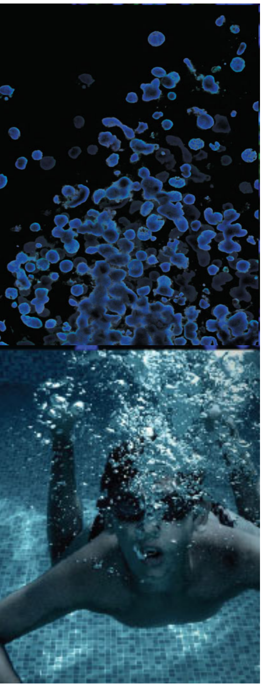
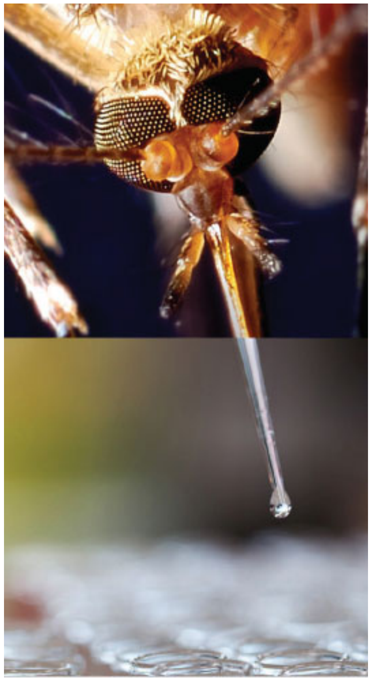
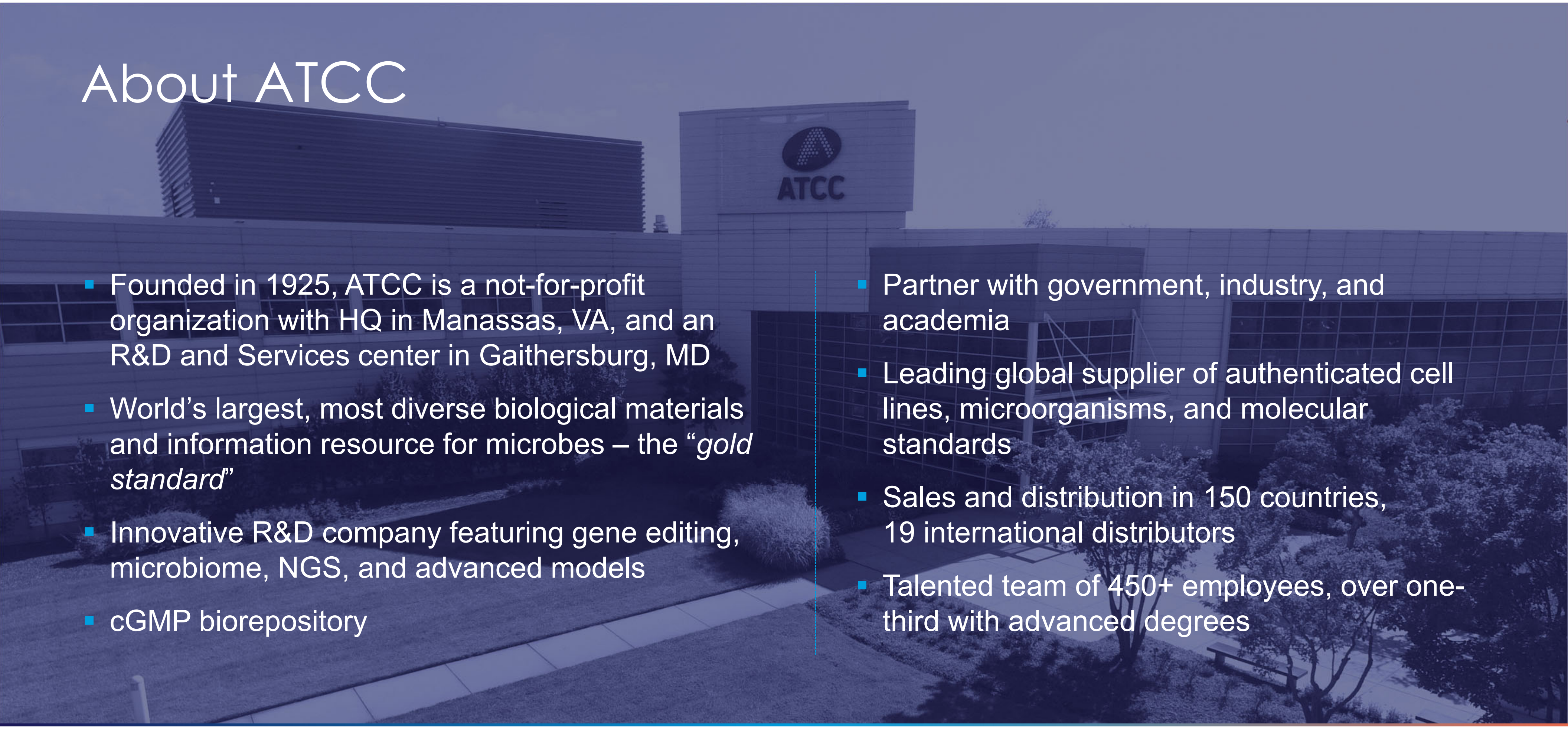


The Importance of Using Next-Generation Sequencing to Further Authenticate the ATCC Microbial Collections

Briana Benton
Technical Manager, ATCC

Credible Leads to Incredible™





About ATCC

- Founded in 1925, ATCC is a not-for-profit organization with HQ in Manassas, VA, and an R&D and Services center in Gaithersburg, MD
- World's largest, most diverse biological materials and information resource for microbes – the "*gold standard*"
- Innovative R&D company featuring gene editing, microbiome, NGS, and advanced models
- cGMP biorepository

- Partner with government, industry, and academia
- Leading global supplier of authenticated cell lines, microorganisms, and molecular standards
- Sales and distribution in 150 countries, 19 international distributors
- Talented team of 450+ employees, over one-third with advanced degrees

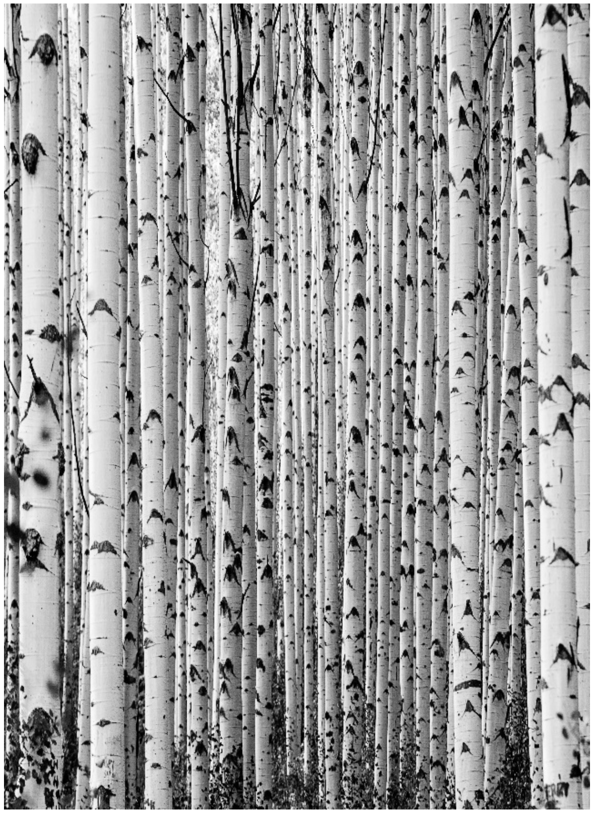
Overview

Using next-generation sequencing to further authenticate the ATCC microbial collections

- Discuss ***why*** ATCC is committed to **providing reference-quality genomes** for items within the microbial collections
- Discuss some of the standardized processes and quality control criteria required for extracting, sequencing, and analyzing our reference-quality genomes
- Explore the ATCC Genome Portal

Providing reference-quality genomes

Why - Challenge # 1

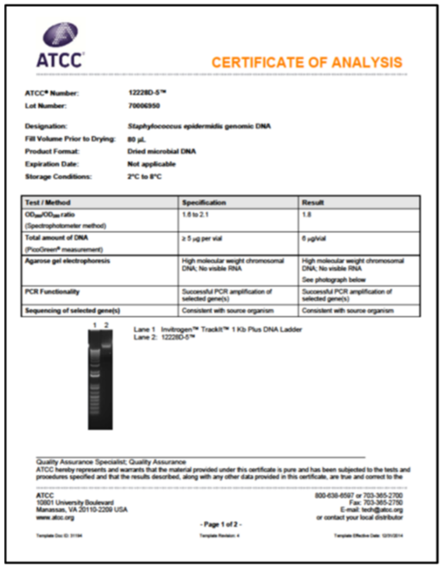


- Public databases routinely host genomic data that is cited as “ATCC,” but there is often no traceability back to genuine ATCC cultures and ATCC doesn’t perform confirmation testing on public data.
  - How do researchers *know* which data set to use?
    - Which is the “correct” one?
    - Close enough?
  - 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 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

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: 703-365-2750
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

