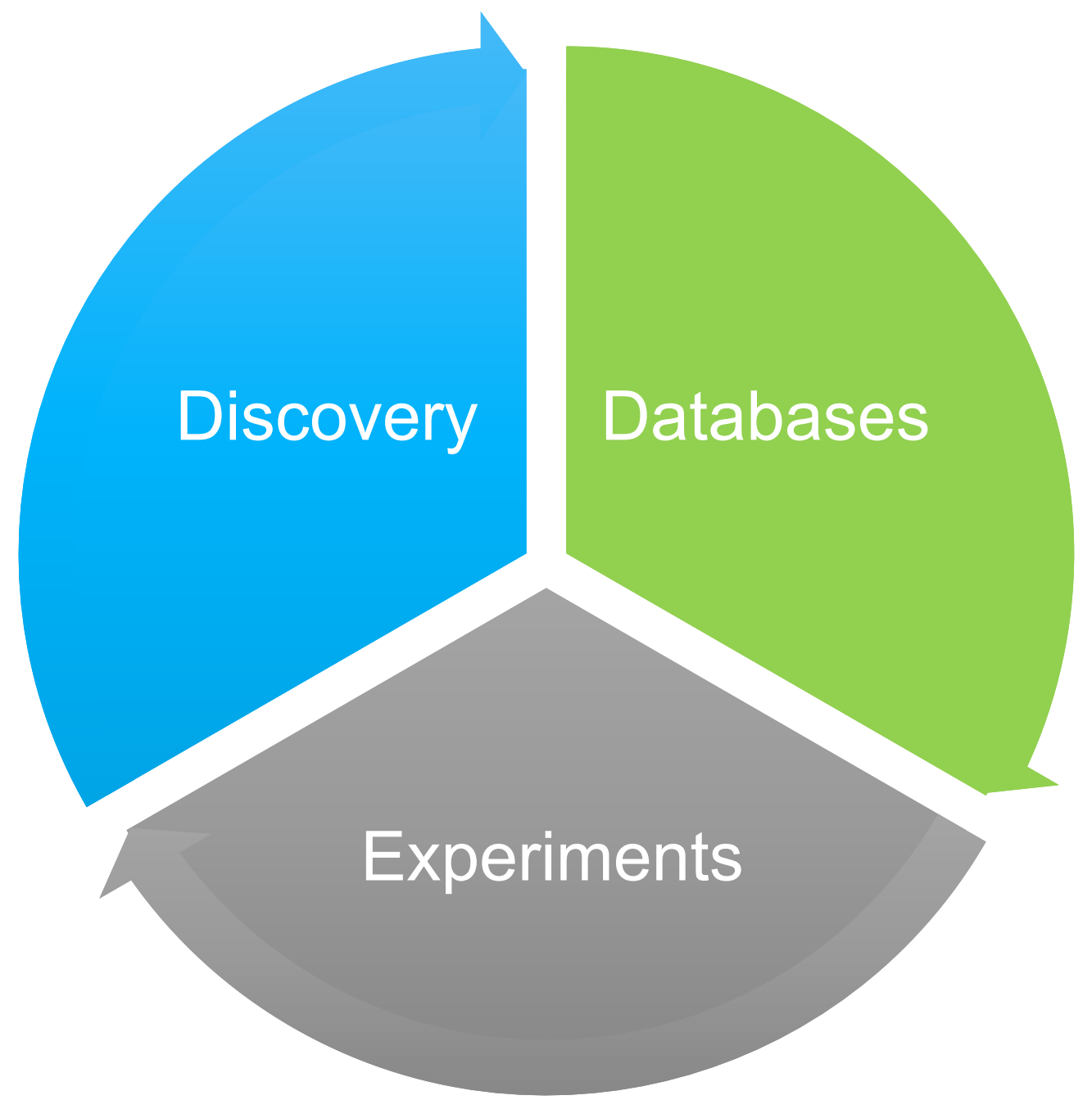
An End-to-End Pipeline for Characterization and Annotation of Traceable Bacterial Material

ATCC®
Credible leads to Incredible™

John Bagnoli, BS; David Yarmosh, MS; Nikhita Puthuveetil, MS; P. Ford Combs, PhD; Amy L. Reese, MS; Briana Benton, BS; Jonathan L. Jacobs, PhD
ATCC , Manassas, VA 20110

Introduction



The tension between genomic data reliability and traceability is a growing area of concern.

