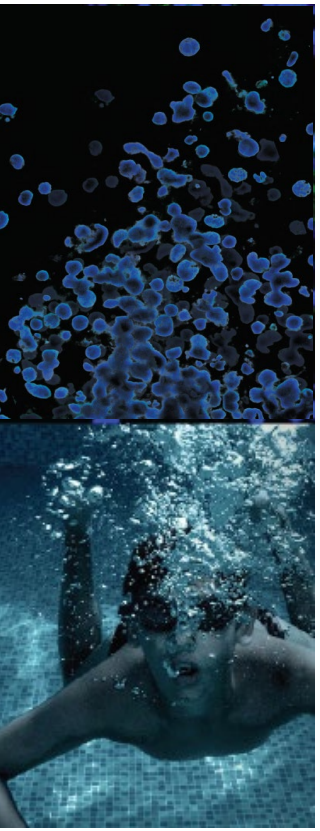
# The ATCC Genome Portal

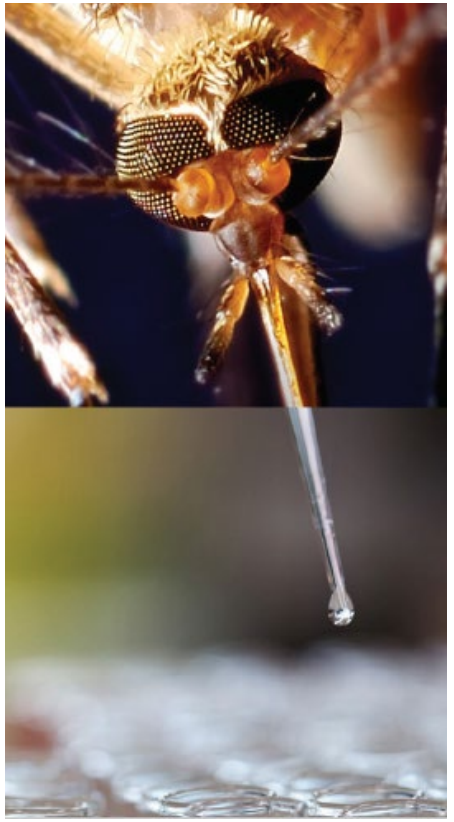
## Authenticity and Traceability for Microbial Genomes



World Microbe Forum 2021
Industry & Science Symposium

Jonathan Jacobs, PhD
Senior Director, Bioinformatics
Principal Scientist
Sequencing and Bioinformatics Center, ATCC

Credible Leads to Incredible™



# Overview

- **The ATCC Genome Portal**
- Traceability and authentication of microbial genomes
- Standards for authenticated reference genomes
- Development roadmap preview

# The ATCC Genome Portal

The ATCC Genome Portal is a cloud-based platform that enables users to easily browse genomic data and metadata by simply logging into the portal

Download whole-genome sequences and annotations of ATCC materials

Search for nucleotide sequences or genes within genomes

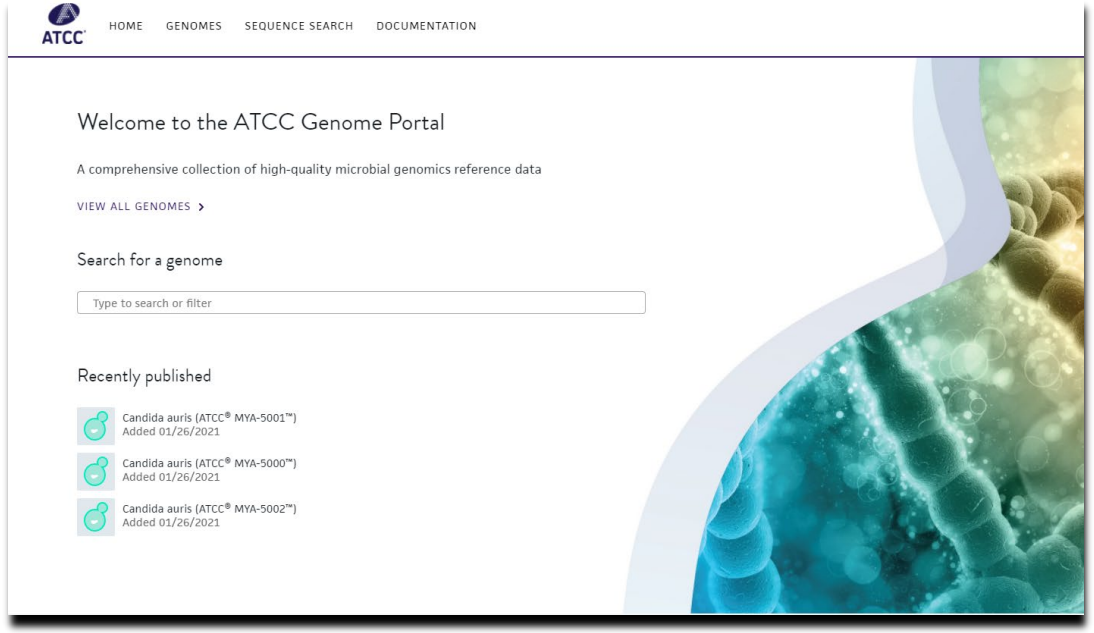
View genome assembly metadata and quality metrics

**genomes.atcc.org**

# The ATCC Genome Portal

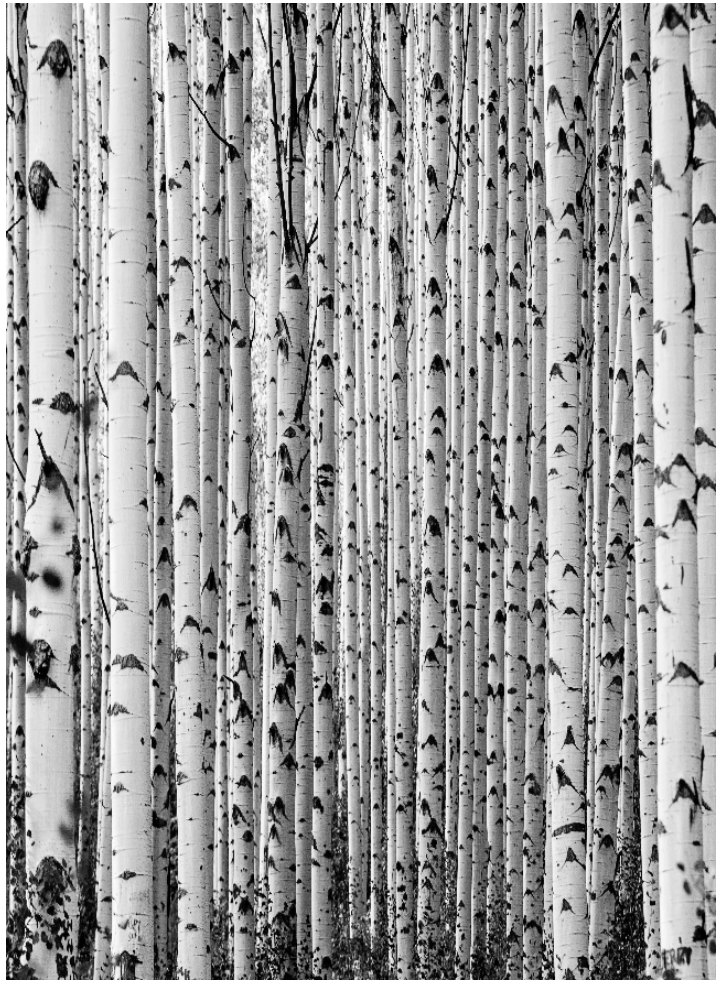
*ATCC's authenticated reference genomes*

- **2017 – 2018** – Planning and proof-of-concept experiments
- **2018** – ATCC Commitment
  - Laboratory, staffing, resources, instrumentation, and bioinformatics development
- **2019** – ATCC Enhanced Authentication Initiative
  - June 2019 – beta launch at ASM Microbe 2019
  - Sept 2019 – formal launch of the ATCC Genome Portal
- **2020+** – Expansion
  - 1200+ authenticated, reference-grade whole-genome assemblies
  - Inclusion of viral, bacterial, and fungal genomes



# Providing reference-quality genomes

*Why - Challenge # 1*



- Public databases routinely host genomic data that is cited as “ATCC,” but…
  - Often no traceability back to genuine ATCC cultures
  - ATCC cannot authenticate 3rd party genomes in public databases
- So, how do researchers *know* which data set to use?
  - Which is the “correct” one?
  - Close enough?
- How do researchers have confidence in their selection?