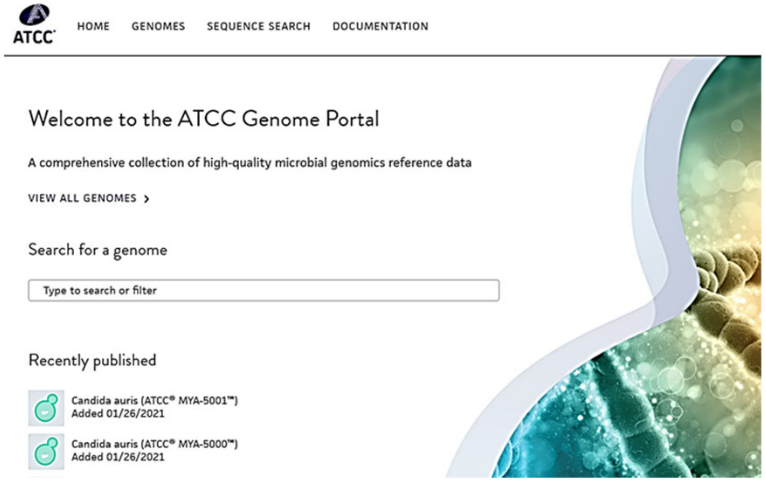
Why - Challenge # 3

- 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?**

The Enhanced Authentication Initiative

ATCC's solution to the authenticated reference genomes

- **2017-2018** – Planning and proof-of-concept experiments
- **2018** – Commitment
  - Laboratory and staffing resources
  - Instrumentation
  - Bioinformatics pipelines
- **2019** – Launch of the Enhanced Authentication Initiative
  - June 2019 – *beta* launch at ASM Microbe
  - Sept 2019 – formal launch of the ATCC Genome Portal
    - Provide our customers with the whole-genome sequences of the specific, authenticated materials researchers need to generate credible data
    - genomes.atcc.org



Welcome to the ATCC Genome Portal

A comprehensive collection of high-quality microbial genomics reference data

Search for a genome

Type to search or filter

Recently published

Candida auris (ATCC® MYA-5001™)
Added 01/26/2021

Candida auris (ATCC® MYA-5000™)
Added 01/26/2021

Overview

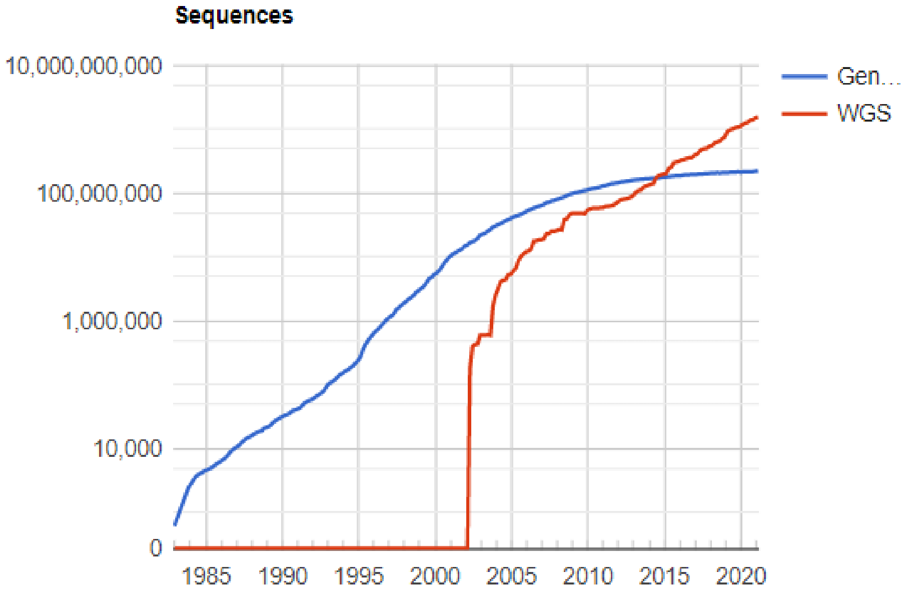
Using next-generation sequencing to further authenticate the ATCC microbial collections

- **Discuss *why* ATCC is committed to providing reference-quality genomes for items within the microbial collections**
- Discuss some of the standardized processes and quality control criteria required for extracting, sequencing, and analyzing our reference-quality genomes
- Explore some of the features of the ATCC Genome Portal

Reference genomes

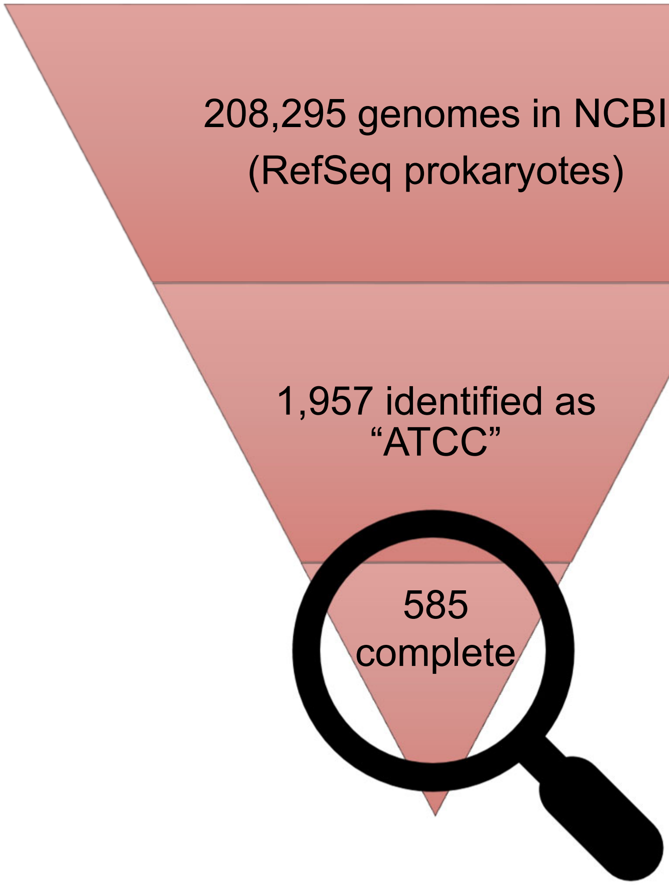
Where can researchers turn to for "reference" genomes?

- 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/

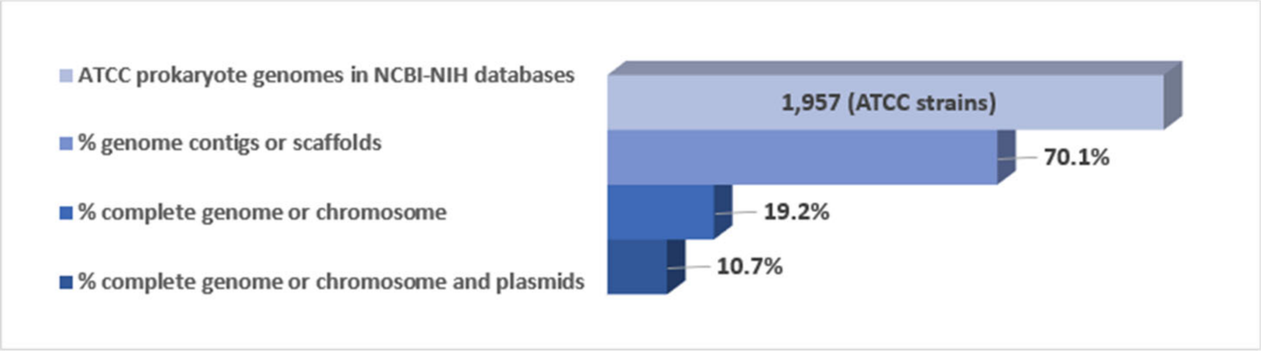
Reference genomes



208,295 genomes in NCBI (RefSeq prokaryotes)

1,957 identified as “ATCC”

585 complete



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

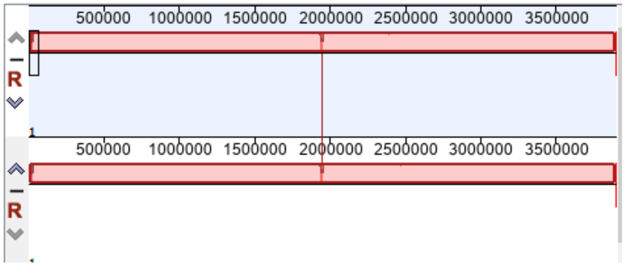
Evaluation of genome sequences from public databases

