AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESEARCH ARTICLE

Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies

David A. Yarmosh,[a] Juan G. Lopera,[a]* Nikhita P. Puthuveetil,[a] Patrick Ford Combs,[a] Amy L. Reese,[a] Corina Tabron,[a] Amanda E. Pierola,[a] James Duncan,[a] Samuel R. Greenfield,[a] Robert Marlow,[a] Stephen King,[a] Marco A. Riojas,[a,b] John Bagnoli,[a] Briana Benton,[a] Jonathan L. Jacobs[a]

[a]American Type Culture Collection (ATCC), Manassas, Virginia, USA
[b]BEI Resources, Manassas, Virginia, USA

ABSTRACT The availability of public genomics data has become essential for modern life sciences research, yet the quality, traceability, and curation of these data have significant impacts on a broad range of microbial genomics research. While microbial genome databases such as NCBI's RefSeq database leverage the scalability of crowd sourcing for growth, genomics data provenance and authenticity of the source materials used to produce data are not strict requirements. Here, we describe the *de novo* assembly of 1,113 bacterial genome references produced from authenticated materials sourced from the American Type Culture Collection (ATCC), each with full genomics data provenance relating to bioinformatics methods, quality control, and passage history. Comparative genomics analysis of ATCC standard reference genomes (ASRGs) revealed significant issues with regard to NCBI's RefSeq bacterial genome assemblies related to completeness, mutations, structure, strain metadata, and gaps in traceability to the original biological source materials. Nearly half of RefSeq assemblies lack details on sample source information, sequencing technology, or bioinformatics methods. Deep curation of these records is not within the scope of NCBI's core mission in supporting open science, which aims to collect sequence records that are submitted by the public. Nonetheless, we propose that gaps in metadata accuracy and data provenance represent an "elephant in the room" for microbial genomics research. Effectively addressing these issues will require raising the level of accountability for data depositors and acknowledging the need for higher expectations of quality among the researchers whose research depends on accurate and attributable reference genome data.

IMPORTANCE The traceability of microbial genomics data to authenticated physical biological materials is not a requirement for depositing these data into public genome databases. This creates significant risks for the reliability and data provenance of these important genomics research resources, the impact of which is not well understood. We sought to investigate this by carrying out a comparative genomics study of 1,113 ATCC standard reference genomes (ASRGs) produced by ATCC from authenticated and traceable materials using the latest sequencing technologies. We found widespread discrepancies in genome assembly quality, genetic variability, and the quality and completeness of the associated metadata among hundreds of reference genomes for ATCC strains found in NCBI's RefSeq database. We present a comparative analysis of *de novo*-assembled ASRGs, their respective metadata, and variant analysis using RefSeq genomes as a reference. Although assembly quality in RefSeq has generally improved over time, we found that significant quality issues remain, especially as related to genomic data and metadata provenance. Our work highlights the importance of data authentication and provenance for the microbial genomics community, and underscores the risks of ignoring this issue in the future.

Editor Garret Suen, University of Wisconsin—Madison

Address correspondence to Jonathan L. Jacobs, jjacobs@atcc.org.

*Present address: Juan G. Lopera, EDAN Diagnostics, Madison, Wisconsin, USA.

The authors declare no conflict of interest.

Received 4 February 2022
Accepted 6 April 2022
Published 2 May 2022

KEYWORDS DNA sequencing, bioinformatics, comparative studies, culture collection, data provenance, genome analysis, genome authentication, genomes, genomics, microbial genomics

The National Center for Biotechnology Information's (NCBI) RefSeq database has become an essential cornerstone of the global genomics research community, but the quality of metadata and the increasing need for manual data curation by end users are growing areas of concern (1–9). As RefSeq continues to expand, so too does the risk for data errors, omission, obfuscation, or falsification to go undetected, especially in large and aggregate bioinformatics analyses, and to potentially damage trust in this enormously important public resource (10, 11). RefSeq contains over 236,000 bacterial genome assemblies spanning more than 67,000 bacterial strains with assigned taxonomic identities. It is the largest collection of nonredundant, annotated genome assemblies available, and it is built exclusively from crowd-sourced data. However, despite extensive efforts to create automated curation pipelines and tools to improve RefSeq data, significant quality issues remain in genome assemblies found within RefSeq (12–14). For example, while all newly deposited prokaryote genome assemblies are automatically annotated, the associated metadata records (i.e., BioSample, BioProject, SRA, and Assembly data) are submitted by depositors who are not required to provide attribution for the biological materials behind each genome (8, 15). In fact, the International Nucleotide Sequence Database Collaboration (INSDC) policy states that "the quality and accuracy of the record are the responsibility of the submitting author, not of the database," which is to say that metadata, which are often crucial for comparative genomics research, are not curated or verified for accuracy (16). This is further complicated by data omissions, poorly controlled sample description terminology, variable taxonomic naming conventions, and competing metadata package formats during submission that require different fields. Indeed, these points have all underpinned several recent studies investigating inconsistencies among "reference genomes" and type strains for a variety of bacterial species (17–21). In many cases, tracing the provenance of an individual assembly to its source material in order to verify its authenticity becomes challenging, and manual curation is frequently required to detect and correct these metadata errors (22). While it is not within the scope for RefSeq to present only the most up-to-date and accurate sequences available, it does place the burden of data accuracy on end users (as opposed to depositors) in recognizing and properly accounting for genome assembly and metadata accuracy issues.

Here, we present the results of an ongoing whole-genome sequencing (WGS) initiative at ATCC to provide end-to-end data provenance from source materials to reference-grade microbial genomes, here referred to as ATCC standard reference genomes (ASRGs). Although over 2,000 ASRGs are currently available from the ATCC Genome Portal, the 1,113 bacterial ASRGs presented here represent those that were available when this study was concluded. We compared these assemblies to those in RefSeq where metadata indicated they were produced by 3rd-party labs that sourced their materials from ATCC. For 366 ASRGs (~33%), we were able use metadata to compare them to one or more assemblies in RefSeq. The remaining 747 ASRGs (~66%) represent assemblies for bacterial strains that do not have clear counterparts for the same strains in RefSeq. All ASRGs described here are available for noncommercial research use via the ATCC Genome Portal (https://genomes.atcc.org) (23).