# Providing reference-quality genomes

*Why - Challenge # 2*

- How do we bring authentication into the genomics era while maintaining our commitment to our customers that we've fully and accurately authenticated our material?
- Typically, authentication* may refer to:
  - Morphology
  - Purity
  - Viability
  - Phenotypic testing
  - Genotypic testing
    - 16S ribosomal gene
    - ITS and D1D2

ATCC

CERTIFICATE OF ANALYSIS

**ATCC® Number:** 12228D-5™
**Lot Number:** 70006950

**Designation:** *Staphylococcus epidermidis* genomic DNA
**Fill Volume Prior to Drying:** 80 µL
**Product Format:** Dried microbial DNA
**Expiration Date:** Not applicable
**Storage Conditions:** 2°C to 8°C

| Test / Method | Specification | Result |
|---|---|---|
| $OD_{260}/OD_{280}$ ratio (Spectrophotometer method) | 1.6 to 2.1 | 1.8 |
| Total amount of DNA (PicoGreen® measurement) | ≥ 5 µg per vial | 6 µg/vial |
| Agarose gel electrophoresis | High molecular weight chromosomal DNA; No visible RNA | High molecular weight chromosomal DNA; No visible RNA<br>See photograph below |
| PCR Functionality | Successful PCR amplification of selected gene(s) | Successful PCR amplification of selected gene(s) |
| Sequencing of selected gene(s) | Consistent with source organism | Consistent with source organism |

1 2

Lane 1 Invitrogen™ TrackIt™ 1 Kb Plus DNA Ladder
Lane 2: 12228D-5™

Quality Assurance Specialist; Quality Assurance
ATCC hereby represents and warrants that the material provided under this certificate is pure and has been subjected to the tests and procedures specified and that the results described, along with any other data provided in this certificate, are true and correct to the

ATCC
10801 University Boulevard
Manassas, VA 20110-2209 USA
www.atcc.org

800-638-6597 or 703-365-2700
Fax: 703-365-2700
E-mail: tech@atcc.org
or contact your local distributor

- Page 1 of 2 -

Template Doc ID: 31194 Template Revision: 4 Template Effective Date: 12/31/2014

*not an inclusive list

# Providing reference-quality genomes

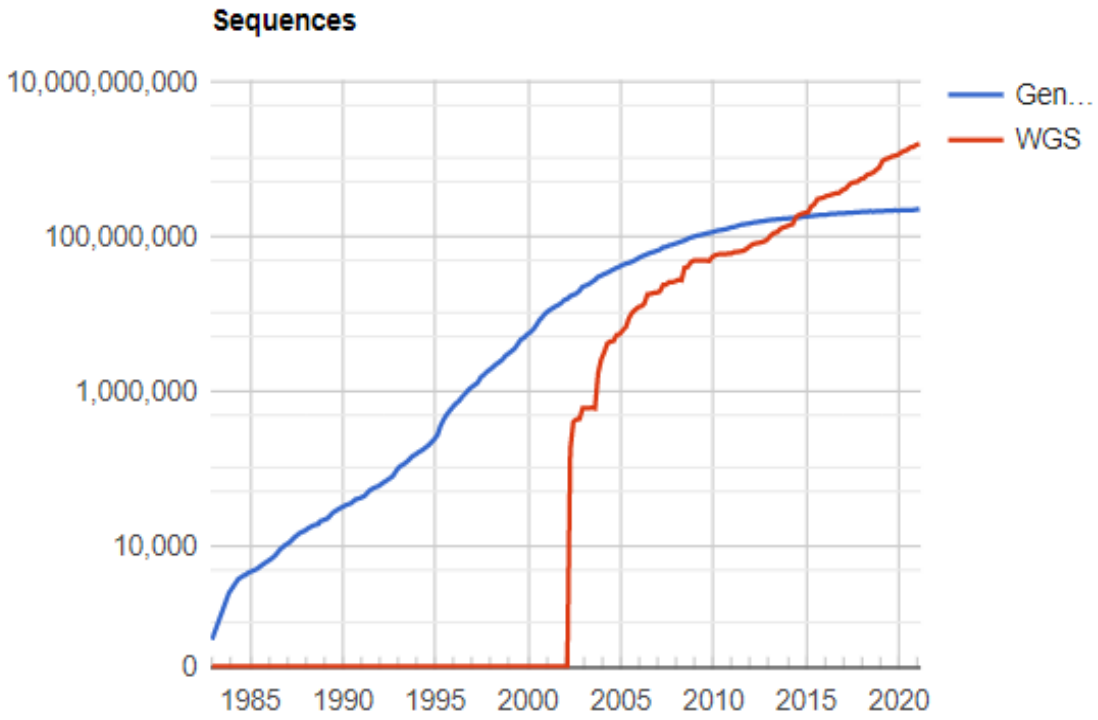
*Why - Challenge # 3*

- Acknowledge there is a problem with reference genomes ✓
- Work through a plan to address the problem ✓
- **How do we effectively and easily provide customers with genomic data while not diluting it or burying it in a public database?**

# Reference genomes

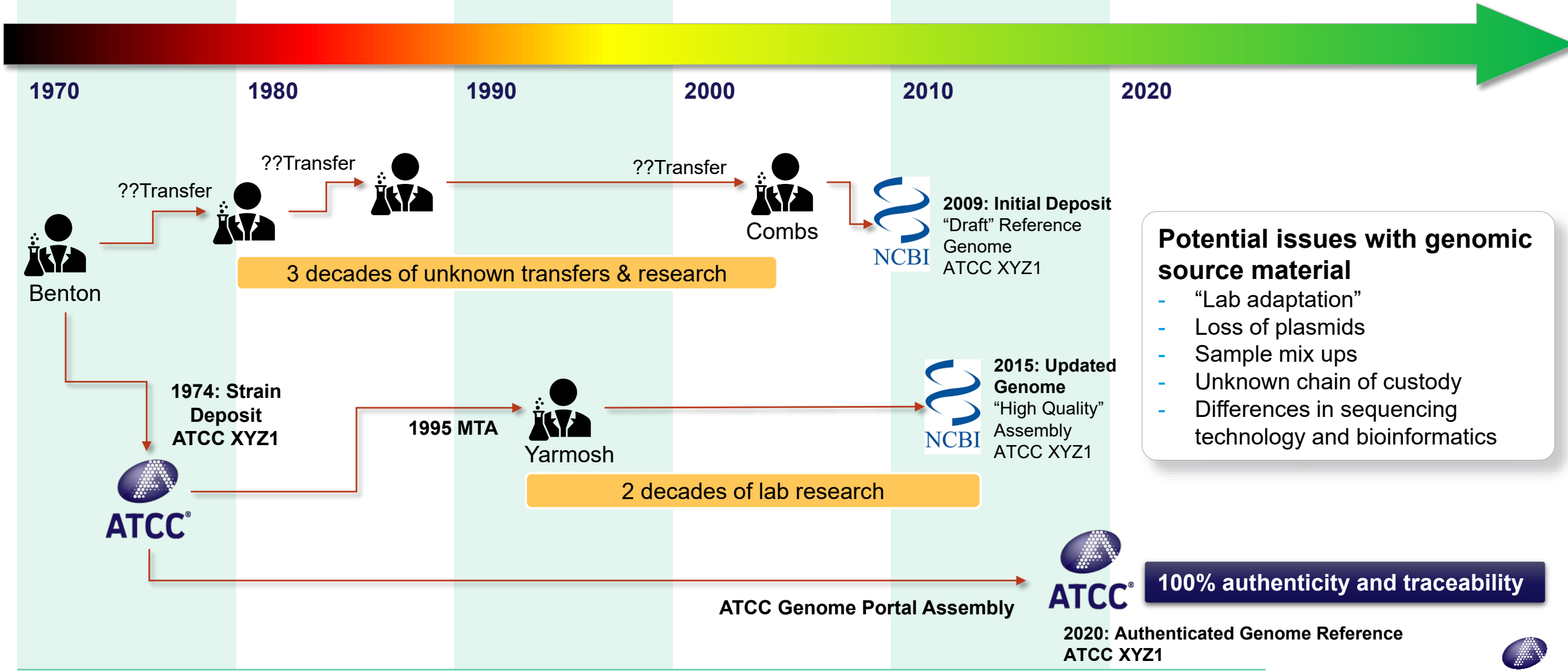
*Where can researchers turn to for "reference" genomes?*

- De facto standard
  - The sequence database for the entire public scientific community
  - Contains numerous genomes
  - Genomes submitted by a variety of labs
- Relatively little curation
- Highly variable quality
- ***NEVER*** authenticated by ATCC