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

Evaluation of public sequences for ATCC 17978

ATCC 17978
External Lab

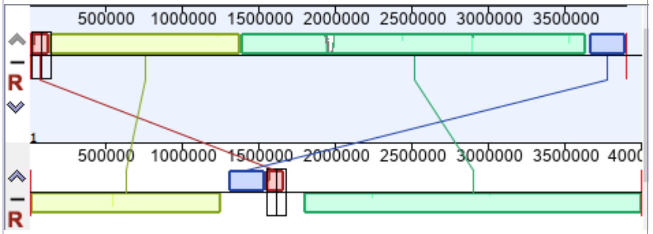
ATCC 17978 Genome Portal



ATCC 17978 External Lab

ATCC 17978
RefSeq GCF_001593425.2

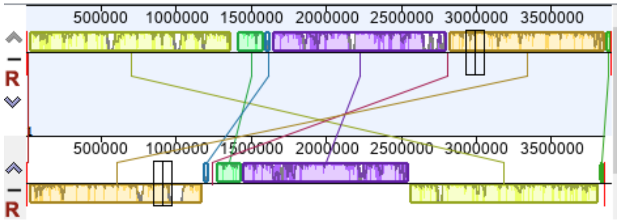
ATCC 17978 Genome Portal



ATCC 17978 RefSeq GCF_001593425.2

ATCC 17978
RefSeq GCF_011067065.1

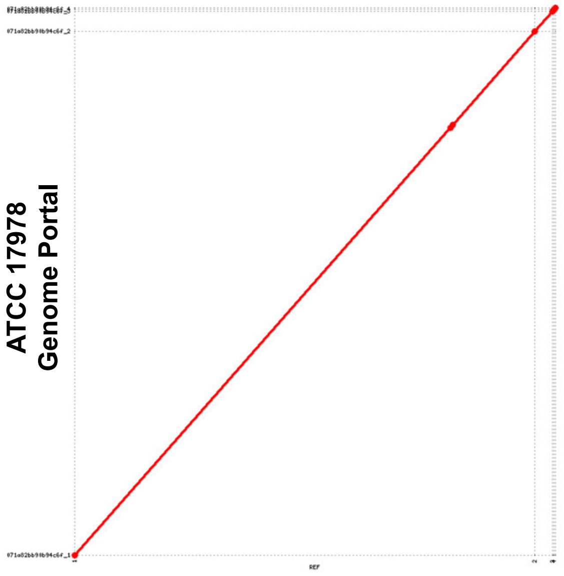
ATCC 17978 Genome Portal



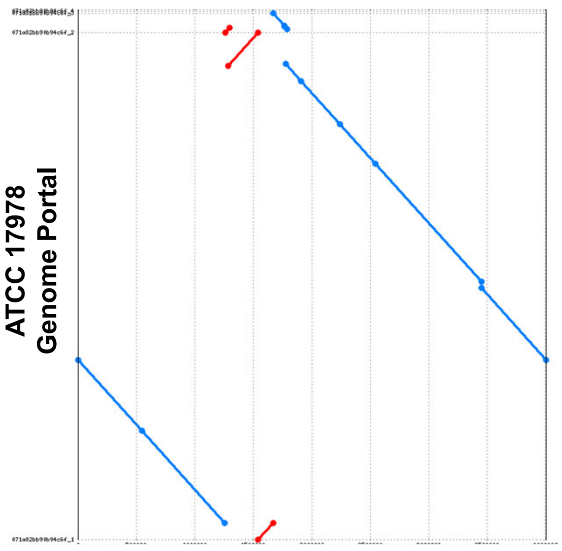
ATCC 17978 RefSeq GCF_011067065.1

Evaluation of public sequences for ATCC 17978

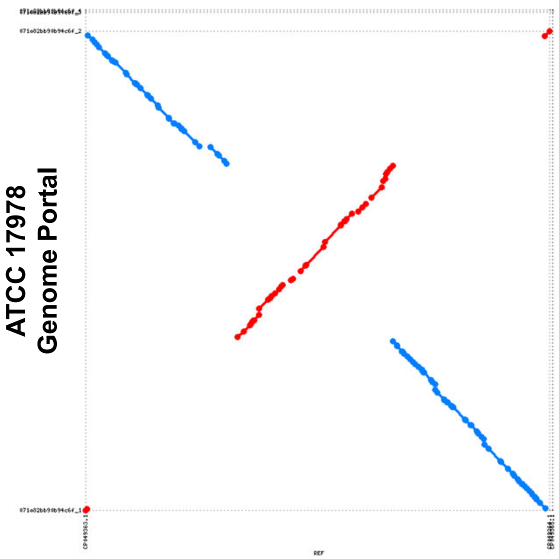
MUMmer alignment with the de novo ATCC 17978 versus GenBank RefSeq genome assemblies GCF_001593425.2 and GCF_011067065.1



ATCC 17978
External Lab



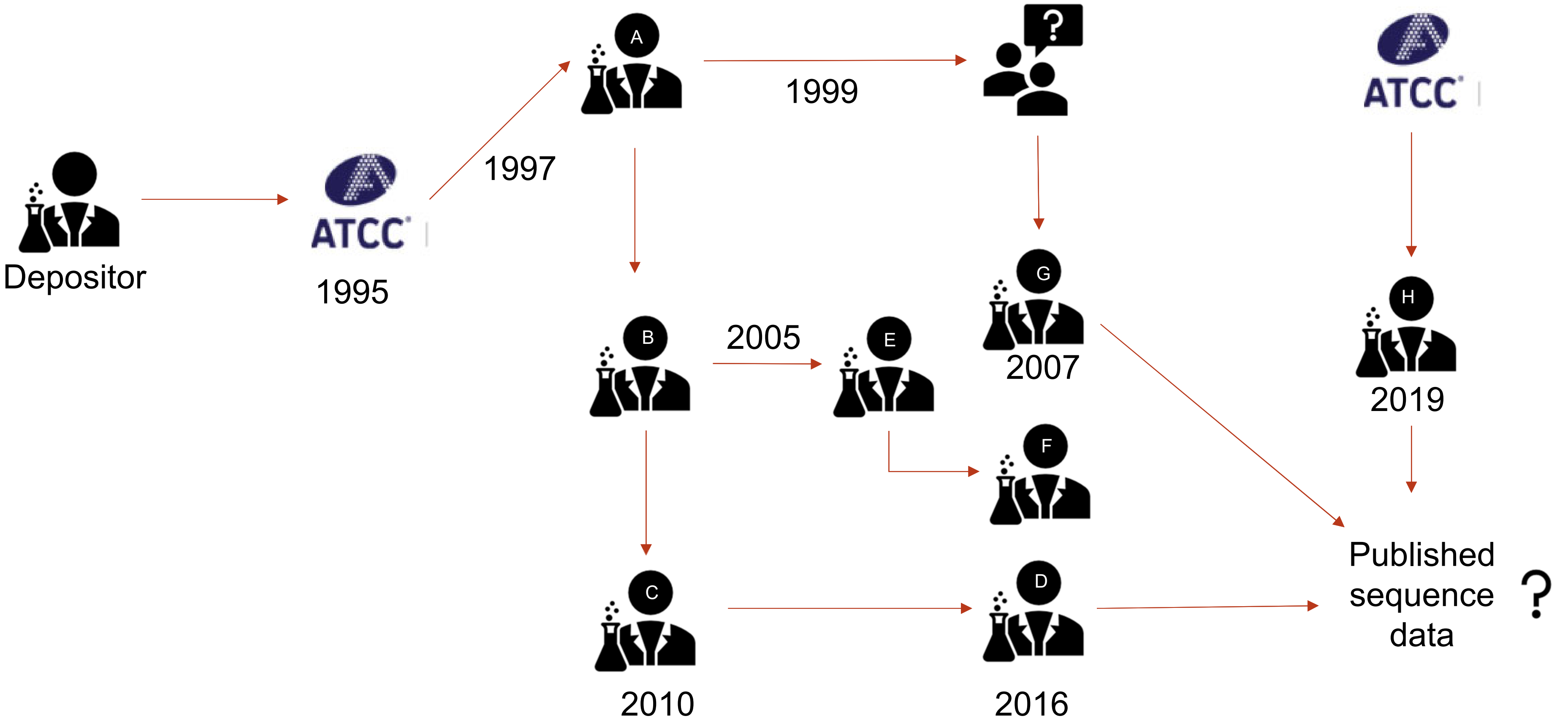
ATCC 17978
RefSeq GCF_001593425.2



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

Genomics data and a traceability and reproducibility crisis

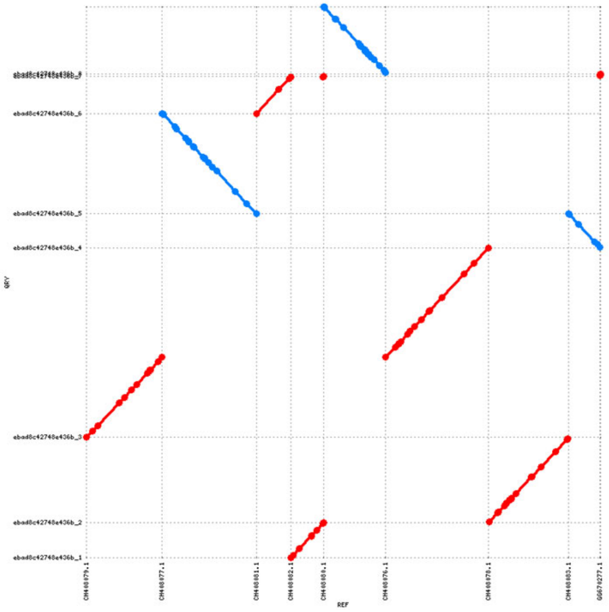


Evaluation of genome sequences from public databases

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

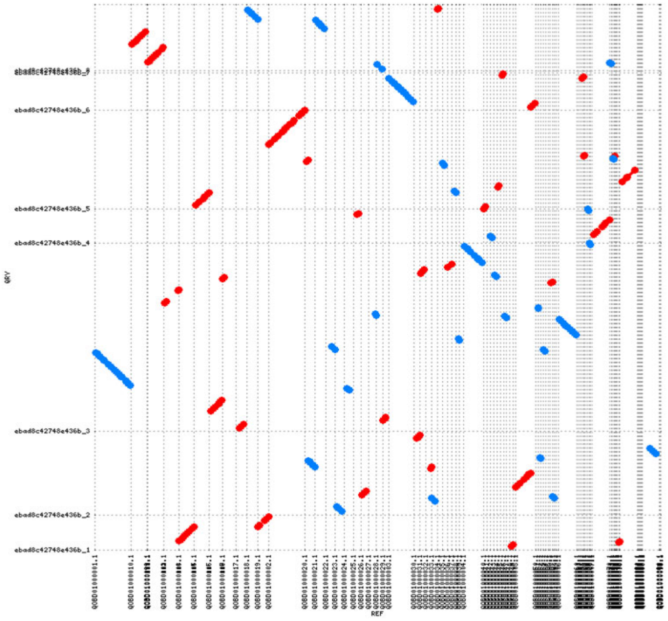
Evaluation of public sequences for ATCC 42720

