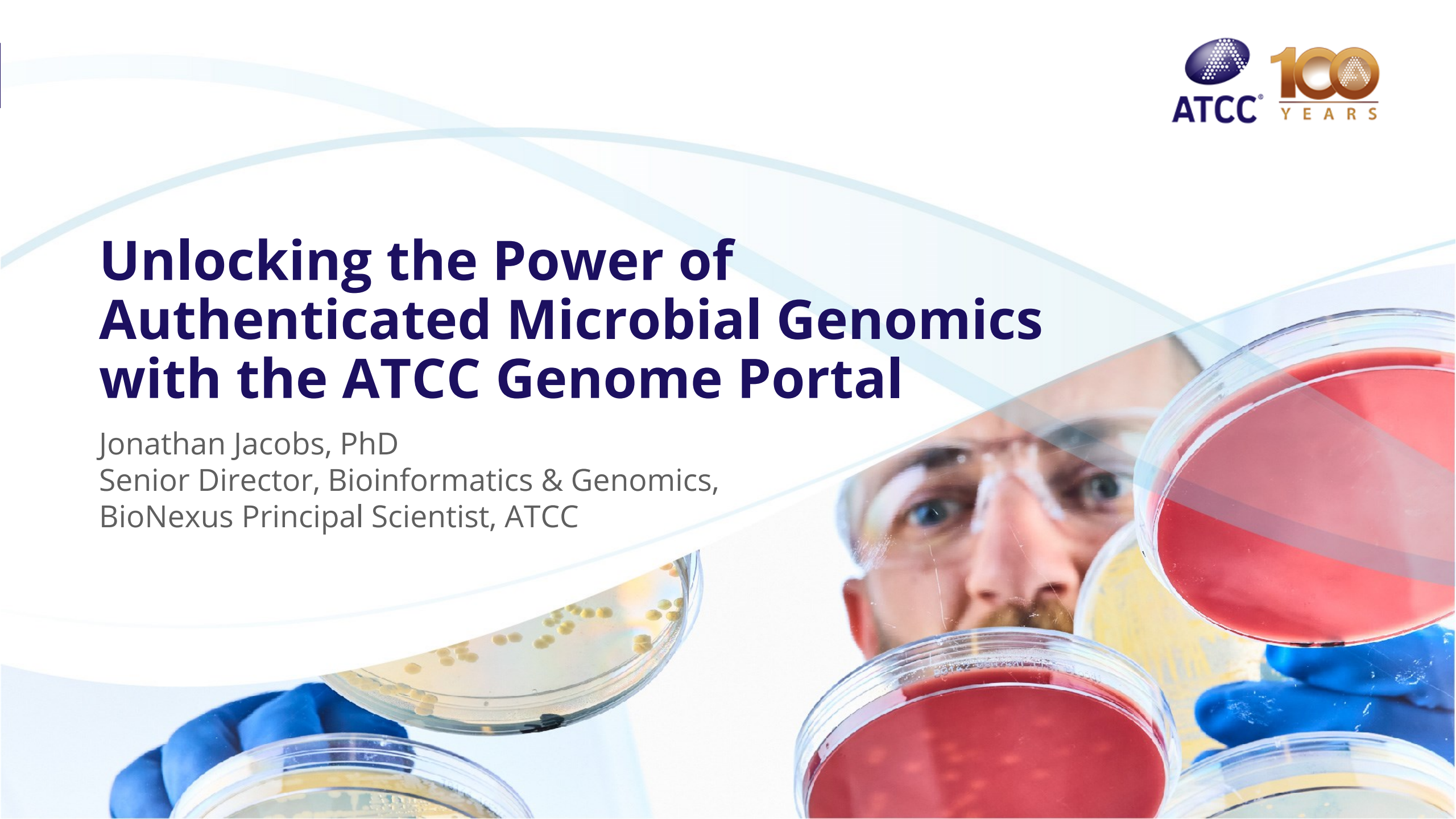


Unlocking the Power of Authenticated Microbial Genomics with the ATCC Genome Portal

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



Introducing today's speaker



Jonathan Jacobs, PhD

Senior Director, Bioinformatics / BioNexus Foundation Principal Scientist, ATCC

Dr. Jonathan Jacobs is the Senior Director of Bioinformatics at ATCC, where he leads the Sequencing & Bioinformatics Center and oversees the development of the ATCC Genome Portal, an authenticated reference genome database for ATCC materials. With over 20 years of experience in molecular genetics, bioinformatics, and microbial genomics, Dr. Jacobs has consistently worked to solve problems and create solutions across academia, government, and industry. His own research has included functional genomics of viral pathogenicity and fungal multi-drug resistance mechanisms, cell line engineering for biopharma and industrial microbiology, and developing metagenomics tools for biosurveillance of emerging infectious diseases.

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



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 Nanopore data used for most assemblies

Standardized

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

Powered by



The discovery loop

How we generally do everything

The (improved) discovery loop

Have an idea



Plan experiments



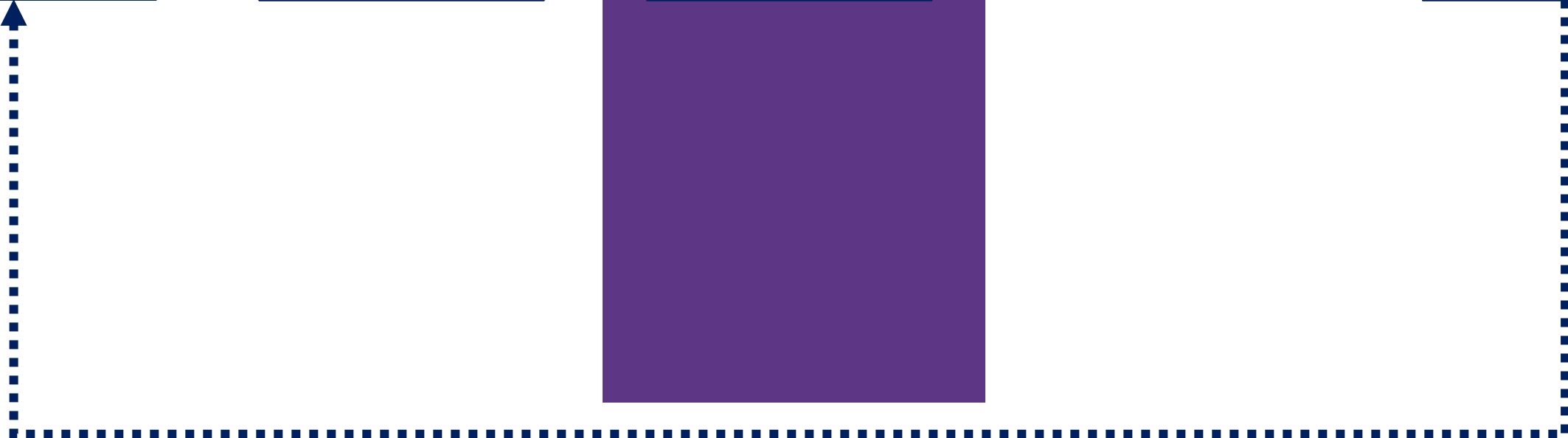
Find data & materials



Do experiment



Analyze results



A common scenario...



Microbiologist

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



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?

A common scenario...



Microbiologist

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

Overlooked assumptions

Get materials, controls, strains, etc.

Find reference data (genomes, genes, etc.)

Research, development, validation

Unexpected results

Why?

Plan your lab research and development

Design assay (bioinformatics)

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



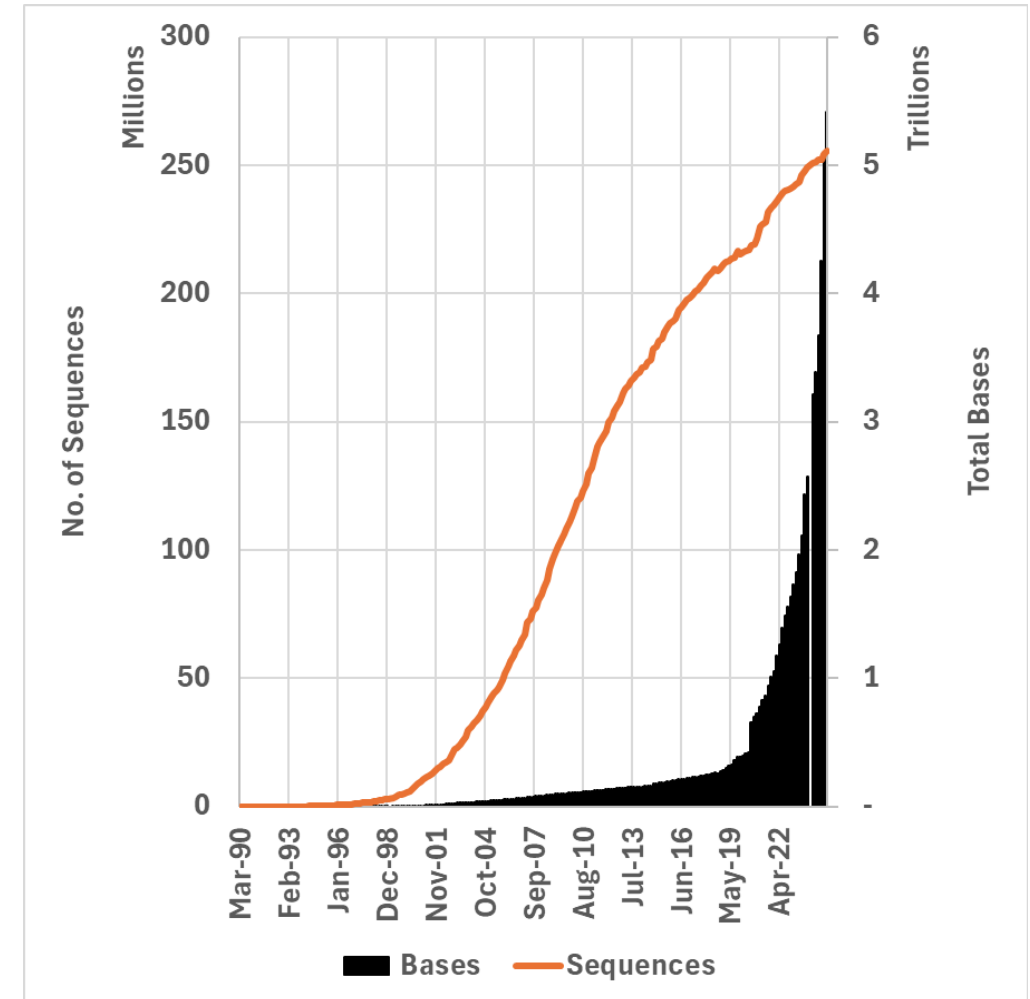
How trustworthy is that data?

