download	View
☆	*Acetivibrio thermocellus*	Type Strain	ATCC® 27405™	August 27, 2019	3.8 Mb	Download	View
☆	*Acetobacter aceti*	Type Strain	ATCC® 15973™	September 29, 2020	3.7 Mb	Download	View
☆	*Acetobacter aceti*		ATCC® 23746™	January 28, 2021	3.7 Mb	Download	View

Fast sequencing search

https://genomes.atcc.org/sequence-search



Search for a genome

```
CTGCTAAGGTTAAAAGAGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCA
AAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAACCA
AATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTT
TGTTTTTATATGGCCCGACAGGTCTAGGTAAGACTCACTTAATGCAAGCAGTTGGTAATGCC
TTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGA
TTTTGTGAGCTCATTACAAAAAGGAAAGGTTGAAGAGTTTAAGAAAAATTGTCGTTCTTTAG
ACTTGTTATTAGTAGATGATATTCATCTTTTGGCAGGGAAAGAAGCAAGTCTTGTTGAATTT
TTCTATACATTTAATGCCTTACTATGCATGATGAATCTAAACAAATTATTTTAACGTCAGAT
CGATATCCTAAAGAATTAACAGAACTTGATCCTCGTTTGGTTTCTCGTTTTTCCTGGGGGCT
ATCAGTAGGTGTTGAACCACCTGATATTGAAACTCGAATCGAAATTCTGCTTAAAAAAGCTG
AAAATAGTGGCGTTGATTTACCTAGAAACTGTGCGTTGTTTATTGCCCAACAAGTCGTAGCG
AACGTACGTGAACTTGAGGGTGCACTGAATAAAGTTGTCGCAATTTCACGTTTTAAAGGTGC
TCCAATTGATCTTGATGTCGTACGGGAATCTTTAAAAGATGTTTTAGCGATCCGTGCTCGTA
CAATTAGTGTAGAAAATATCCAGCGTGTAGTGAGTGAATATTTCCGAATTCCATTAAAAGAG
CTGGTAGGTCCAAAGCGTACCCGAATTTATGCTCGACCACGTCAGTTGGCGATGGGGCTTGC
CCGTGAATTAACGGGGGATAGTTTTCCTGAAATTGGAATGGCTTTTGGTGGGCGTGATCACA
GTACAGTGATGCATGCTTGTGAAAAAGTCGTCAGTTTACGGGAAGAAGACCCAATCTTTGAT
GAAGATTATAAGAACTTATTACGTTTGCTTCAAAGTTAA
```

Bases that match the genomic sequence of a genome published on the portal are highlighted in gray. Upon rollover, bases that match the genomic sequence of a genome in the search results are highlighted in an additional color.

Results on 1403 bases

Acinetobacter baumannii **(ATCC® 19606™)** 1398 bases matched (100.00%)

3 contigs

4.0 Mb

Acinetobacter baumannii **(ATCC® 19187™)** 1398 bases matched (100.00%)

4 contigs

4.0 Mb

Acinetobacter baumannii **(ATCC® 17961™)** 1398 bases matched (100.00%)