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 Genome Portal

ATCC 42720
RefSeq GCF_000003835.1



ATCC 42720 Genome Portal

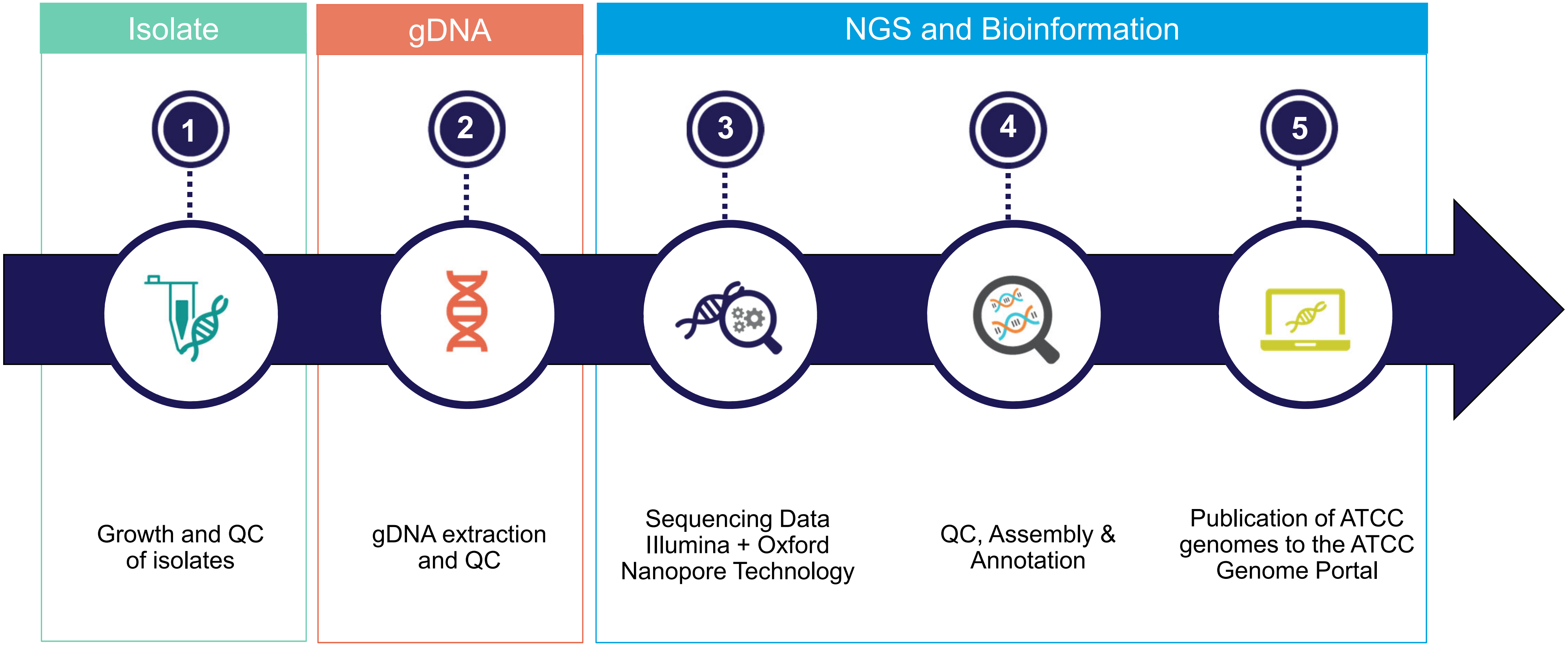
ATCC 42720
RefSeq GCA_003675505.1

Overview

Using next-generation sequencing to further authenticate the ATCC microbial collections

- Discuss *why* ATCC is committed to providing reference-quality genomes for items within the microbial collections
- **Discuss some of the standardized processes and quality control criteria required for extracting, sequencing, and analyzing our reference-quality genomes**
- Explore some of the features of the ATCC Genome Portal

Authenticated physical material coupled with reference-quality genome sequences



Growth and QC of isolates

gDNA extraction and QC

Sequencing Data Illumina + Oxford Nanopore Technology

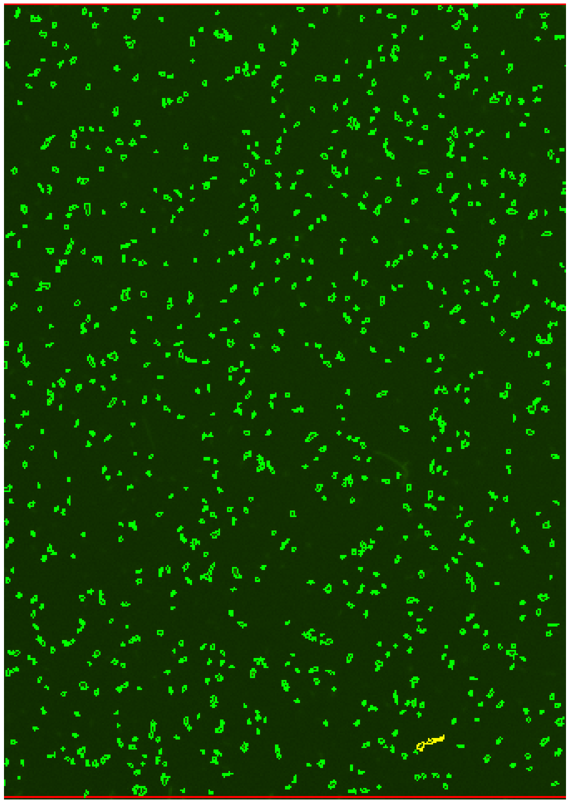
QC, Assembly & Annotation

Publication of ATCC genomes to the ATCC Genome Portal

Processes for producing reference-quality genomes

Extraction of gDNA

- Start with a fresh culture grown according to ATCC's item-specific manufacturing process
- Determine the cell count
  - Typically start with ≥10^9 cells/mL
- The "best" extraction method depends on the organism
- Simply recovering DNA is not good enough
  - Concentration
    - Measured by Qubit™ or Picogreen®
  - Purity
    - Measured with NanoDrop™
    - $A_{260/280}$ ≥1.7 to ≤ 2.1
  - Quality and Integrity
    - Fragment size is measured by Fragment Analyzer™



