Utilizing the Scaffolding Ability of Ultra-Long Oxford Nanopore Reads to Assemble Reference-Grade Genomes Originating from Authenticated ATCC Materials

ATCC®

Credible leads to Incredible™

Briana Benton,[1] Andrew Frank,[1] Anna McCluskey,[1] Steve King,[1] Nick Greenfield,[2] Juan Lopera[1]
[1]ATCC, Manassas, VA 20110; [2]One Codex, San Francisco, CA 94110

Background

The advancement and accessibility of next-generation sequencing (NGS) technologies have rapidly transformed microbiological research by providing the ability to analyze and profile microbial communities via metagenomics analyses. These sequencing-based applications have relied on the availability of fully assembled reference genomes for bioinformatics analyses, particularly for variant calling in diagnostic and clinical microbiology. However, despite the availability of existing genome sequences in public databases, the quality, completeness, authenticity, accuracy, and traceability of genomic data are inadequate; the lack of standards for genome quality exacerbates these underlying problems. To address this, ATCC has implemented a robust NGS and genome assembly workflow to advance authentication of bacterial strains in the ATCC collection. Our workflow is accompanied by rigorous quality control methods and criteria to ensure that the data proceeding to the next step are the highest quality. Only data that pass all quality control criteria are published to the ATCC Genome Portal, an online database of reference-grade bacterial genomes.

Whole-Genome Sequencing Workflow

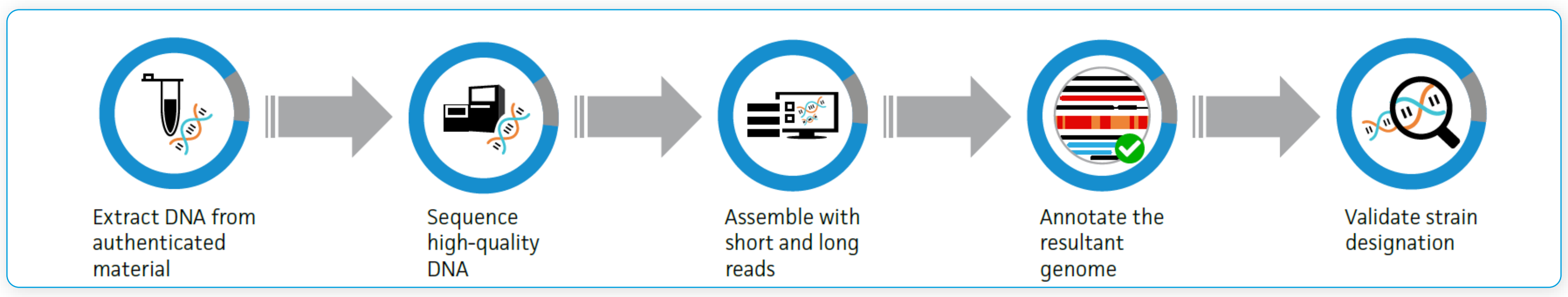

Extract DNA from authenticated material → Sequence high-quality DNA → Assemble with short and long reads → Annotate the resultant genome → Validate strain designation

Figure 1. Bacterial whole-genome sequencing workflow.

Extraction of Authenticated Material

ATCC® No.	Organism	PicoGreen® (ng/μL)	A_{260}/A_{280} Ratio	Mean Fragment size
8739™	*Escherichia coli*	101.9	1.92	>60kb
12228™	*Staphylococcus epidermidis*	76.0	1.80	48kb
13048™	*Klebsiella aerogenes*	98.1	1.86	47kb
14028™	*Salmonella enterica* subsp. *enterica*	88	1.84	59kb
17978™	*Acinetobacter baumannii*	133.3	1.91	46kb