There is risk trusting individual and aggregate genome assemblies in genomic databases.

The need for well-characterized high-quality genomics data is crucial for life science research.

To address the above problems, ATCC has developed the ATCC Genome Portal[1] and an ongoing whole-genome sequencing (WGS) initiative to produce genomics data that can be traced back to the source material.

Bacterial and Archaea Pipeline

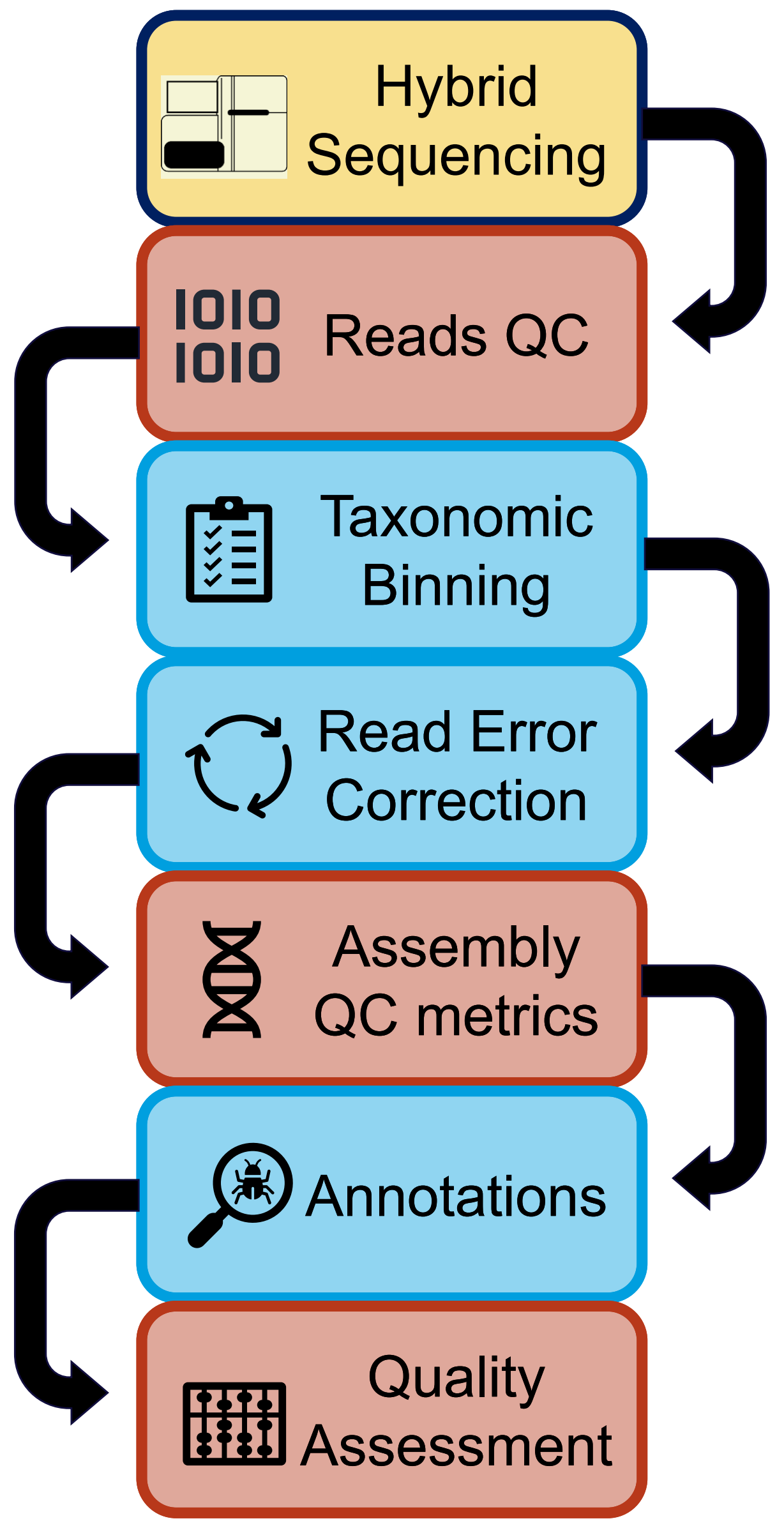
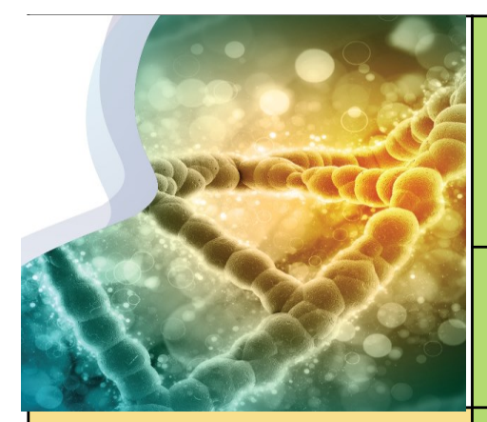