Fusobacterium nucleatum ATCC® 25586™
6.58 x 10^8 cells /mL

Processes for producing reference-quality genomes

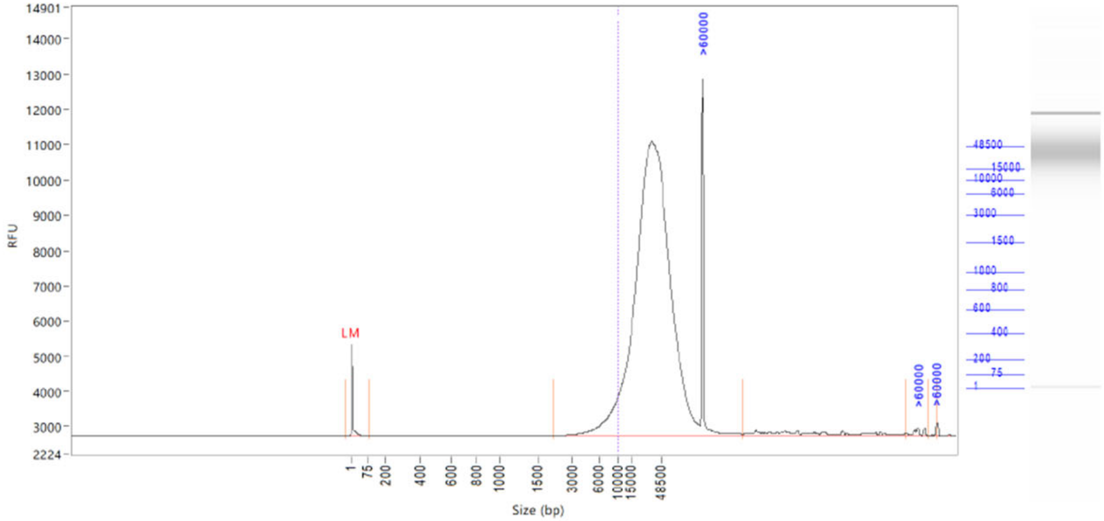
ATCC extraction quality control

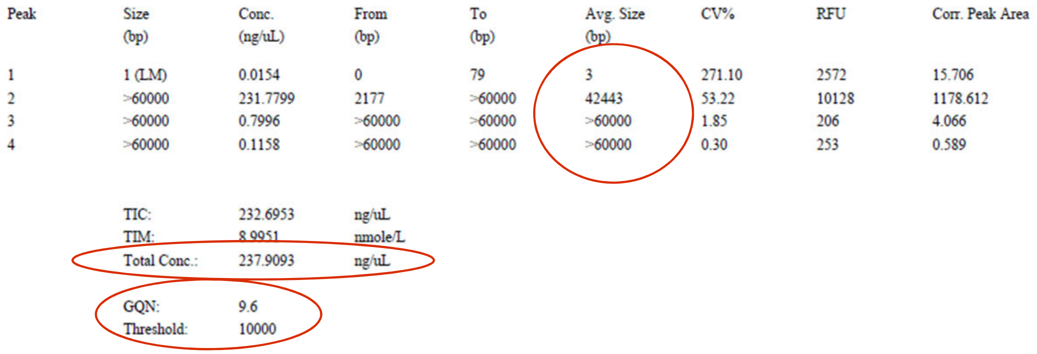
ATCC® no.	Species	Qubit (ng/µL)	$A_{260/}A_{280}$	DNA fragment size (range)**
8739™	*Escherichia coli*	101.9	1.92	49.5 kb (1.5 – >60 kb)
13048™	*Klebsiella aerogenes*	98.1	1.86	49.5 kb (1.6 – >60 kb)
11828™	*Cutibacterium acnes*	197.7	1.84	29.8 kb (0.8 – >60 kb)
6538™	*Staphylococcus aureus*	97.8	1.85	32.9 kb (2.7 – >60 kb)
BAA-2797™	*Pseudomonas aeruginosa*	153.3	1.99	44.1 kb (1.1 – >60 kb)
824™	*Clostridium acetobutylicum*	73.8	2.05	12.5 kb (4.6 – 57.8 kb)
6538™	*Staphylococcus aureus*	37.1	2.00	26.2 kb (6.9 – >60 kb)
27774™	*Desulfovibrio desulfuricans*	69.2	1.99	58.5 kb (13.3 – >60 kb)
11842™	*Lactobacillus delbrueckii*	64.8	2.02	41.9 kb (6.1 – >60 kb)
15697™	*Bifidobacterium longum*	76.2	1.95	51.3 kb (10.5 – >60 kb)

** Main peak reported

Processes for producing reference-quality genomes

Fragment analysis of gDNA





Peak	Size (bp)	Conc. (ng/uL)	From (bp)	To (bp)	Avg. Size (bp)	CV%	RFU	Corr. Peak Area
1	1 (LM)	0.0154	0	79	3	271.10	2572	15.706
2	>60000	231.7799	2177	>60000	42443	53.22	10128	1178.612
3	>60000	0.7996	>60000	>60000	>60000	1.85	206	4.066
4	>60000	0.1158	>60000	>60000	>60000	0.30	253	0.589

TIC:	232.6953	ng/uL
TIM:	8.9951	nmole/L
Total Conc.:	237.9093	ng/uL
GQN:	9.6	
Threshold:	10000	

- *Corynebacterium tuberculostearicum* (ATCC® 35692™)
- Total concentration: 234 ng/µL
- Average fragment size: ≥42,000bp
- GQN: 9.6 with a threshold of 10,000bp
  - “Genomic Quality Number”
  - 96% of the sample contains fragments larger than 10,000 bp

Processes for producing reference-quality genomes

Library preps for both Illumina® and Oxford Nanopore Technologies®

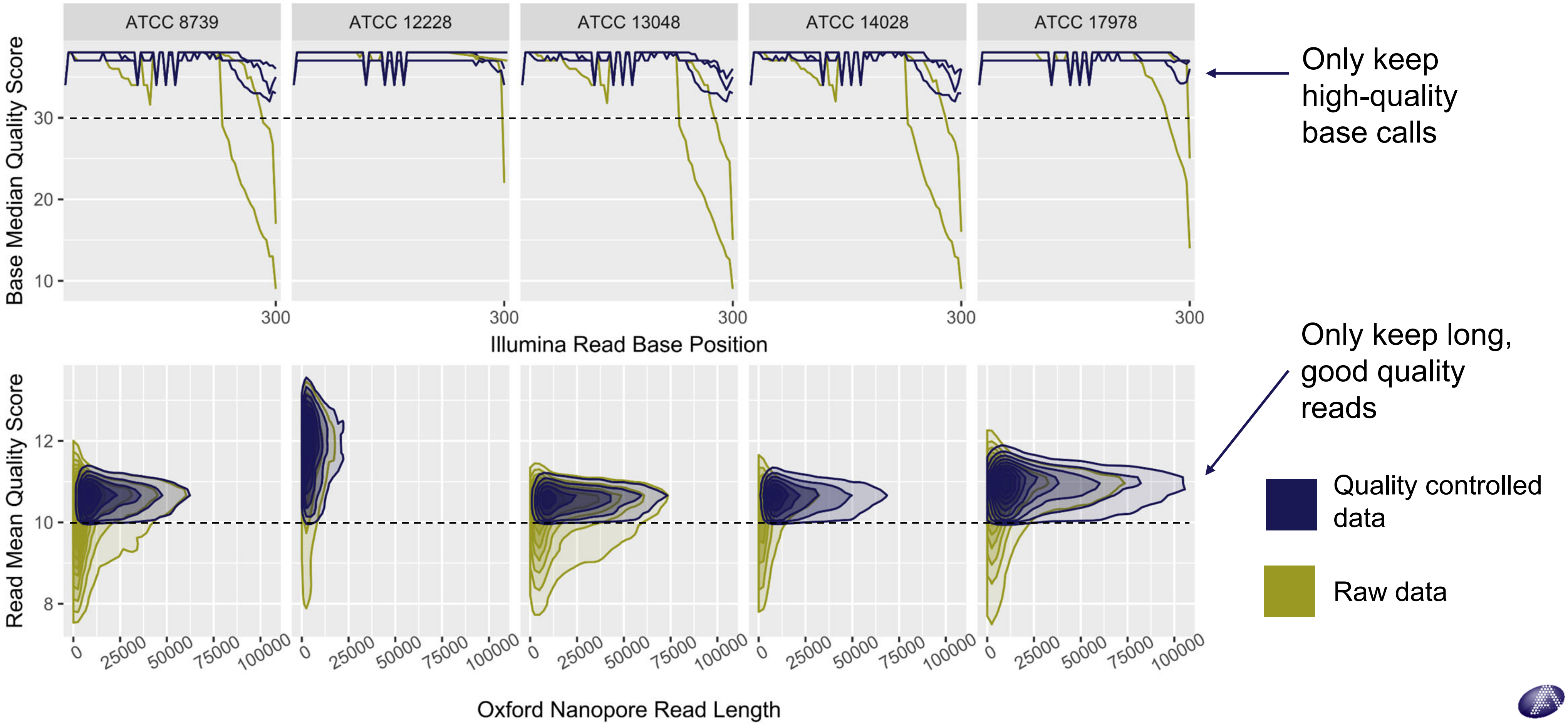
Illumina

- DNA libraries are prepared using Illumina's DNA Prep kit and unique dual indexes (Cat. # 20018705)
- RNA libraries are prepared using NEBNext Ultra II RNA Library Prep Kit (Cat # E7770S)
- Sequenced on the MiSeq® or NextSeq® instrument
  - Paired-end read set per sample
  - Multiplexing is based on the estimated genome size
  - Data necessary to generate at least 100X coverage of the genome
- Reads are adapter trimmed using the adapter trimming option on the Illumina instrument

Oxford Nanopore Technologies

- Libraries are prepared using ONT's Ligation Sequencing Kit (SQK-LSK109) with the Native Barcoding Expansion kit (EXP-NBD104 or EXP-NBD114)
- Sequenced on the GridION using the version 9.4.1 flow cell
- The quantity of samples typically multiplexed is based on the estimated genome size of the given organism.
- Flow cells are run for 48-72 hours
- Barcode detection, demultiplexing, and barcode trimming are completed on the instrument, parallel to the run