5 contigs

4.0 Mb

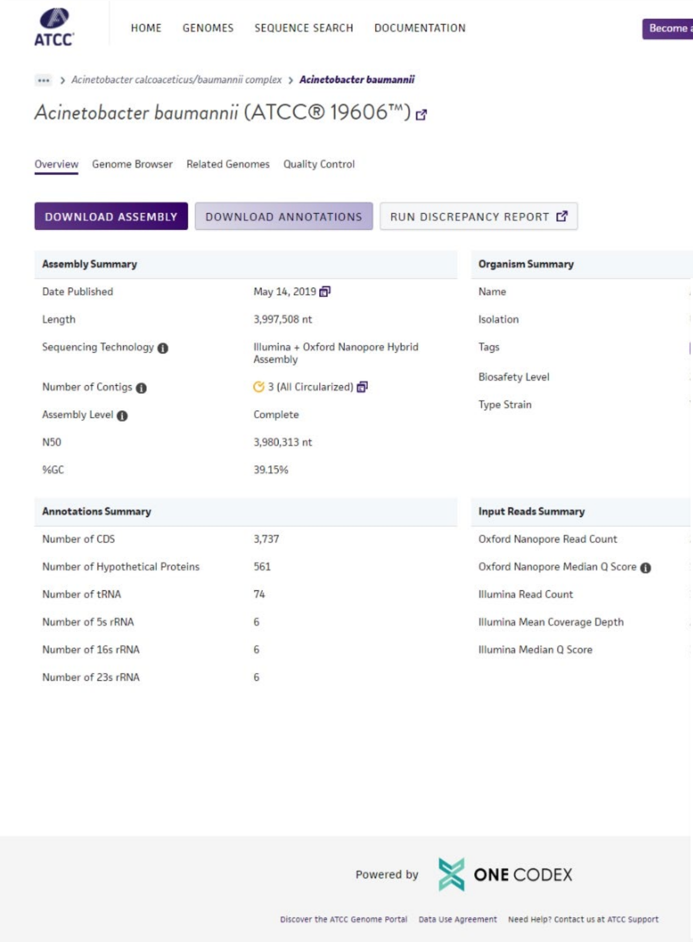
Acinetobacter baumannii **(ATCC® BAA-2887™)** 1395 bases matched (99.00%)

Search results are almost instantaneous!

ATCC® Genome Portal: Reference Genome Details

Example: *Acinetobacter baumannii* (ATCC® 19606™)

Overview page



••• > *Acinetobacter calcoaceticus/baumannii complex* > ***Acinetobacter baumannii***

Acinetobacter baumannii (ATCC® 19606™)

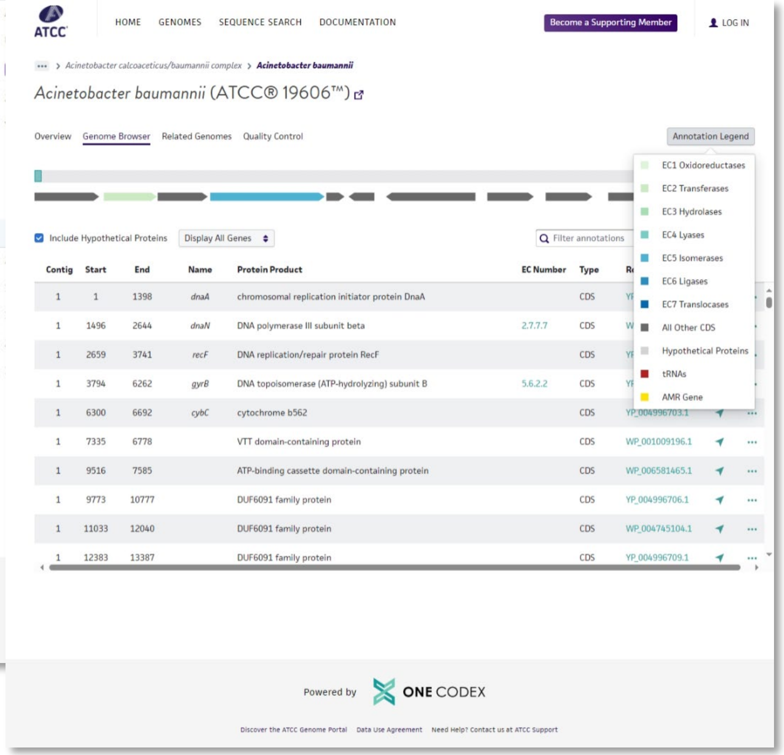
Assembly Summary	
Date Published	May 14, 2019
Length	3,997,508 nt
Sequencing Technology	Illumina + Oxford Nanopore Hybrid Assembly
Number of Contigs	3 (All Circularized)
Assembly Level	Complete
N50	3,980,313 nt
%GC	39.15%

Organism Summary
Name
Isolation
Tags
Biosafety Level
Type Strain

Annotations Summary	
Number of CDS	3,737
Number of Hypothetical Proteins	561
Number of tRNA	74
Number of 5s rRNA	6
Number of 16s rRNA	6
Number of 23s rRNA	6

Input Reads Summary
Oxford Nanopore Read Count
Oxford Nanopore Median Q Score
Illumina Read Count
Illumina Mean Coverage Depth
Illumina Median Q Score