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);
4. Table of Contents;

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.24268

REVIEW

Human Mutation

Pathogenic noncoding variants in the neurofibromatosis type 1 schwannomatosis predisposition genes

PEREZ-BECERRIL ET AL.

Cristina Perez-Becerril

Division of Evolution and Genomics, Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester Academic Health Science Centre, School of Biological Sciences, University of Manchester, 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

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

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);
4. Table of Contents

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 CBF β -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 CBF β -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

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

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/nucleotide/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>



COMMENTARY

Open Access

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

Supratim Mukherjee^{1*}, Marcel Huntemann¹, Natalia Ivanova¹, Nikos C. Kyrpides¹

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 been 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 effi measures. The presence of PhiX contamination in several publicly available isolat errors when such data are used in comparative genomics analyses. Such contan far-reaching consequences in the form of erroneous data interpretation and ana measures to proofread raw sequences before releasing them to the broader sci

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 synthesis technology [6] championed by Illumina [7] is

Despite its high platform does come need to be address One such challenge used as a quality ar runs. PhiX is an i with a single-stranc 5386 nucleotides a sequenced by Fred defined genome se used as a control fi majority of its librai using PhiX at a low raised up to 40% fi on the concentrati the same line along with the sample or used as a conser.

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

METHOD

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@nu.ac.kr
¹School of Biological Sciences, Seoul National University, Seoul, 08826, South Korea
²Center 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 Perteu,^{1,2} Aleksey V. Zimin,^{1,3}

¹Johns Hopkins 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; if Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA

Genome Biology

Open Access



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 B. The absence of the sequences from the human assembly offers a likely expla- bles. 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 e 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 contam- genuine genomic sequences.

article.]

ily available ge- es to well over tal resources for ing microbiome mplex samples rence databases), but for practi- today are still tigs or scaffolds o chromosomes e or “finished” ry chromosome the human ge- ther animal ge- m assembly, ffects that con- sequence has tive regions are ng to problems cies vary widely g thousands of ir.

Contamination of genomic sequences can be particularly problematic for metagenomic studies. For example, if a genome la- beled as species X contains fragments of the human genome, then any sample containing human DNA might erroneously be identi- fied as also containing species X. Since human DNA is virtually al- ways present in the environment of sequencing laboratories, human contamination is very common in sequencing experi- ments 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, Perteu 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. Severin² and Hridesh Rajan¹

¹Department of Computer Science and ²Genome 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 sequencing data being deposited into public repositories is increasing rapidly. Public databases are becoming a primary source of data for each submission that is prone to user error. Unfortunately, most public databases rely on user input and do not have methods for identifying errors in the data. This can lead to error propagation. Previous research on a small subset of the NR database found that the error rate was 7.8% at the species level. To the best of our knowledge, the amount of misclassification in the NR database is much higher. We propose a heuristic method to detect potentially misclassified taxa in the NR database. We use a technique and quality control to find the most probable misclassification for each annotation.

Results: We found more misclassified taxa in the NR database. Using simulated data, we show that the proposed method can detect misclassified proteins. The proposed approach can be used to identify misclassified proteins.

Availability and implementation: The code and Docker container are available at <https://github.com/hbagheri/misclassification>.

Contact: hbagheri@iastate.edu

Supplementary information: Supplementary information is available at <https://doi.org/10.1093/bioinformatics/btaa586>.

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 soybean cross, then all sequences derived from that cross will be labeled

~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 Kapon¹, 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

Abstract

Information theoretic approaches have been used in a wide variety of bioinformatics applications. In comparative genomics, methods, based on short DNA words, or *k*-mers, are used to analyze the informational properties of genomes. *k*-mers of varying lengths for genome comparison have been used in a variety of applications, including genome classification, genome clustering, and genome distance estimation. We analyze the informational properties of genomes, comparative genomics and taxonomy. We use a 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>



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

Received: April 30, 2021

Accepted: October 2, 2021

Published: October 14, 2021

Copyright: © 2021 Bussi et al. This is an open

~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

Filters

Download Select columns 9 Genomes Rows per page 100 1-9 of 9

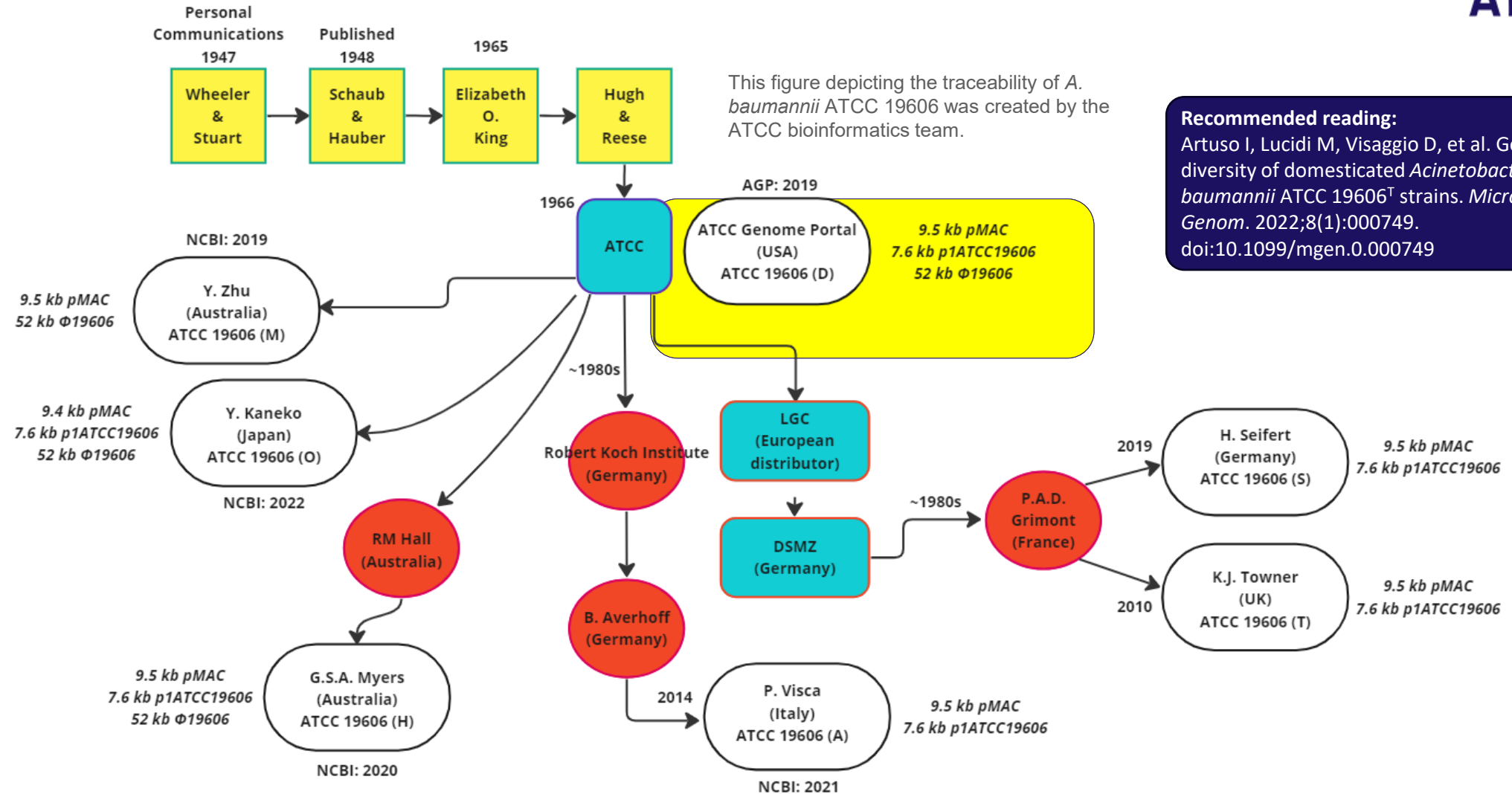
<input type="checkbox"/> Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action
<input type="checkbox"/> ASM1337208v1	GCA_013372085.1	GCF_013372085.1	Acinetobacter baumannii ATCC...	ATCC 17978 (strain)	NCBI RefSeq Submitter	⋮
<input type="checkbox"/> ASM479715v2	GCA_004797155.2	GCF_004797155.2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

- 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?*

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>

Which reference is the “right” reference?