Sequencing QC – Read trimming/filtering



Only keep high-quality base calls

Only keep long, good quality reads

■ Quality controlled data

■ Raw data

Processes for producing reference-quality genomes

QC Metrics for both Illumina and Oxford Nanopore Technologies

Illumina

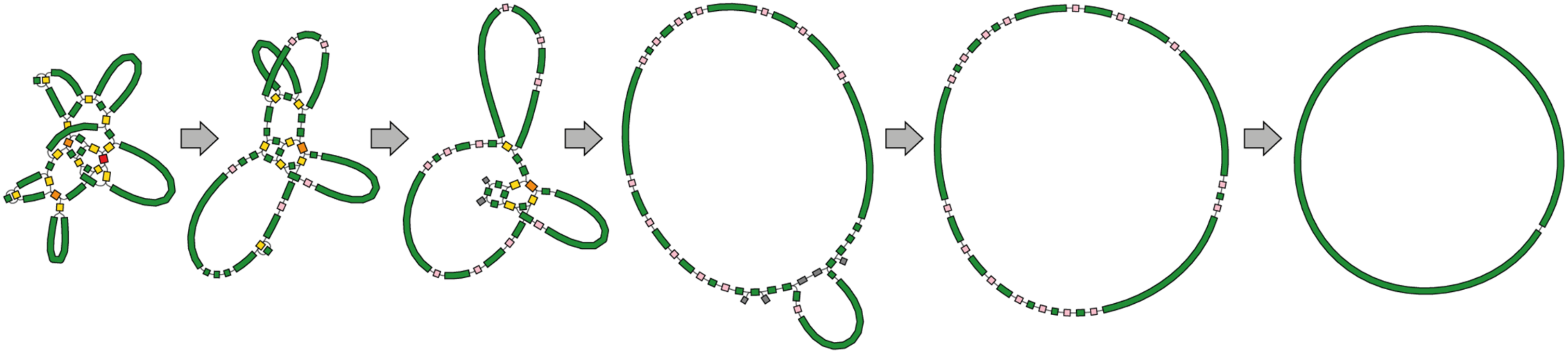
- Remove low-quality regions and adapter sequences
- This also ensures removal of adapter sequences otherwise missed by Illumina software
- Assess the quality of the read set by using FastQC
- Illumina reads must pass the following quality control:
  - Median Q score, all bases > 30
  - Median Q score, per base > 25
  - Ambiguous content (% N bases) < 5%

➕

Oxford Nanopore Technologies

- ONT ultra-long reads are critical for scaffolding over the low-complexity regions of bacterial or fungal genomes during hybrid assembly, but they have limited influence in determining base identity given enough Illumina coverage.
- All data is trimmed and filtered for low-quality regions
- The quality control metrics used across all ONT read sets produced are:
  - Minimum mean Q score, per read > 10
  - Minimum read length > 5000
- To perform this quality control step, we employ NanoFilt on demultiplexed ONT read sets in addition to barcode sequence removal during demultiplexing

Hybrid genome assembly



Illumina-only genome assembly
150 bp reads

Long reads mapped to a tangled region creates a resolved bridge
Successively applying bridges resolves the structure of the genome

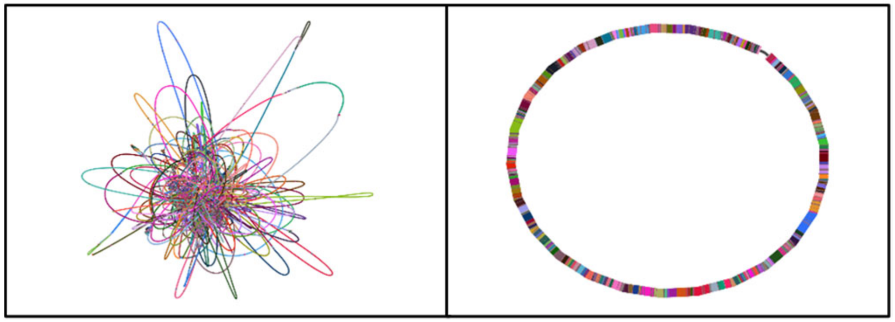
Completed hybrid assembly

Image reproduced from https://github.com/rrwick/Unicycler

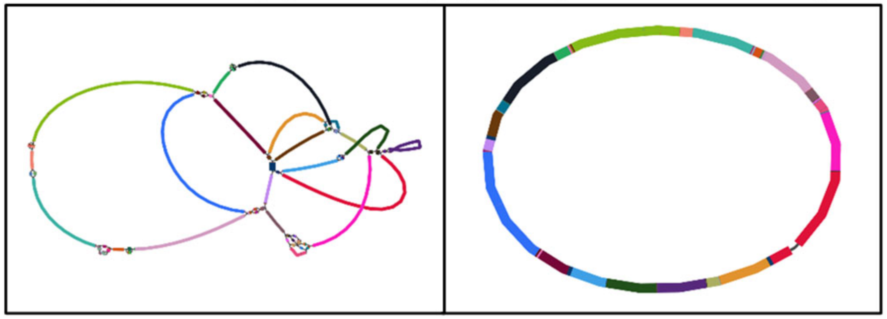
Advantage of hybrid assemblies

Illumina-only assembly	Hybrid assembly

Neisseria meningitidis (ATCC® 53417™)

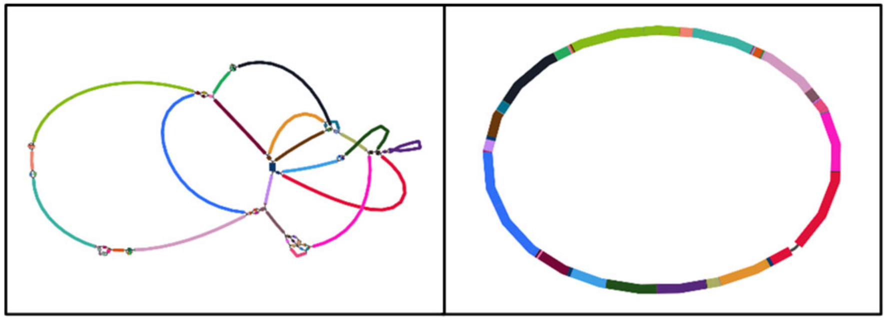


Campylobacter jejuni subsp. *jejuni* (ATCC® 43446™)

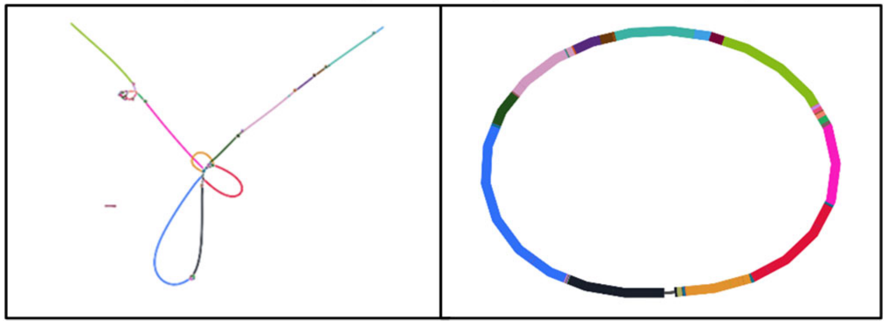


Illumina-only assembly	Hybrid assembly

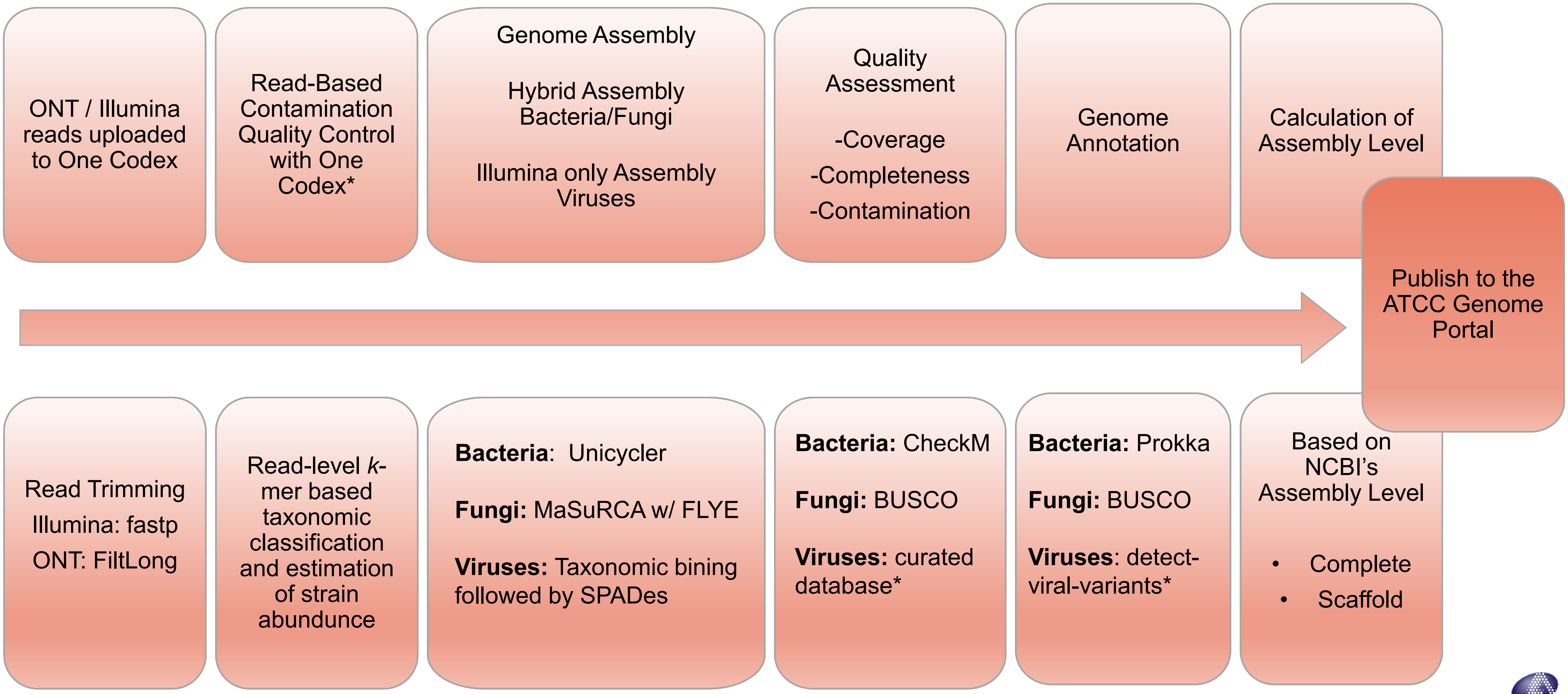
Campylobacter jejuni subsp. *jejuni* (ATCC® 43446™)