RESULTS

Whole-genome sequencing of 1,113 ATCC bacterial strains. High-molecular-weight genomic DNA (HMW-gDNA) was extracted from 1,113 bacterial strains obtained from ATCC's biorepository and subjected to hybrid whole-genome sequencing (WGS), assembly, and deposition to the ATCC Genome Portal. Briefly, strain selection was based on a combination of the number factors, including frequency of requests from ATCC's repository over the last 5 years, inventory availability, biosafety level, and

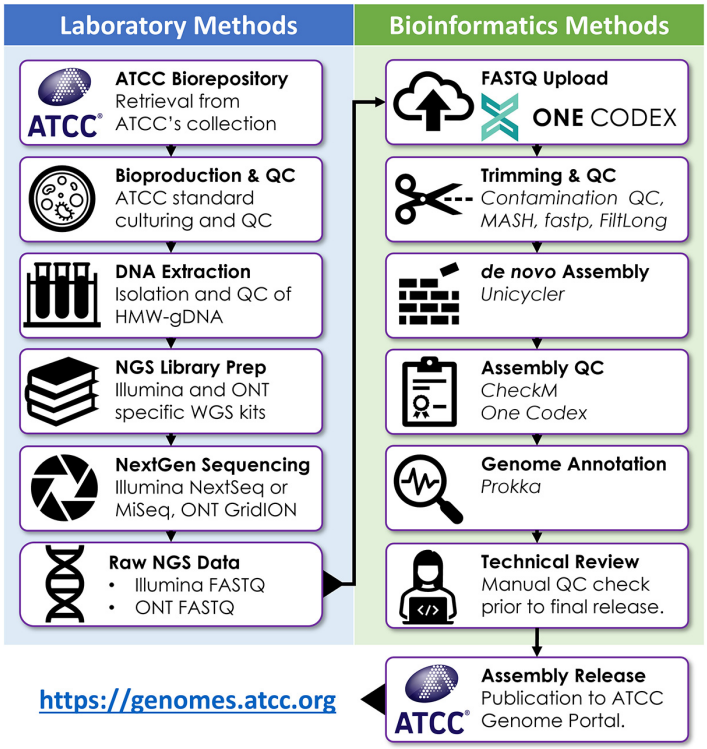

FIG 1 Pipeline for end-to-end genomic data provenance. Source materials were obtained directly from the ATCC biorepository and tracked through to the final assembly and genome annotation. Upfront culture conditions varied depending on the species cultured, but downstream process steps were performed using standardized protocols for DNA extraction, library prep, sequencing, and bioinformatics. Each pipeline is hosted on One Codex's cloud infrastructure.

specific researcher requests. This resulted in 11 different bacterial phyla being included in this study. Briefly, each strain was cultured using strain-specific protocols and subjected to quality control (QC) for contamination, viability, purity, phenotype, and taxonomic identity (Fig. 1). For WGS, HMW-gDNA was split and subjected to sequencing using both Illumina and Oxford Nanopore Technologies (ONT) next-generation sequencing (NGS) platforms (Fig. 1). Next, reads were taxonomically classified using the One Codex platform to assess the purity of each NGS library prior to *de novo* assembly (24). Read sets were then down-sampled to predetermined coverage depths (Illumina, 100×; ONT, 60×) expected to be optimal for bacterial genome assemblies (25–28). Lastly, a hybrid assembly pipeline incorporating reads from both platforms produced *de novo* assemblies for each strain using *Unicycler* (28). High-level summary metrics for each ASRG are shown in Fig. 2 and Table S1 in the supplemental material. All 1,113 ASRG assemblies were estimated to be over 95% complete by *CheckM*; 1,015 were found to be over 99% complete and 329 are 100% complete (29). A total of 617 are considered high-quality, closed genome references.

Survey of bacterial genome assemblies in RefSeq. We compared the ASRG assemblies to those in NCBI's RefSeq bacterial database (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/) labeled as representing ATCC bacterial strains, i.e., assemblies where the ATCC strain name (or a synonymous name) was indicated in the title, description, or organism name fields in the GenBank assembly record. We intentionally did not search RefSeq using a traditional comparative genomics approach (i.e., by sequence homology, BLAST, etc.) since this would require arbitrary thresholds for determining strain identity, and metadata descriptors are intended to be useful for these types of queries. Using this approach, we found 2,701 genome assemblies in RefSeq, which collectively comprised 1,960 different ATCC strains (Fig. 3A and Table S2).

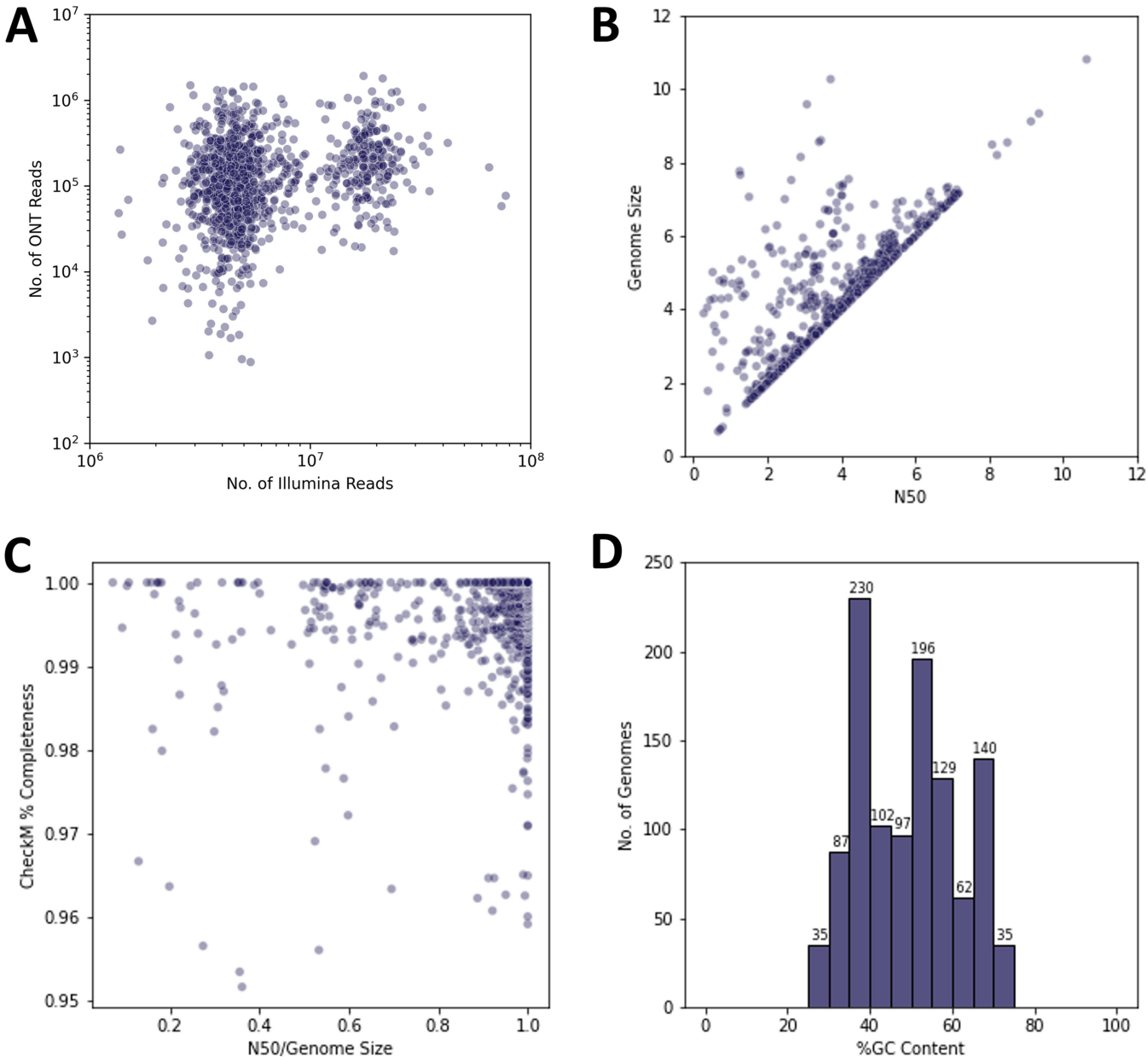

FIG 2 Sequencing and quality metrics for 1,113 bacterial genome assemblies. (A) Illumina versus ONT reads for ASRGs before down-sampling; (B) N_{50} metrics versus genome size; (C) N_{50} normalized by genome size versus *CheckM* genome completion estimates; (D) diversity of GC content for all 1,113 ASRG assemblies.