Genome browser



••• > *Acinetobacter calcoaceticus/baumannii complex* > ***Acinetobacter baumannii***

Acinetobacter baumannii (ATCC® 19606™)

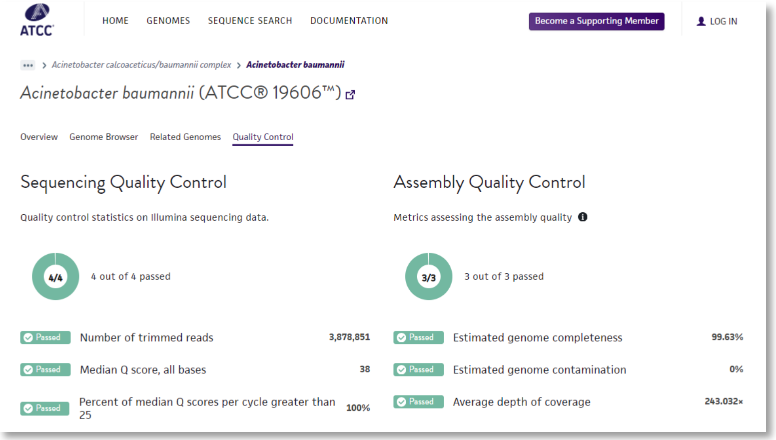
Annotation Legend:

- EC1 Oxidoreductases
- EC2 Transferases
- EC3 Hydrolases
- EC4 Lyases
- EC5 Isomerases
- EC6 Ligases
- EC7 Translocases
- All Other CDS
- Hypothetical Proteins
- tRNAs
- AMR Gene

☑ Include Hypothetical Proteins

Contig	Start	End	Name	Protein Product	EC Number	Type	R
1	1	1398	*dnaA*	chromosomal replication initiator protein DnaA		CDS	YF
1	1496	2644	*dnaN*	DNA polymerase III subunit beta	2.7.7.7	CDS	W
1	2659	3741	*recF*	DNA replication/repair protein RecF		CDS	YF
1	3794	6262	*gyrB*	DNA topoisomerase (ATP-hydrolyzing) subunit B	5.6.2.2	CDS	YF
1	6300	6692	*cybC*	cytochrome b562		CDS	YP_004996703.1
1	7335	6778		VTT domain-containing protein		CDS	WP_001009196.1
1	9516	7585		ATP-binding cassette domain-containing protein		CDS	WP_006581465.1
1	9773	10777		DUF6091 family protein		CDS	YP_004996706.1
1	11033	12040		DUF6091 family protein		CDS	WP_004745104.1
1	12383	13387		DUF6091 family protein		CDS	YP_004996709.1

View quality control data



••• > *Acinetobacter calcoaceticus/baumannii complex* > ***Acinetobacter baumannii***

Acinetobacter baumannii (ATCC® 19606™)

Sequencing Quality Control

Quality control statistics on Illumina sequencing data.

4 out of 4 passed

✓ Passed	Number of trimmed reads	3,878,851
✓ Passed	Median Q score, all bases	38
✓ Passed	Percent of median Q scores per cycle greater than 25	100%

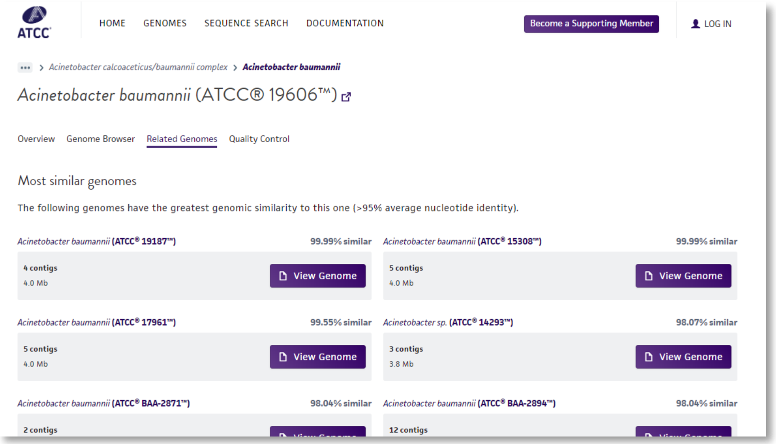
Assembly Quality Control

Metrics assessing the assembly quality

3 out of 3 passed

✓ Passed	Estimated genome completeness	99.63%
✓ Passed	Estimated genome contamination	0%
✓ Passed	Average depth of coverage	243.032×