Streptococcus gordonii (ATCC® 35105™)



ATCC genome assembly process



ONT / Illumina reads uploaded to One Codex

Read Trimming

Illumina: fastp

ONT: FiltLong

Read-Based Contamination Quality Control with One Codex*

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

Genome Assembly

Hybrid Assembly Bacteria/Fungi

Illumina only Assembly Viruses

Bacteria: Unicycler

Fungi: MaSuRCA w/ FLYE

Viruses: Taxonomic bining followed by SPADes

Quality Assessment

-Coverage

-Completeness

-Contamination

Bacteria: CheckM

Fungi: BUSCO

Viruses: curated database*

Genome Annotation

Bacteria: Prokka

Fungi: BUSCO

Viruses: detect-viral-variants*

Calculation of Assembly Level

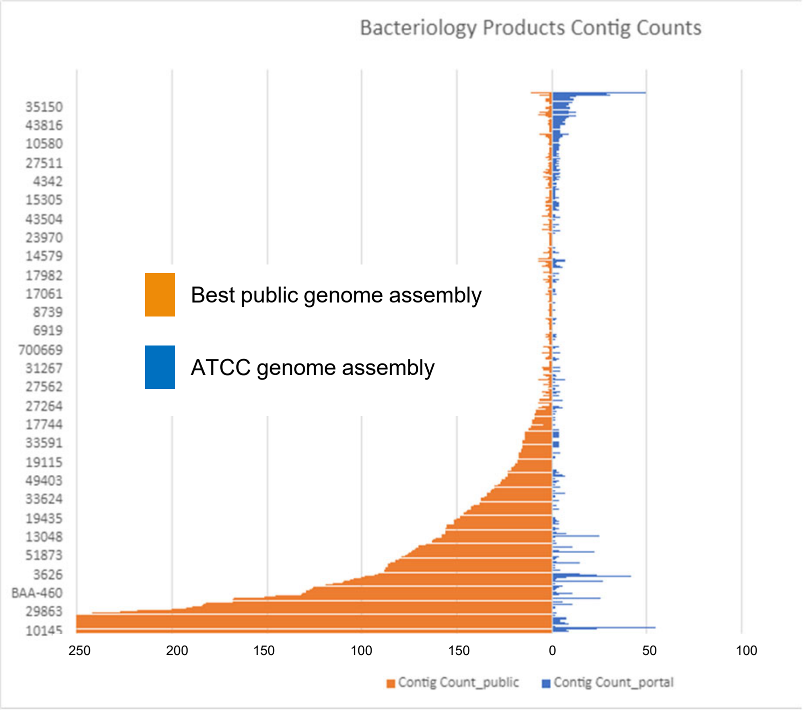
Based on NCBI's Assembly Level

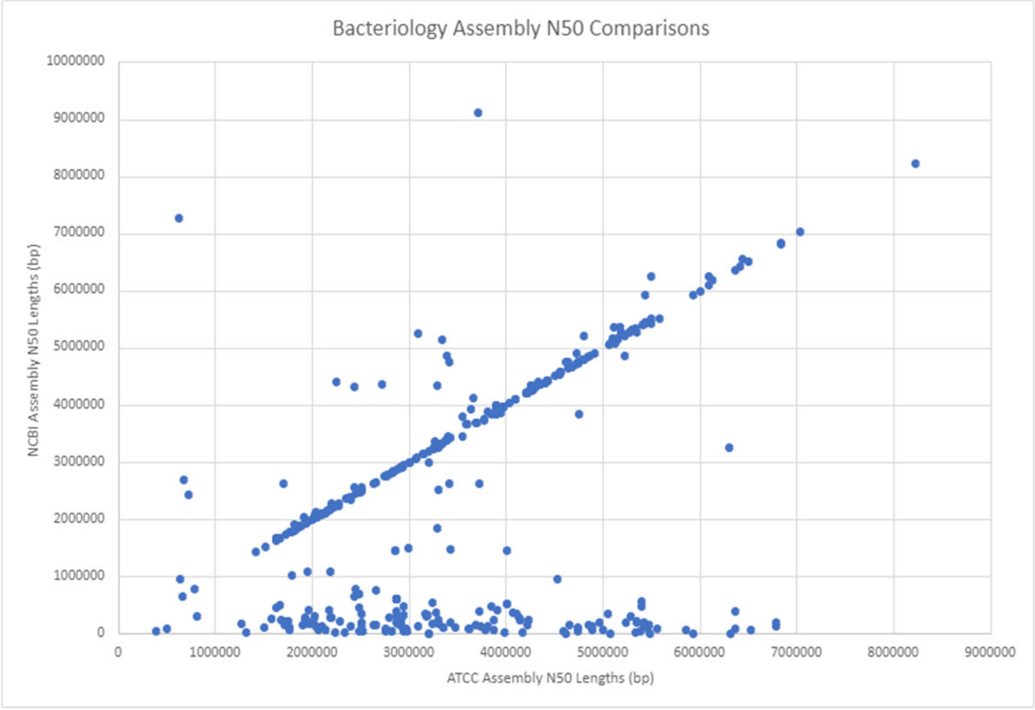
- Complete
- Scaffold

Publish to the ATCC Genome Portal

* One Codex proprietary software. Details and references are available on our technical document.

ATCC assemblies improve upon public assemblies





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

Overview

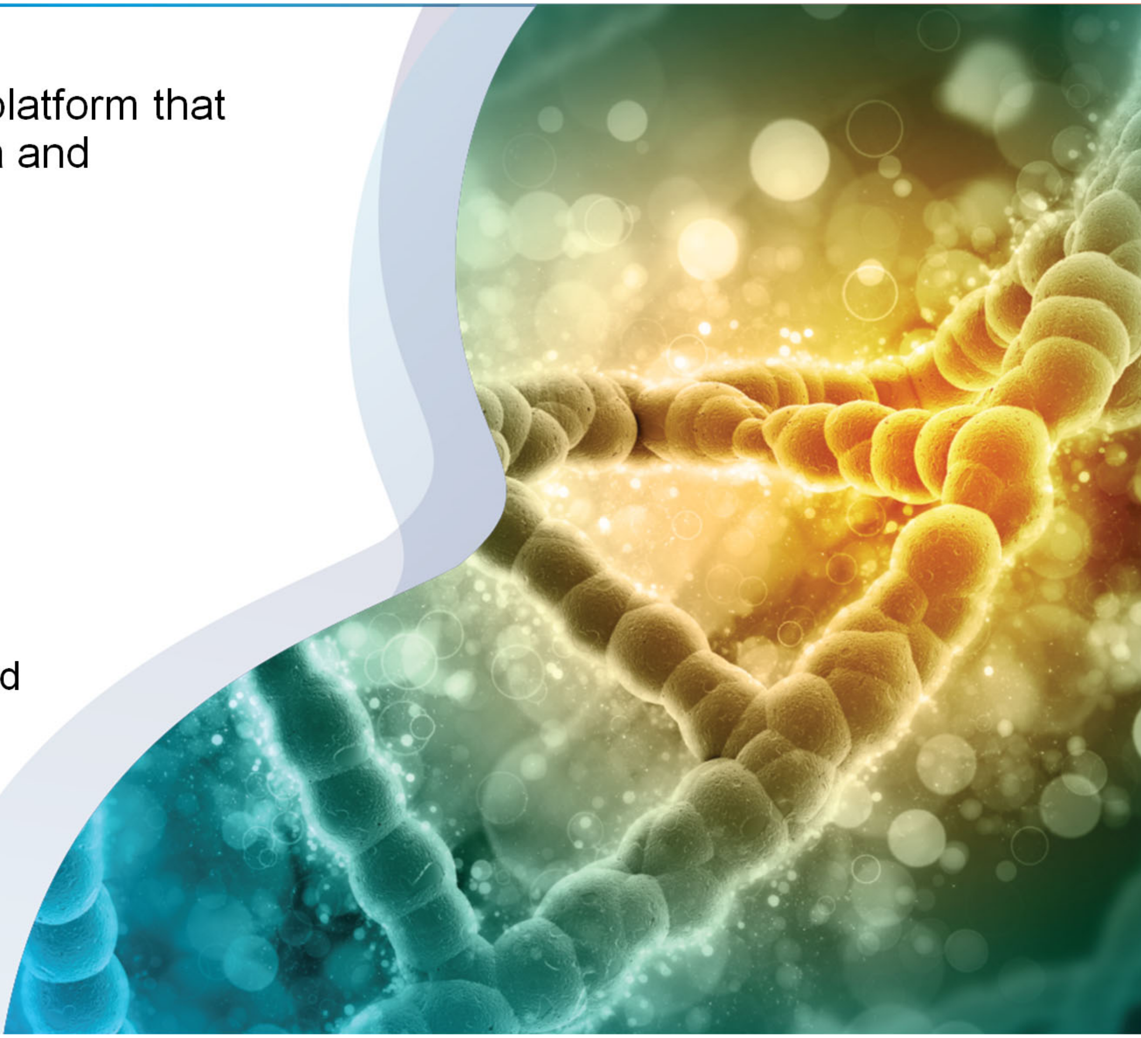
Using next-generation sequencing to further authenticate the ATCC microbial collections