Figure 1. An overview of the pipeline. Hybrid assembly uses Illumina and Oxford Nanopore Technologies (ONT) to generate the FASTQs. Reads are trimmed and filtered using fastp followed by taxonomic classification and binning into kingdoms using kraken2. Long reads are error-corrected using FMLRC before going into assembly. Unicycler is used for assembly and contigs go through polishing via polypolish. Contigs are checked against QC criteria and those that pass are collected as part of the assembly. Annotation is performed by the NCBI Prokaryotic Genome Annotation Pipeline (PGAP). Lastly, a series of checks are done to flag an assembly for any potential problems. These are then manually reviewed: taxonomic IDs are checked against the current designations and are evaluated for completeness, similarity to the reference, PGAP confidence, and contamination score.

Pipeline Metrics Comparison

Table 1. Examples of pipeline differences in assembly metrics. NCBI refers to assemblies designated with the ATCC product catalog number.

	*Bacillus licheniformis* (ATCC® 14580™)		*Pseudomonas fluorescens* (ATCC® 13525™)		*Vibrio natriegens* (ATCC® 14048™)		*Coprococcus eutactus* (ATCC® 27759™)	
	ATCC	NCBI	ATCC	NCBI	ATCC	NCBI	ATCC	NCBI
Length	4,214,933	4,222,597	6,505,843	6,511,547	5,177,329	5,175,153	3,096,507	3,102,987
Contigs	5	1	3	1	2	2	5	23
N50	3,015,942	4,222,597	6,181,175	6,511,547	3,250,180	3,248,023	2,289,537	624,153
N50/Total	0.72	1.0	0.95	1.0	0.63	0.63	0.74	0.20
GC %	46.2%	46.2%	60.0%	60.0%	45.1%	45.1%	43.1%	43.1%
Completeness	98.8%	98.8%	99.9%	99.9%	100%	100%	99.2%	99.3%
Contamination	0.0%	0.0%	0.52%	0.52%	2.8%	2.8%	0.0%	0.0%

