

The Importance of Data Provenance & Tracking Strains to Sequence

Why culture collections are the foundation of trustworthy digital biology

Jonathan L Jacobs, PhD

jjacobs@atcc.org

About ATCC

- **Mission-focused biological materials & standards organization**

- **Founded in 1925**

- **Independent 501(c)(3) not-for-profit org**

- Quality Accreditation by multiple industry standards

- ISO 9001 – Quality Management System & Assurance
- ISO 13485 – QMS for Diagnostics / Medical Devices
- ISO/IEC 17025 – Certified Testing / Calibration Lab
- ISO 17034 – Biological Reference Materials Provider

- Standards development partner

- ANSI / ISO / IEC / AOAC Standards Working Groups
- NIST Biological Standards working groups

World's largest, most diverse living biorepository

- >200,000+ cell materials in our collection
- 20,000+ metazoan cell lines
 - Incl. >3k human and animal cell lines, primary cells, and organoids
- 3,000+ viruses
- 19,500+ bacteria
- 56,000+ fungi and protists
- The primary US Patent repository for biological materials.

In the age of AI biology, what is a culture collection actually for?

We are the bridge between physical materials and trustworthy data.

From Preservation to Innovation

The ***last*** 100
years :



Laboratory Scientists

In the context of Digital Biology...

The **next** 100
years :



Laboratory scientists

+



Data scientists

How do we best **pursue our *Mission*** in this context?

What **unmet needs** can we **address today**?

How do we **prepare for the future** today?

The next 100 years for culture collections...

The Last 100 Years



Materials Sourced from **Culture Collections**

Top Priorities

- Reproducibility
- Stewardship and sustainable access
- Enablement of scientific discovery and innovation

An Unseen Gap



Data / Results from **Third Party Sources**

Variable or Unknown Methods
Materials Not Attributable
Results often Not Reproducible

Many Sources
Variable Levels of Trust

The Next 100 Years



Trusted Sources

Authenticated **Materials & Data**
From **One Source**

End-to-End Data Provenance
Maximum Scientific Reproducibility

The Same Mission – just expanded
to support an AI-enabled future

Foundational Reference Data for Authenticated Materials Should Come from the Source

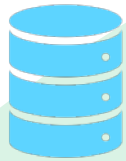
The Quality Arc in Digital Biology



GenBank
SRA
GEO
Dryad
Zenodo

Uncontrolled Public Data

- Moderate to high risk
- Little to no HITL curation
- Massive size
- Broad scope
- Highly variable quality
- Missing / non-standard metadata
- Often no data provenance



Community Reference Data

- Moderate risk
- Automated data QC
- Non-redundant
- Standardized metadata formats
- Some HITL curation
- Little to no provenance

RefSeq (automated)
BV-BRC
UniProtKB (TrEMBL)
GBIF
Cellosaurus



Expert Curated Data

- "AI READY"
- Lower risk
- HITL curation
- Standardized
 - data structures
 - metadata / ontologies
- Better reproducibility

RefSeq (curated)
SwissProt (UniProtKB)
ClinVar
KEGG
TAIR
HGMD
Ingenuity Pathway Analysis



Authenticated Reference Data

- "AI READY"
- Least risk
- HITL curation
- End to End Data Provenance
- Standardized laboratory methods
- Standardized bioinformatics methods
- Quality Assurance (i.e. ISO9001)
- Traceable to authenticated materials
- Very high reproducibility (goal)