- Discuss *why* ATCC is committed to providing reference-quality genomes for items within the microbial collections
- Discuss some of the standardized processes and quality control criteria required for extracting, sequencing, and analyzing our reference-quality genomes
- **Explore some of the features of the ATCC Genome Portal**

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

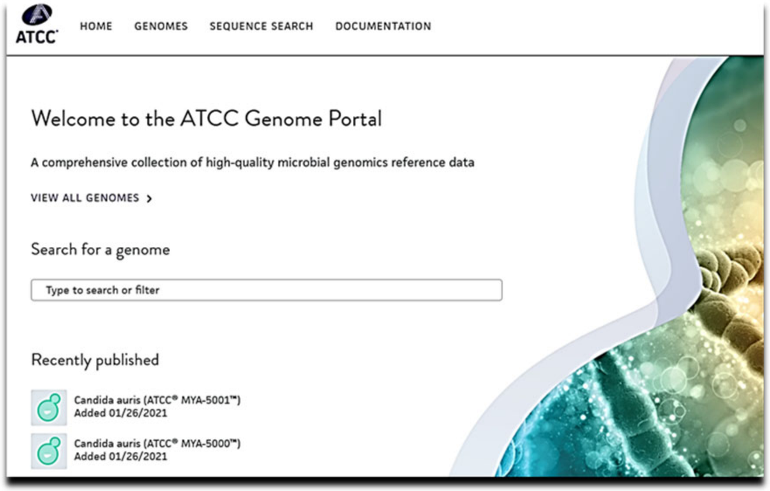


genomes.atcc.org

Summary

Using next-generation sequencing to further authenticate the ATCC microbial collections

- Discussed ***why*** ATCC is committed to **providing reference-quality genomes** for items within the microbial collections
  - traceability and reproducibility crisis
  - authentication in the genomics era
  - provide customers with easily accessible genomic data
- Discussed some of the standardized processes and quality control criteria required for extracting, sequencing, and analyzing our reference-quality genomes
  - gDNA extraction and QC
  - NGS library preps
  - Data QC
  - genome assembly process
- Explored the ATCC Genome Portal



Welcome to the ATCC Genome Portal

A comprehensive collection of high-quality microbial genomics reference data

Search for a genome

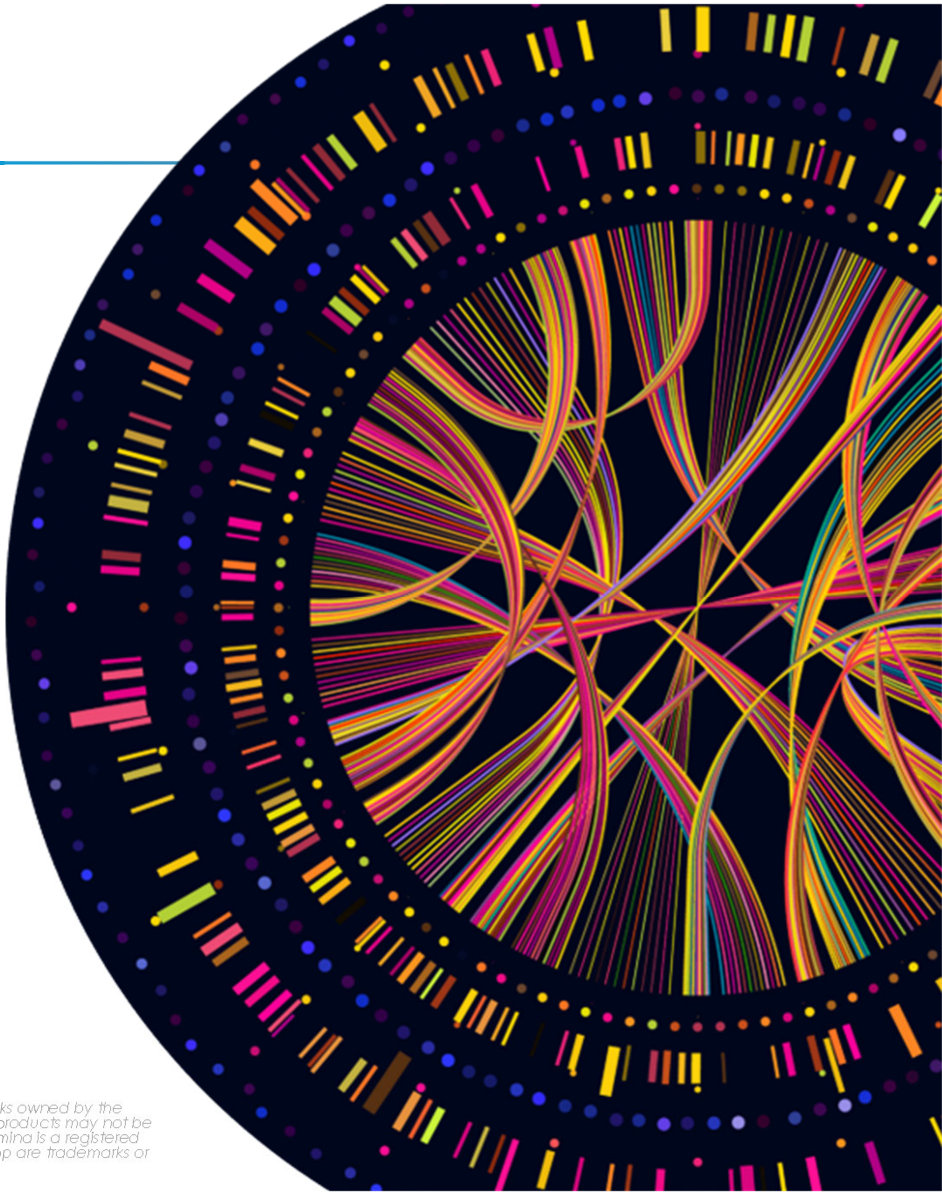
Recently published

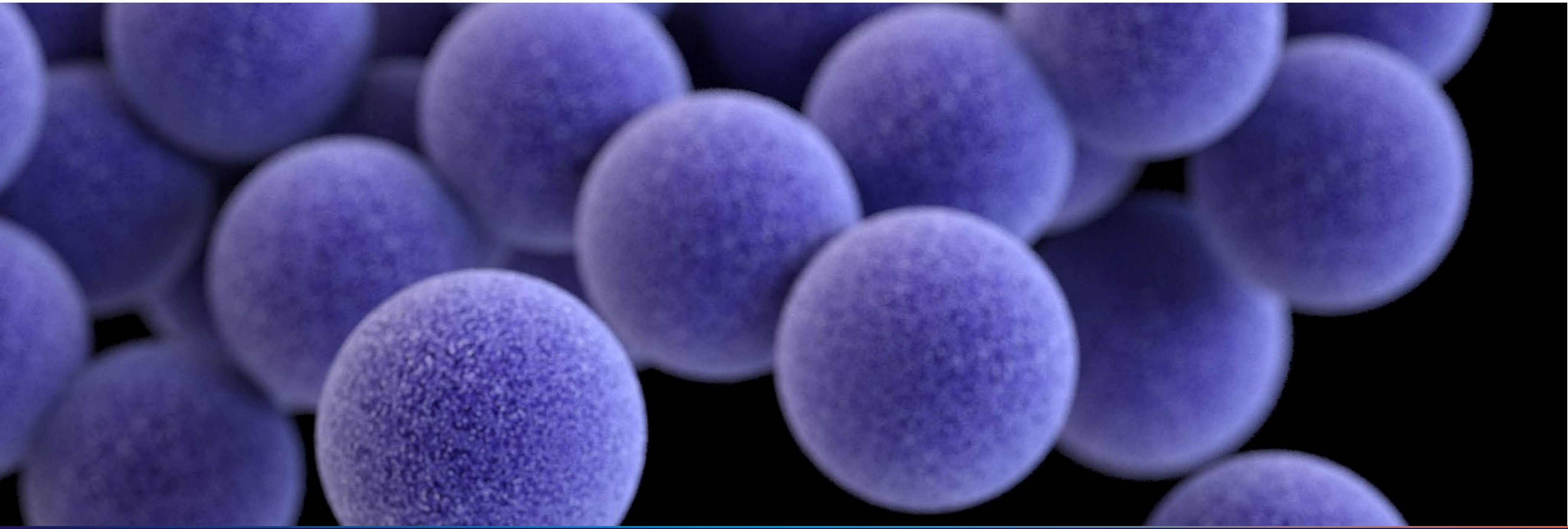
Candida auris (ATCC® MYA-5001™)
Added 01/26/2021

Candida auris (ATCC® MYA-5000™)
Added 01/26/2021

Acknowledgements

- Amanda Pierola, BS
- Brian Shapiro, PhD
- David Yarmosh, MS
- John Bagnoli, BS
- Juan Lopera, PhD
- Jonathan Jacobs, PhD
- P. Ford Combs, MS
- Nikhita Putheveetil, MS
- Samuel Greenfield, BS
- Stephen King, MS
- Our partners at One Codex





Thank you