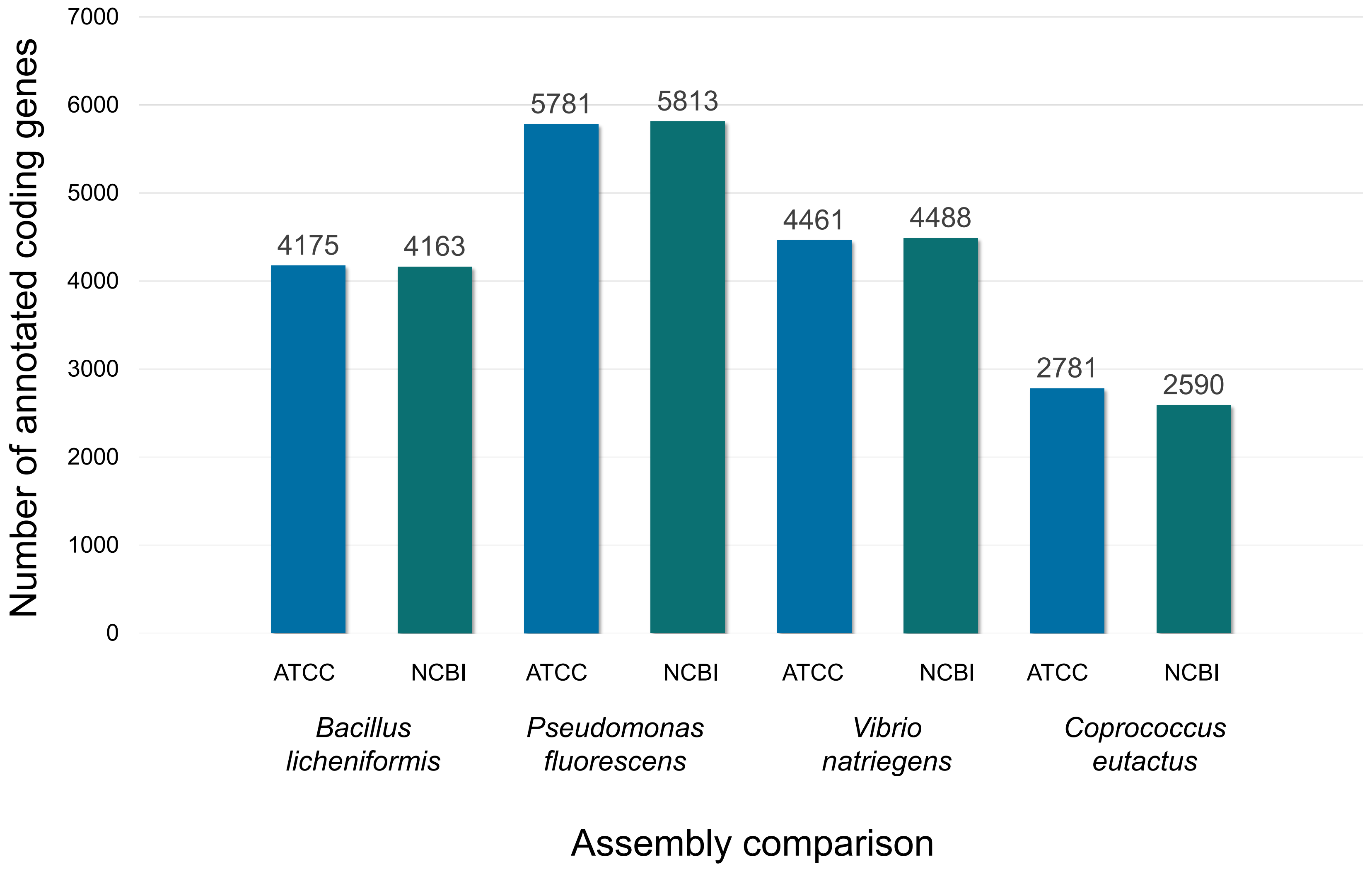


Figure 2. Both ATCC and NCBI assemblies were annotated with PGAP (https://www.ncbi.nlm.nih.gov/genome/annotation_prok/) not including hypothetical proteins. Here, we again see differences in the authenticated material versus public databases.