Interestingly, RefSeq had numerous examples of bacterial strains represented by multiple assemblies or submitted by different groups, and it often included "type strains" resulting from intentional genetic modification (e.g., there are 33 different RefSeq assemblies for *Serratia marcescens* subsp. *marcescens* ATCC 13880). This is despite it representing a "nonredundant" database (although the specific meaning of this is not clearly defined) (9). Moreover, while each of these 33 assemblies have fields in the assembly record describing them as genetically modified genomes, each is also labeled as "assembled from type material." Overall, we found one or more duplicate assemblies in RefSeq for 158 strains for which we also produced an ASRG, including instances of assemblies for genetically modified strains mislabeled as representing type strains (see Table S2). These errors and strain duplications create risks for researchers who may unwittingly use these data in their own research yet remain unaware of these issues.

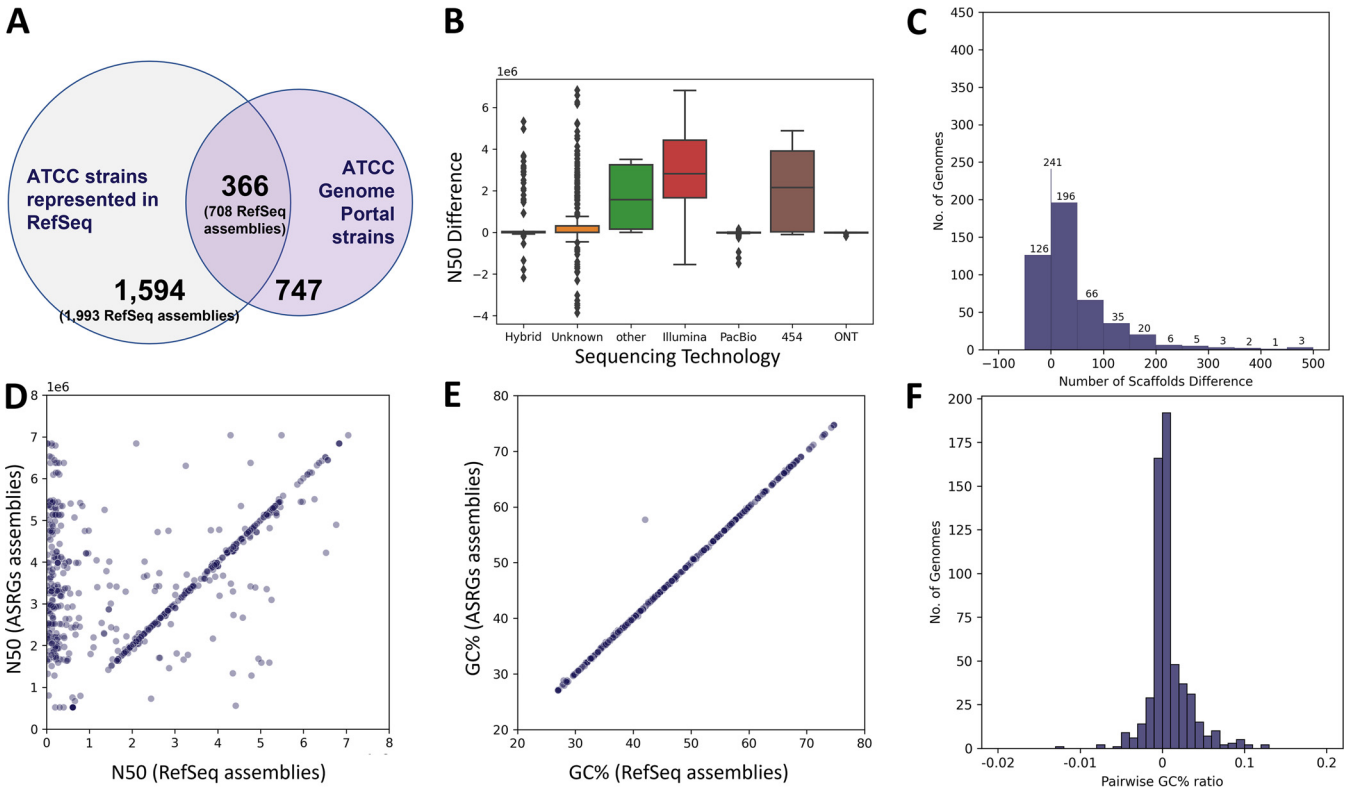

FIG 3 Comparative metrics for 1,113 ASRGs versus RefSeq Assemblies. (A) Intersection of ASRGs versus RefSeq for strains labeled as being from ATCC. In parentheses are the total numbers of RefSeq assemblies, allowing for strain redundancy. (B) N_{50} variability of RefSeq versus ASRGs by sequencing technology. Note that the scale is 1E6. (C) Differences in contig counts for ASRG versus RefSeq assemblies. Positive values indicate that the RefSeq assembly had more contigs. (D) Ratios of ASRG N_{50} values (*y* axis) to RefSeq N_{50} values ("public," *x* axis). Density along the diagonal indicates that many assemblies are similar, while density along the *y* axis indicates ASRGs with higher N_{50} values. (E) GC content for ASRGs (*y* axis) versus RefSeq (*x* axis). Nearly all assemblies have less than 0.1% difference in GC content. (F) Pairwise GC content differences between ASRGs and comparable RefSeq assemblies for the same strain.

Further examination of the metadata for the 2,701 RefSeq assemblies labeled as ATCC strains also revealed numerous records with incomplete, missing, or obscured descriptor fields (Fig. S1). For example, "assembly type" is present in every assembly record but the entry is "na" for all. "sequencing technology" is not included or has an entry of "unknown" for 1,088 assemblies (~40% [Table S2]), and spelling and nonstandard abbreviations further complicate the rest. **"**Assembly method" is not included for 1,082 assemblies, contains the entry "unknown**"** for 88 assemblies or "other" for four assemblies, and has numerous misspellings for various bioinformatics tools (i.e., "Velevt" or "Velveth" for the Velvet assembler). One example (GCF_015708605.1) simply indicates the "assembly method" as "several assembly pipelines, manual curation v. 2018-09-27." Underutilized fields included "description," "isolate," and "relation to type material," which had no entries in 99%, 98%, and 38% of the assembly records, respectively. The damaging impact that inconsistent depositor metadata has on scientific research and reproducibility has been extensively covered elsewhere (1, 3, 30). Within the context of this study, these metadata gaps reveal difficulties in adequately identifying the appropriate RefSeq sequences for further research.