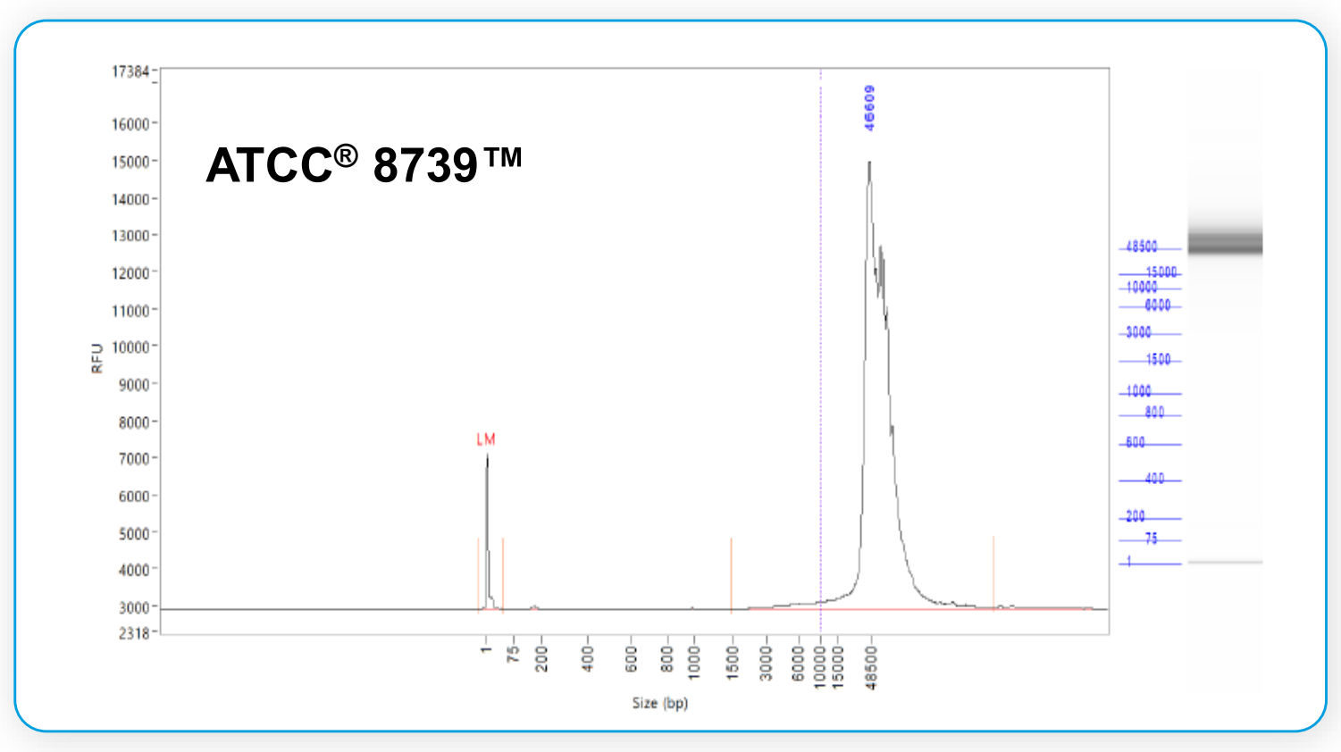


Figure 2. Assessment of quality and quantity of extracted genomic DNA. Fragment size graph obtained from the Agilent Fragment Analyzer platform.