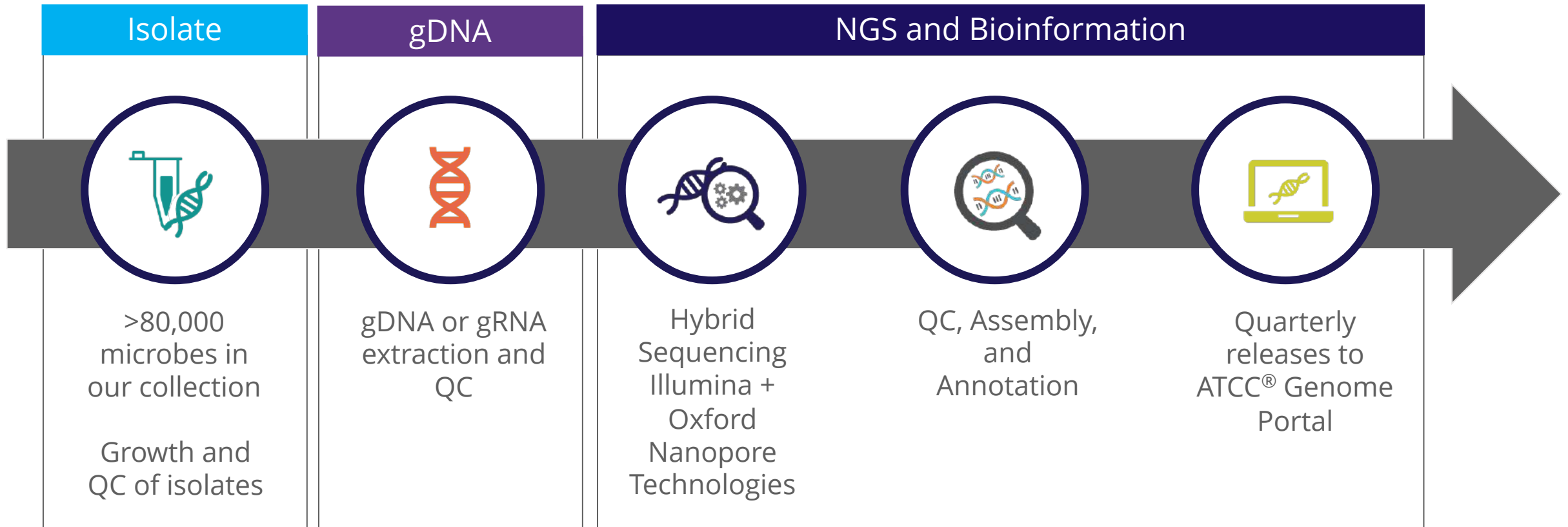
ATCC Genome Portal
ATCC Cell Line Land



“Authenticated Genome”

- 1. Full end to end data provenance (i.e. traceable from materials to data)***
- 2. Uniformly applied, documented and standardized laboratory methods, bioinformatics methods, terminologies, ontologies, and quality metrics (i.e. ISO).***

Authenticated material + Reference-quality data + Standardized bioinformatics



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

ATCC Genome Portal

<https://genomes.atcc.org>

Our Microbial Genomics Data...

7,000 authenticated genomes



- 5,819 bacteria, 591 viruses, 583 fungi, and 7 protists
- 250 new genomes every quarter
- ISO 9001 & ISO 17025 quality standards
- Sequenced on both ILMN & ONT platforms
- Assembled and annotated in-house
- DNA methylation data on all ONT data
- Full end-to-end data provenance

FEATURES

Search & browse AGP catalog

Find related genomes by ANI / MASH

View assembly statistics and quality control data.