https://www.ncbi.nlm.nih.gov/genbank/statistics/

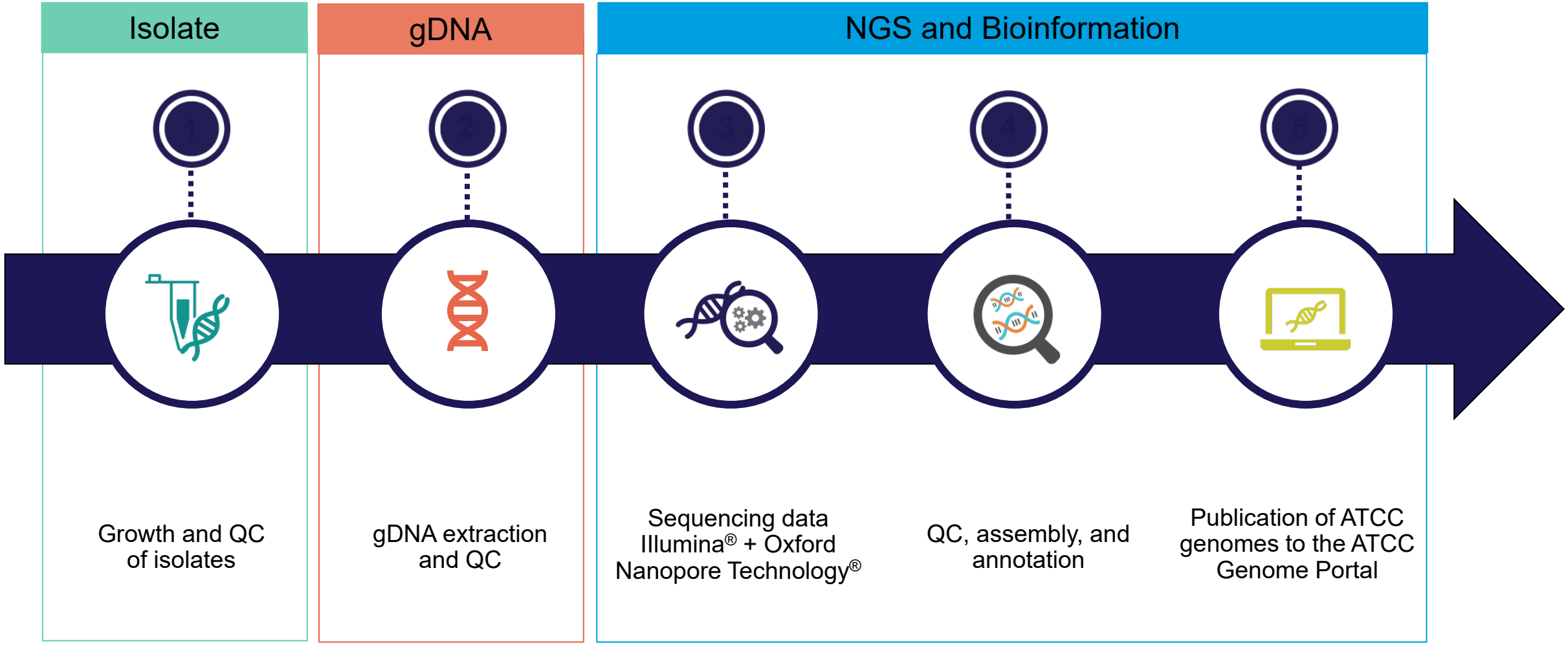
# The elephant in the room: Authenticated reference genomes




**Potential issues with genomic source material**

- "Lab adaptation"
- Loss of plasmids
- Sample mix ups
- Unknown chain of custody
- Differences in sequencing technology and bioinformatics

# Our process: Authenticated physical material coupled with reference-quality genome sequences



# ATCC genome assembly process

ONT / Illumina reads uploaded to One Codex

Read-Based Contamination Quality Control with One Codex*

Genome Assembly

Hybrid Assembly (Bacteria/Fungi)

Illumina-only Assembly (Viruses)

Quality Assessment

- Coverage
- Completeness
- Contamination

Genome Annotation

Calculation of Assembly Level

Publish to the ATCC Genome Portal

Read Trimming

Illumina: fastp

ONT: FiltLong

Read-level *k*-mer based taxonomic classification and estimation of strain abundance

**Bacteria**: Unicycler

**Fungi:** MaSuRCA w/ FLYE

**Viruses:** Taxonomic bining followed by SPADes

**Bacteria:** CheckM

**Fungi:** BUSCO

**Viruses:** curated database*

**Bacteria:** Prokka

**Fungi:** BUSCO

**Viruses**: detect-viral-variants*

Based on NCBI's Assembly Level

- Complete
- Scaffold

* One Codex proprietary software

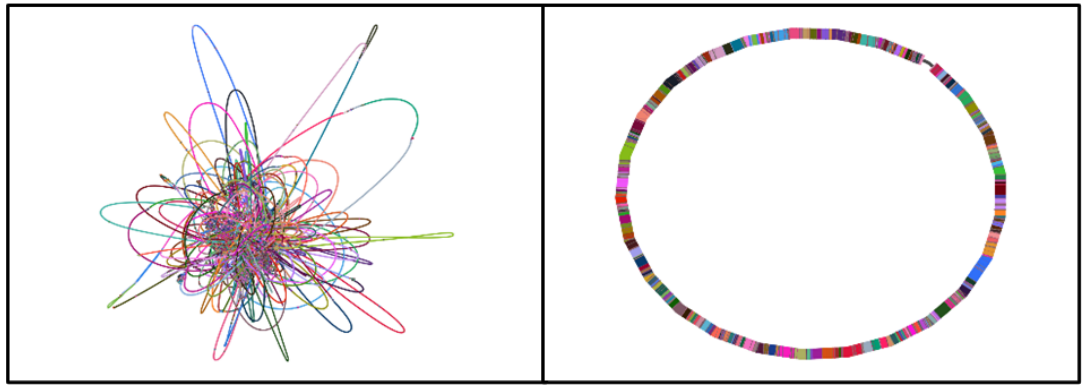


ATCC

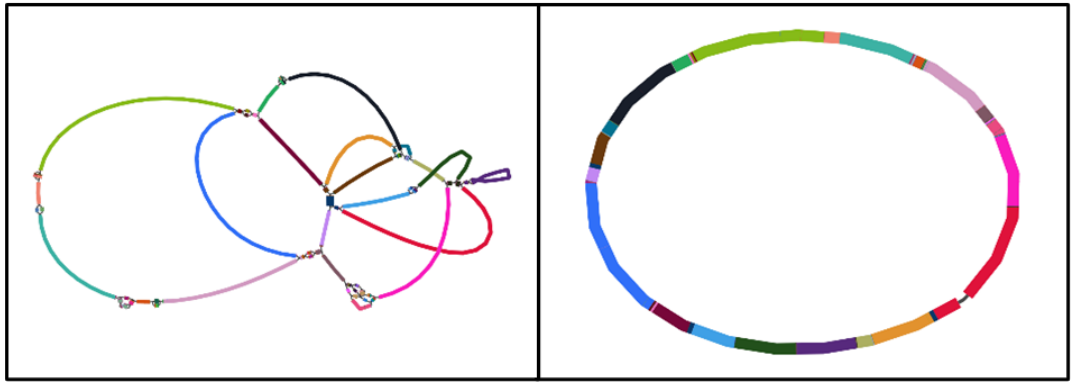
# Advantage of hybrid assemblies

| Illumina-only assembly | Hybrid assembly |
|---|---|

*Neisseria meningitidis* (ATCC® 53417™)

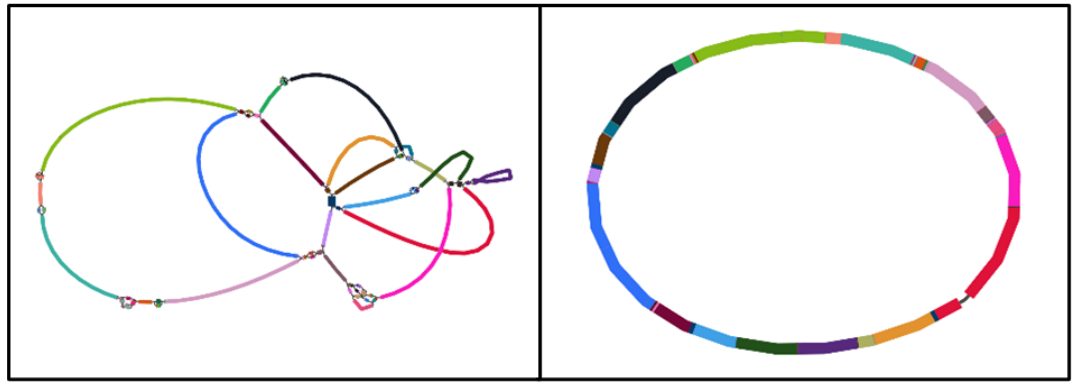


*Campylobacter jejuni* subsp. *jejuni* (ATCC® 43446™)