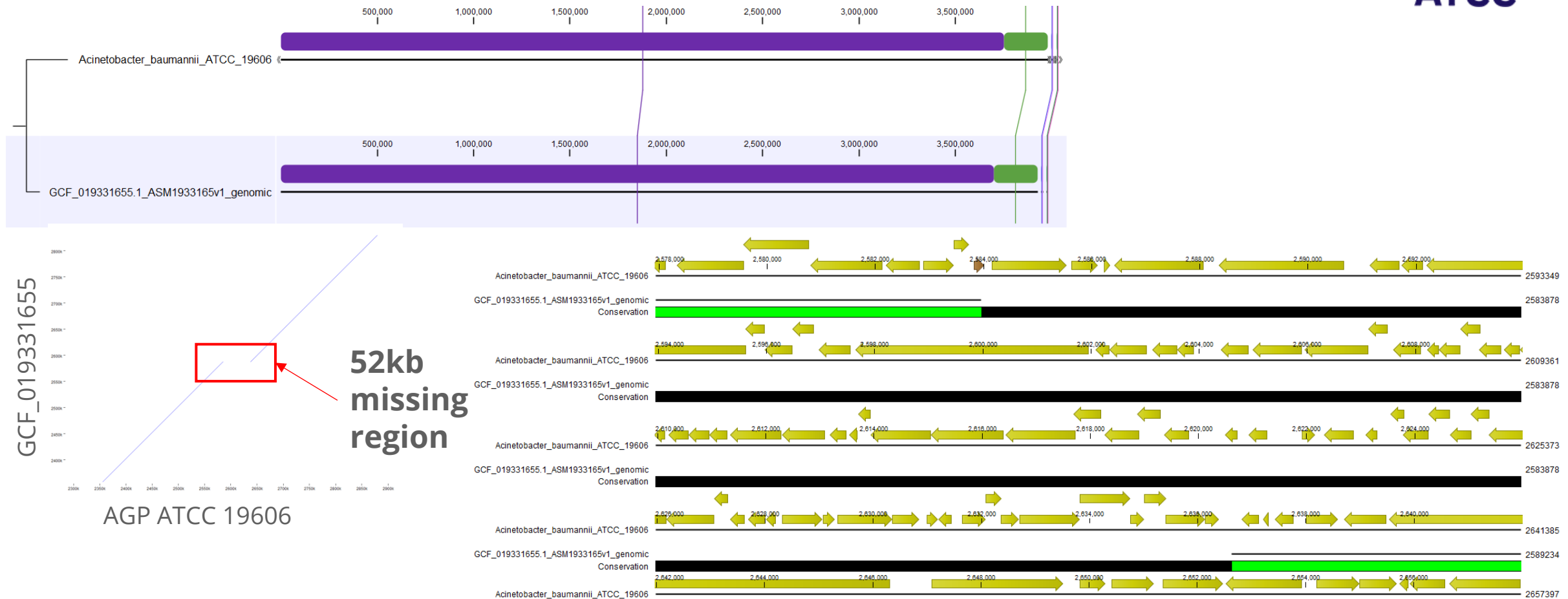


Recommended reading:

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 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.



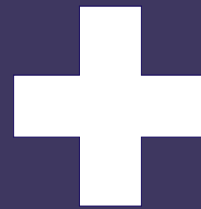
The ATCC® Genome Portal

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

Our solution

Physical Repository

- Strains
- Derivatives
- Standards
- Reference materials



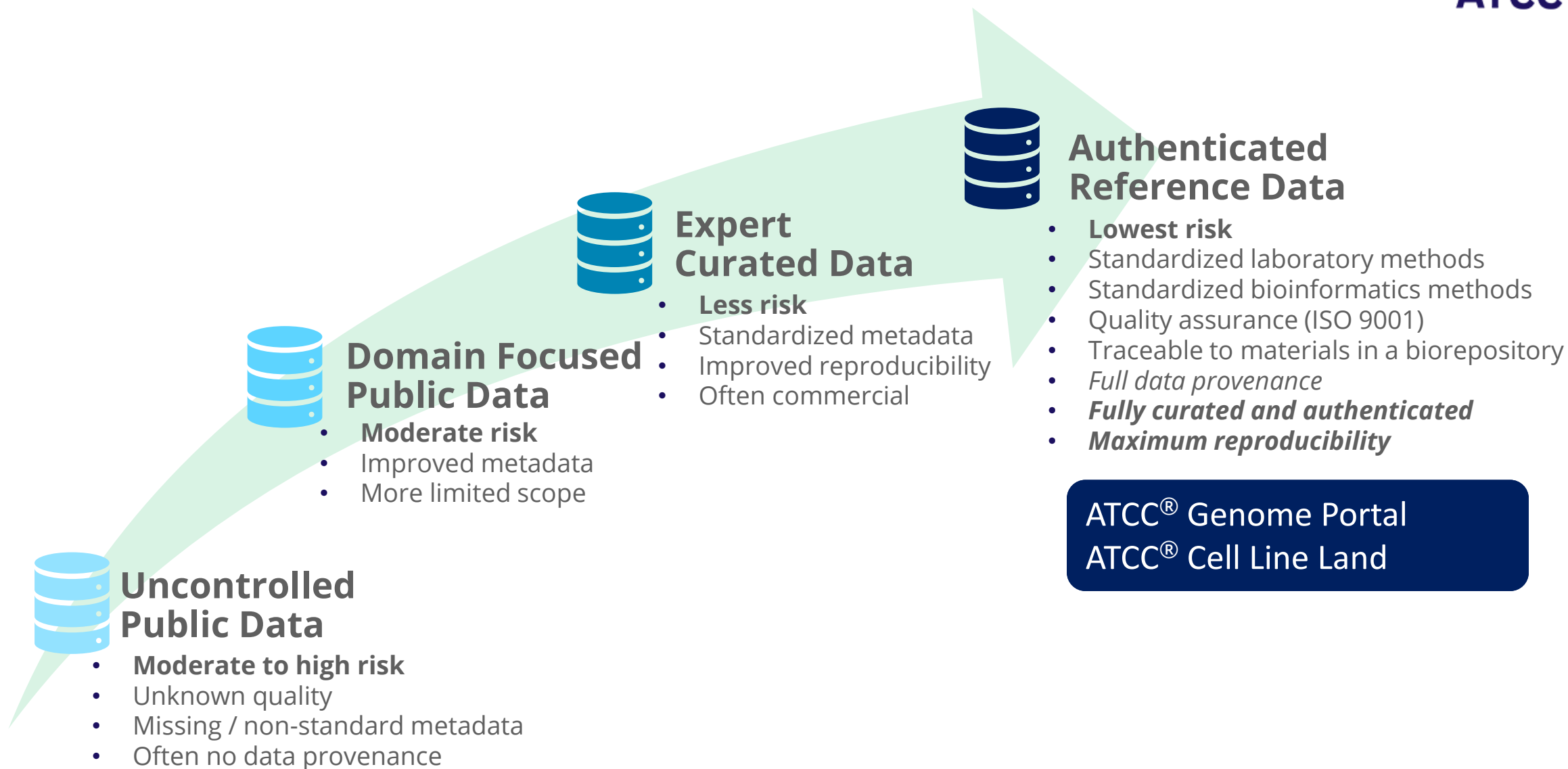
Authenticated Reference Data

- Sequencing data
- Assembled genomes
- Annotated genes

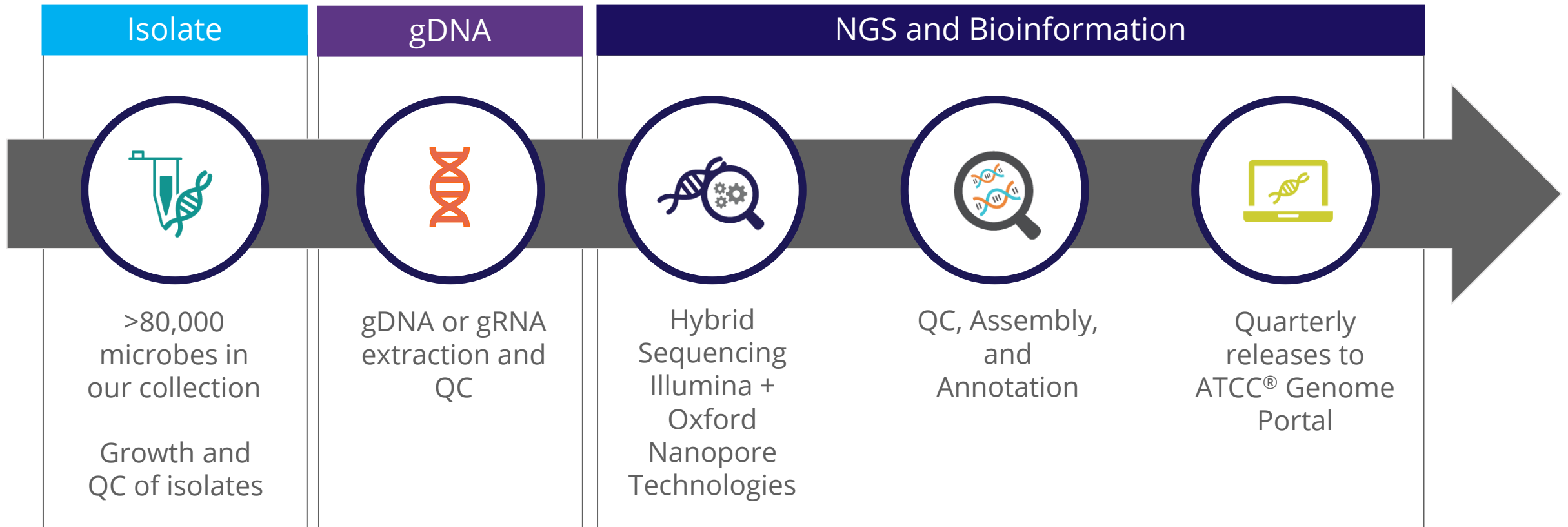


Drive scientific advancement by providing the scientific community with high-quality, annotated whole-genome sequence (WGS) information traceable to ATCC®'s authenticated biological materials.