Sequencing High-Quality DNA

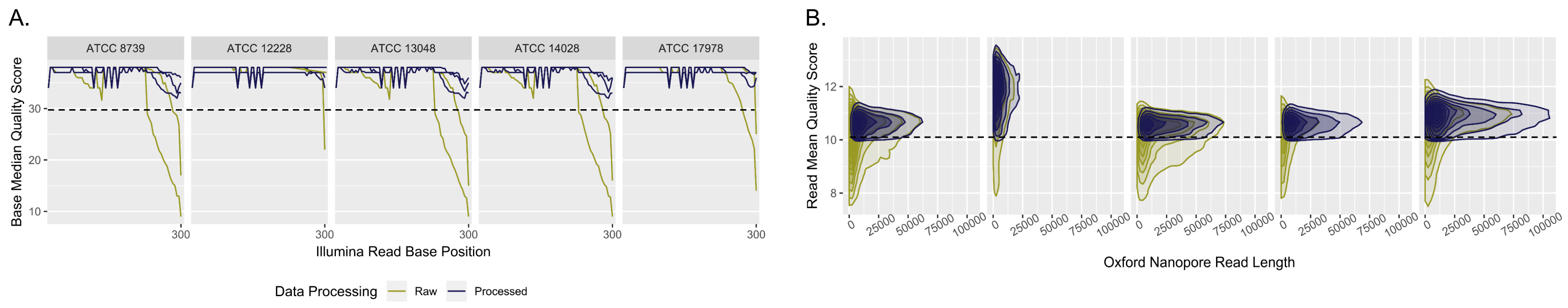
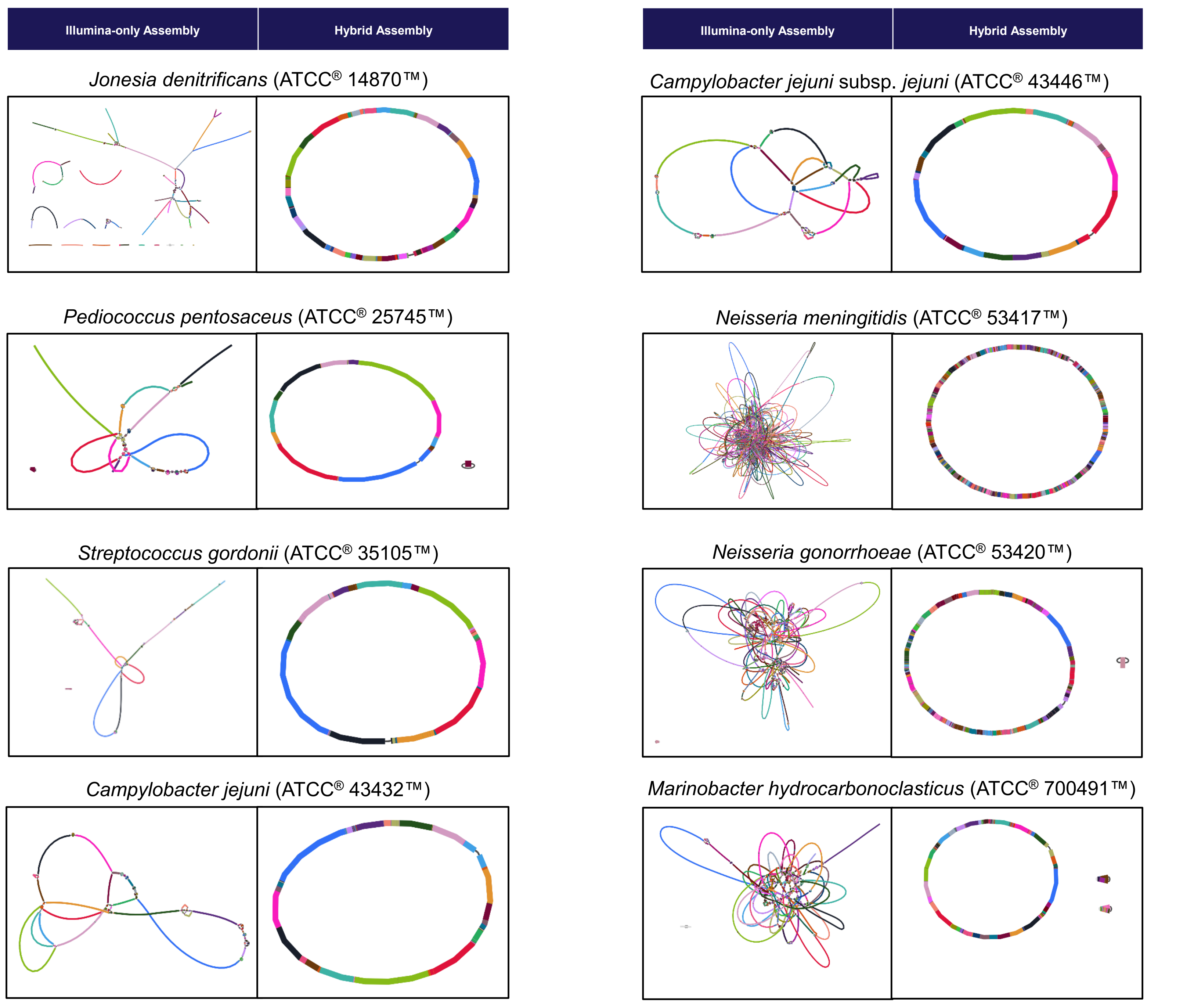


Figure 3. ATCC's bacterial genome sequencing quality control (A) substantially improves the quality of Illumina® reads, and (B) improves the length distribution of reads from the Oxford Nanopore Technologies® (ONT) platform. This approach ensures the longest, highest-quality reads are used for assembly. The dashed line indicates the quality score cutoff used for each sequencing technology.