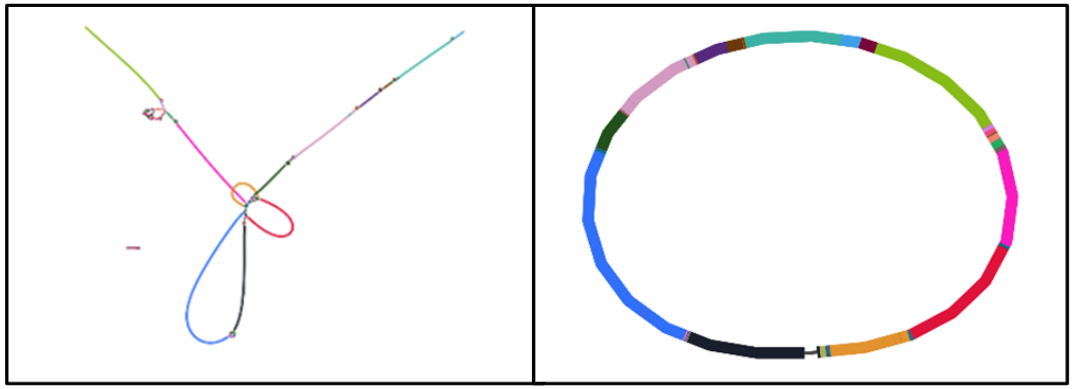


| Illumina-only assembly | Hybrid assembly |
|---|---|

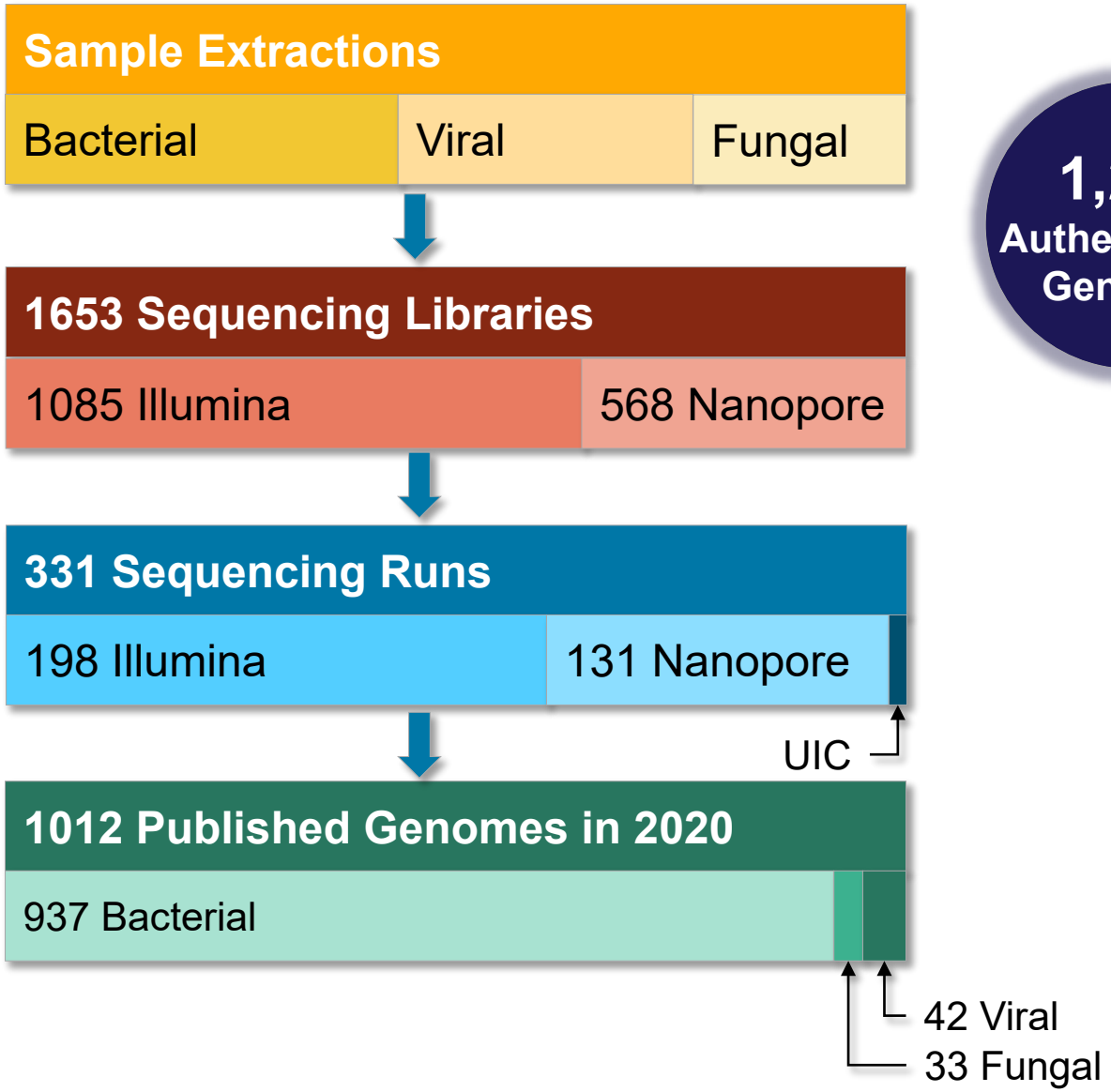
*Campylobacter jejuni* subsp. *jejuni* (ATCC® 43446™)



*Streptococcus gordonii* (ATCC® 35105™)



# The ATCC Genome Portal




**1,251 Authenticated Genomes** ← Today

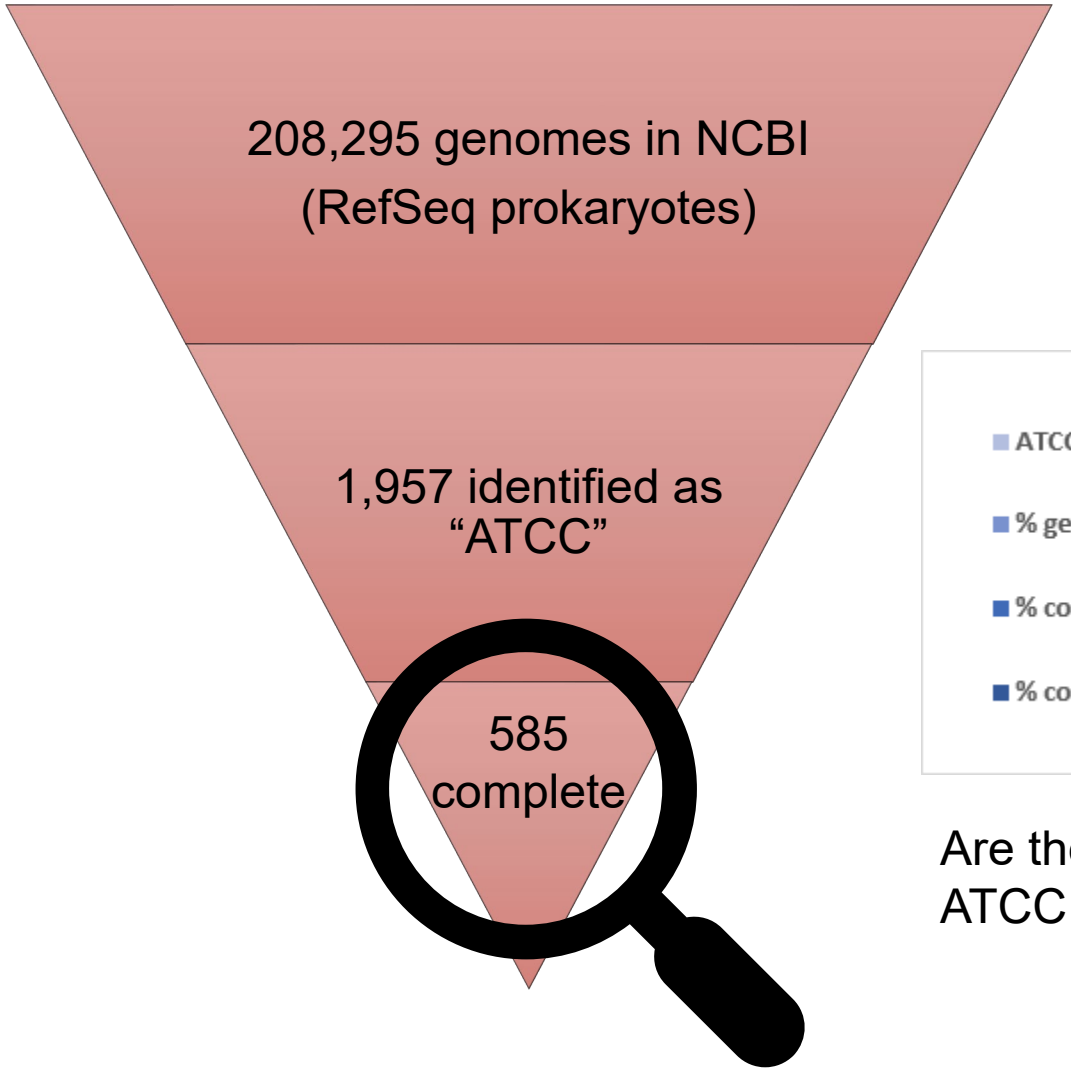
- 1,118 bacterial genomes (739 complete circularized, 391 type strains)
- 59 viral genomes
- 74 mycology genomes

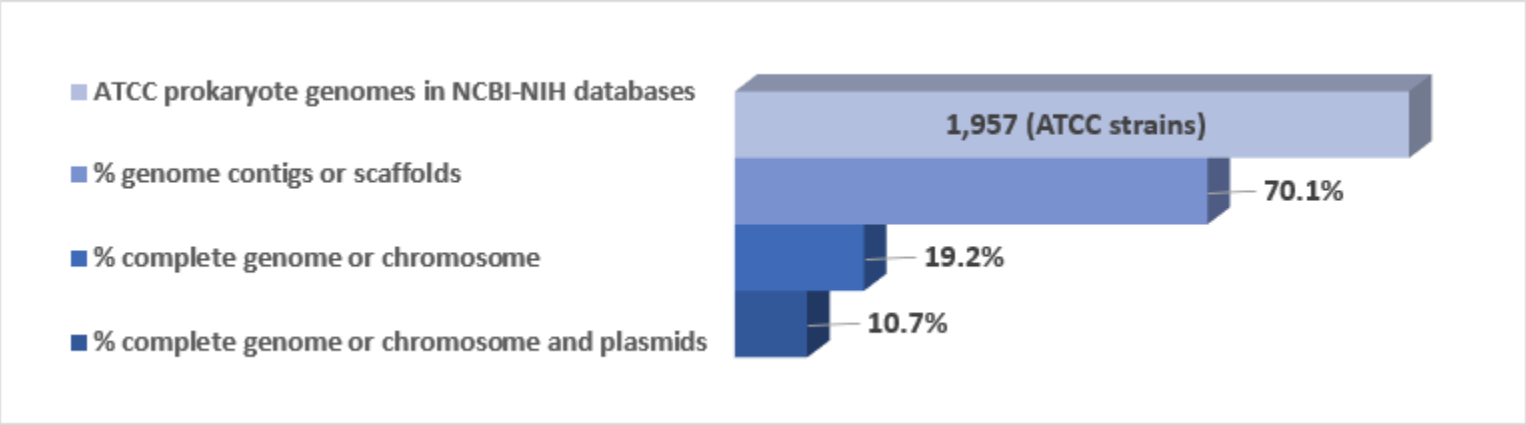
- Monthly updates
- All genomes are traceable to ATCC's biomaterials
- Hybrid assemblies for all bacterial & fungal genomes
- All genomes annotated
- Additional improvements to fungal and viral genome annotations coming

