June 21, 2019
FRIDAY - MBP-71

Microbial Whole Genome Sequencing and Assembly: Existing Challenges and the Need for Authentic Reference Genomes

ATCC®
Credible leads to Incredible™

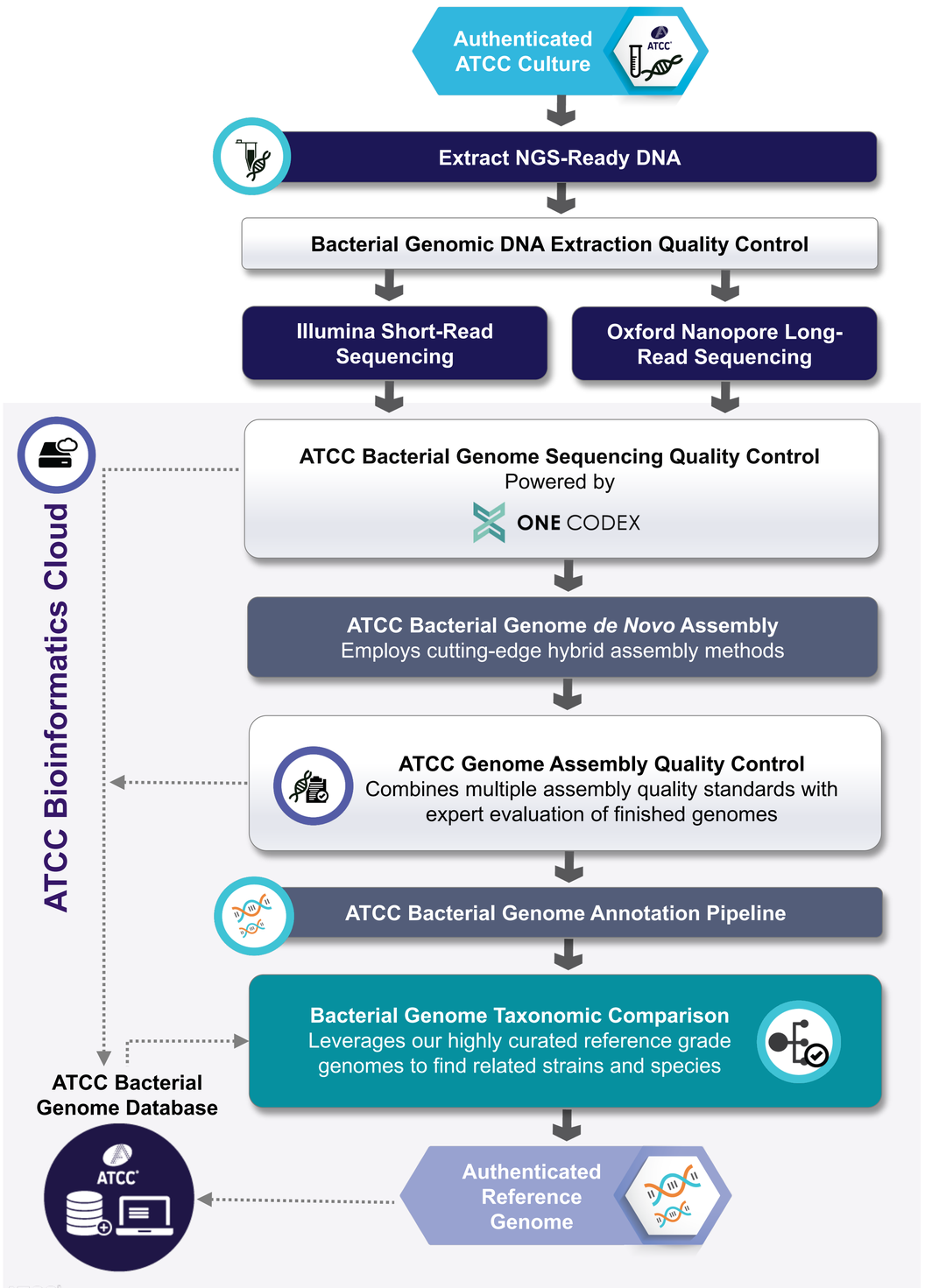
Andrew Frank, MS,[1] Anna McCluskey,[1] Steve King, MS,[1] Nick Greenfield, MA,[2] Juan Lopera, PhD[1]
[1]ATCC, Manassas, VA 20110; [2]One Codex, San Francisco, CA 94110

