September 12, 2022

Translational Ramifications of Crowd-Sourced Genomics Data

ATCC®

Credible leads to Incredible®

David A. Yarmosh, MS; Nikhita P. Puthuveetil, MS; P. Ford Combs, PhD; Amy L. Reese, MS; Corina Tabron, MS; James Duncan, BS; Samuel R. Greenfield, BS; Robert Marlow, BS; Stephen King, MS; Marco A. Riojas, PhD; Ana Fernandes, BS; John Bagnoli, BS; Briana Benton, BS; Jonathan L. Jacobs, PhD

ATCC, Manassas, VA 20110

Background

Public genomics databases serve a critical role in the life science research community. Despite existing guidelines that require the inclusion of metadata associated with a given genome assembly, other relevant data (e.g., sequencing platform, assembly method) are often incomplete or missing. Ultimately, this gap renders the assembly data itself questionable from the perspective of reliability, traceability, and accuracy. Previously, Yarmosh et al.[1] illustrated the impact of poor data provenance by comparing several publicly available assemblies to assemblies that have complete traceability—ATCC Standard Reference Genomes (ASRGs). It was found that some public assemblies, which were labeled as derivatives of ATCC source material, had a tendency toward fewer relative variants between these assemblies and their ASRG counterparts. However, several of these assemblies still contained a large quantity of variants, including those inducing translational changes. Here, we investigate these translational changes in terms of best matching gene identity, annotated gene name, and gene multimapping.

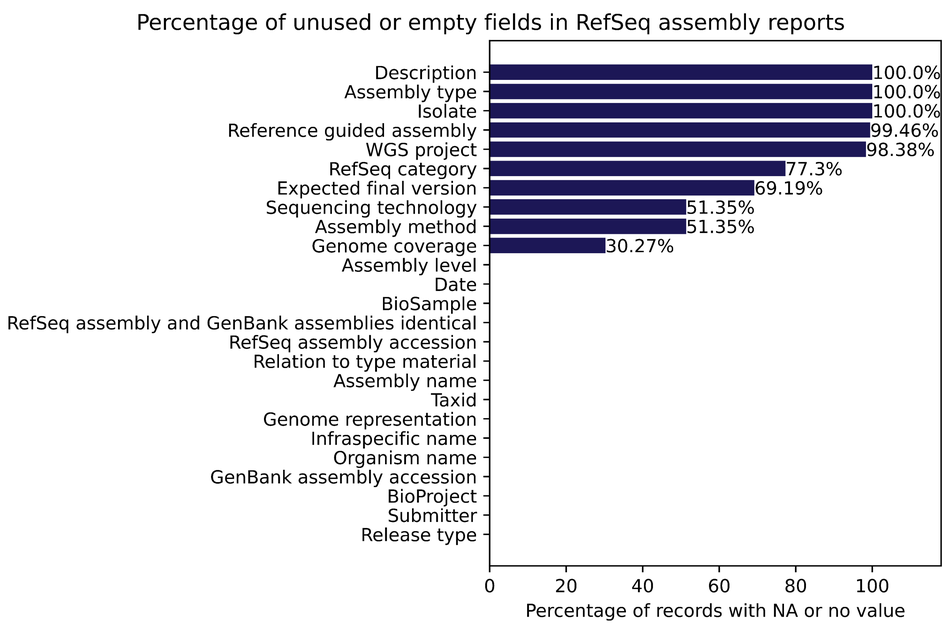

Figure 1: Metadata usage breakdown of the dataset included in this study. Despite guidelines suggesting the usage of several metadata fields, these fields are inconsistently filled out by submitters. Several of these fields represent highly relevant information that is necessary to appropriately determine the inclusion of pertinent data in analyses.

Methods

To better understand the consequences outlined in a previous study,[1] the Prokaryotic Genome Annotation Pipeline (PGAP)[2] was run on 185 public assemblies labeled as ATCC type material and their 125 corresponding ASRGs. Annotations from both sets were compared in terms of amino acid identity, gene count, gene identity, and the gain/loss of stop codons using reciprocal BLAST searches[3] of genes. Analytical code is available upon request.