Assembling with Short and Long Reads



Jonesia denitrificans (ATCC® 14870™)

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

Pediococcus pentosaceus (ATCC® 25745™)

Neisseria meningitidis (ATCC® 53417™)

Streptococcus gordonii (ATCC® 35105™)

Neisseria gonorrhoeae (ATCC® 53420™)

Campylobacter jejuni (ATCC® 43432™)

Marinobacter hydrocarbonoclasticus (ATCC® 700491™)

Figure 4. Hybrid assembly is a state-of-the-art technique that uses both highly accurate Illumina short reads and ultra-long scaffolding ONT reads. In general, this technique begins with an optimized Illumina assembly. The longest of these resultant contigs are then assembled alongside the ONT reads; this combined assembly then undergoes multiple rounds of both long-read and short-read polishing.

Analysis and Comparisons

Table 1. Reproducibility of ATCC *de-novo* assemblies. Using the ATCC genome assembly workflow, we are able to consistently replicate datasets.

Organism	ATCC® No.	Sequence dataset	total consensus (Mbp)	# of contigs (circular)	N50 (Mbp)	Largest contig (Mbp)	Smallest contig (Mbp)	# of Ns	%GC
Salmonella enterica	14028™	dataset 1	4.96	3	4.78	4.78	0.08	0	52.2
		dataset 2	4.96	3	4.78	4.78	0.08	0	52.2
Escherichia coli	8739™	dataset 1	4.75	1	4.75	4.75	4.75	0	50.9
		dataset 2	4.75	1	4.75	4.75	4.75	0	50.9
Acinetobacter baumannii	17978™	dataset 1	4.08	4	3.90	3.90	0.01	0	38.9
		dataset 2	4.08	4	3.90	3.90	0.01	0	38.9

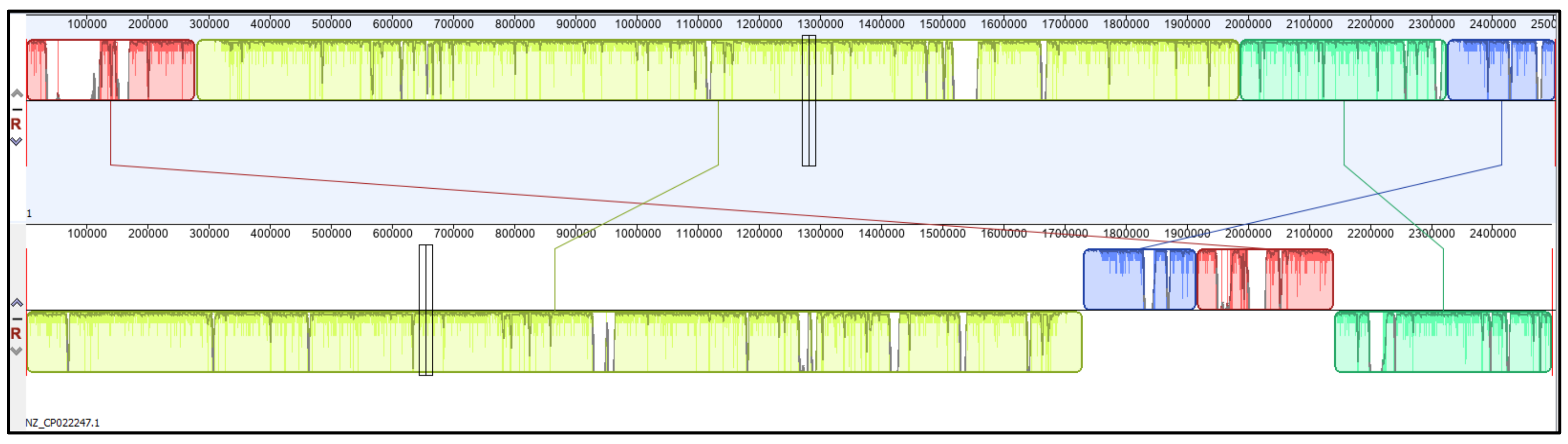


Figure 5. Often times, publicly available data is incomplete, lacks in depth traceability, or is incorrect. Here, we show an assembly of ATCC® 12228™ from NCBI compared with our own assembly of ATCC® 12228™. Clear differences are evident in our assembly (top) verses the RefSeq PacBio-only assembly.

