The Elephant in the Room: Microbial Reference Genome Authentication and Traceability

David Yarmosh, MS; Ford Combs, MS; Nikhita Puthuveetil, MS; Juan Lopera, PhD; Samuel Greenfield, BS; Stephen King, MS; Marco Riojas, PhD; Amanda Pierola, BS; Briana Benton, BS; John Bagnoli, BS; Jonathan Jacobs, PhD

ATCC, Manassas, VA 20110

ATCC®

Credible leads to Incredible™

Background

Public databases, such as NCBI's RefSeq, lack complete traceability from assembly sequences to the physical materials they derived from. Consequently, genome assembly submissions to these databases often lack sufficiently complete metadata surrounding the sequences, leading to the potential for incongruencies between the provided data and the original material. This is by no fault of these databases but results from the nature of publicly sourced genomics data.

Despite this gap, the broader microbial genomics community continues to place a tremendous amount of trust in these databases.. Research is constantly being performed with the assumption that these sequences are not only accurate but represent what the researcher expects. Those with the diligence to authenticate these sequences upon download find an undue burden to eliminate genomes that do not belong in a given collection. ATCC has released over 1,200 genomes, directly traceable back to source material to answer this issue.

Of the 100 strains evaluated, 35 were found to have more than 50 variants, 8 with over 100 variants, and 2 with over 50kb in differences relative to those in RefSeq. Evidence for the need of superior traceability can be found in many strains available on NCBI with multiple assemblies. The difference between any two of these indicates that at least one of them is not precisely describing the strain. Sources of error are legion. Here, we investigate the relationship between metadata and assembly quality.

ATCC and NCBI Assembly Quality

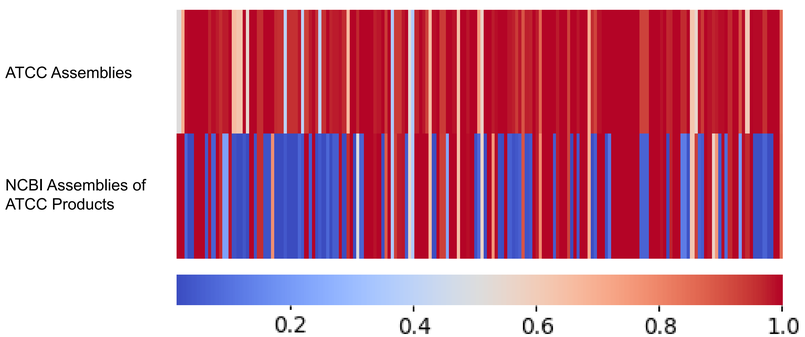

Figure 1: Heatmap representing 486 ratios of contig N50 to Assembly Length for internal ATCC assemblies (top) and publicly available assemblies from GenBank (bottom). All GenBank assemblies are labeled as "complete genome" or "chromosome" level assemblies.

Assembly Variance

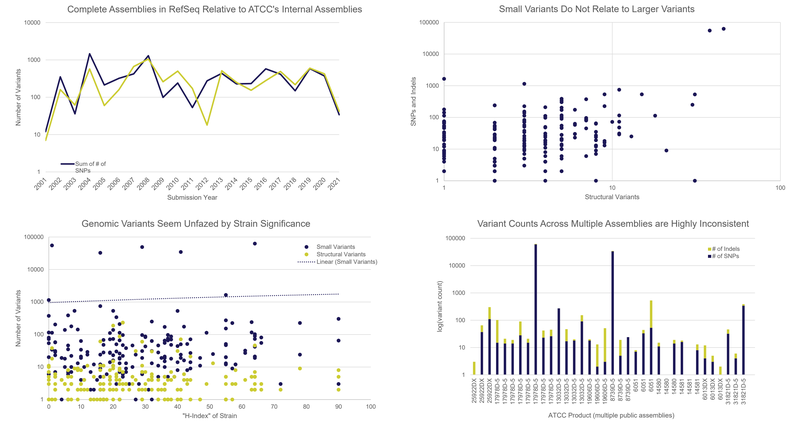

Figure 2: All SNP/InDel/Structural variant comparisons are made from GenBank assemblies relative to their ATCC genome counterparts. (A) trends of SNPs and InDels across all qualifying assemblies by year. (B) Relationship between smaller variants (SNPs/InDels) and structural variants. (C) ATCC uses a value that is inspired by an author's H-index to determine the community significance of a given strain or organism. It does not appear that this value has a clear relationship with either variant type. (D) Of the ATCC assemblies that have 3 or more GenBank "complete" counterparts, there are many differences between the GenBank assemblies from the perspective of variation relative to ATCC assemblies. This indicates a lack of coherence among publicly derived assembly data.