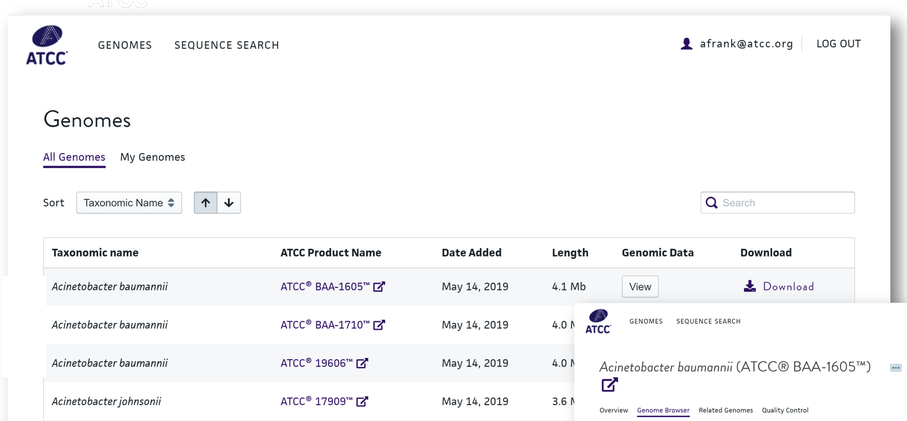
Background

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 is 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.

Whole Genome Sequencing Workflow

1. Proprietary Extraction Process for Authenticated Material
2. Sequencing High Quality DNA
3. Assembling with Short and Long Reads
4. Validating Strain Designation and Publication





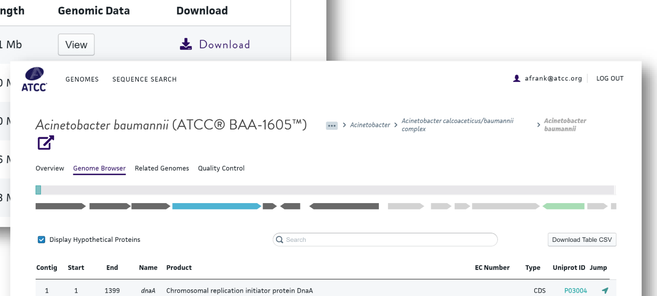


Figure 1. Comprehensive ATCC bacterial whole genome sequencing workflow, and screenshots from the new ATCC genome portal.

Extraction of Authenticated Material

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

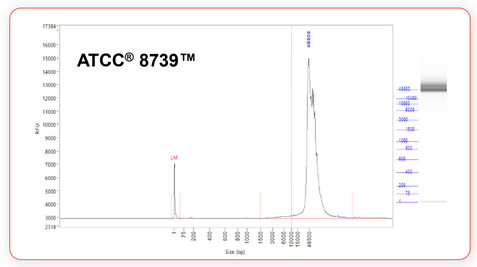


Figure 2. Assessment of quality and quantity of extracted genomic DNA. Fragment size graph obtained from the Agilent Fragment Analyzer platform.