Integrated genome browser

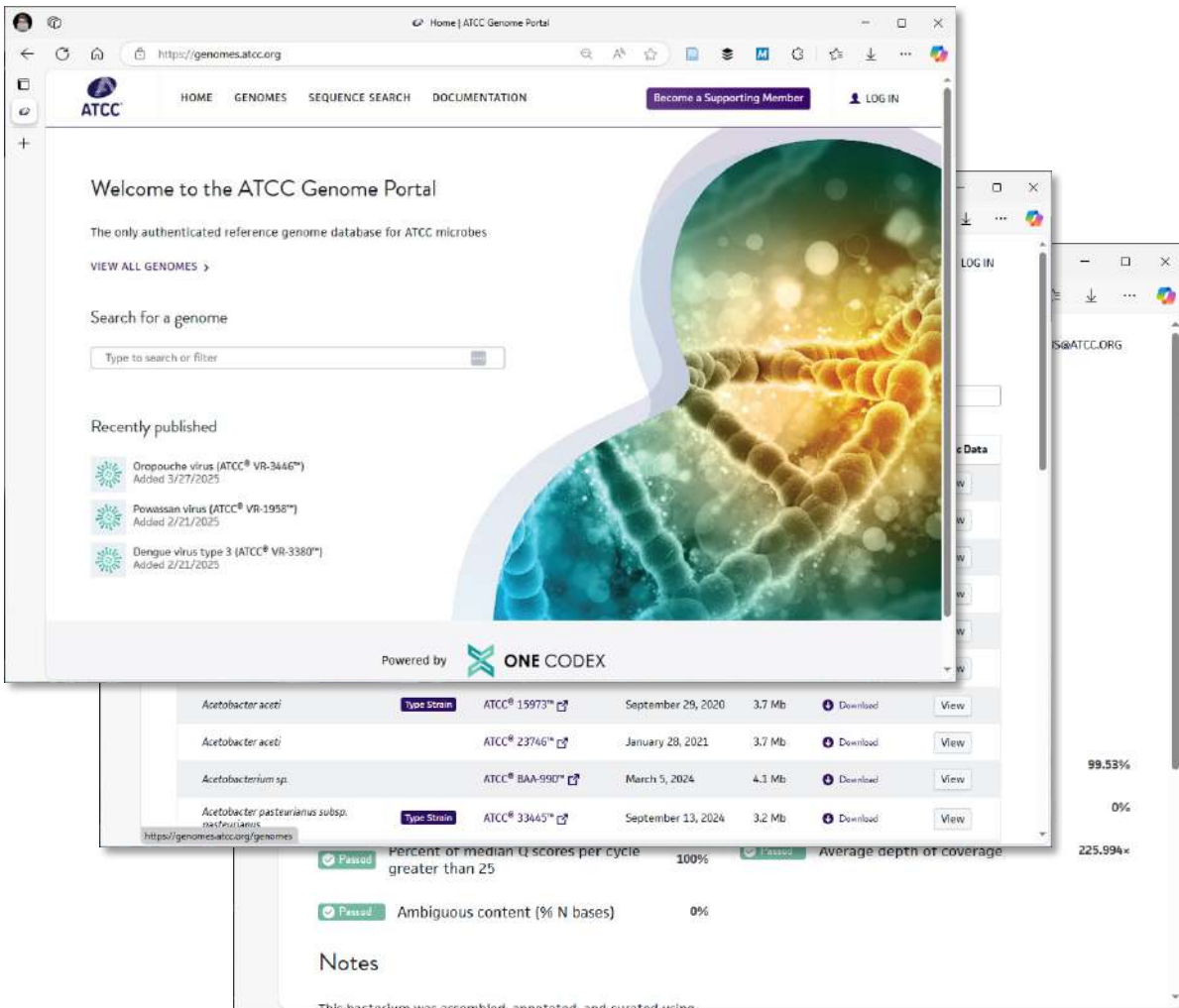
Download genomics data and genome annotations

Genome comparison & discrepancy report tool

Programmatic REST-API

GitHub & Google Colab Tutorials

Full documentation



ATCC Genome Portal

<https://genomes.atcc.org>



Our Eukaryotic Cell Biology Data...

1,034 cell lines & organoids

- >5,000 RNAseq and >900 exome datasets
- 5x biological replicates per transcriptome
- 654 human cell lines
- 323 human organoids
- 57 mouse cell lines
- ~1,000 new datasets per year
- ISO 9001 & ISO 17025 quality standards

FEATURES

Search & browse AGP catalog

Find related genomes by ANI / MASH

View assembly statistics and quality control data.