Of the 2,701 RefSeq assemblies for ATCC bacterial strains, 708 had a counterpart ASRG (Fig. 3A and Table S2). Of these, 303 (43%) are labeled "complete genome" or "chromosome" level assemblies. Despite this, N_{50} values were largely inferior to their ASRG counterparts (Fig. 3B). While 241 RefSeq assemblies had the same number of scaffolds as their corresponding ASRGs, 341 were more fragmented. Altogether, 662 ASRGs had N_{50} values equivalent or superior to their RefSeq counterparts (ATCC N_{50}/RefSeq $N_{50} \geq 0.95$), while 46 ASRG assemblies were more fragmented (Fig. 3D). The greatest difference was observed for a RefSeq assembly for *Pseudomonas aeruginosa* ATCC 700888 (GCF_000297315.1), which comprised 600 contigs, while the ASRG

equivalent is a closed and finished genome, containing a single replicon. Because this *Pseudomonas* strain was submitted to RefSeq in 2012 after having been sequenced on an Illumina GAIIx, it is no surprise that our assembly (produced with both Illumina and Nanopore data) produced superior assembly. Nonetheless, arbitrarily excluding this genome from our comparisons also would not befit the study, as one of our primary goals was to investigate the overall quality of assemblies and their corresponding metadata independent of the context, timing, or relative perceived importance of each of them.

Comparative genomics of 303 RefSeq assemblies. Next, we compared the 303 complete RefSeq assemblies to their corresponding ASRGs for the same strains (represented by 212 ASRGs). First, we found that the pairwise average nucleotide identity (ANI) ranged from 97% to 100% for identical strains, which at first glance suggested a high level of similarity (31). Although large differences in the high-level assembly metrics were previously observed (e.g., N_{50} and GC content), after conducting pairwise whole-genome alignments with *MUMmer4* for all 303 RefSeq assemblies against ASRGs for the same strain, we found that 292 had over 95% of their sequence aligned. Next, we examined pairwise structural variations and found significant differences in sequence repeats, inversions, insertions/deletions (indels), and translocations between RefSeq assemblies and ASRGs for the same strains (Tables S3 and S4) (32). Analysis with *dnadiff* of all 303 RefSeq assemblies revealed an average 6.73 structural rearrangements in comparison to ASRGs, the worst of which was GCF_000160895.1 for *Bacillus cereus* ATCC 10876, with 232 structural differences (despite both assemblies having over 99% reciprocally aligned bases). Structural relocations were the most common, with 256 RefSeq assemblies having at least 1 per assembly (average, 4.3 per assembly). Structural inversions were found in 74 RefSeq assemblies (average, 2.2). Translocations were relatively rare, with only 9 RefSeq assemblies having structural translocations relative to the ASRG assembly for the same strain (Table S4). We also found that RefSeq assemblies with the greatest number of structural differences from the ATCC assemblies corresponded to those submitted to NCBI prior to 2010 and for which "sequencing technology" or "assembly method" was not indicated in the RefSeq metadata. The distribution of structural variations in the 303 complete RefSeq assemblies compared to their corresponding ASRGs is shown in Fig. S2.

Variants in 303 RefSeq assemblies. Next, we sought to investigate the prevalence of single-nucleotide polymorphisms (SNPs) and indels that would arise by using RefSeq assemblies as a reference genome against which Illumina sequencing data would be mapped—a common approach used by labs without the resources or expertise for *de novo* assembly and annotation. For each of the 303 complete RefSeq assemblies described above, we mapped the same Illumina reads used in creating the corresponding ASRGs for the same strain. Variant calling from the resulting consensus genomes was carried out on all 303 references to detect SNPs and indels in each (see Materials and Methods). Overall, the number of SNPs and indels per assembly ranged from 0 (none detected) to as many as 60,064 SNPs (*Acinetobacter baumannii* ATCC 17978, GCF_011067065.1) and 2,699 indels for a given assembly (*Parabacteroides distasonis* ATCC 8503, GCF_900683725.1) (Table S5). The median level of SNPs and indels was 7 SNPs and 8 indels per assembly, with 7 of the 303 mappings having no detectable SNPs and indels. These results were promising overall, yet significant outliers were detected, and 26 strains had SNPs and indels beyond an extreme-outlier boundary, i.e., greater than 3 times the interquartile range (IQR) above the median, with 9 of them having over 1,000 SNPs and indels each (Fig. S4a and b and Fig. S5).

A total of 111 assemblies had fewer than 10 variants (SNPs and indels), while 15 assemblies had more than 500 variants (SNPs and indels). Not surprisingly, as the number of SNPs increased, so too did the number of indels (Fig. S3). Of these, 52 of the 303 assemblies had no expected nonsynonymous mutations, but 87 had at least 10 nonsynonymous variants per genome (Fig. S4b). Importantly, 52 RefSeq assemblies identified as "assembled from type material" were found to have at least 10 nonsynonymous

variants, and 7 assemblies had over 100; this could have potentially deleterious impacts on future comparative genomics studies utilizing those reference assemblies (Table S5).

We found that RefSeq assemblies without the label "reference genome" or "representative genome" (250 genomes) were enriched for SNPs (7.6-fold) and indels (9.6-fold) compared to complete reference RefSeq genomes (53 assemblies). Furthermore, type strain assemblies in RefSeq (i.e., labeled as "assembly designated as neotype," "assembly from synonym type material," or "assembly from type material") had generally had fewer SNPs and indels than other assemblies overall, but some significant exceptions to this were also observed (see above). No statistically significant enrichment for SNPs or indels was detectable by taxonomic clade or GC content. Collectively, these results underscore the importance of either manual curation (e.g., "reference genome" or "representative genome") or data provenance of the originating materials (e.g., "assembled from type strain material") and that they are both important drivers in reducing variability and improving published genome assembly quality.

DISCUSSION

Over the last 20 years, several noncommercial and government initiatives have specifically tried to address issues relating to the quality and standardization of metadata for microbial genomics, which has had some benefit for end users, but substantial work remains to be done (15, 33, 34). As the unmet need for curated, high-quality microbial genomics data continues to grow, we will no doubt continue to see commercial initiatives designed to address gaps in quality, content, and reliability, such as Qiagen's CLC Microbial Reference Database, ARES Genetics' ARESdb, and the One Codex platform (24, 35, 36). While these public and private efforts have been relatively successful, others have raised concerns about declines in public microbial genomics metadata (2, 3, 5, 11, 14, 37). We assert that widespread gaps in the traceability of genome assemblies to their originating biological materials, lab protocols, and bioinformatics methods represent a fundamental weakness in these data that hinders research reproducibility and progress. Further, databases such as RefSeq, that do not aim for completely attributable data, are potentially being misused in research studies where provenance and authenticity are assumed. We are attempting to address these gaps for ATCC strains by reestablishing the provenance of these data to physical materials held within ATCC's biorepository and making these data available for research use purposes via the ATCC Genome Portal (23).