# Overview

- The ATCC Genome Portal
- **Traceability and authentication of reference genomes**
- Standards for authenticated reference genomes
- Development roadmap preview

# Reference genomes

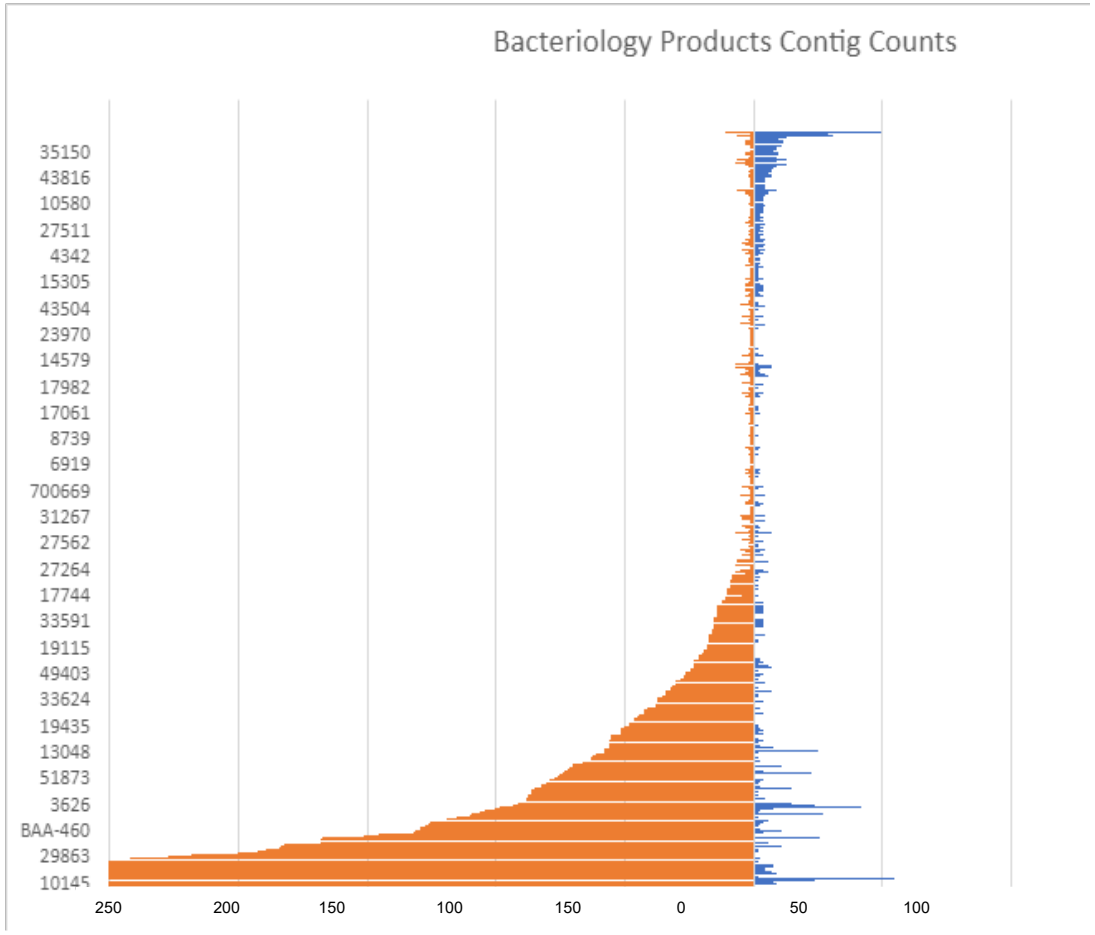



Are these 585 RefSeq genomes traceable back to authenticated ATCC cultures with well-documented growth and storage conditions?

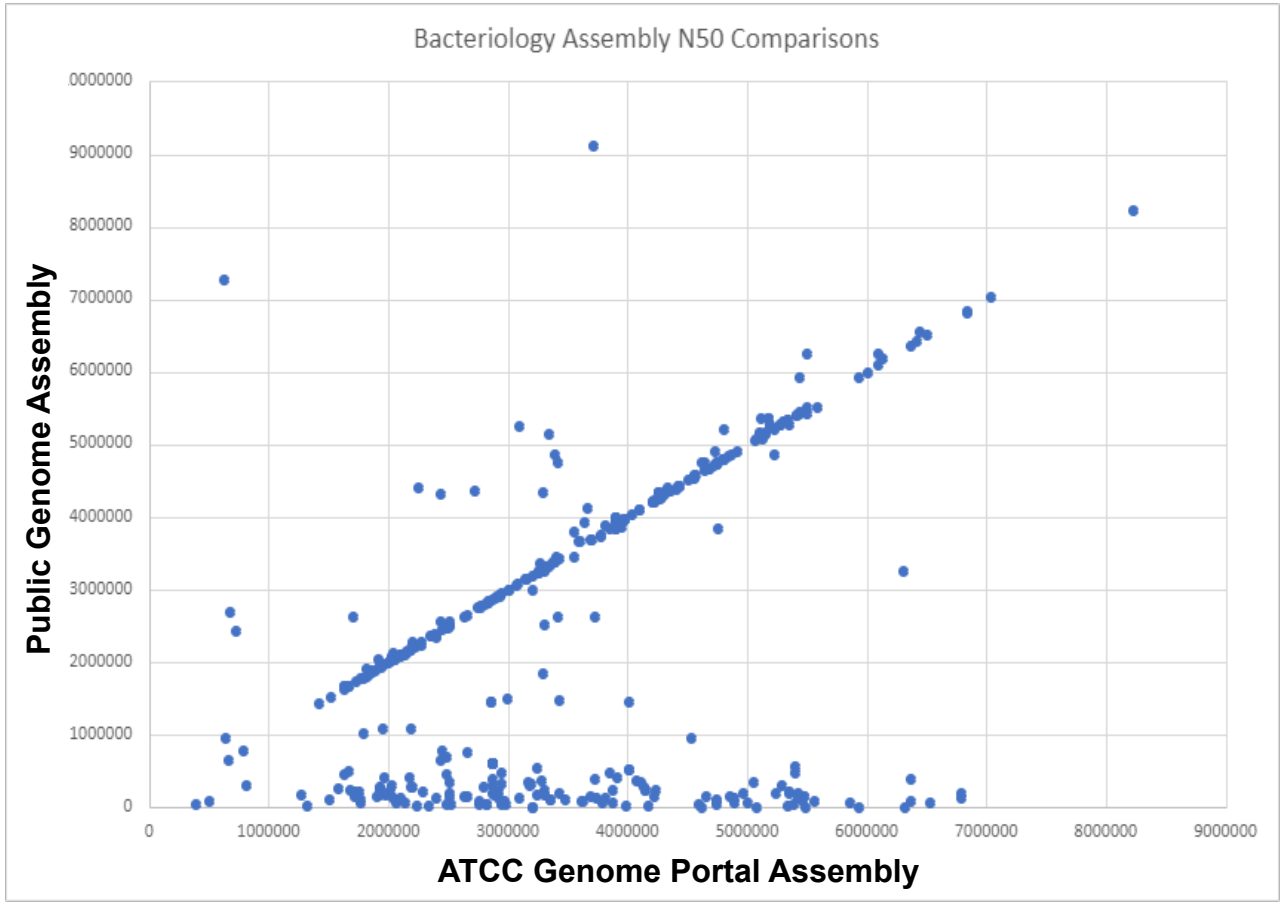
# Genome assembly quality

*Equivalency analysis of ATCC Genome Portal assemblies vs. those from public databases*




Number of Contigs per assembly

**#PUBLIC DATA** **ATCC GENOME PORTAL**




The **downward** trend in contig count and the **upward** trend in N50 indicate the ATCC produced genomes are of higher quality

# Evaluation of genome sequences from public databases

| Product | NCBI existing reference genomes | NCBI assembly level (plasmids) | Sequencing technology and coverage | # of SNPs | # of indels | Average coverage (variants) |
|---|---|---|---|---|---|---|
| *Acinetobacter baumannii* (ATCC® 17978™) | GCA_001593425.2 | Complete Genome | Illumina (300.0x) | 14 | 5 | 210.1 |
| | GCA_000015425.1* | Complete Genome (2) | Not available | 118 | 656 | 152.7 |
| | GCA_014672775.1 | Complete Genome (1) | PacBio (399.24x) | 15 | 87 | 170.4 |
| | GCA_013372085.1 | Complete Genome (2) | Illumina, Nanopore (80x) | 14 | 2 | 210.2 |
| | GCA_004797155.2 | Complete Genome (2) | PacBio (247.19x) | 28 | 62 | 162.1 |
| | GCA_001077675.1 | Complete Genome (1) | Illumina, PacBio (153x) | 15 | 6 | 135.9 |
| | GCA_011067065.1 | Complete Genome (2) | PacBio (231.08x) | 60227 | 2486 | 165.6 |
| *Candida albicans* (ATCC® 10231™) | GCA_015227795.1 | 3, 081 Contigs | NovaSeq (16x) | 10174 | 1573 | 265.6 |
| | GCA_002276455.1 | 2,219 Scaffolds | HiSeq (95x) | 13408 | 2390 | 274.6 |
| *Meyerozyma guilliermondii* (ATCC® 6260™) | GCF_000149425.1 | 9 RefSeq Scaffolds | Not available | 505 | 1973 | 278.2 |
| | GCA_006942155.1 | 9 Contigs | ONT+MiSeq (240x) | 74 | 386 | 223.3 |
| *Clavispora lusitaniae* (ATCC® 42720™) | GCF_000003835.1 | 9 RefSeq Scaffolds | Not available | 587 | 2336 | 265.6 |
| | GCA_003675505.1 | 109 Scaffolds | NextSeq (182x) | 102 | 5142 | 236.9 |