Sequencing and Assembly Technologies

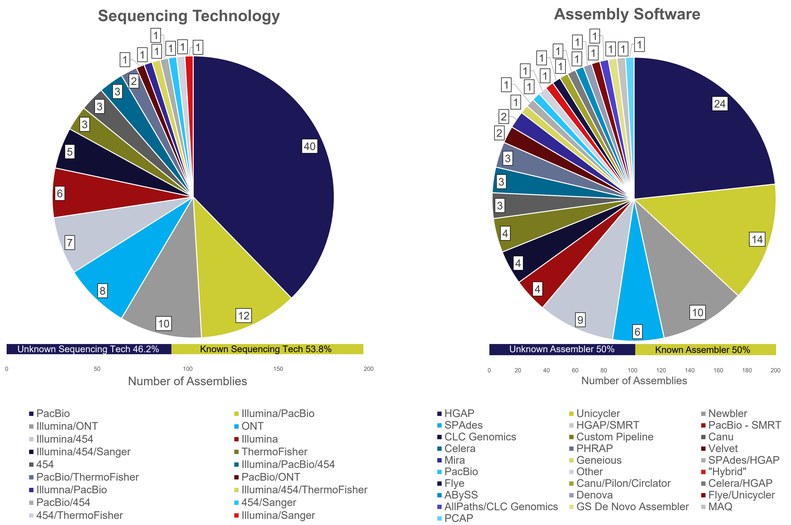

Figure 3: Illustration of listed sequencing technology (left) and assembly software (right). These are compiled into broad categories to simplify from the differences in specific instruments (*i.e.,* MinION and GridION are both "ONT"), version number, access dates, and data entry inconsistencies.

The Materials Behind the Genomes

Table 1: What is the traceability of the materials behind these genome references?

Categories	Count	Example Comment
Not Listed	11	GCF_004794235.2: *Acinetobacter baumannii* ATCC17978 was grown for more than 8 years in the laboratory environment. We called Lab-WT strain. Whole genome sequencing was performed to find mutation genes in comparison with the laboratory-evolved polymyxin B-resistant strains
assembly from type material - no source	13	
assembly from type material - Non-ATCC Source	4	GCF_014672755.1: Bacteria and source DNA available from ATCC19606-BlokeschLab (original ATCC19606 from ATCC)
Non-ATCC Lab	8	GCF_013372085.1: Bacteria and source DNA available from Nick Dillon, ATCC
ATCC	1	
Treated with mutagen	3	GCF_003591595.1: Silver resistant *Escherichia coli* isolates were obtained by stepwise adaptation of *Escherichia coli* ATCC 8739 to increasing silver nitrate concentrations. The complete Genome of a mutant capable of growing in the presence of 9000 mg/L silver nitrate was sequenced and assembled.
Grown outside of ATCC	10	GCF_001593425.2: Using In-house wild type strain for *A. baumannii* 17978, strain AB042 and AB043 were generated.

In GenBank assembly metadata, the material attribution is sometimes listed. In cases where it is listed, the attribution frequently specifies a source that is distinct from ATCC and often indicates that the material had undergone some process to alter its genome prior to sequencing.

Conclusion

ATCC's equivalency analysis has found that while genomic assembly technologies and methods are ever improving, myriad assembly quality issues remain in the public domain, including incoherence between multiple assemblies of a given strain. The impact of any given assembly difference is currently unclear, though the critical absence of sample metadata is a further confounding factor to interassembly discrepancy.

With this in mind, ATCC launched the ATCC Genome Portal, which aims to provide complete traceability from their sample stock through high-quality published genome sequences. ATCC documentation details every processing step, from sample retrieval to assembly. Because ATCC is the repository as well as the sequencer, they are uniquely positioned to ensure genomics data correspond to the correct material.

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