At the outset of the work described here, we sought to systematically sequence ATCC's bacterial collection—which has become an ongoing initiative that has recently expanded to also include viruses and fungi. However, during the course of our work, we found that bacterial genome assemblies in RefSeq labeled as representing ATCC strains poorly compared against the ASRGs we produced in-house, both in terms of genome assembly metrics and as they related to strain and assembly metadata. Although sequencing technologies have improved dramatically over the last 2 decades, what surprised us in our initial analysis were the disparities in the quality, accuracy, traceability, and completeness of metadata associated with RefSeq assemblies—which are largely technology independent. It is out of the scope of this study to suggest mechanisms by which NCBI could further control for these concerns, but it does highlight the need for users to consider databases that seek to accurately and consistently capture this information. We found that gaps in data provenance are playing a role in the poor data quality overall. The number of incomplete records increases over time and records are not regularly replaced with more complete versions, which we feel underscores the importance of the genomics initiatives at ATCC. As an example, over 33% (1,087) of the RefSeq assemblies included in our study completely lacked any description for how they were sequenced or assembled in the first place. Furthermore, among the 584 institutions listed among the BioProjects containing assemblies for ATCC bacterial strains (Table S2), only 85 (14.5%) of those institutions definitively obtained these strains directly from ATCC, which no doubt has a negative impact on the traceability

and quality of genome assemblies found within RefSeq. Although it is an estimate, due to institutional name changes or potentially other issues, this nonetheless highlights gaps in our understanding of the source of the strains used to produce these assemblies and the historical provenance of the data associated with them, despite being labeled as a representative or reference genomes for ATCC strains.

There are myriad reasons for why the ASRGs presented in this study generally outperform their counterparts in RefSeq. Obviously, next-generation sequencing technology has improved significantly over the last decade on all points. In addition, *de novo* assembly software is continuously improving, and ATCC developed a standardized workflow for all ASRGs—which is typically not done by research-focused groups. Further, ASRGs are produced directly from biomaterials traceable to physical inventory lots held within ATCC's biorepository in an ISO 9000- and ISO 17025-compliant laboratory, which collectively serves to reduce the opportunity for lab-acquired adaptations, strain domestication, incorrect sample labeling, sample or library contamination, breaks in provenance, and other disruptions in authenticity. It is not within RefSeq's mission to account for these sources of error, nor is it within the mission of most publicly accessible genome databases, but it does represent a gap that is often overlooked by many and that may contribute significantly to issues related to scientific reproducibility. In a broader context, the microbial research community would benefit from establishing a standard for documenting the provenance of physical cultures, isolates obtained from them, and data derived from those isolates (e.g., genomics data). Prior efforts by other groups, such as AOAC International's Stakeholder Panel on Agent Detection Assays (SPADA) Working Group for Bacterial Strain Verification, have attempted to establish provenance standards for physical strains and isolates. In contrast, no formal standard has been proposed for genomics data provenance and authenticity. Specifically, the "source material" attribute of the MIGS genomics metadata standard is defined as being "optional," which perhaps represents a missed opportunity to improve the quality and traceability of genomics research data (15, 38).

In general, researchers should be cautious about the data they obtain from public databases and avoid blindly ingesting reference genome data without first being curious about the origins of the data, the methods used to produce them, and the completeness of the associated metadata. We suggest that for assemblies that are named after and represent culture collection strains or type strains, the highest level of assembly quality, metadata completeness, and data provenance should be expected from depositors; otherwise, many of the issues we have described above will continue to persist. Simply assuming that domestication or laboratory adaptation of strains accounts for the variability observed in RefSeq and GenBank assemblies is insufficient, as these differences can often result in real-world phenotypic changes not reflected in the strains held within culture collections themselves (17–19, 21). We want to encourage data depositors and researchers to continue to use culture collection identifiers whenever submitting new assemblies for strains held within these collections, but we also urge the administrators of public genome databases to place a higher level of accountability on the completeness and quality of these submissions. Doing so would serve to improve the overall reliability of these public resources and reduce the amount of postsubmission curation done by end users wishing to use these data directly in their research. Lastly, we suggest that further systematic studies are needed to better understand the risks and prevalence of inauthentic and inaccurate data found in microbial genomics databases, and the impact they have on basic and applied research. It is our hope that initiatives focused on genomic data provenance, such as the work presented above, will serve to highlight the value of establishing higher standards for traceability and accountability in public microbial genomics databases.

MATERIALS AND METHODS

Sample acquisition and culture conditions. All the bacterial cell cultures and genomic DNA used in this study met or exceeded ATCC's quality standards (https://www.atcc.org/about-us/quality-commitment), underwent extensive phenotypic and genotypic characterization to ensure accurate strain identification, and

were extensively tested for contamination before being accepted for use in this study. ATCC is certified by the ANSI National Accreditation Board (ANAB) to meet both ISO 17034:2016 standards as a reference material producer and ISO/IEC 17025:2017 as a testing and calibration reference laboratory. Each bacterial strain included in this study is available from ATCC's biorepository and was authenticated according to protocols executed in accordance with ATCC's quality management system (see above). The specific protocols for each strain varied depending on the specific species in question. In general, strain identification and authentication included assessment of colony morphology, Gram staining, culture purity, metabolic profiling, antibiotic susceptibility testing (AST), broad-spectrum biochemical reactivity testing, 16S rRNA gene sequencing, ribotyping, matrix-assisted laser desorption ionization–time of flight mass spectrometry (e.g., bioMérieux Vitek MS system), and whole-genome next-generation sequencing (NGS). Additional details used for culturing, growth conditions, and authentication of each bacterial strain are available online in each bacterial strain's catalog page at https://www.atcc.org/ and by visiting ATCC's Bacterial Cell Culture portal (39).

DNA templates and quality control. To facilitate the successful NGS library preparation for multiple sequencing platforms (long- and short-read sequences), both high-quality and high-quantity input DNA was obtained from authenticated genomic DNA (gDNA) available in ATCC bacterial nucleic acids repository (40). ATCC uses several commercially available extraction kits and in-house-validated protocols to obtain pure high-molecular-weight DNA depending on the biological characteristics of the organism undergoing extraction. For strains with no preexisting genomic DNA in ATCC's repository, total high molecular weight genomic DNA (HMW-gDNA) was extracted from thawed or resuspended frozen cultures with 10^7 to 10^9 cells/mL using the Qiagen Genomic-Tip 20/g or 100/g kit and analyzed for purity, concentration, and fragment size. HMW-gDNA samples meeting or exceeding the following criteria were subjected to sequencing: median fragment size larger than 20 kb, optical density A_{260}/A_{280} between 1.75 and 2.00, and a final elution concentration over 20 ng/μL per extraction.