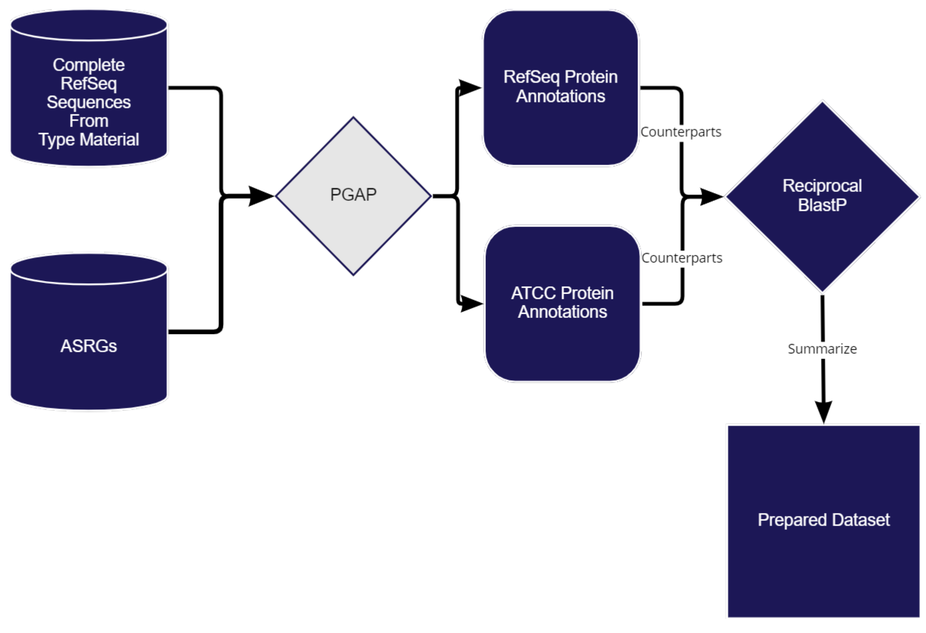

Figure 2: Analysis of public assemblies and corresponding ASRGs. RefSeq sequences were included in this dataset if they were labeled as “Complete Genome” or “Chromosome” level assemblies, they were labeled as “Assembled from type material” or similar phrasing, and there existed a bacterial assembly in RefSeq labeled with an ATCC catalog number that had been published to the ATCC Genome Portal.[3] Both the 185 RefSeq sequences satisfying these criteria and their matching ASRGs were analyzed for predicted genes using PGAP build-6021, and each set of protein annotations were further analyzed using a Reciprocal Best Hits analysis that is based on the BlastP algorithm with no lower-bound threshold for a match. Further tuning and parsing of the resultant dataset was performed using a combination of custom scripts in bash and python, establishing the prepared dataset. Additionally, supplementary tables S2, S4, and S5 from our previously published work[1] were included for more detailed comparison.