Hybrid Assembly

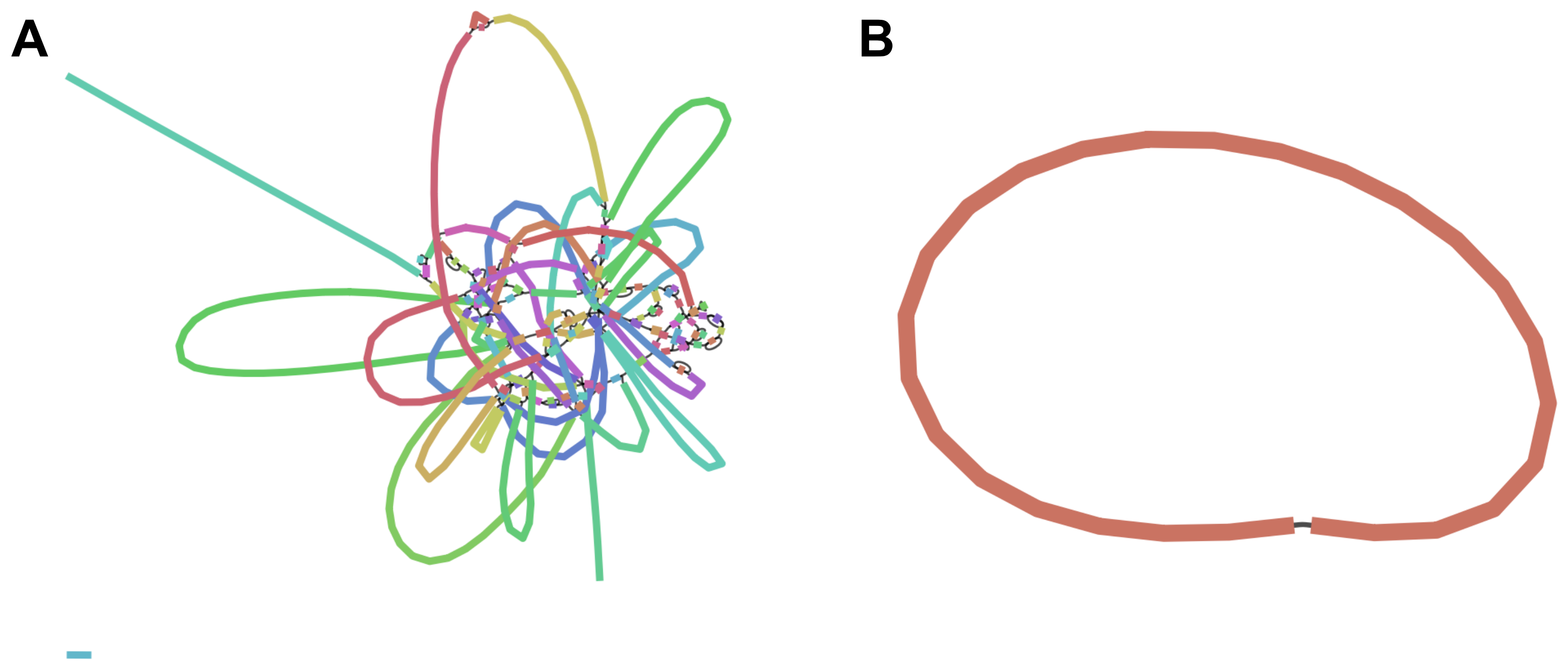


Figure 3. The potential for long-read technologies to generate high-quality assemblies has improved greatly. However, the ability to generate assemblies in the absence of other sequencing methods that are high quality has not yet been achieved. ATCC employs a hybrid assembly technique using accurate but highly fragmented Illumina reads with ONT long reads as a "best of both worlds" approach. Here is a comparison of a Graphical Fragment Assembly (GFA) of *Mycoplasma bovis* (ATCC® 25523™) using (A) Illumina only versus (B) Hybrid with ONT.

Conclusion

The 'omics data of ATCC products generated directly from the source material often differs from the data found in public databases. To ensure we are providing accurate and reliable data, we will continue to develop and improve our assembly methods by leveraging tested bioinformatic approaches.

References

1. Benton B, King S, et al. The ATCC Genome Portal: Microbial Genome Reference Standards with Data Provenance. Microbiol Resour Announc 10 (47), e00818-21, 2021. https://doi.org/10.1128/MRA.00818-21
2. Yarmosh DA, Lopera JG, et al. Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies. mSphere e00077-22, 2022. https://doi.org/10.1128/msphere.00077-22

Visit the ATCC Genome Portal at genomes.atcc.org

ATCC 10801 University Boulevard, Manassas, Virginia 20110-2209 **Phone:** 800.638.6597 **Email:** sales@atcc.org **Web:** www.atcc.org