The arc of database quality in genomics



Authenticated physical material coupled with reference-quality genome sequences

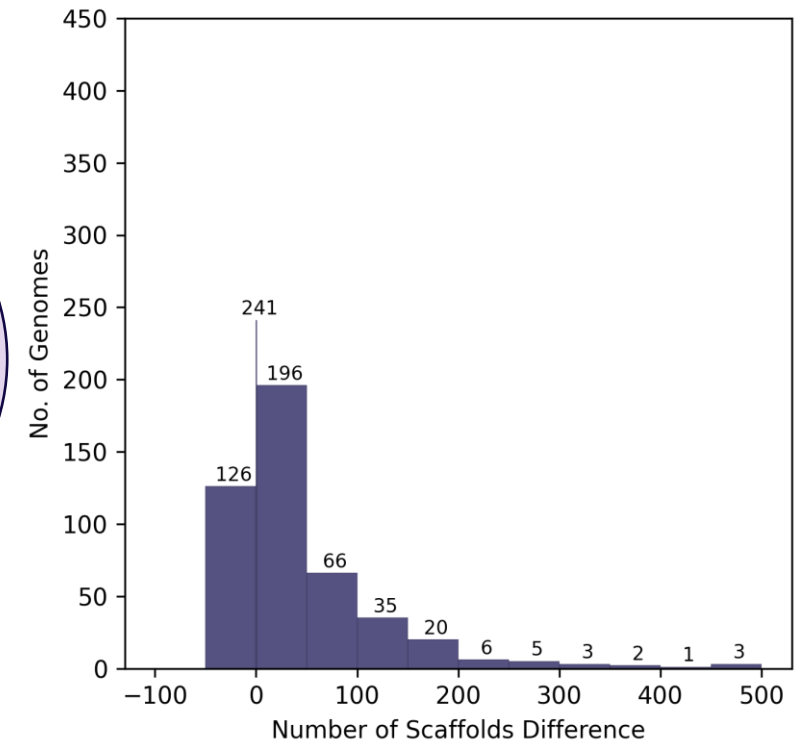
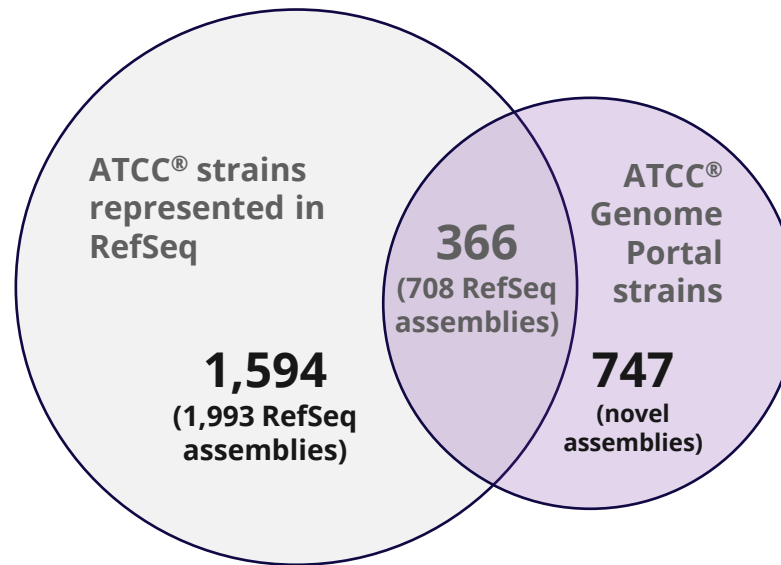
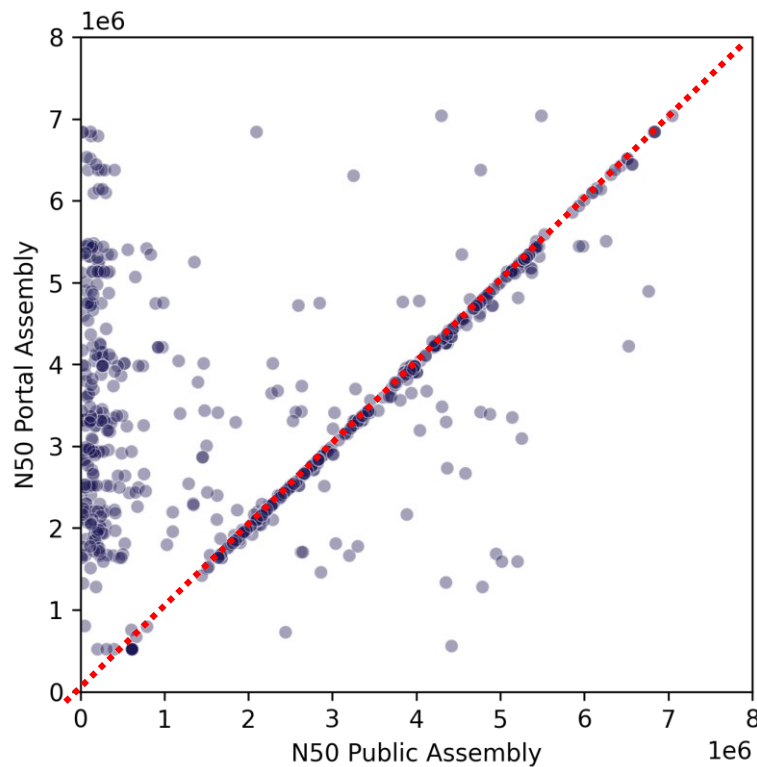


- Fully traceable and authenticated to ATCC® materials
- All genome assemblies produced in-house at ATCC® in an ISO-certified laboratory

Comparison of ATCC® vs. NCBI RefSeq bacterial assemblies



>98% of our assemblies were more complete and of higher quality than RefSeq



Yarmosh DA, Lopera JG, Puthuveetil NP, et al. Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies. mSphere 7(3): e0007722, 2022. PubMed: 35491842

Overview of our process

ATCC Repository



ATCC Repository
Retrieval from ATCC collection



Bioproduction & QC
ATCC® Standards culturing and qc
■ VITEK 2 ■ 16S analysis



Data Provenance
ISO-9001 batch number and metadata authentication of vials

Sequencing & Bioinformatics Center (SBC)



Nucleic Acid Extraction
Automated extraction and QC of HMW genomic DNA



NGS Library Prep
Illumina and ONT specific WGS kits
■ dsDNA, dsRNA, ssRNA, ssDNA



NextGen Sequencing
Illumina and ONT specific WGS kits



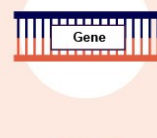
Raw NGS Data
Run output QC and demultiplexing
■ Illumina FASTQ
■ ONT FASTQ/POD5



Trimming / Filtering / QC
Contamination QC, read classification, fastp, NanoFilt



de novo Assembly
Unicycler (Bacteria)
SPAdes (Viruses)
MaSuRCA (Fungi & Protists)



Functional Annotation
■ PGAP, funannotate, VIGA



Assembly QC
■ CheckM / CheckV / Busco
■ Coverage, Phred, and N Bases
■ Length, contigs, circularization



Technical Review
■ Taxonomic analysis consensus
■ Manual QC check

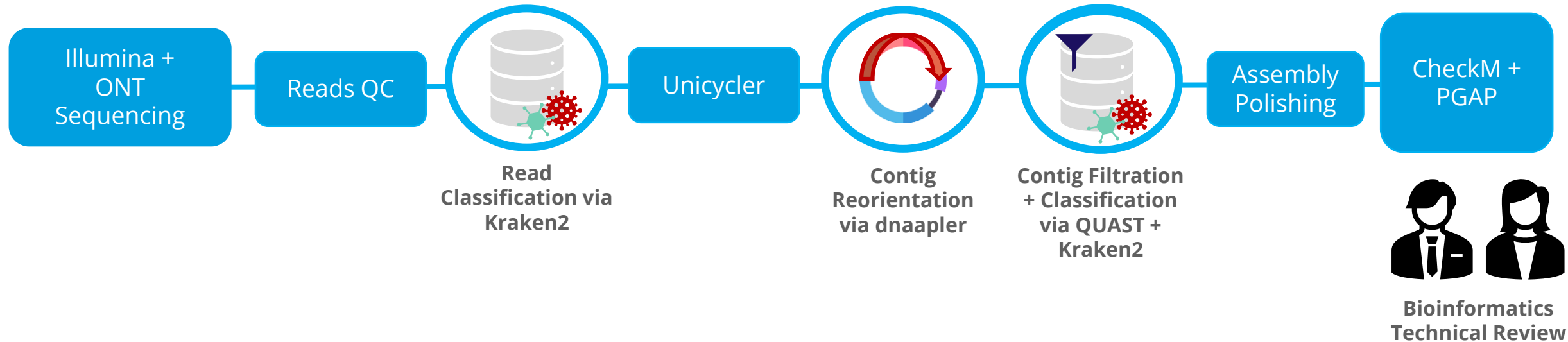
Refer the expanded list of references in the Appendix at the end of the presentation:

1. Simão FA, et al. Bioinformatics 2015;31:3210–3212.
2. Parks DH, et al. Genome Res 2015;25:1043–1055.
3. Nayfach S, et al. Nat Biotechnol 2021;39:578–585.

4. Palmer JM, Stajich J. Zenodo 2020.
5. Tatusova T, et al. Nucleic Acids Res 2016;44:6614–6624.
6. Bankevich A, et al. J Comput Biol 2012;19:455–477.

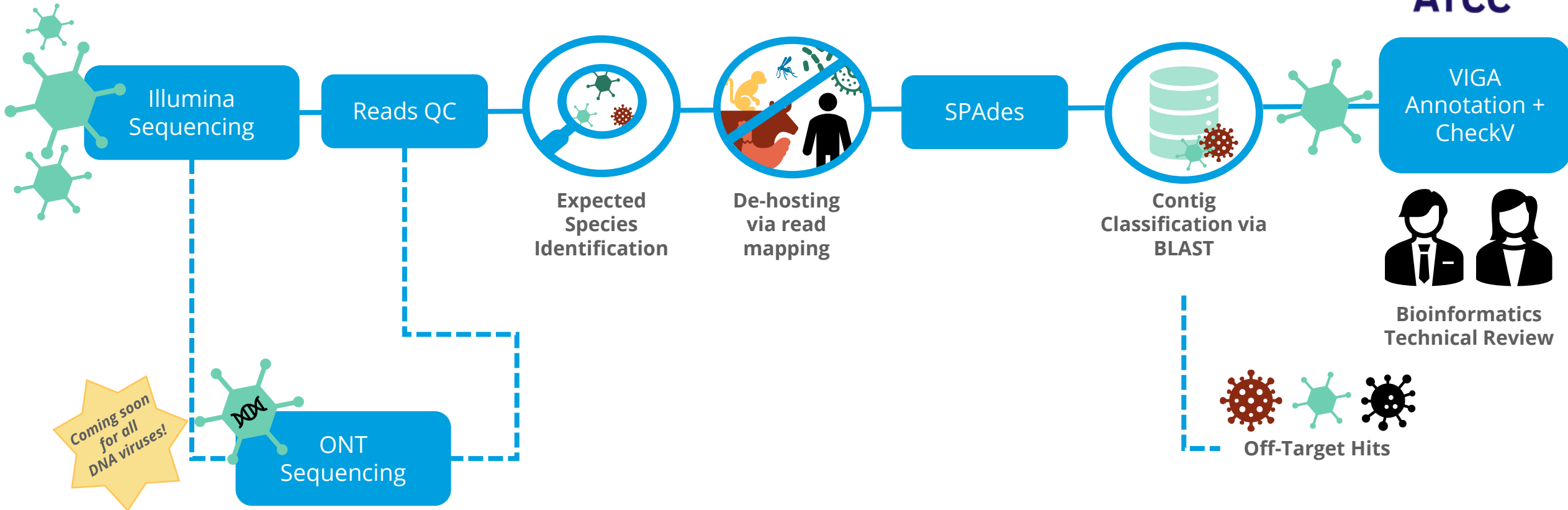
7. Zimin AV, et al. Bioinformatics 2013;29:2669–2677.
8. Wick RR, et al. PLoS Comput Biol 2017;13:e1005595.
9. Fu P, et al. Brief Bioinform 2023;25:bbad444.