Find related genomes



••• > *Acinetobacter calcoaceticus/baumannii complex* > ***Acinetobacter baumannii***

Acinetobacter baumannii (ATCC® 19606™)

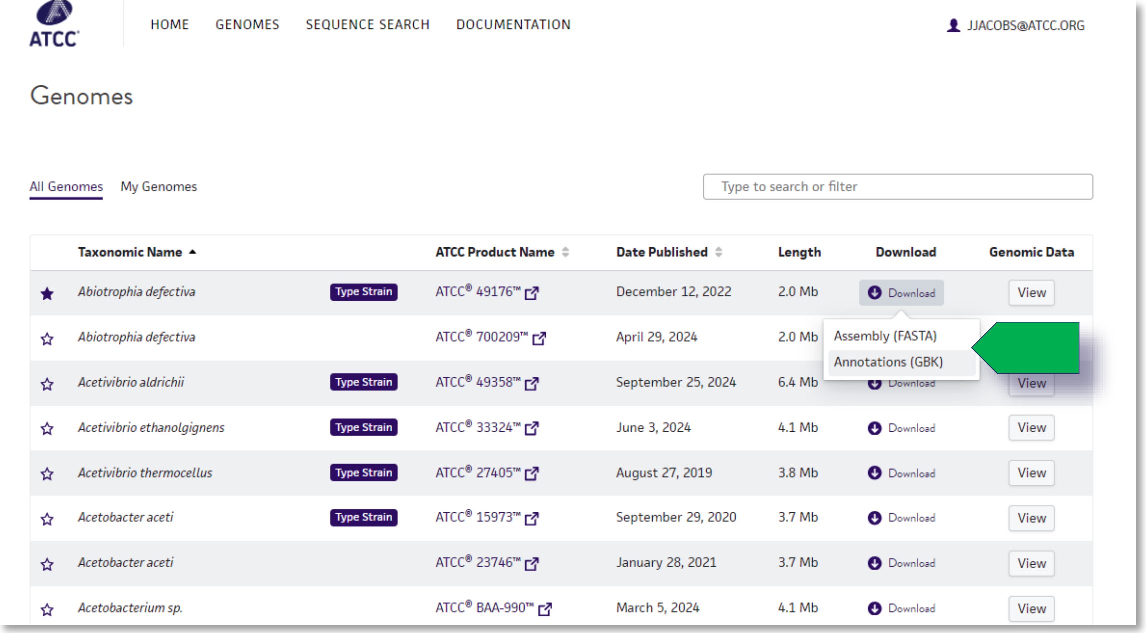
Most similar genomes

The following genomes have the greatest genomic similarity to this one (>95% average nucleotide identity).

Genome	Similarity	Contigs	Size
Acinetobacter baumannii **(ATCC® 19187™)**	**99.99% similar**	4 contigs	4.0 Mb
Acinetobacter baumannii **(ATCC® 15308™)**	**99.99% similar**	5 contigs	4.0 Mb
Acinetobacter baumannii **(ATCC® 17961™)**	**99.55% similar**	5 contigs	4.0 Mb
Acinetobacter sp. **(ATCC® 14293™)**	**98.07% similar**	3 contigs	3.8 Mb
Acinetobacter baumannii **(ATCC® BAA-2871™)**	**98.04% similar**	2 contigs	
Acinetobacter baumannii **(ATCC® BAA-2894™)**	**98.04% similar**	12 contigs	