Integrated genome browser

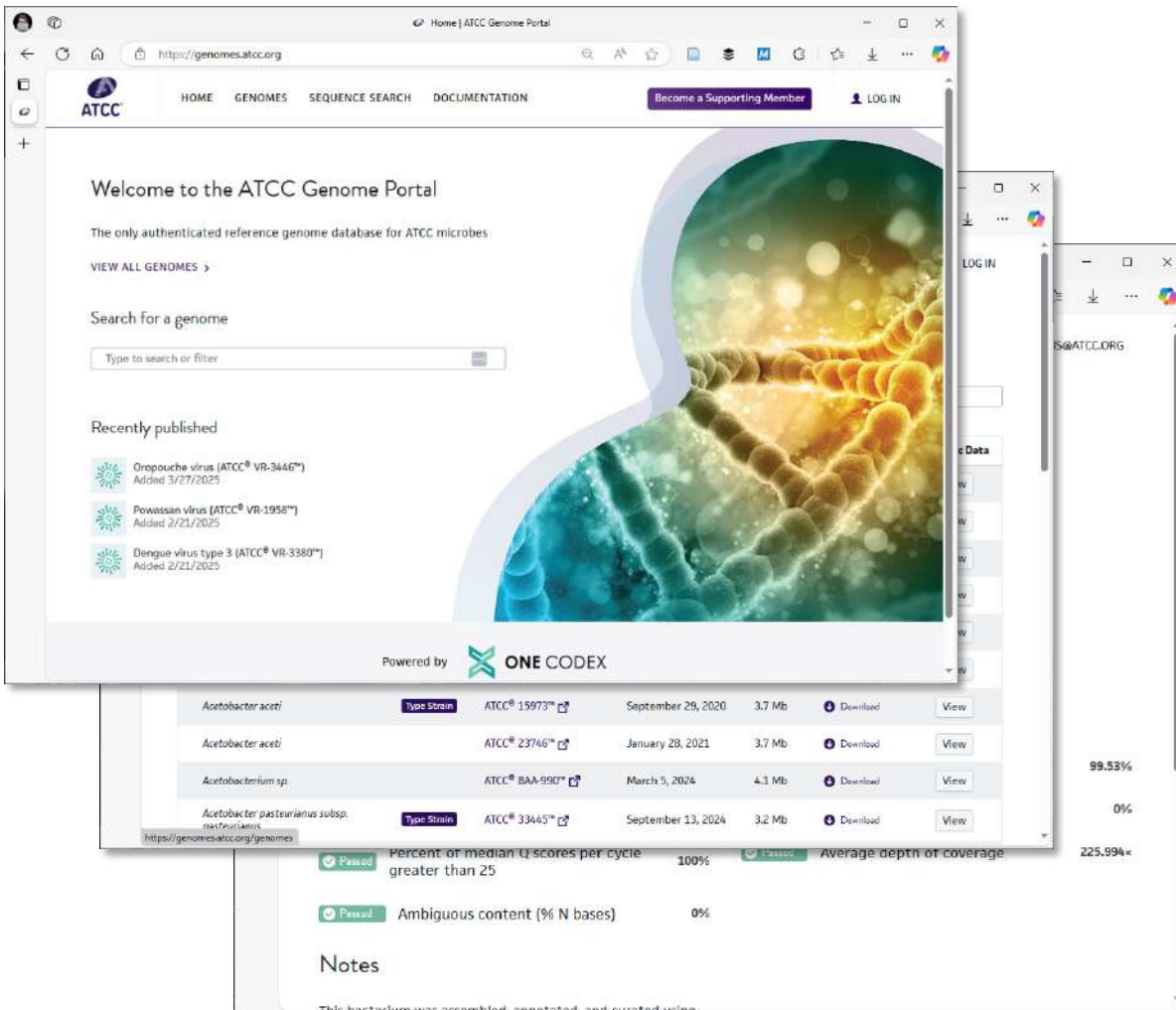
Download genomics data and genome annotations

Genome comparison & discrepancy report tool

Programmatic REST-API

GitHub & Google Colab Tutorials

Full documentation



When provenance is lost...

One illustrative example

Which genome assembly is the *right* one?



Acinetobacter baumannii – 9 assemblies for the reference strain...

Genome

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

Selected taxa: Enter one or more taxonomic names

Filters

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

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

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

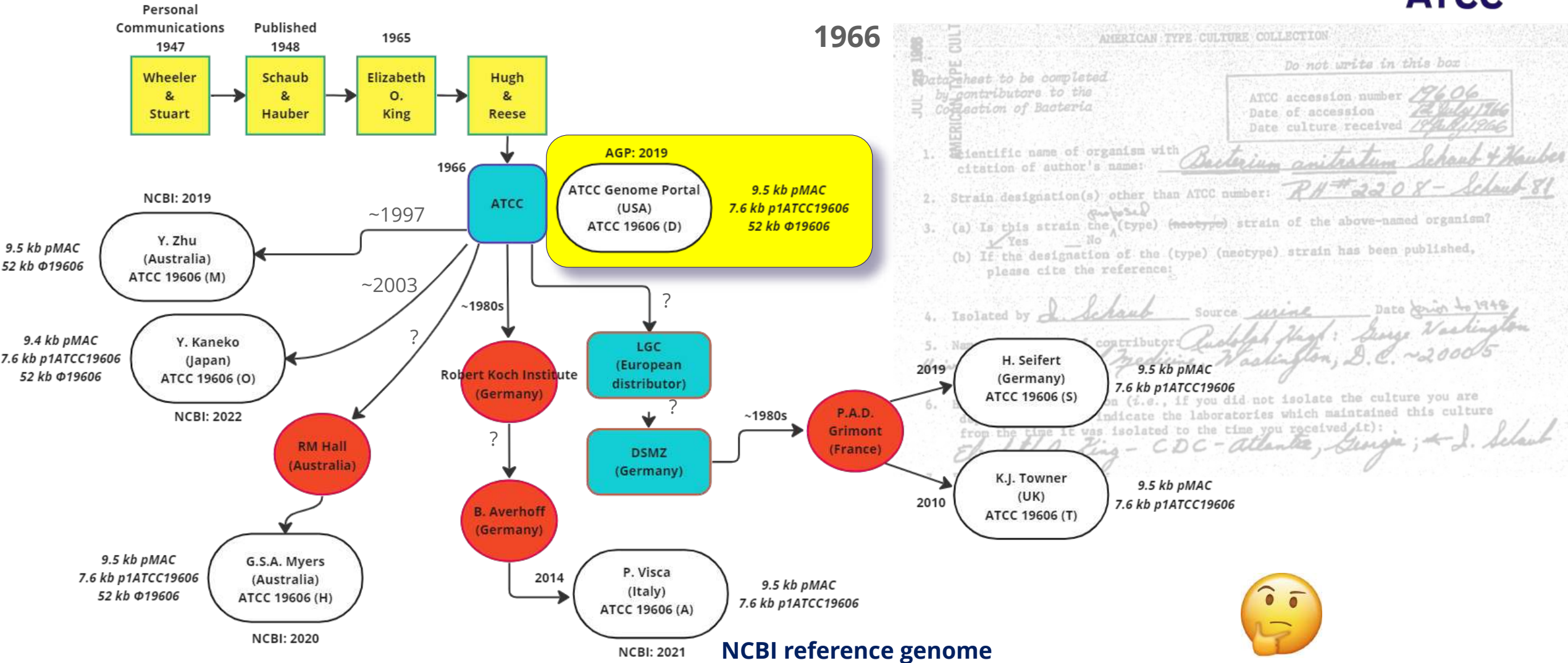
*How do researchers *know* which assembly reflects the actual (physical) reference material?*