1 strain

7 assemblies

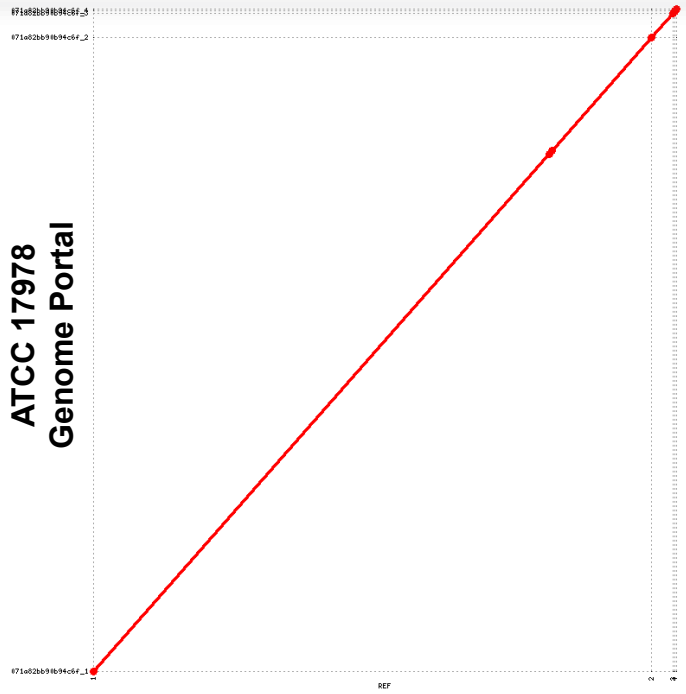
Unknown origin of materials

ATCC

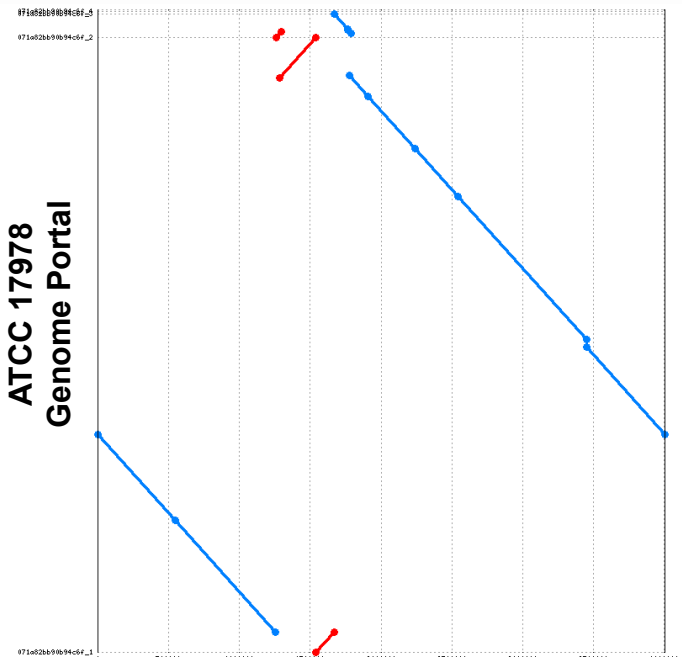
# Evaluation of public sequences for ATCC 17978

*MUMmer alignment with the de novo ATCC 17978 versus GenBank RefSeq genome assemblies GCF_001593425.2 and GCF_011067065.1*

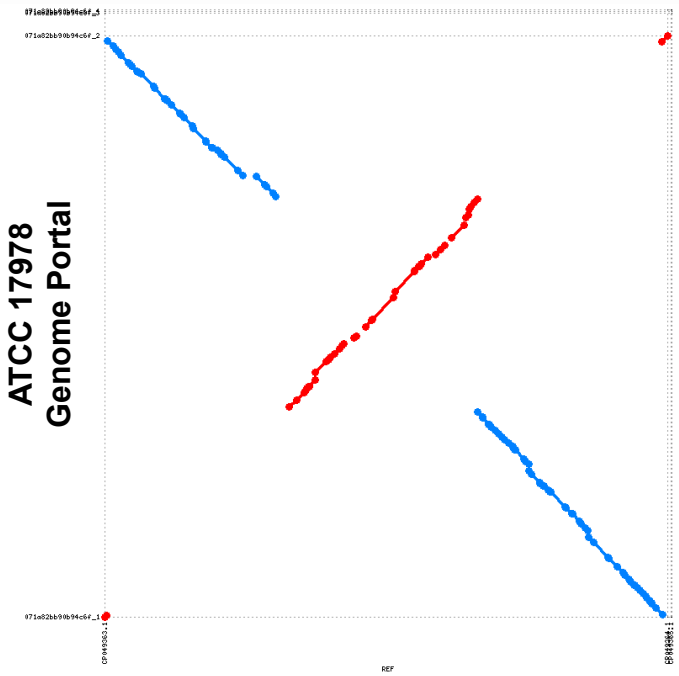
*Acinetobacter baumannii* strain 5377 (ATCC 17978)




**ATCC 17978**
**External Lab**




**ATCC 17978**
**RefSeq GCF_001593425.2**




**ATCC 17978**
**RefSeq GCF_011067065.1**



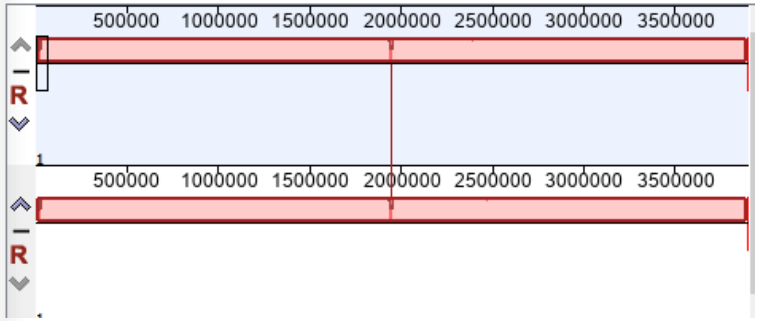
Marçais G, *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* 14(1):e1005944, 2018

# Evaluation of public sequences for ATCC 17978

*Acinetobacter baumannii* strain 5377 (ATCC 17978)