Download genome references

Directly from search results



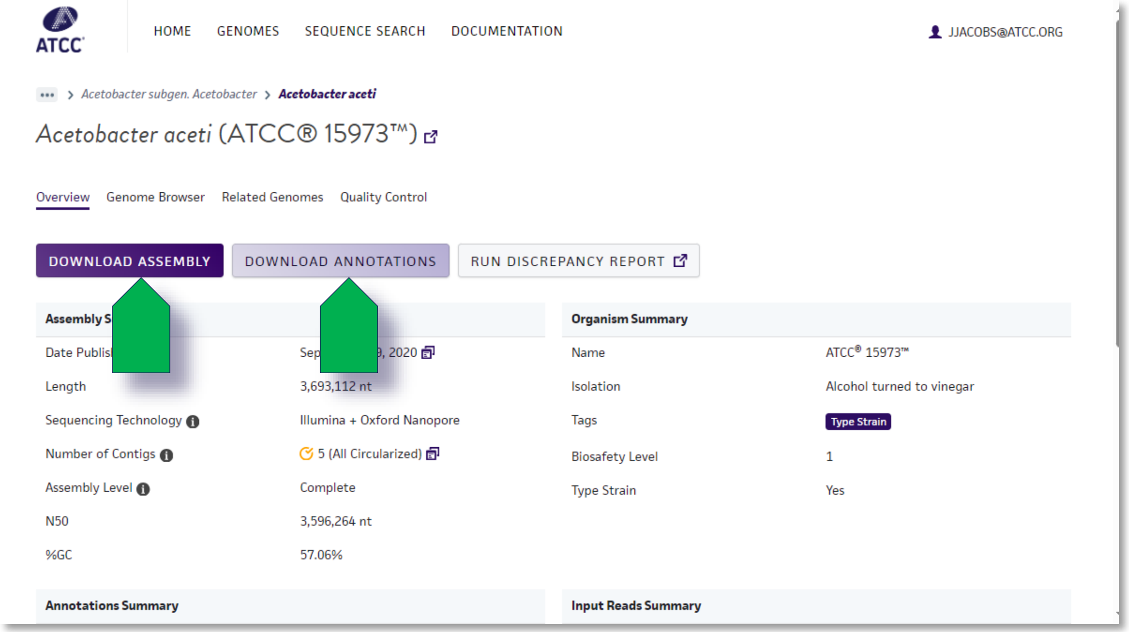
Genomes

	Taxonomic Name ▲		ATCC Product Name	Date Published	Length	Download	Genomic Data
★	*Abiotrophia defectiva*	Type Strain	ATCC® 49176™	December 12, 2022	2.0 Mb	Download	View
☆	*Abiotrophia defectiva*		ATCC® 700209™	April 29, 2024	2.0 Mb	Assembly (FASTA) Annotations (GBK)	
☆	*Acetivibrio aldrichii*	Type Strain	ATCC® 49358™	September 25, 2024	6.4 Mb	Download	View
☆	*Acetivibrio ethanolgignens*	Type Strain	ATCC® 33324™	June 3, 2024	4.1 Mb	Download	View
☆	*Acetivibrio thermocellus*	Type Strain	ATCC® 27405™	August 27, 2019	3.8 Mb	Download	View
☆	*Acetobacter aceti*	Type Strain	ATCC® 15973™	September 29, 2020	3.7 Mb	Download	View
☆	*Acetobacter aceti*		ATCC® 23746™	January 28, 2021	3.7 Mb	Download	View
☆	*Acetobacterium sp.*		ATCC® BAA-990™	March 5, 2024	4.1 Mb	Download	View

Two file formats for download:

1. **FASTA files**
   - **.fasta** files are smaller and include only basic information and unannotated DNA sequence for entire genome.
2. **GenBank files**
   - **.gbk** files are larger but include annotations for all known genes and rich metadata for the organism.

Download from details page



Acetobacter aceti (ATCC® 15973™)

Assembly S	
Date Publis	Sep 9, 2020
Length	3,693,112 nt
Sequencing Technology	Illumina + Oxford Nanopore
Number of Contigs	5 (All Circularized)
Assembly Level	Complete
N50	3,596,264 nt
%GC	57.06%

Organism Summary	
Name	ATCC® 15973™
Isolation	Alcohol turned to vinegar
Tags	Type Strain
Biosafety Level	1
Type Strain	Yes

Annotations Summary

Input Reads Summary

While both formats can be opened with a plain-text editor (i.e., Notepad), these files will often be thousands of lines long. They are intended to be imported into 3rd party bioinformatics software for data visualization.