Answer: None of them.

Unfortunately, we have no records of any of these groups obtaining the strain directly from ATCC...

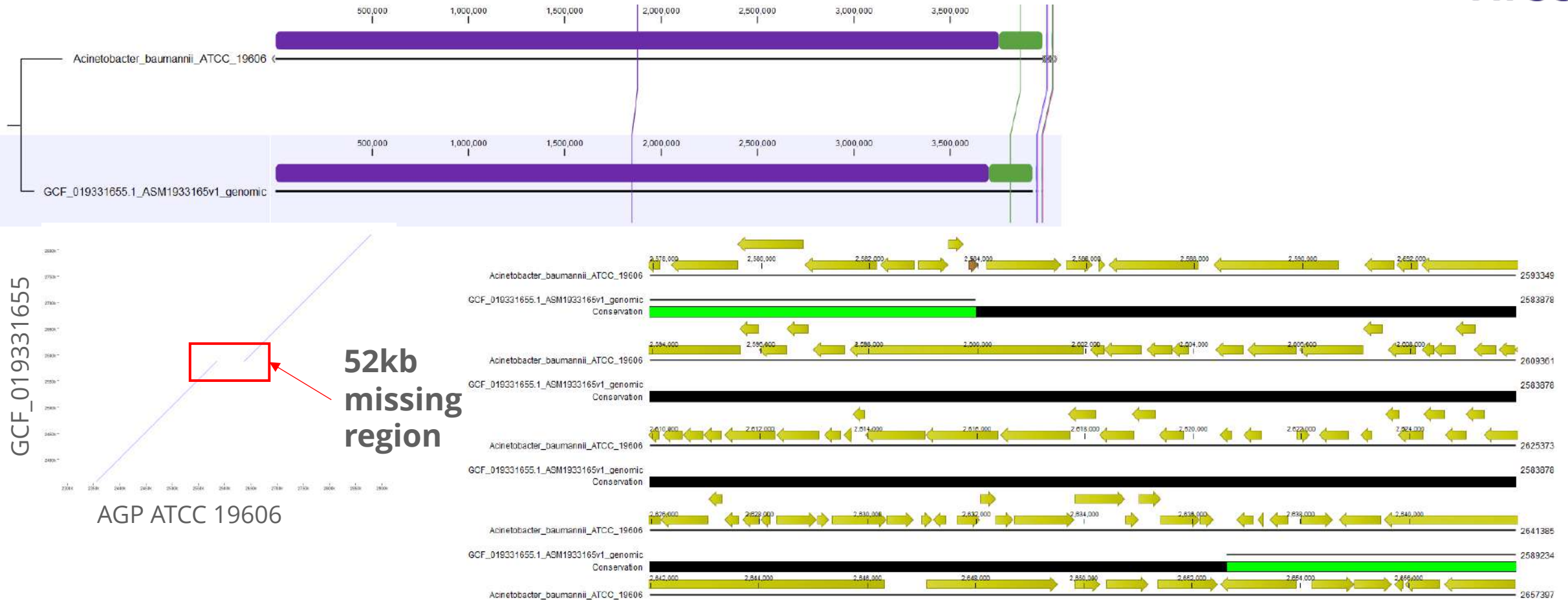
41 bioprojects
113 biosamples
132 SRA datasets

Which reference is the “right” reference?