Bacterial genome assembly



- All bacterial strains are sequenced on both sequencing platforms.
- NGS reads are trimmed and filtered, assembled using a pipeline built around *Unicycler*, and finally polished.
- Assembly QC is based on CheckM.
- Annotation is based on NCBI's PGAP pipeline.

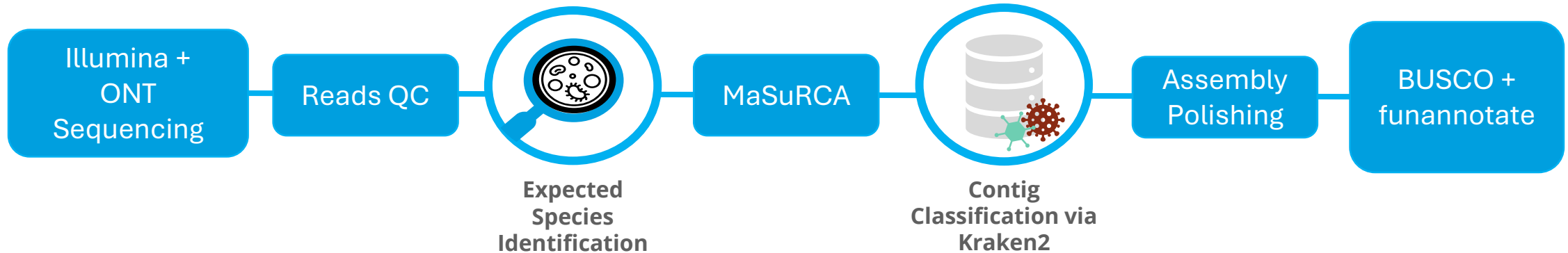
Virology genome assembly



- All viral strains are sequenced on Illumina.
- (soon) DNA viruses are also sequenced on ONT.
- Samples go through an *in silico* de-hosting against each respective genomes for the host cell line used to produce the virus.
- De novo assembly is based around the SPAdes assembler.
- BLAST, VIGA, and CheckV are used for post-assembly QC.

Microbial eukaryotes genome assembly

Fungi and Protists

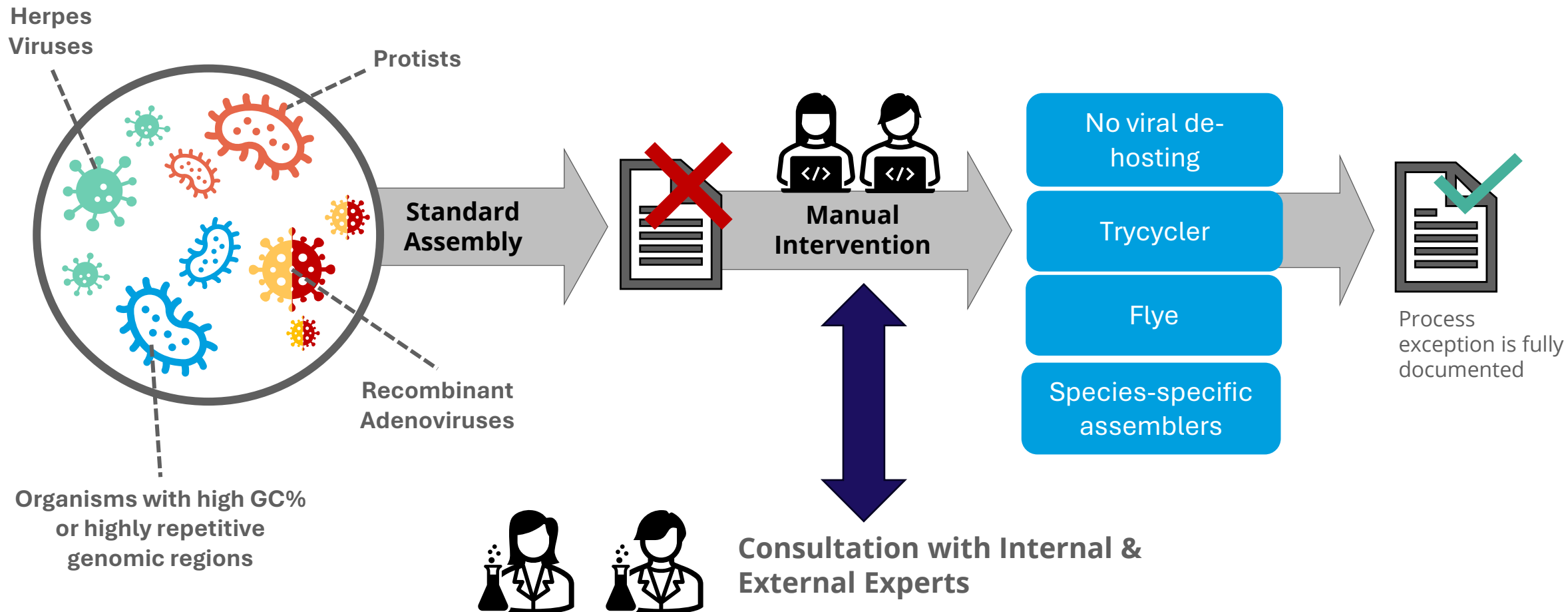


- All microbial eukaryotic samples are sequenced on both Illumina and Oxford Nanopore
- Assembly pipeline is built around the MaSuRCA assembler
- Annotation is done using BUSCO and FunAnnotate

Our “white glove” genome assembly process

Sometimes, genome assemblies need some TLC, but for our group even this is documented and standardized.

Some recent examples:





“Authenticated Genome”

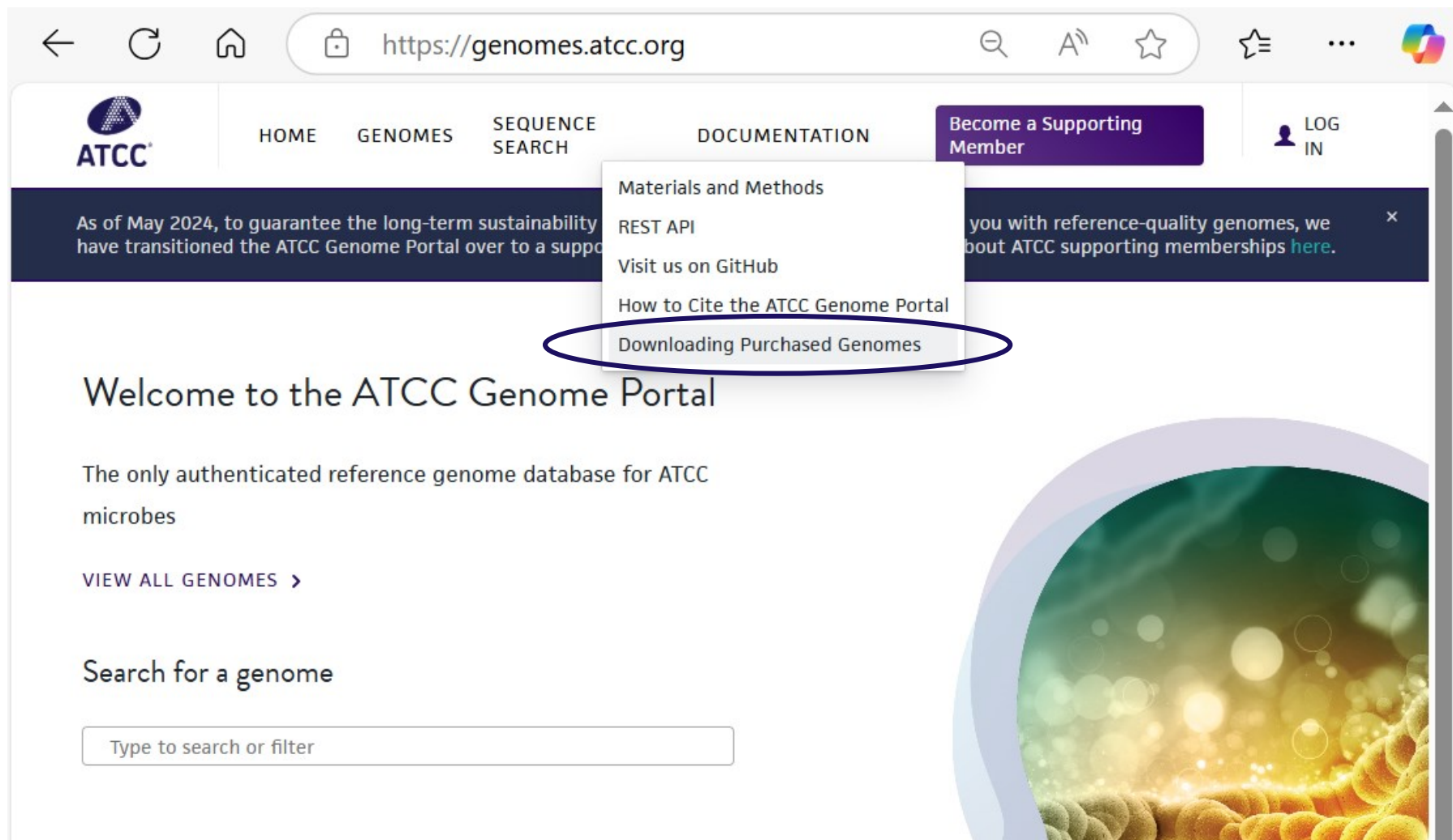
1. ***Traceable*** authenticated materials
2. ***Documented*** methods (i.e., ISO)
3. Exceeds ***standardized*** quality metrics
4. Full end-to-end ***data provenance*** (from materials to data)

The ATCC® Genome Portal

Accessing our data

If customers have purchased the physical product...