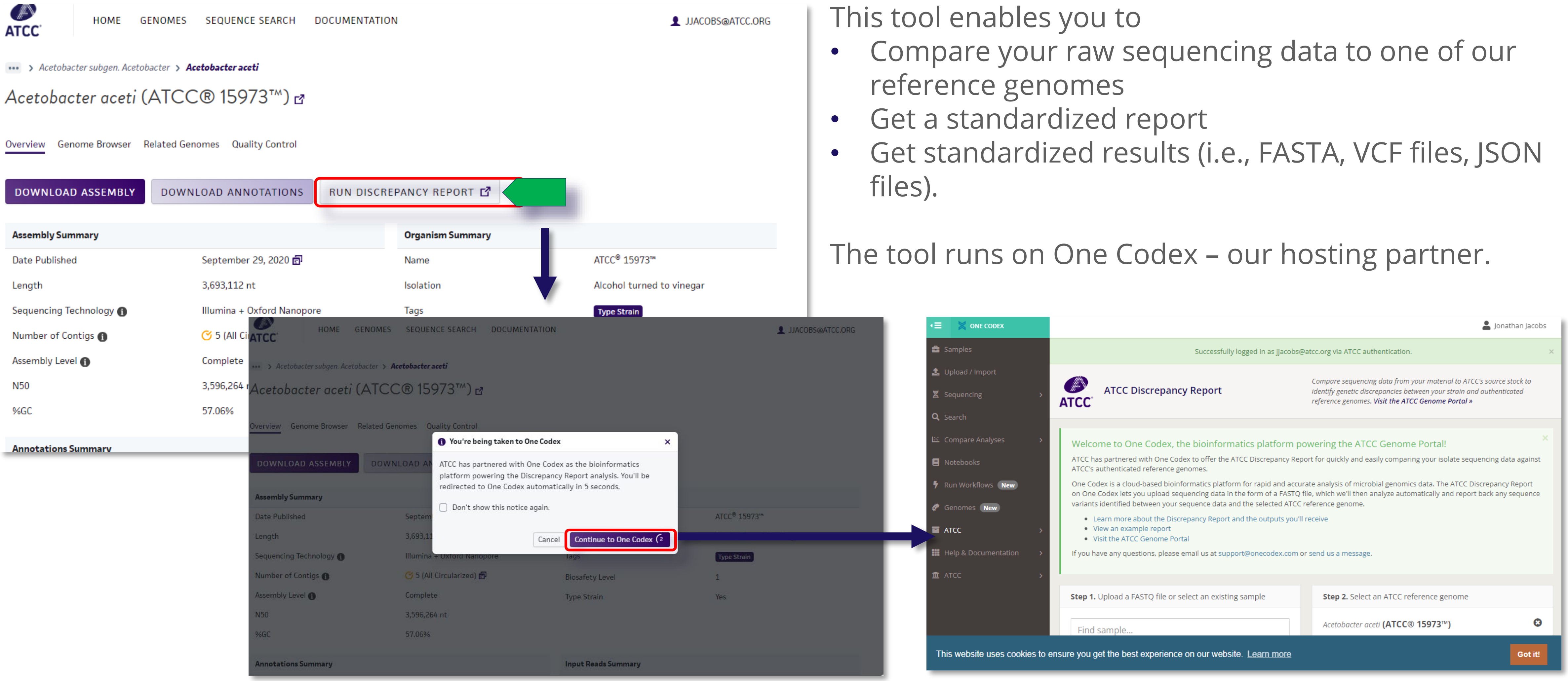
We also have a REST-API for programmatic access to the ATCC® Genome Portal, including downloads. This is often the preferred approach by data scientists and bioinformaticians.

Run a Discrepancy Report

This tool enables you to

- Compare your raw sequencing data to one of our reference genomes
- Get a standardized report
- Get standardized results (i.e., FASTA, VCF files, JSON files).

The tool runs on One Codex – our hosting partner.