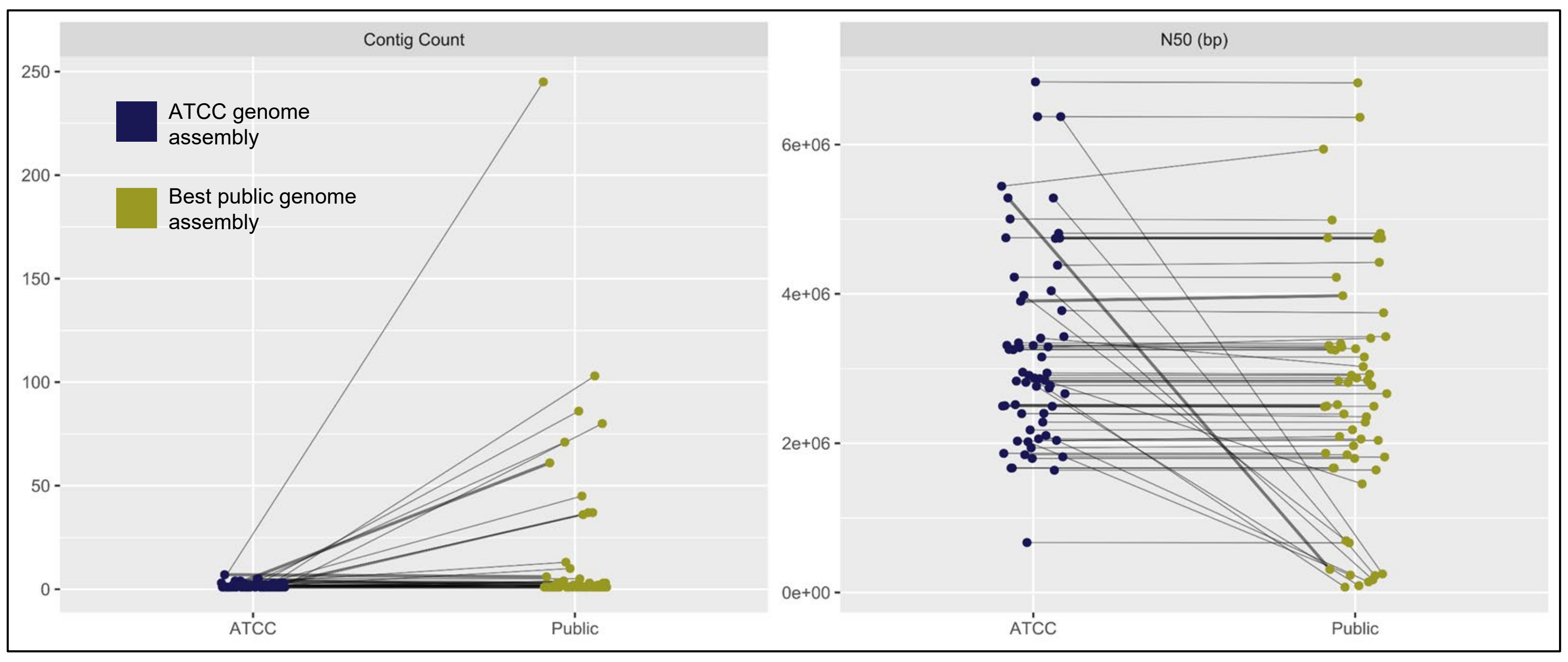