Download genome(s) with the lot number



If customers have purchased the physical product...

Download genome(s) with the lot number



The screenshot shows the ATCC Genome Portal website. The URL in the browser is <https://genomes.atcc.org/genomes/79f43b45f79b4abc>. The navigation bar includes links for HOME, GENOMES, SEQUENCE SEARCH, and DOCUMENTATION, along with a 'Become a Supporting Member' button and a 'LOG IN' button (circled in red). A banner message states: '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](#).' The main content area is for *Staphylococcus aureus* (ATCC® 6538™). It includes tabs for Overview, Genome Browser, Related Genomes, and Quality Control. Below the tabs are three buttons: 'DOWNLOAD ASSEMBLY', 'DOWNLOAD ANNOTATIONS', and 'RUN DISCREPANCY REPORT'. The page is divided into two summary sections: 'Assembly Summary' and 'Organism Summary'.

Assembly Summary		Organism Summary	
Date Published	August 27, 2019	Name	ATCC® 6538™
Length	2,800,485 nt	Isolation	Lesion
Sequencing Technology	Illumina + Oxford Nanopore Hybrid Assembly	Biosafety Level	2
		Type Strain	No

If customers have purchased the physical product...

Download genome(s) with the lot number

A screenshot of the ATCC Genome Portal website. The browser address bar shows "https://genomes.atcc.org". The website header includes the ATCC logo, navigation links (HOME, GENOMES, SEQUENCE SEARCH, DOCUMENTATION), a "Become a Supporting Member" button, and a "LOG IN" link. A dark blue banner below the header contains a message about the portal's transition to a supporting membership service as of May 2024. The main content area is titled "Welcome to the ATCC Genome Portal" and describes it as "The only authenticated reference genome database for ATCC microbes". Below this is a "VIEW ALL GENOMES >" link and a "Search for a genome" section. In the search section, the input field contains "2001". A dropdown menu is open, showing a table with the following data:

Name
2001
Catalog Number
2001
Tag
2001
Taxonomy

The "Catalog Number" row is circled in blue. To the right of the search section is a large, stylized image of a DNA double helix structure with a glowing yellow and orange center.

If customers have purchased the physical product...

Download genome(s) with the lot number



← ↻ 🏠 <https://genomes.atcc.org/genomes?text=candida%20...> 🔍 🗨️ ☆ ☆ ⋮ 🌈

HOME GENOMES SEQUENCE SEARCH DOCUMENTATION [Become a Supporting Member](#) 👤 LOG IN

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

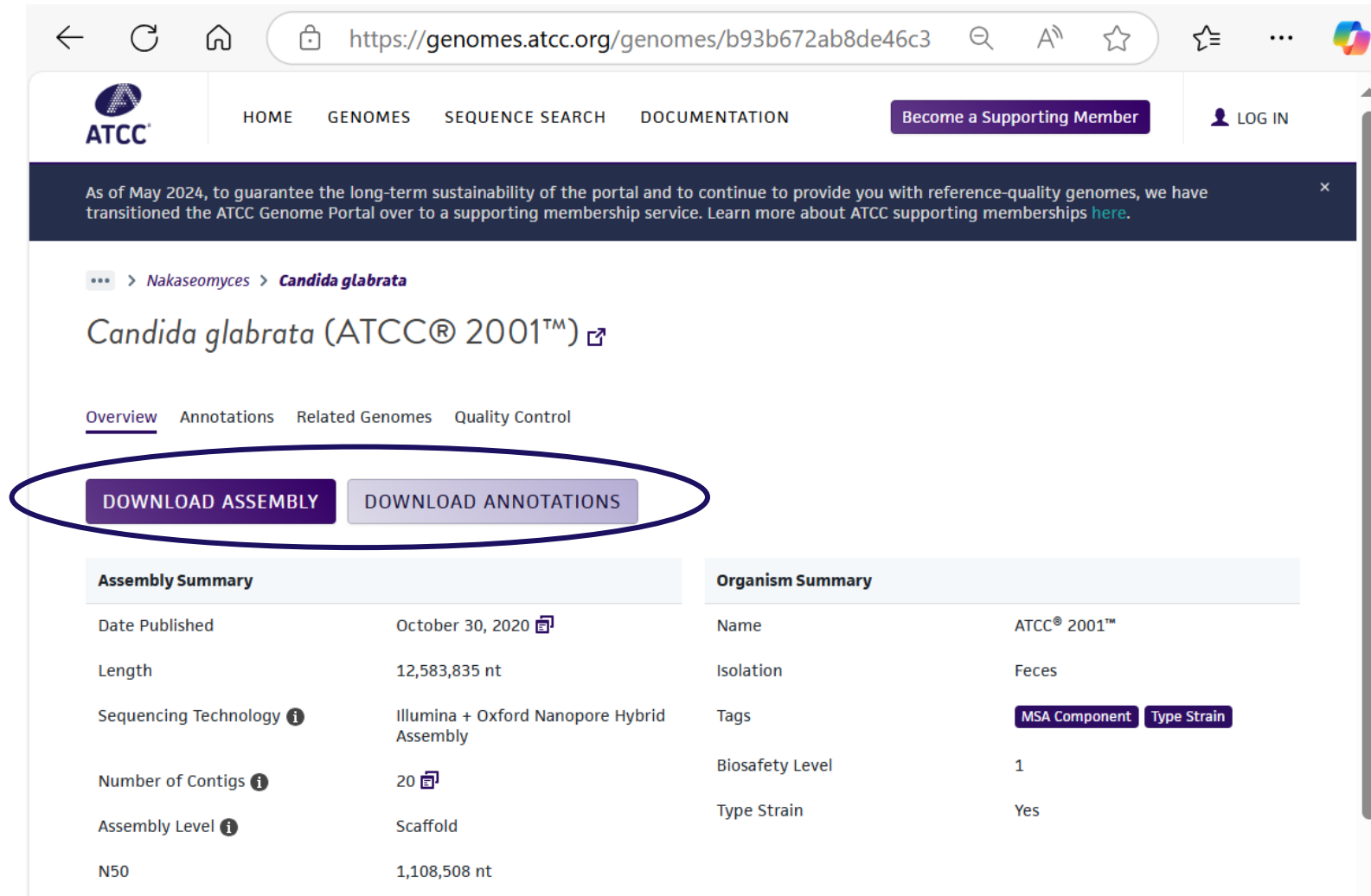
All Genomes [Name: candida glabrata ✕](#)

Taxonomic Name ▲	ATCC Product Name ⇅	Date Published ⇅	Length	Download	Genomic Data
<i>Candida glabrata</i>	ATCC® 15545™ 🔗	December 22, 2020	12.6 Mb	⬇ Download	View
<i>Candida glabrata</i>	ATCC® 36909™ 🔗	October 30, 2020	12.6 Mb	⬇ Download	View
<i>Candida glabrata</i> MSA Component Type Strain	ATCC® 2001™ 🔗	October 30, 2020	12.6 Mb	⬇ Download	View
<i>Candida glabrata</i>	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



The screenshot shows the ATCC Genomes Portal interface. The browser address bar displays the URL <https://genomes.atcc.org/genomes/b93b672ab8de46c3>. The ATCC logo is in the top left, and navigation links for HOME, GENOMES, SEQUENCE SEARCH, and DOCUMENTATION are in the top center. A purple button labeled 'Become a Supporting Member' and a 'LOG IN' link are on the right. A dark blue banner at the top contains a message about the portal's transition to a supporting membership service as of May 2024.

The main content area shows the breadcrumb path: [...](#) > [Nakaseomyces](#) > [Candida glabrata](#). Below this, the title 'Candida glabrata (ATCC® 2001™)' is displayed with an external link icon. A tabbed interface shows 'Overview' as the active tab, with other tabs for 'Annotations', 'Related Genomes', and 'Quality Control'.

Two buttons are visible: 'DOWNLOAD ASSEMBLY' (highlighted with a blue circle) and 'DOWNLOAD ANNOTATIONS'. Below these buttons are two summary tables.


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






*** > Nakaseomyces > *Candida glabrata*

Candida glabrata (ATCC® 2001™) 

Overview Annotations Related Genomes Quality Control


DOWNLOAD ASSEMBLY DOWNLOAD ANNOTATIONS

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

Cancel View Plans



Access to the entire database...purchase a *Supporting Membership*



Explore our annual Supporting Membership opportunities

	Free	Individual	Research Group	Institution
View organism and genome metadata, assemblies, and annotations	✓	✓	✓	✓
Search for genomes of interest	✓	✓	✓	✓
Purchase the corresponding authenticated ATCC source materials	✓	✓	✓	✓
Download genome assemblies and annotations	Only for purchased products	All products	All products	All products
Access our secure REST-API	Not available	✓	✓	✓
Analyze isolates with Discrepancy Reports	Fee for each report	12 free reports per year	60 free reports per year	Inquire
Members with full access	0	1	5	Unlimited

\$600 / \$1,800


\$2,400 / \$7,200

Inquire

✉ nextgen@atcc.org

Access to the entire database...