Artuso I, et al. **Genome diversity of domesticated *Acinetobacter baumannii* ATCC 19606^T strains.** Microbial Genomics 8(1): 000749, 2022.

Comparison to RefSeq reference assembly



- 52kb region missing, includes 74 genes, 51 are single-copy genes

Future Directions We Can Take

Advocate for Authenticated Reference Data



Culture Collection Reference Genomes

ATCC Genome Portal, NCTC3000, JCM

- Sequenced directly from authenticated, living material
- Chain-of-custody defined at the outset
- Full end-to-end provenance
- Reproducible by design
- A ***“living database”*** that improves as methods improve

Community Reference Genomes

Community-aggregated efforts

- Data first; provenance reconstructed later (if ever)
- Physical source often unavailable or ambiguous
- Quality metrics can be misleading
 - WFCC 10K: ~20% of assemblies >5% contamination
 - FDA-ARGOS: strains renamed; ~5.8% flagged atypical

***Just because a genome is “Complete”
doesn’t mean it’s “Correct”.***

Treat *Metadata as Infrastructure*

AI-enabled biology benefits from structured, well curated metadata just as much as the experimental data it describes.

- **Standards already exist. The gap is enforcement, not invention.**
MlxS/MIGS, Darwin Core (GGBN), PHA4GE, BioCompute, FAANG, ENCODE
- **Provenance fields are (unfortunately) optional**, so they are the first thing lost.
Source material, chain-of-custody, and methods are rarely required or validated
- **A sequence is only reliable and trustworthy when it is connected to:**
 - (1) a traceable physical sample
 - (2) how the sample was collected, preserved, handled and processed
 - (3) the computational methods that produced the sequencing data
- **Treat experimental and sample metadata as shared scientific infrastructure.**
 - Preserved, persistently identified, and curated with the same intent as the sequence itself.

What Can Our Community Can Do?

- **Sequence data authenticated by culture collections.** Make “collection-backed reference genomes” the norm so every genome traces to authenticated, living material.
- **Capture provenance information at accession.** Treat metadata as infrastructure – this is the highest-impact step any culture collection can take.
- **Link records to living vouchers.** Use persistent identifiers (i.e. RRIDs, DOIs, etc.) to connect every sequence back to a retrievable strain.
- **Standardize together, don’t reinvent alone.** The USCCN could serve as a platform to convene the development of shared tooling and data standards for interoperability.
- **Advocate for sustainable funding to support a shared digital infrastructure** that makes all of the above usable and durable.

The next 100 years in biology will place as much importance on the quality and trustworthiness of the data as the physical materials from which it was produced.

Future Work on the ATCC Genome Portal



Coming Soon

- More whole genome assemblies of microbial strains
- More RNAseq & exome data for cell lines and organoid models
- New authenticated data packages
 - Antimicrobial Susceptibility Testing (AST)
 - Biolog phenotypic array data
- Improved AMR annotation pipelines
- Biosynthetic gene cluster prediction
- MIXS compliant data structures
- External database linkages
- Curated externally sourced data
- Improved comparative assembly tools
- Containerized pipelines for end-users

Thank you!

Sequencing & Bioinformatics Center nextgen@atcc.org



The ATCC Genome Portal
<https://genomes.atcc.org>

Jonathan Jacobs, PhD

Senior Director, Bioinformatics
BioNexus Principal Investigator

✉ jjacobs@atcc.org

Genomics Team

Briana Benton, PMP

Ana Fernandes
Stephen King, M.Sc.
James Duncan, M.Sc.
Robert Marlow
Corina Tabron, M.Sc.
Jade Kirkland
Noah Wax, M.Sc.
Rula Khairi
Hannah McConnell
Kaitlyn Gaffney

Bioinformatics Team

John Bagnoli

David Yarmosh, M.Sc.
Nikhita Puthaveetil, M.Sc.
Joseph Petron, Ph.D.
Amy Reese, M.Sc.

Scientific Leads

Ajeet Singh, Ph.D.
Scott V. Nguyen, Ph.D.
Prech Uapinyoying, Ph.D.