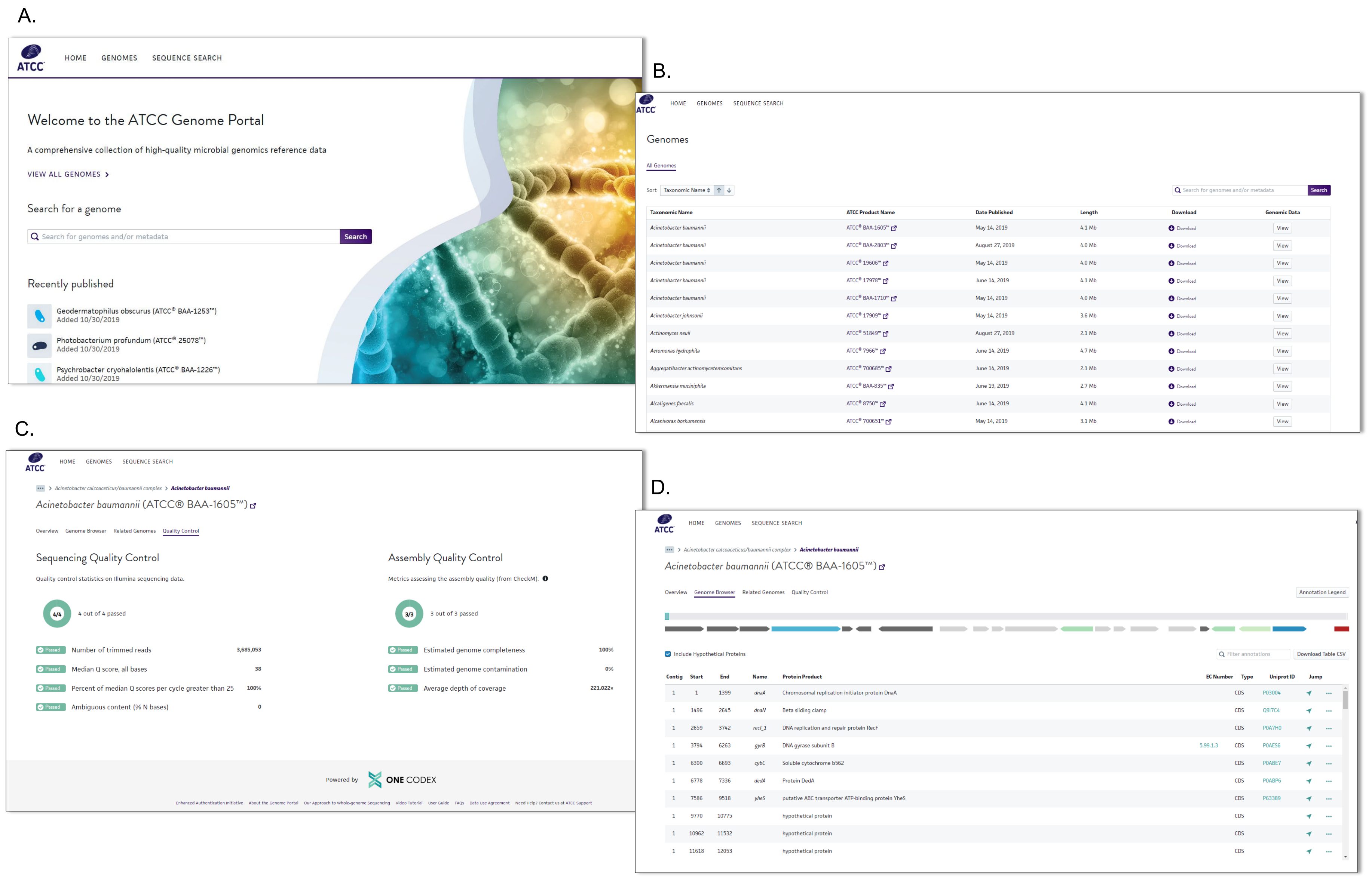