Short-read next-generation sequencing. High-quality gDNA from each strain was subjected to whole-genome sequencing using a short-read NGS workflow. Briefly, sequencing libraries from each extraction were prepared using the DNA prep kit, indexed using DNA/RNA UD indexes (Illumina), and subsequently subjected to paired-end sequencing on either an Illumina MiSeq or NextSeq 2000 instrument. Sample multiplexing was based on achieving a minimum 100× average depth of coverage for each genome. Base-calling and adapter trimming were initially done using onboard Illumina instrument software and followed by an additional round of trimming and quality score filtering using *fastp* with default settings and *FastQC* (26, 41). Illumina reads accepted for further use passed the following quality control thresholds: median Q score, >30 for all bases; median Q score, >25 per base; and ambiguous content (N bases), <5%. The versions of bioinformatics software used throughput this study are listed in Table S6.

Long-read next-generation sequencing. Long-read sequencing was carried out using the Oxford Nanopore Technologies (ONT) GridION platform. ONT ligation sequencing kit (SQK-LSK109) sequencing libraries were prepared from the same physical samples of HMW-gDNA as used for Illumina sequencing described above, multiplexed using the ONT native barcoding expansion kit (EXP-NBD104 or EXP-NBD114), and sequenced using GridION flow cells (ONT; R9.4.1). As with Illumina sequencing, the number of samples multiplexing was based on the estimated genome size of a given organism and sequencing was performed for a minimum of 48 h per flow cell. Using the most up-to-date version of *MinKNOW*, reads were base-called, using the high-accuracy settings, demultiplexed, and barcode trimmed. Furthermore, ONT sequencing reads were quality trimmed and filtered using *Filtlong* to meet the following minimum acceptance criteria: minimum mean Q score per read of >10 and minimum read length of >5,000 bp (42). *Filtlong* was run with the following settings: min_length 1,000, target_bases = genome size × Oxford Nanopore sequencing targeted depth.

Assembly of ATCC standard reference genomes. For genome references deposited to the ATCC Genome Portal, genome assembly size was first estimated from raw reads using *MASH*, and this estimate was used to down-sample the Illumina and ONT raw sequencing libraries to maximum 100× and 40× coverages, respectively (43). These coverage requirements were selected to maximize accuracy for individual consensus base calls in the final assemblies (25, 27). After down-sampling each sequencing library, a hybrid *de novo* assembly approach was taken using *Unicycler* with default settings (28). Briefly, Illumina libraries were first assembled individually into contigs. The longest contigs in the initial set were then scaffolded with reads from the ONT library. The combined hybrid assembly was then iteratively polished using both long and short reads from both input libraries, resulting in highly contiguous or closed reference genomes. Sequencing and assembly artifacts of less than 1,000 bp that also had significantly different coverage depth (e.g., "chaff" contigs) were removed from the final draft reference (44). These draft assemblies were subsequently checked using One Codex to confirm the species (24). Finally, each draft assembly was assessed for completeness and potential contamination with *CheckM lineage_wf*, which is based on orthologous gene copy numbers present in an assembly (29). Assemblies which were determined to have a *CheckM* "completeness" score above 95% and a contamination value below 5% were deemed final assemblies. Each final assembly was subsequently annotated using *Prokka* with default settings for coding DNA sequence (CDS), rRNA, tRNA, signal leader peptide, and noncoding RNA identification (45). Parameters for the various tools are shown in Table S6. Finally, each complete and annotated genome was deposited into the ATCC Genome Portal and is referred to here as an ATCC standard reference genome (ASRG) (23).

Characterization of public genome assemblies. To gather the public assemblies of ATCC bacterial strains, the "assembly_summary_refseq.txt" file was downloaded from NCBI via FTP (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/), accessed July 2019. This file contains accession numbers and metadata, such as "isolate," "assembly level," and "tax ID," for every assembly in NCBI bacterial RefSeq. First,

this file was filtered to keep all records that contained either the "ATCC" or "NCTC" keyword. This was done because many strains have synonymous ATCC and NCTC identifiers (IDs), though often only one of the two is present in a record. Of the records containing "ATCC" or "NCTC," all that included "ATCC" were kept, but records containing "NCTC" were filtered to keep only those with a synonymous ATCC ID. This final set of records contained the 2,701 public assemblies of ATCC strains. While "assembly_summary_refseq.txt" does contain metadata, the complete set of metadata was collected by downloading the "assembly_report.txt" for each assembly from the NCBI ftp site. Metadata comparisons were performed using the *compare.all.levels.py* script after appending the RefSeq assembly data with a GC content column, calculated by *bbnorm_stats.sh*, all of which was paralleled with GNU Parallel (46). ATCC's Genome Portal does not distinguish between contigs and scaffolds, which RefSeq defines as contigs that are connected across gaps. For this, all data comparison of ASRGs in terms of contiguity uses RefSeq scaffold information.

Comparisons of NCBI and ATCC genome assembly metrics. For each of the bacterial strains included in the ATCC Genome Portal, we identified and downloaded all 2,701 genome assemblies that had the same name or similar names from NCBI's RefSeq and Genome Assembly databases. For the 303 NCBI assemblies with a finished assembly status of "complete" or "chromosome" and representation in ATCC's Genome Portal, we carried out pairwise whole-genome alignments for each NCBI and ASRG using *MUMmer4* and its associated suite of tools for comparative genomics (32). In some cases, due to duplications in RefSeq and NCBI's Genome Assembly database, multiple NCBI assemblies were compared against the same ASRG assembly. Following the creation of the alignments, we identified genome-wide variants for each NCBI assembly compared to the ASRG assembly, including single-nucleotide polymorphisms (SNPs), insertions and deletions (indels), and structural variants (SVs). Genome-wide comparisons using *dnadiff* included assembly length, number of contigs, pairwise percent aligned, and N_{50} values (SVs_and_ANI.sh) (47). Furthermore, *MUMmer4*'s *dnadiff* tool was run with default settings using the ASRG assemblies against each NCBI RefSeq assembly, and relocations, translocations, and inversions are reported alongside total and aligned bases (32). Prior to running *MUMmer4*'s *dnadiff* tool on these assemblies, each was filtered to remove contigs of <1kb in length to prevent short sequences from exaggerating SVs between assemblies. Structural variants included breakpoints, relocations, translocations, and inversions, and summarized as rearrangements.

Read-Mapping and Variant Calling with NCBI Assemblies. For reference based variant calling, 303 "complete genome" assemblies were downloaded from NCBI's RefSeq that had corresponding ASRG assemblies and used as references for read-mapping and variant calling. Only the corresponding Illumina sequencing reads from each ASRG assembly (see above) was used as input. For each RefSeq genome, we mapped reads from a corresponding ASRG assembly using BWA-mem v0.7.17 with the default parameters (48). Quality metrics were recorded using Qualimap bamqc v2.2.1 (49). Variants were called using GATK Haplotype caller (v.4.1.8.1, standard minimum confidence threshold = 30) (50). The Ensembl Variant Effect Predictor (VEP) was used to identify synonymous and non-synonymous variants in each reference mapping (51). The complete pipeline is available on our GitHub repository.