HOME GENOMES SEQUENCE SEARCH DOCUMENTATIONLOG IN

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

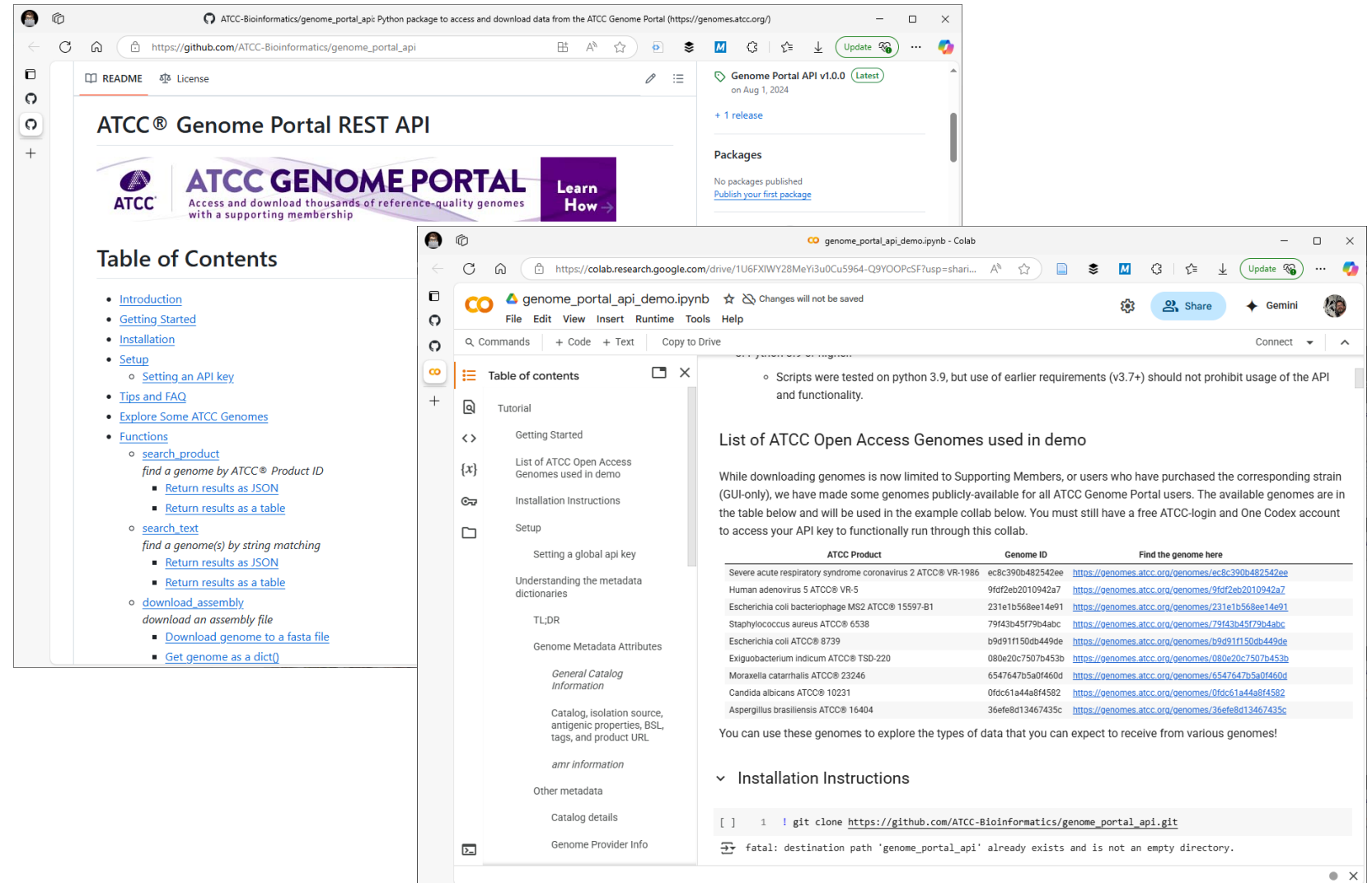
ATCC has partnered with [One Codex](#) to provide access to premium ATCC Genome Portal content and easy to use analyses on One Codex. [Learn more about the One Codex and ATCC partnership](#). If you wish to use a Purchase Order for your Supporting Membership, please contact nextgen@atcc.org.

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
Sign Up	Log In To View Pricing	Log In To View Pricing	Log In To View Pricing
<ul style="list-style-type: none">✓ View organism and genome metadata, assemblies, and annotations✓ Search for genomes of interest	<ul style="list-style-type: none">✓ View organism and genome metadata, assemblies, and annotations✓ Search for genomes of interest✓ Download genome assemblies and annotations✓ Access the REST API✓ Analyze isolates with 12 Discrepancy Reports included✓ 1 seat	<ul style="list-style-type: none">✓ View organism and genome metadata, assemblies, and annotations✓ Search for genomes of interest✓ Download genome assemblies and annotations✓ Access the REST API✓ Analyze isolates with 60 Discrepancy Reports included✓ 5 seats	<ul style="list-style-type: none">✓ View organism and genome metadata, assemblies, and annotations✓ Search for genomes of interest✓ Download genome assemblies and annotations✓ Access the REST API✓ Analyze isolates with Discrepancy Reports (Inquire for details)✓ 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.



The image shows two side-by-side browser windows. The left window displays the GitHub repository for the ATCC Genome Portal REST API. The right window shows a Google Colab notebook titled 'genome_portal_api_demo.ipynb' which includes a table of ATCC Open Access Genomes used in the demo.

ATCC® Genome Portal REST API

Access and download thousands of reference-quality genomes with a supporting membership

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\(\)](#)

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 colab below. You must still have a free ATCC-login and One Codex account to access your API key to functionally run through this colab.