**ATCC 17978**
**External Lab**

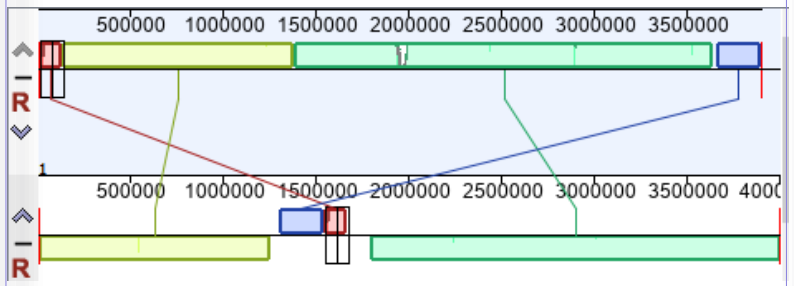
**ATCC 17978 Genome Portal**




**ATCC 17978 External Lab**

**ATCC 17978**
**RefSeq GCF_001593425.2**

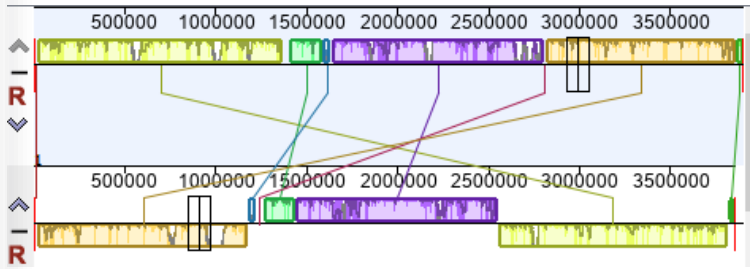
**ATCC 17978 Genome Portal**




**ATCC 17978 RefSeq GCF_001593425.2**

**ATCC 17978**
**RefSeq GCF_011067065.1**

**ATCC 17978 Genome Portal**




**ATCC 17978 RefSeq GCF_011067065.1**

# Evaluation of genome sequences from public databases

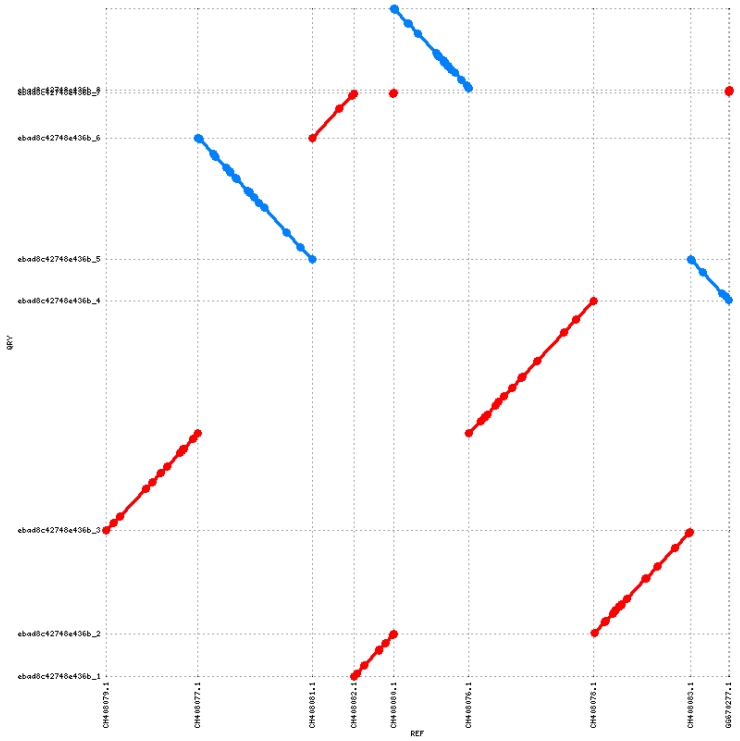
| Product | NCBI existing reference genomes | NCBI assembly level (plasmids) | Sequencing technology and coverage | # of SNPs | # of indels | Average coverage (variants) |
|---|---|---|---|---|---|---|
| *Acinetobacter baumannii* (ATCC 17978) | GCA_001593425.2 | Complete Genome | Illumina (300.0x) | 14 | 5 | 210.1 |
| | GCA_000015425.1* | Complete Genome (2) | Not available | 118 | 656 | 152.7 |
| | GCA_014672775.1 | Complete Genome (1) | PacBio (399.24x) | 15 | 87 | 170.4 |
| | GCA_013372085.1 | Complete Genome (2) | Illumina, Nanopore (80x) | 14 | 2 | 210.2 |
| | GCA_004797155.2 | Complete Genome (2) | PacBio (247.19x) | 28 | 62 | 162.1 |
| | GCA_001077675.1 | Complete Genome (1) | Illumina, PacBio (153x) | 15 | 6 | 135.9 |
| | GCA_011067065.1 | Complete Genome (2) | PacBio (231.08x) | 60227 | 2486 | 165.6 |
| *Candida albicans* (ATCC 10231) | GCA_015227795.1 | 3, 081 Contigs | NovaSeq (16x) | 10174 | 1573 | 265.6 |
| | GCA_002276455.1 | 2,219 Scaffolds | HiSeq (95x) | 13408 | 2390 | 274.6 |
| *Meyerozyma guilliermondii* (ATCC 6260) | GCF_000149425.1 | 9 RefSeq Scaffolds | Not available | 505 | 1973 | 278.2 |
| | GCA_000942155.1 | 9 Contigs | ONT+MiSeq (240x) | 74 | 386 | 223.3 |
| *Clavispora lusitaniae* (ATCC 42720) | GCF_000003835.1 | 9 RefSeq Scaffolds | Not available | 587 | 2336 | 265.6 |
| | GCA_003675505.1 | 109 Scaffolds | NextSeq (182x) | 102 | 5142 | 236.9 |



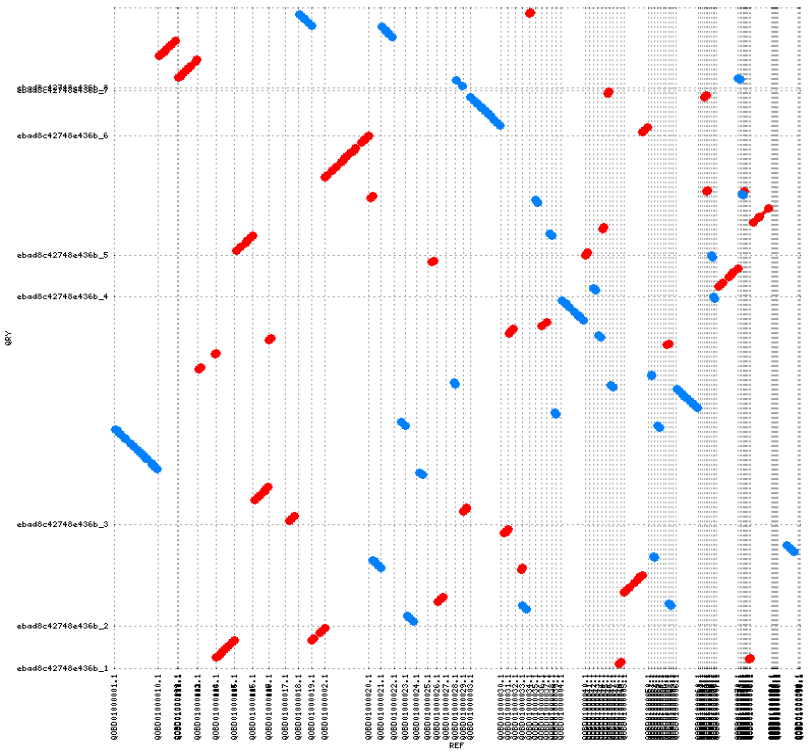
ATCC

# Evaluation of public sequences for ATCC 42720

*MUMmer whole genome alignments of ATCC de-novo genome assembly of ATCC 42720 versus GenBank RefSeq genome assemblies GCF_000003835.1 and GCA_003675505.1*




**ATCC 42720**
**RefSeq GCF_000003835.1**




**ATCC 42720**
**RefSeq GCA_003675505.1**

# Overview

- The ATCC Genome Portal
- Traceability and authentication of reference genomes
- **Standards for authenticated reference genomes**
- Development roadmap preview

# Selected timeline for (microbial) genomics standards

| 1970s | 1980s | 1990s | 2000s | 2010s | 2020s |
|---|---|---|---|---|---|
| | **1982 –** GenBank and ENA created | | **2005 –** Genomic Standards Consortium established<br><br>**2008** – Minimal Information on Genome Sequence (MIGS) specification<br><br>**2009 –** Genome Project Standards published by GSC | **2012** – CDC NGS Standards for Clinical Testing (Nex-StoCT)<br><br>**2014 –** Viral Genome Reference Standards<br><br>**2016 –** FDA Draft Guidance on NGS for Pathogen Identification | **2020** – ATCC Enhanced Authentication Initiative<br><br>**2020** – ATCC Genome Portal Launch |
| **1974 –** Complete RNA genome bacteriophage MS2 | **1984 –** Complete Epstein Barr virus genome | **1995 –** Complete genome of *Haemophilus influenzae* | **2001** – Draft Human Genome<br>**2007** - Genomic Encyclopedia of Bacteria and Archaea (GEBA) and Human Microbiome Project (HMP) launch. | **2011** – GEBA II Launched | **2020 –** First “end to end” gapless genome for Human Chr. X |

# Recognition of the importance of traceability to biomaterials

nature biotechnology

PERSPECTIVE

## The minimum information about a genome sequence (MIGS) specification

Dawn Field*[1], George Garrity[2], Tanya Gray[1], Norman Morrison[3,4], Jeremy Selengut[5], Peter Sterk[6], Tatiana Tatusova[7], Nicholas Thomson[8], Michael J Allen[9], Samuel V Angiuoli[5,10], Michael Ashburner[11], Nelson Axelrod[5], Sandra Baldauf[12], Stuart Ballard[13], Jeffrey Boore[14], Guy Cochrane[6], James Cole[2], Peter Dawyndt[15], Paul De Vos[16,17], Claude dePamphilis[18], Robert Edwards[19,20], Nadeem Faruque[6], Robert Feldman[21], Jack Gilbert[9], Paul Gilna[22], Frank Oliver Glöckner[23], Philip Goldstein[24], Robert Guralnick[24], Dan Haft[5], David Hancock[3,4], Henning Hermjakob[6], Christiane Hertz-Fowler[8], Phil Hugenholtz[25], Ian Joint[9], Leonid Kagan[5], Matthew Kane[26], Jessie Kennedy[27], George Kowalchuk[28], Renzo Kottmann[23], Eugene Kolker[29–31], Saul Kravitz[5], Nikos Kyrpides[32], Jim Leebens-Mack[33], Suzanna E Lewis[34], Kelvin Li[5], Allyson L Lister[35,36], Phillip Lord[35], Natalia Maltsev[20], Victor Markowitz[37], Jennifer Martiny[38], Barbara Methe[5], Ilene Mizrachi[7], Richard Moxon[39], Karen Nelson[5,40], Julian Parkhill[8], Lita Proctor[26], Owen White[10], Susanna-Assunta Sansone[6], Andrew Spiers[42], Robert Stevens[3], Paul Swift[1], Chris Taylor[6], Yoshio Tateno[43], Adrian Tett[1], Sarah Turner[1], David Ussery[44], Bob Vaughan[6], Naomi Ward[45], Trish Whetzel[46], Ingio San Gil[41], Gareth Wilson[1] & Anil Wipat[35,36]



With the quantity of genomic data increasing at an exponential rate, it is imperative that these data be captured electronically, in a standard format. Standardization activities must proceed within the auspices of open-access and international working bodies. To tackle the issues surrounding the development of better descriptions of genomic investigations, we have formed the Genomic Standards Consortium (GSC). Here, we introduce the minimum information about a genome sequence (MIGS) specification with the intent of promoting participation in its development and discussing the resources that will be required to develop improved mechanisms of metadata capture and exchange. As part of its wider goals, the GSC also supports

can manipulate it to provide new solutions to critical problems. Such solutions include therapies and cures for disease, industrial products, approaches for biodegradation of xenobiotic compounds and renewable energy sources. With improvements in sequencing technologies, the growing interest in metagenomic approaches and the proven power of comparative analysis of groups of related genomes, we can envision the day when it will be commonplace to sequence tens to hundreds of genomes or more as part of a single study. At current rates of genome sequencing, it has been estimated that >4,000 bacterial genomes will be available soon after 2010 (ref. 1).