Results

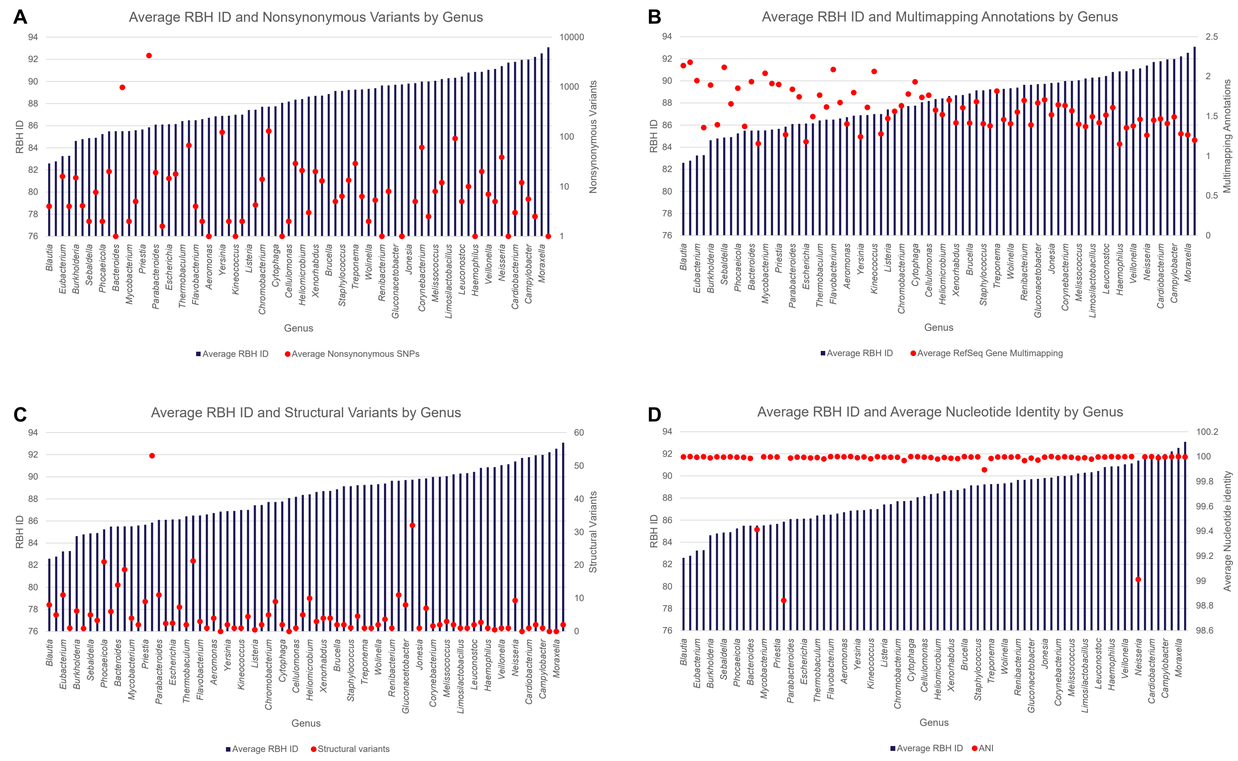

Figure 3: Relationships between RBH ID and common sequence comparison metrics, summarized by genus. (A) Nonsynonymous variants indicate amino acid changes between sequences but show little correlation to RBH values. (B) Multimapping annotations according to repeated results from RBH analysis. There is a slight trend toward higher RBH values having lower multimapping events. (C) Structural variants are weakly correlated with RBH values. (D) Average Nucleotide Identity between RefSeq and ATCC sequences are consistently very high, regardless of RBH values, with few outlier sequences.

Table 1: Statistical summary of several sequence comparison parameters

	Average RBH ID	Average Gene Name Match	ANI	Structural variants	SNPS	Indels	Synonymous	Non-synonymous	Gain Stop	Lose Start	Lose Stop	Average RefSeq Gene Multimapping
STDev	2.52	0.06	0.012	14.56	104.13	70.81	20.71	51.16	0.78	0.45	0.41	0.21
Mean	87.70	0.88	99.99	4.15	28.61	26.74	5.93	13.12	0.22	0.06	0.09	1.54
Accessions 1 STDev from Mean	41	2	11	176	171	169	171	175	174	176	171	149

Discussion

Despite the label “Assembly from type material,” several records contain substantial genomic and proteomic differences relative to their corresponding ASRGs. While the average Reciprocal Best Hit identity (RBH ID) across all genomes is 87.7%, over 19.5% of annotated genes have less than 75% identity to their best hit. As there is at best a weak trend relating RBH ID to ANI, small or large variant occurrence, the gain/loss of start and stop codons, or the multimapping frequency of these annotations, the source of all discrepancy between the RefSeq and ASRG datasets is unclear.

What should be interpreted from these data is that great care must be taken when selecting a dataset appropriate for usage in downstream analytics and research. Inaccurate data carry implications in PCR design, database construction, comparative genomics and proteomics, biomarker discovery, taxonomic classification, and simple alignment or assembly validation, among other common needs.

The absence of significant metadata fields, such as sequencing or assembly method and source location or purpose, present insurmountable challenges when it comes to such a selection. It remains outside the mission of most genomics databases (including RefSeq) to account for these fields.

Currently, there exists no protocol that will definitively produce a perfect description of genomic content. Thus, selection of genomic information to include in research is best based on a database that represents consistent methodology and is routinely updated, preferably from a single source with reliable traceability.

References

1. Yarmosh DA, et al. Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies. mSphere 7(3):e0007722, 2022. PubMed: 35491842
2. Tatusova T, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44(14):6614-24, 2016. PubMed: 27342282
3. Ward N, Moreno-Hagelsieb G. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? PLoS One 9(7):e101850, 2014. PubMed: 25013894

Learn more about our Enhanced Authentication Initiative

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