••• > *Acetobacter subgen. Acetobacter* > ***Acetobacter aceti***

Acetobacter aceti (ATCC® 15973™)

Assembly Summary

Date Published	September 29, 2020
Length	3,693,112 nt
Sequencing Technology	Illumina + Oxford Nanopore
Number of Contigs	5 (All Circularized)
Assembly Level	Complete
N50	3,596,264 nt
%GC	57.06%

Organism Summary

Name	ATCC® 15973™
Isolation	Alcohol turned to vinegar
Tags	Type Strain
Biosafety Level	1
Type Strain	Yes

Annotations Summary

Input Reads Summary

You're being taken to One Codex

ATCC has partnered with One Codex as the bioinformatics platform powering the Discrepancy Report analysis. You'll be redirected to One Codex automatically in 5 seconds.

☐ Don't show this notice again.

Successfully logged in as jjacobs@atcc.org via ATCC authentication.

ATCC Discrepancy Report

Compare sequencing data from your material to ATCC's source stock to identify genetic discrepancies between your strain and authenticated reference genomes. ***Visit the ATCC Genome Portal »***

Welcome to One Codex, the bioinformatics platform powering the ATCC Genome Portal!

ATCC has partnered with One Codex to offer the ATCC Discrepancy Report for quickly and easily comparing your isolate sequencing data against ATCC's authenticated reference genomes.

One Codex is a cloud-based bioinformatics platform for rapid and accurate analysis of microbial genomics data. The ATCC Discrepancy Report on One Codex lets you upload sequencing data in the form of a FASTQ file, which we'll then analyze automatically and report back any sequence variants identified between your sequence data and the selected ATCC reference genome.

- Learn more about the Discrepancy Report and the outputs you'll receive
- View an example report
- Visit the ATCC Genome Portal

If you have any questions, please email us at support@onecodex.com or send us a message.

Step 1. Upload a FASTQ file or select an existing sample

Find sample...

Step 2. Select an ATCC reference genome

Acetobacter aceti **(ATCC® 15973™)**

This website uses cookies to ensure you get the best experience on our website. Learn more

Recent updates and coming soon

- **More genomes:** On June 18, 2025, ATCC® released an additional 250 genomes to the AGP. We are now at **5,750 released genomes** and counting!

- **Methylation data:** We released methylation data for **over 3,500 items** in the ATCC® bacteriology collection. All sequenced items in the bacteriology collection will have methylation data available by the end of Q3, 2025.

- **AMR annotations:** We have expanded the process for annotating AMR genes and will release updated AMR annotations in Q3, 2025.

- **Interface updates:** The AGP user interface will include several improvements to the overall user experience in Q3 2025.
  - Browse the *entire* ATCC® microbial collection by taxonomy—quickly see near neighbors, relationship to the type strain, request NGS, etc.
  - Add all data associated with a product to “My Data Bundle” for download
  - Search for organisms with similar sequences or by **gene name**

- **Expanded content:** ATCC® will provide WES and RNA-Seq files for over 600 **ATCC® human and mouse cell lines** in Q4, 2025.

Summary

- Current publicly available, consensus-driven genomic sequences often lack data authenticity, quality, completeness, or traceability, which can compromise research findings. This gap in data provenance potentially complicates downstream bioinformatics applications and research objectives.
- The ATCC® Genome Portal is a rapidly growing ISO 9001–compliant database of high-quality reference genomes from authenticated microbial strains in the ATCC® collection.
- Through this cloud-based platform, scientists can easily access and download meticulously curated whole-genome sequences from their browser or our secure API.

Thank you!

Sequencing & Bioinformatics Center

✉ nextgen@atcc.org

Jonathan Jacobs, PhD

Senior Director, Bioinformatics

BioNexus Principal Investigator

✉ **jjacobs@atcc.org**

Genomics Lab

Briana Benton, PMP
Ana Fernandes
Ajeet Singh, PhD
Stephen King, MSc
James Duncan, MSc
Robert Marlow
Corina Tabron, MSc
Jade Kirkland
Noah Wax, MSc
Rula Khairi
Hannah McConnell
Kaitlyn Gaffney

Bioinformatics Lab

John Bagnoli
David Yarmosh, MSc.
Nikhita Putheveetil, MSc.
Joseph Petron, PhD.
Amy Reese, MSc

Scott V Nguyen, PhD
Senior Biocuration Scientist

Our Partner

ONE CODEX

ATCC®

CREDIBLE LEADS TO INCREDIBLE

Thank You