Given the importance of the growing genome collection, the capital investment in its creation and the benefits of leveraging its value

**"Source material identifier" is an exception; the *GSC recommends this be a core descriptor*, *but as of yet, physical archives are not yet routinely created for all cases or types of biological material subjected to genome sequencing …***

**This was in 2008.**

We agree.

But, 12 years later "physical archives are [still] not yet routinely created" by groups doing whole genome sequencing.

Chain of custody of biomaterials is rarely or poorly documented.

Field, D. *et al.* (2008) 'The minimum information about a genome sequence (MIGS) specification', *Nature Biotechnology*, 26(5), pp. 541–547. doi: 10.1038/nbt1360.

ATCC

# Overview

- The ATCC Genome Portal
- Traceability and authentication of reference genomes
- Importance of authenticated reference genomes
- **Development roadmap preview**

# ATCC Genome Portal development goals


**2,500** Authenticated Microbial Genomes

**5,000** Authenticated Microbial Genomes

**7,500** Authenticated Microbial Genomes

**10,000** Authenticated Microbial Genomes

**ATCC Genome Portal 2.0**

- Fungal and viral genome annotation improvements
- Batch downloads
- **API for bioinformatics**
- Lab Protocols

- Interactive taxonomy browser
- Biosynthetic gene cluster analysis
- Genome Forum

- Microbial phenotypic data, including AMR

**Authenticated Human & Animal Cell Line Data**

- RNAseq, WES, STR, cell imaging, flow cytometry, and more

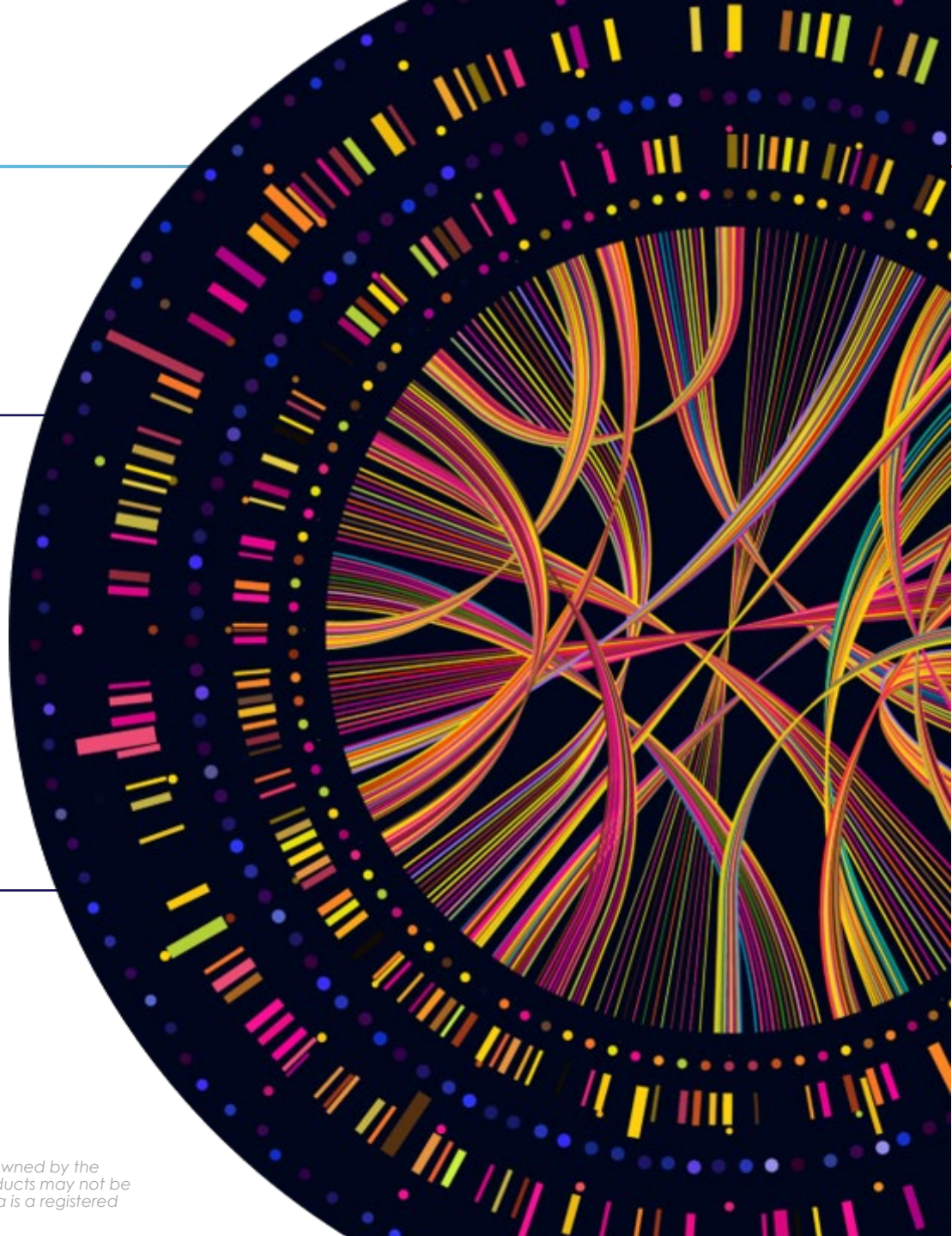
# The ATCC Genome Portal Team

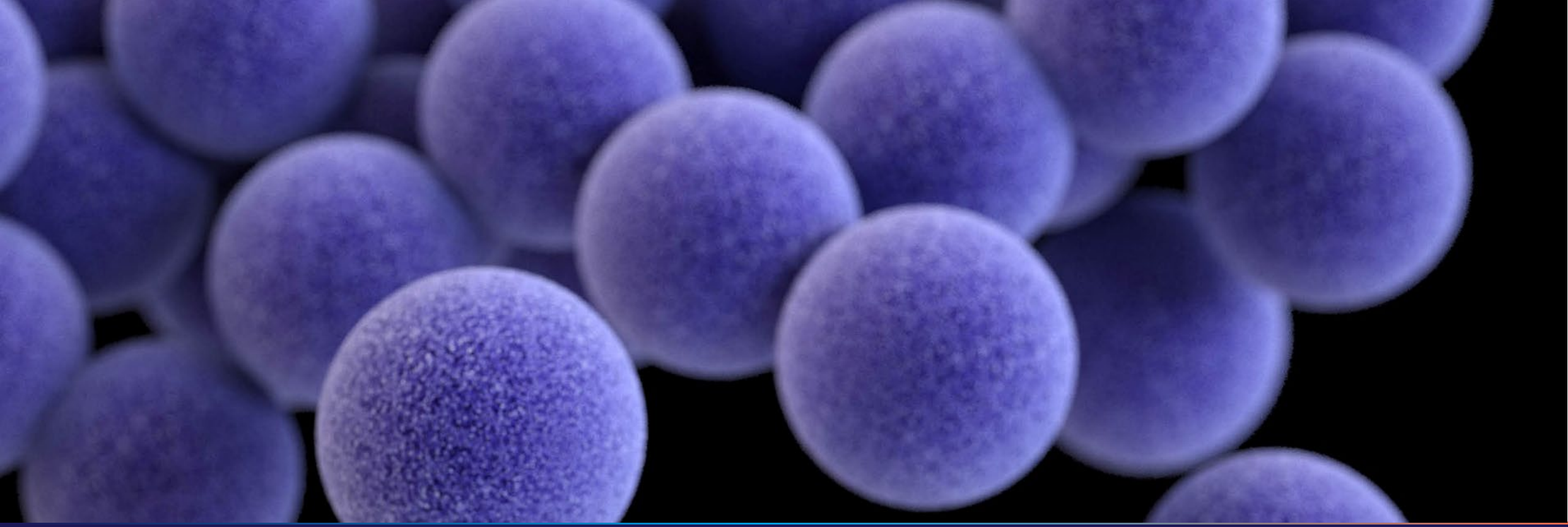
**Jonathan Jacobs, PhD**
jjacobs@atcc.org

| Genomics Lab | Bioinformatics Lab | Partners |
|---|---|---|
| **Briana Benton** | **John Bagnoli** | Juan Lopera, PhD |
| Stephen King, MSc | David Yarmosh, MSc | Marco Riojas, PhD |
| James Duncan, MSc | Nikhita Putheveetil, MSc | |
| Robert Marlow | P. Ford Combs, MSc | |
| Samuel Greenfield | | |
| Corina Tabron | | |
| Fabio Martinez | | |
| Amanda Pierola | | |

*… and One Codex!*

***JOIN OUR TEAM!*** *We're hiring!*





# Thank you

ATCC