Data availability. Source code for ATCC scripts is available at https://github.com/ATCC-Bioinformatics/Equivalency_Analysis. Raw NGS data (FASTQ files) are available for research use only from https://github.com/ATCC-Bioinformatics/AGP-Raw-Data/blob/main/AGP_Raw-Data-Access.txt. ASRGs and associated metadata are available directly from the ATCC Genome Portal (https://genomes.atcc.org) or via our REST-API (access details available upon request).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, TIF file, 1.4 MB.

FIG S2, TIF file, 0.9 MB.

FIG S3, TIF file, 1.6 MB.

FIG S4, TIF file, 1.8 MB.

TABLE S1, XLSX file, 0.1 MB.

TABLE S2, XLSX file, 0.5 MB.

TABLE S3, XLSX file, 0.1 MB.

TABLE S4, XLSX file, 0.05 MB.

TABLE S5, XLSX file, 0.04 MB.

TABLE S6, XLSX file, 0.02 MB.

ACKNOWLEDGMENTS

We thank One Codex for contributing to the development of the ATCC Genome Portal. We also thank Raymond Cypess, Mindy Goldsborough, and Rebecca Bradford for critical comments and review prior to submission.

Conceptualization, J.G.L., B.B., J.B., and J.L.J.; Data Curation, D.A.Y., J.G.L., N.P.P., P.F.C., A.L.R., and M.A.R.; Formal Analysis, D.A.Y., N.P.P., P.F.C., and A.L.R.; Investigation, C.T., A.E.P., J.D., S.R.G., S.K., R.M., and B.B.; Project Administration, B.B. and J.B.; Software, D.A.Y., N.P.P., P.F.C., A.L.R., and J.B.; Supervision, B.B., J.B., and J.L.J.; Visualization, P.F.C.,

B.B., and J.L.J.; Writing – Original Draft, D.A.Y., J.G.L., and J.L.J.; Writing – Review & Editing, D.A.Y., N.P.P., P.F.C., B.B., J.B., M.A.R., and J.L.J.

The work described here was financially supported entirely by the American Type Culture Collection.

All authors are employees of the American Type Culture Collection, which solely funded the work presented here and provided all the bacterial strain materials needed for the research. No other competing interests are claimed.

REFERENCES

1. Gonçalves RS, Musen MA. 2019. The variable quality of metadata about biological samples used in biomedical experiments. Sci Data 6:190021. https://doi.org/10.1038/sdata.2019.21.
2. Pettengill JB, Beal J, Balkey M, Allard M, Rand H, Timme R. 2021. Interpretative labor and the bane of non-standardized metadata in public health surveillance and food safety. Clin Infect Dis ciab615.
3. Rajesh A, Chang Y, Abedalthagafi MS, Wong-Beringer A, Love MI, Mangul S. 2021. Improving the completeness of public metadata accompanying omics studies. Genome Biol 22:106. https://doi.org/10.1186/s13059-021-02332-z.
4. Vangay P, Burgin J, Johnston A, Beck KL, Berrios DC, Blumberg K, Canon S, Chain P, Chandonia J-M, Christianson D, Costes SV, Damerow J, Duncan WD, Dundore-Arias JP, Fagnan K, Galazka JM, Gibbons SM, Hays D, Hervey J, Hu B, Hurwitz BL, Jaiswal P, Joachimiak MP, Kinkel L, Ladau J, Martin SL, McCue LA, Miller K, Mouncey N, Mungall C, Pafilis E, Reddy TBK, Richardson L, Roux S, Schriml LM, Shaffer JP, Sundaramurthi JC, Thompson LR, Timme RE, Zheng J, Wood-Charlson EM, Eloe-Fadrosh EA. 2021. Microbiome metadata standards: report of the National Microbiome Data Collaborative's workshop and follow-on activities. mSystems 6:e01194-20. https://doi.org/10.1128/mSystems.01194-20.
5. Toczydlowski RH, Liggins L, Gaither MR, Anderson TJ, Barton RL, Berg JT, Beskid SG, Davis B, Delgado A, Farrell E, Ghoojaei M, Himmelsbach N, Holmes AE, Queeno SR, Trinh T, Weyand CA, Bradburd GS, Riginos C, Toonen RJ, Crandall ED. 2021. Poor data stewardship will hinder global genetic diversity surveillance. Proc Natl Acad Sci U S A 118:e2107934118. https://doi.org/10.1073/pnas.2107934118.
6. Leipzig J, Nüst D, Hoyt CT, Soiland-Reyes S, Ram K, Greenberg J. 2021. The role of metadata in reproducible computational research. arXiv 2006.08589 [cs.DL]. https://arxiv.org/abs/2006.08589.
7. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire MK, Durkin AS, Gonzales NR, Gwadz M, Lanczycki CJ, Song JS, Thanki N, Wang J, Yamashita RA, Yang M, Zheng C, Marchler-Bauer A, Thibaud-Nissen F. 2021. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res 49:D1020–D1028. https://doi.org/10.1093/nar/gkaa1105.
8. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. 2014. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res 42:D553–D559. https://doi.org/10.1093/nar/gkt1274.
9. Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:D501–504. https://doi.org/10.1093/nar/gki025.
10. Gopalakrishna G, Wicherts JM, Vink G, Stoop I, Van den Akker O, Riet G, Bouter L. 2021. Prevalence of responsible research practices and their potential explanatory factors: a survey among academic researchers in The Netherlands. MetaArXiv https://doi.org/10.31222/osf.io/vk9yt.
11. Caswell J, Gans JD, Generous N, Hudson CM, Merkley E, Johnson C, Oehmen C, Omberg K, Purvine E, Taylor K, Ting CL, Wolinsky M, Xie G. 2019. Defending our public biological databases as a global critical infrastructure. Front Bioeng Biotechnol 7:58. https://doi.org/10.3389/fbioe.2019.00058.
12. Steinegger M, Salzberg SL. 2020. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biol 21:115. https://doi.org/10.1186/s13059-020-02023-1.
13. Segerman B. 2020. The most frequently used sequencing technologies and assembly methods in different time segments of the bacterial surveillance and RefSeq genome databases. Front Cell Infect Microbiol 10:527102. https://doi.org/10.3389/fcimb.2020.527102.
14. Smits THM. 2019. The importance of genome sequence quality to microbial comparative genomics. BMC Genomics 20:662. https://doi.org/10.1186/s12864-019-6014-5.
15. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, dePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, et al. 2008. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 26:541–547. https://doi.org/10.1038/nbt1360.
16. Karsch-Mizrachi I, Takagi T, Cochrane G, International Nucleotide Sequence Database Collaboration. 2018. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 46:D48–D51. https://doi.org/10.1093/nar/gkx1097.
17. Dorman MJ, Thomson NR. 2020. "Community evolution"—laboratory strains and pedigrees in the age of genomics. Microbiology (Reading) 166:233–238. https://doi.org/10.1099/mic.0.000869.
18. Artuso I, Lucidi M, Visaggio D, Capecchi G, Lugli GA, Ventura M, Visca P. 2022. Genome diversity of domesticated *Acinetobacter baumannii* ATCC 19606^T strains. Microb Genom 8:000749. https://doi.org/10.1099/mgen.0.000749.
19. Pascoe B, Williams LK, Calland JK, Meric G, Hitchings MD, Dyer M, Ryder J, Shaw S, Lopes BS, Chintoan-Uta C, Allan E, Vidal A, Fearnley C, Everest P, Pachebat JA, Cogan TA, Stevens MP, Humphrey TJ, Wilkinson TS, Cody AJ, Colles FM, Jolley KA, Maiden MCJ, Strachan N, Pearson BM, Linton D, Wren BW, Parkhill J, Kelly DJ, van Vliet AHM, Forbes KJ, Sheppard SK. 2019. Domestication of Campylobacter jejuni NCTC 11168. Microb Genom 5:e000279. https://doi.org/10.1099/mgen.0.000279.
20. Draper JL, Hansen LM, Bernick DL, Abedrabbo S, Underwood JG, Kong N, Huang BC, Weis AM, Weimer BC, van Vliet AHM, Pourmand N, Solnick JV, Karplus K, Ottemann KM. 2017. Fallacy of the unique genome: sequence diversity within single Helicobacter pylori strains. mBio 8:e02321-16. https://doi.org/10.1128/mBio.02321-16.
21. Klockgether J, Munder A, Neugebauer J, Davenport CF, Stanke F, Larbig KD, Heeb S, Schöck U, Pohl TM, Wiehlmann L, Tümmler B. 2010. Genome diversity of Pseudomonas aeruginosa PAO1 laboratory strains. J Bacteriol 192:1113–1121. https://doi.org/10.1128/JB.01515-09.
22. Schmedes SE, King JL, Budowle B. 2015. Correcting inconsistencies and errors in bacterial genome metadata using an automated curation tool in Excel (AutoCurE). Front Bioeng Biotechnol 3:138. https://doi.org/10.3389/fbioe.2015.00138.
23. Benton B, King S, Greenfield SR, Puthuveetil N, Reese AL, Duncan J, Marlow R, Tabron C, Pierola AE, Yarmosh DA, Combs PF, Riojas MA, Bagnoli J, Jacobs JL. 2021. The ATCC Genome Portal: microbial genome reference standards with data provenance. Microbiol Resour Announc 10:e00818-21. https://doi.org/10.1128/MRA.00818-21.
24. Minot SS, Krumm N, Greenfield NB. 2015. One Codex: a sensitive and accurate data platform for genomic microbial identification. bioRxiv 027607. https://doi.org/10.1101/027607.
25. Desai A, Marwah VS, Yadav A, Jha V, Dhaygude K, Bangar U, Kulkarni V, Jere A. 2013. Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. PLoS One 8:e60204. https://doi.org/10.1371/journal.pone.0060204.
26. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

27. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15:121–132. https://doi.org/10.1038/nrg3642.
28. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.
29. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.
30. Marc DT, Beattie J, Herasevich V, Gatewood L, Zhang R. 2016. Assessing metadata quality of a federally sponsored health data repository. AMIA Annu Symp Proc 2016:864–873.
31. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9.
32. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol 14:e1005944. https://doi.org/10.1371/journal.pcbi.1005944.
33. Yilmaz P, Gilbert JA, Knight R, Amaral-Zettler L, Karsch-Mizrachi I, Cochrane G, Nakamura Y, Sansone S-A, Glöckner FO, Field D. 2011. The Genomic Standards Consortium: bringing standards to life for microbial ecology. ISME J 5:1565–1567. https://doi.org/10.1038/ismej.2011.39.
34. Sichtig H, Minogue T, Yan Y, Stefan C, Hall A, Tallon L, Sadzewicz L, Nadendla S, Klimke W, Hatcher E, Shumway M, Aldea DL, Allen J, Koehler J, Slezak T, Lovell S, Schoepp R, Scherf U. 2019. FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. Nat Commun 10:3313. https://doi.org/10.1038/s41467-019-11306-6.
35. Couto N, Schuele L, Raangs EC, Machado MP, Mendes CI, Jesus TF, Chlebowicz M, Rosema S, Ramirez M, Carriço JA, Autenrieth IB, Friedrich AW, Peter S, Rossen JW. 2018. Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens. Sci Rep 8:13767. https://doi.org/10.1038/s41598-018-31873-w.
36. Ferreira I, Beisken S, Lueftinger L, Weinmaier T, Klein M, Bacher J, Patel R, von Haeseler A, Posch AE. 2020. Species identification and antibiotic resistance prediction by analysis of whole-genome sequence data by use of ARESdb: an analysis of isolates from the Unyvero lower respiratory tract infection trial. J Clin Microbiol 58:e00273-20. https://doi.org/10.1128/JCM.00273-20.
37. Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. 2019. Human contamination in bacterial genomes has created thousands of spurious proteins. Genome Res 29:954–960. https://doi.org/10.1101/gr.245373.118.
38. AOAC International. 2020. Appendix R: guidelines for verifying and documenting the relationships between microbial cultures. J AOAC Int 103:900–903. https://doi.org/10.1093/jaoacint/qsaa046.
39. ATCC. 2021. ATCC bacteriology culture guide. https://www.atcc.org/resources/culture-guides/bacteriology-culture-guide. Accessed 22 July 2021.
40. ATCC. 2021. Nucleic acids from bacteria and archaea. https://www.atcc.org/microbe-products/bacteriology-and-archaea/nucleic-acids. Accessed 22 July 2021.
41. Wingett SW, Andrews S. 2018. FastQ Screen: a tool for multi-genome mapping and quality control. F1000Res 7:1338. https://doi.org/10.12688/f1000research.15931.2.
42. Wick R, Menzel P. 2019. Filtlong—quality filtering tool for long reads. https://github.com/rrwick/Filtlong.
43. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132. https://doi.org/10.1186/s13059-016-0997-x.
44. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marcais G, Pop M, Yorke JA. 2012. GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Res 22:557–567. https://doi.org/10.1101/gr.131383.111.
45. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
46. Tange O. 2011. GNU parallel—the command-line power tool. login 36:42–47.
47. Khelik K, Lagesen K, Sandve GK, Rognes T, Nederbragt AJ. 2017. NucDiff: in-depth characterization and annotation of differences between two sets of DNA sequences. BMC Bioinformatics 18:338. https://doi.org/10.1186/s12859-017-1748-z.
48. Li H, Durbin R. 2009. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. Bioinformatics, 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
49. Okonechnikov K, Conesa A, García-Alcalde F. 2015. Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data. Bioinformatics, btv566. https://doi.org/10.1093/bioinformatics/btv566.
50. Auwera G, van der O'Connor BD. 2020. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra, First edition. O'Reilly Media: Sebastopol, CA.
51. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie G. R. S., Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. Genome Biol, 17(1):122. https://doi.org/10.1186/s13059-016-0974-4.