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	9fd2eb2010942a7	https://genomes.atcc.org/genomes/9fd2eb2010942a7
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	0f0c61a44a8f4582	https://genomes.atcc.org/genomes/0f0c61a44a8f4582
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

```
[ ] 1 ! 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

The ATCC® Genome Portal



The screenshot shows the ATCC Genome Portal website in a web browser. The address bar displays "https://genomes.atcc.org". The navigation bar includes the ATCC logo, links for HOME, GENOMES, SEQUENCE SEARCH, and DOCUMENTATION, a "Become a Supporting Member" button, and a "LOG IN" link. A dark blue banner at the top contains a message about the portal's transition to a supporting membership service as of May 2024. The main content area features a "Welcome to the ATCC Genome Portal" heading, a description of the portal as the only authenticated reference genome database for ATCC microbes, a "VIEW ALL GENOMES" link, and a "Search for a genome" section with a search input field. Below this, a "Recently published" section lists three entries: Oropouche virus (ATCC® VR-3446™) added 3/27/2025, Powassan virus (ATCC® VR-1958™) added 2/21/2025, and Dengue virus type 3 (ATCC® VR-3380™) added 2/21/2025. Each entry is accompanied by a small circular icon. The footer of the page states "Powered by" followed by the ONE CODEX logo.

← ↻ 🏠 <https://genomes.atcc.org> 🔍 🔊 ☆ ⌵ ⋮

ATCC® HOME GENOMES SEQUENCE SEARCH DOCUMENTATION [Become a Supporting Member](#) 👤 LOG IN

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

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**

Browse for your data



[HOME](#) [GENOMES](#) [SEQUENCE SEARCH](#) [DOCUMENTATION](#)

Complete Collection

Bacteriology Collection

Mycology Collection

Protistology Collection

Virology Collection

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

[HOME](#) [GENOMES](#) [SEQUENCE SEARCH](#) [DOCUMENTATION](#)

[Become a Supporting Member](#)

JJACOBS@ATCC.ORG

Genomes

[All Genomes](#) [My Genomes](#)

Type to search or filter

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

Fast sequencing search

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



HOME GENOMES SEQUENCE SEARCH DOCUMENTATION

Become a Supporting Member

LOG IN

Search for a genome

CTGCTAAGGTTAAAAGAGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCA
AAAAAGAACTATTAAATCCACAATTTACTTTTTCTACTATTTGTTGAAGGCCGTTCTAACCA
AATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCCTT
TGTTTTTATATGGCCCGACAGGTCTAGGTAAGACTCAGTTAATGCAAGCAGTTGGTAATGCC
TTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAAGTTTGTACAAGA
TTTTGTGAGCTCATTACAAAAAGGAAAGTTGAAGAGTTTAAAGAAAAATGTCGTTCTTTAG
ACTTGTATTAGTAGATGATATTCATCTTTTGGCAGGGAAAGCAAGCTTGTGTAATTT
TTCTATACATTTAATGCCTTACTATGCAATGATGAATCTAAACAAATTTTAAACGTCAGAT
CGATATCCTAAAGAATTAACAGAACTTGATCCTCGTTTGGTTTCTCGTTTTCTGCGGGGCT
ATCAGTAGGTGTTGAACCACTGATATTGAAACTCGAATCGAAATTCGCTTAAAAAGCTG
AAAATAGTGGCGTTGATTACCTAGAACTGTGCGTTGTTTATTGCCCAACAGTCGTAGCG
AACGTACGTGAACCTGAGGGTGCACTGAATAAAGTTGTCGCAATTTACGTTTTAAAGGTGC
TCCAATTGATCTTGATGTCGTACGGGAATCTTTAAAGATGTTTTAGCGATCCGTGCTCGTA
CAATTAGGTAGAAAAATATCCAGCGTGTAGTGAGTGAATATTTCCGAATTCATTAAGAGAG
CTGGTAGGTCCAAAGCGTACCGAATTTATGCTCGACCACTGAGTTGGCGATGGGGCTTGC
CCGTGAATTAACGGGGATAGTTTCTGAAATGGAATGGCTTTTGGTGGCGTGATCACA
GTACAGTGATGATGCTTGTGAAAAAGTCGTAGTTTACGGGAAGAAAGACCAATCTTTGAT
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.

Q Search

Results on 1403 bases

Acinetobacter baumannii (ATCC® 19606™)

1398 bases matched (100.00%)

3 contigs

4.0 Mb

View Genome

Acinetobacter baumannii (ATCC® 19187™)

1398 bases matched (100.00%)

4 contigs

4.0 Mb

View Genome

Acinetobacter baumannii (ATCC® 17961™)

1398 bases matched (100.00%)

5 contigs

4.0 Mb

View Genome

Acinetobacter baumannii (ATCC® BAA-2887™)

1395 bases matched (99.00%)

3 contigs

4.0 Mb

View Genome

Search results are almost instantaneous!

ATCC® Genome Portal: Reference Genome Details

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



Overview page

HOME GENOMES SEQUENCE SEARCH DOCUMENTATION

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

Acinetobacter baumannii (ATCC® 19606™)

Overview Genome Browser Related Genomes Quality Control

DOWNLOAD ASSEMBLY

DOWNLOAD ANNOTATIONS

RUN DISCREPANCY REPORT

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%
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

Organism Summary	
Name	Acinetobacter baumannii
Isolation	
Tags	
Biosafety Level	
Type Strain	
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

HOME GENOMES SEQUENCE SEARCH DOCUMENTATION

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

Acinetobacter baumannii (ATCC® 19606™)

Overview Genome Browser Related Genomes Quality Control

Include Hypothetical Proteins

Display All Genes

Filter annotations

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

Annotation Legend

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

View quality control data

HOME GENOMES SEQUENCE SEARCH DOCUMENTATION

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

Acinetobacter baumannii (ATCC® 19606™)

Overview Genome Browser Related Genomes Quality Control

Sequencing Quality Control

Quality control statistics on illumina sequencing data.

4/4

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

3/3

3 out of 3 passed

Passed

Estimated genome completeness

99.63%

Passed

Estimated genome contamination

0%

Passed

Average depth of coverage

243.032x

Find related genomes

HOME GENOMES SEQUENCE SEARCH DOCUMENTATION

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

Acinetobacter baumannii (ATCC® 19606™)

Overview Genome Browser Related Genomes Quality Control

Most similar genomes

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

Acinetobacter baumannii (ATCC® 19187™)

99.99% similar

4 contigs
4.0 Mb

View Genome

Acinetobacter baumannii (ATCC® 15308™)

99.99% similar

5 contigs
4.0 Mb

View Genome

Acinetobacter baumannii (ATCC® 17961™)

99.55% similar

5 contigs
4.0 Mb

View Genome

Acinetobacter sp. (ATCC® 14293™)

98.07% similar

3 contigs
3.8 Mb

View Genome

Acinetobacter baumannii (ATCC® BAA-2871™)

98.04% similar

2 contigs

View Genome

Acinetobacter baumannii (ATCC® BAA-2894™)

98.04% similar

12 contigs

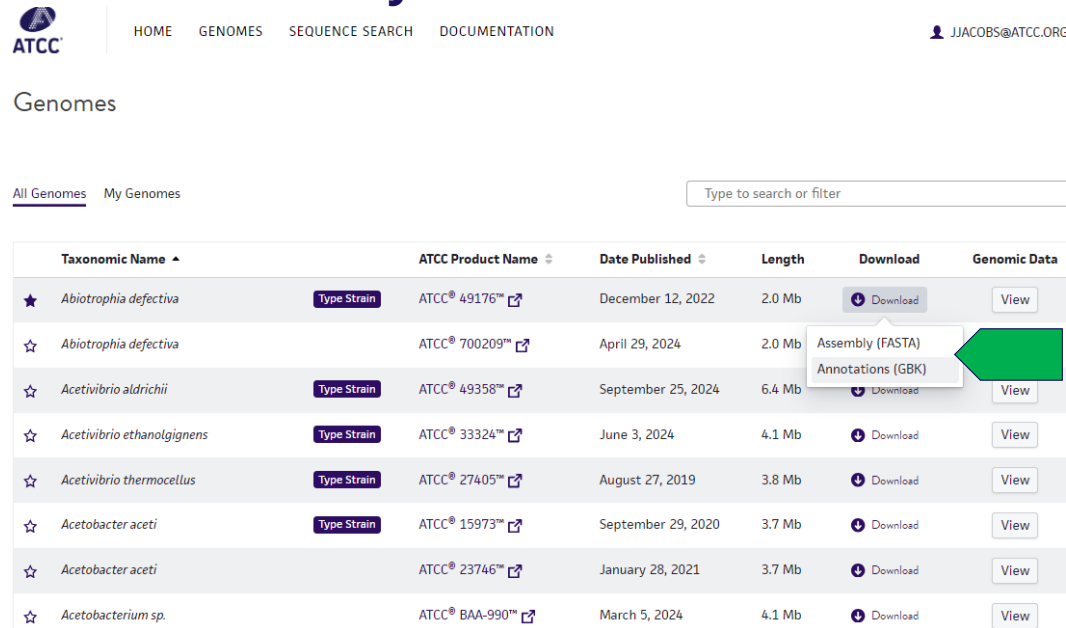
View Genome

© 2025 American Type Culture Collection. ATCC product identifiers marked with the TM symbol are trademarks owned by the American Type Culture Collection.

49

Download genome references

Directly from search results



Taxonomic Name	ATCC Product Name	Date Published	Length	Download	Genomic Data
★ <i>Abiotrophia defectiva</i>	Type Strain ATCC® 49176™	December 12, 2022	2.0 Mb	Download	View
☆ <i>Abiotrophia defectiva</i>	ATCC® 700209™	April 29, 2024	2.0 Mb	Download	View
☆ <i>Acetivibrio aldrichii</i>	Type Strain ATCC® 49358™	September 25, 2024	6.4 Mb	Download	View
☆ <i>Acetivibrio ethanoligignens</i>	Type Strain ATCC® 33324™	June 3, 2024	4.1 Mb	Download	View
☆ <i>Acetivibrio thermocellus</i>	Type Strain ATCC® 27405™	August 27, 2019	3.8 Mb	Download	View
☆ <i>Acetobacter acetii</i>	Type Strain ATCC® 15973™	September 29, 2020	3.7 Mb	Download	View
☆ <i>Acetobacter acetii</i>	ATCC® 23746™	January 28, 2021	3.7 Mb	Download	View
☆ <i>Acetobacterium sp.</i>	ATCC® BAA-990™	March 5, 2024	4.1 Mb	Download	View

Two file formats for download:

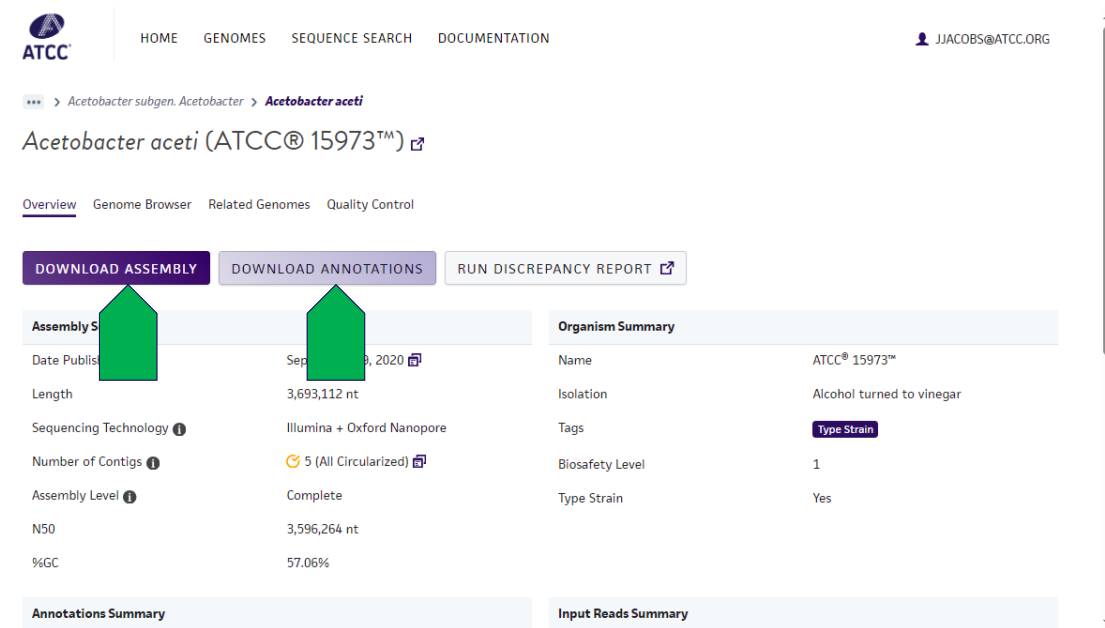
1. FASTA files

- **.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 acetii (ATCC® 15973™)	
Overview	Genome Browser Related Genomes Quality Control
DOWNLOAD ASSEMBLY DOWNLOAD ANNOTATIONS RUN DISCREPANCY REPORT	
Assembly Summary	Organism 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%
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



ATCC

HOME GENOMES SEQUENCE SEARCH DOCUMENTATION JJACOBS@ATCC.ORG

Acetobacter subgen. Acetobacter > **Acetobacter acetii**

Acetobacter acetii (ATCC® 15973™)

Overview Genome Browser Related Genomes Quality Control

DOWNLOAD ASSEMBLY DOWNLOAD ANNOTATIONS **RUN DISCREPANCY REPORT**

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%

Annotations Summary

Organism Summary

Name	ATCC® 15973™
Isolation	Alcohol turned to vinegar
Tags	Type Strain

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.

Cancel **Continue to One Codex**

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.

ONE CODEX

Jonathan Jacobs

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 acetii (ATCC® 15973™)

This website uses cookies to ensure you get the best experience on our website. [Learn more](#) **Got it!**

Summary

Recent publications



Benton B, et al. **The ATCC® Genome Portal: Microbial Genome Reference Standards with Data Provenance.** Microbiology Resource Announcements 10(47): e00818-21, 2021



Yarmosh DA, et al. **Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies.** mSphere 7(3): e0007722, 2022.



Nguyen SV, et al. **The ATCC® Genome Portal: 3,938 authenticated microbial reference genomes.** Microbiology Resource Announcements Epub ahead of print e0104523, 2024.

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 Puthaveetil, MSc.
Joseph Petron, PhD.
Amy Reese, MSc

Scott V Nguyen, PhD
Senior Biocuration Scientist

Our Partner





ATCC[®]

CREDIBLE LEADS TO INCREDIBLE

Questions

Appendix

Open-source tools referenced on slide 29

1. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.
2. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25:1043–1055.
3. Nayfach S, Camargo AP, Schulz F, Elie-Fadrosh E, Roux S, Kyrpides NC. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 39:578–585.
4. Palmer JM, Stajich J. 2020. Funannotate v1.8.1: Eukaryotic genome annotation (v1.8.1). Zenodo.
5. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* 44:6614–6624.
6. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* 19:455–477.
7. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* 29:2669–2677.
8. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 13:e1005595.
9. Fu P, Wu Y, Zhang Z, Qiu Y, Wang Y, Peng Y. 2023. VIGA: a one-stop tool for eukaryotic virus identification and genome assembly from next-generation-sequencing data. *Briefings in Bioinformatics* 25:bbad444.