Sequencing High-Quality DNA

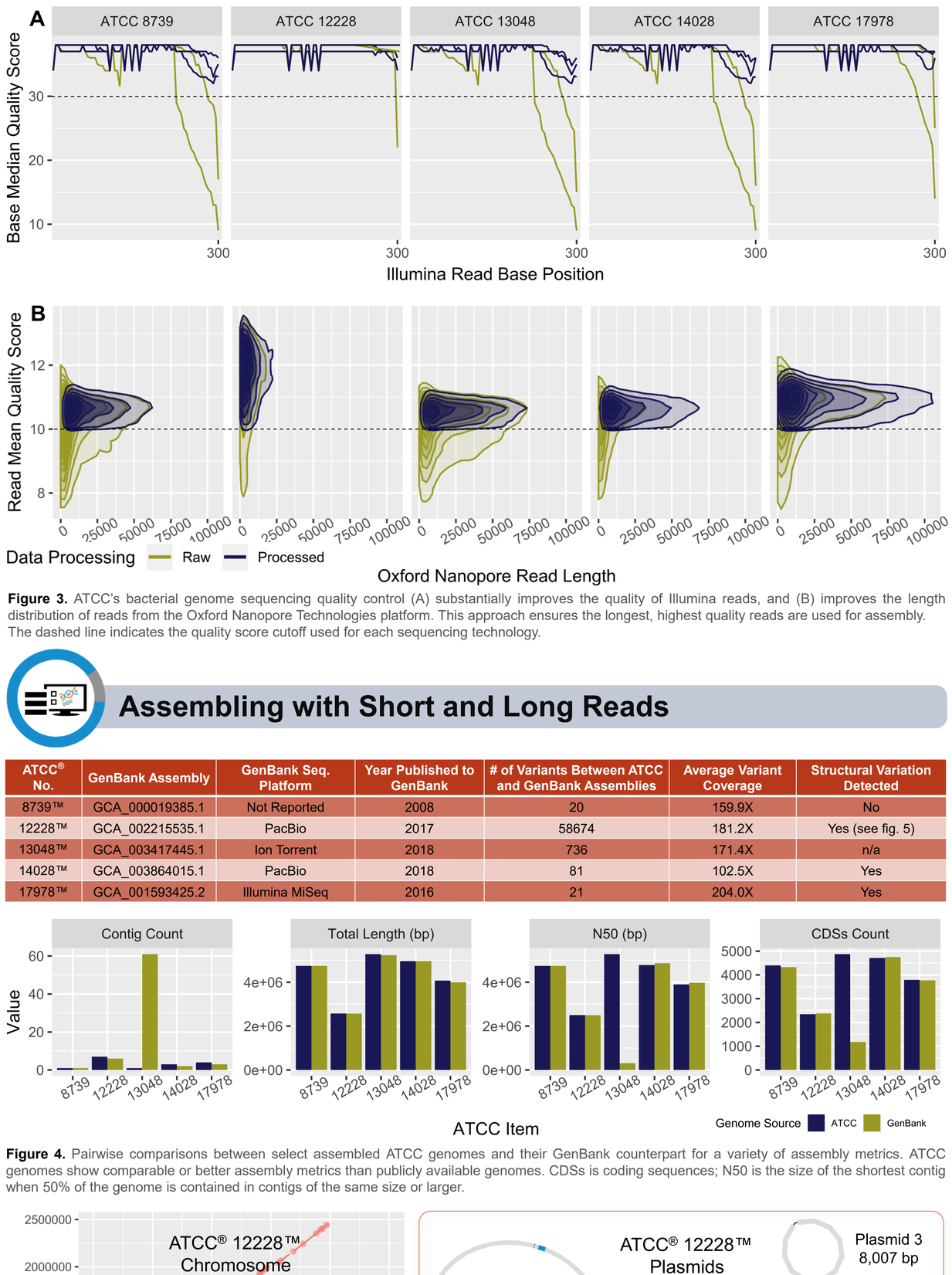


Figure 3. ATCC's bacterial genome sequencing quality control (A) substantially improves the quality of Illumina reads, and (B) improves the length distribution of reads from the Oxford Nanopore Technologies platform. This approach ensures the longest, highest quality reads are used for assembly. The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

Assembling with Short and Long Reads

ATCC® No.	GenBank Assembly	GenBank Seq. Platform	Year Published to GenBank	# of Variants Between ATCC and GenBank Assemblies	Average Variant Coverage	Structural Variation Detected
8739™	GCA_000019385.1	Not Reported	2008	20	159.9X	No
12228™	GCA_002215535.1	PacBio	2017	58674	181.2X	Yes (see fig. 5)
13048™	GCA_003417445.1	Ion Torrent	2018	736	171.4X	n/a
14028™	GCA_003864015.1	PacBio	2018	81	102.5X	Yes
17978™	GCA_001593425.2	Illumina MiSeq	2016	21	204.0X	Yes

Figure 4. Pairwise comparisons between select assembled ATCC genomes and their GenBank counterpart for a variety of assembly metrics. ATCC genomes show comparable or better assembly metrics than publicly available genomes. CDSs is coding sequences; N50 is the size of the shortest contig when 50% of the genome is contained in contigs of the same size or larger.

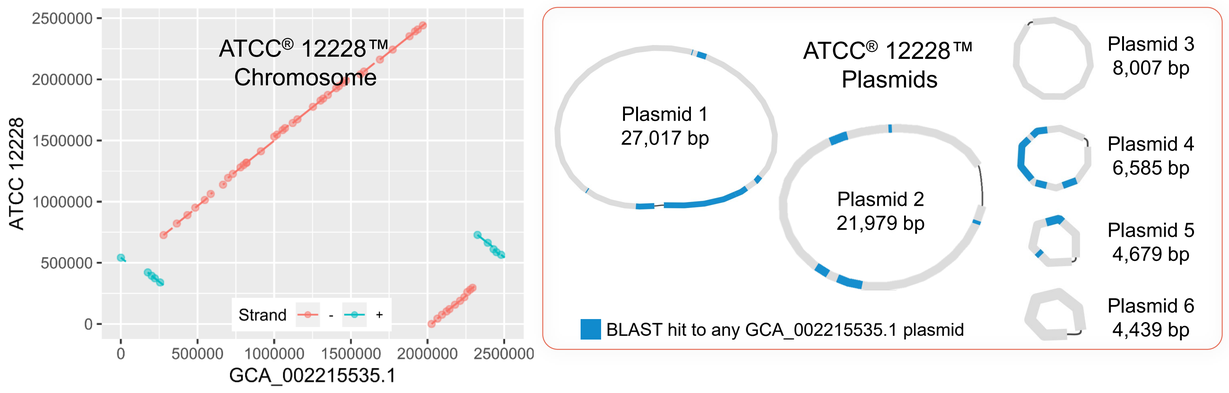


Figure 5. MUMmer alignment of ATCC *de novo* genome assembly of ATCC® 12228™ versus GenBank RefSeq genome assembly GCA_002215535.1, and plasmid alignments. Results are indicative of substantial structural variation and no complete matching plasmids between assemblies.

Summary

- We have found numerous genomes that contain substantial variations as compared to their public database counterparts; variations may be attributable to differences in strain propagation, DNA extraction, sequencing, and downstream analysis that influence the overall quality of data in historical sequencing databases.
- Our standardized and reproducible genome sequencing, assembly, and annotation workflow allows researchers to access higher-quality genomes that are fully authenticated and matched with ATCC strains.

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