Figure 6. **Downward** trend in contig count and **upward** trend in N50 indicate ATCC is producing more complete genomes than the best publicly available alternative genome.



A.

Welcome to the ATCC Genome Portal

A comprehensive collection of high-quality microbial genomics reference data

VIEW ALL GENOMES ›

Search for a genome

Search for genomes and/or metadata [Search]

Recently published

Geodermatophilus obscurus (ATCC® BAA-1253™)
Added 10/30/2019

Photobacterium profundum (ATCC® 25078™)
Added 10/30/2019

Psychrobacter cryohalolentis (ATCC® BAA-1226™)
Added 10/30/2019

B.

Genomes

All Genomes

Sort: Taxonomic Name

Search for genomes and/or metadata [Search]

Taxonomic Name	ATCC Product Name	Date Published	Length	Download	Genomic Data
Acinetobacter baumannii	ATCC® BAA-1605™	May 14, 2019	4.1 Mb	Download	View
Acinetobacter baumannii	ATCC® BAA-2803™	August 27, 2019	4.0 Mb	Download	View
Acinetobacter baumannii	ATCC® 19606™	May 14, 2019	4.0 Mb	Download	View
Acinetobacter baumannii	ATCC® 17978™	June 14, 2019	4.1 Mb	Download	View
Acinetobacter baumannii	ATCC® BAA-1710™	May 14, 2019	4.0 Mb	Download	View
Acinetobacter johnsonii	ATCC® 17909™	May 14, 2019	3.6 Mb	Download	View
Actinomyces neuii	ATCC® 51849™	August 27, 2019	2.1 Mb	Download	View
Aeromonas hydrophila	ATCC® 7966™	June 14, 2019	4.7 Mb	Download	View
Aggregatibacter actinomycetemcomitans	ATCC® 700685™	June 14, 2019	2.1 Mb	Download	View
Akkermansia muciniphila	ATCC® BAA-835™	June 19, 2019	2.7 Mb	Download	View
Alcaligenes faecalis	ATCC® 8750™	June 14, 2019	4.1 Mb	Download	View
Alcanivorax borkumensis	ATCC® 700651™	May 14, 2019	3.1 Mb	Download	View

C.

... › *Acinetobacter calcoaceticus/baumannii complex* › ***Acinetobacter baumannii***

Acinetobacter baumannii (ATCC® BAA-1605™)

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	3,685,053
Passed	Median Q score, all bases	38
Passed	Percent of median Q scores per cycle greater than 25	100%
Passed	Ambiguous content (% N bases)	0

Assembly Quality Control

Metrics assessing the assembly quality (from CheckM).

3/3 — 3 out of 3 passed

Passed	Estimated genome completeness	100%
Passed	Estimated genome contamination	0%
Passed	Average depth of coverage	221.022×

Powered by ONE CODEX

Enhanced Authentication Initiative · About the Genome Portal · Our Approach to Whole-genome Sequencing · Video Tutorial · User Guide · FAQs · Data Use Agreement · Need Help? Contact us at ATCC Support

D.

... › *Acinetobacter calcoaceticus/baumannii complex* › ***Acinetobacter baumannii***

Acinetobacter baumannii (ATCC® BAA-1605™)

Overview | Genome Browser | Related Genomes | Quality Control | Annotation Legend

☑ Include Hypothetical Proteins

Filter annotations | Download Table CSV

Contig	Start	End	Name	Protein Product	EC Number	Type	Uniprot ID	Jump
1	1	1399	*dnaA*	Chromosomal replication initiator protein DnaA		CDS	P03004	
1	1496	2645	*dnaN*	Beta sliding clamp		CDS	Q9I7C4	
1	2659	3742	*recF_1*	DNA replication and repair protein RecF		CDS	P0A7H0	
1	3794	6263	*gyrB*	DNA gyrase subunit B	5.99.1.3	CDS	P0AES6	
1	6300	6693	*cybC*	Soluble cytochrome b562		CDS	P0ABE7	
1	6778	7336	*dedA*	Protein DedA		CDS	P0ABP6	
1	7586	9518	*yheS*	putative ABC transporter ATP-binding protein YheS		CDS	P63389	
1	9770	10775		hypothetical protein		CDS		
1	10962	11532		hypothetical protein		CDS		
1	11618	12053		hypothetical protein		CDS		

Figure 7. Screenshots from the new ATCC Genome Portal. The top left figure illustrates the "Home" page followed by the "Genomes" page, which lists all genomes that pass our rigorous QC criteria. Here, users are able to download the respective fasta and/or gbk files for each organism. (C) The "Quality Control" page, which displays the QC values for each respective genome, and (D) the "Genome Browser" page, which illustrates the annotation of each respective genome. All data is available at genomes.atcc.org.

Summary

- Our hybrid-assembly method allows us to leverage the power of highly accurate Illumina short-reads with the scaffolding ability of ultra-long ONT reads to generate high-quality reference genomes that are more complete and accurate than what can be generated with each individual technology alone.
- Our standardized and reproducible genome sequencing, assembly, and annotation workflow allows researchers to access higher-quality genomes that are fully authenticated and matched with authenticated ATCC strains.
- ATCC will continue to sequence and assemble the genomes of organisms in our bacteriology collection and make them publicly available to the scientific research community via monthly additions to the ATCC Genome Portal. Data are accessible at genomes.atcc.org.
- We are currently working to expand our applications of the ONT sequencing platform to further enrich the characterization of our virology and mycology collections.

ATCC 10801 University Boulevard, Manassas, Virginia 20110-2209 **Phone:** 703.263.8277 **Email:** SalesRep@atcc.org **Web:** www.atcc.org