ARGOS-QC: A Comparative Quality Assessment Tool for Genome Assemblies

Yi Yan and Heike Sichtig
FDA Center for Devices and Radiological Health (CDRH)

Introduction

Assessment of genome assembly quality remains an open problem despite the rapid advances in sequencing and assembly technologies. Currently, multiple metrics are needed for the assessment of genome assembly quality, and the interpretation of these metrics depends on the intended use of the assemblies [1]. Here we present ARGOS-QC, a software tool that provides a generalized assessment of genome assembly quality through a comparative approach. ARGOS-QC calculates commonly used assembly quality metrics (N50, L50, ANI, et al.) of target assemblies and compares these metrics to user-defined backgrounds, such as all GenBank microbial assemblies. ARGOS-QC also provides meta-data comparison of target assemblies based on user-defined parameters. The results of ARGOS-QC are captured in both static and interactive reports to aid in the interpretation of the assembly assessment result. In this study, we present an overview of the ARGOS-QC workflow and use the ARGOS-QC for the quality assessment of the NCTC 3000 sequencing project.

Quality Control Pipeline Description:

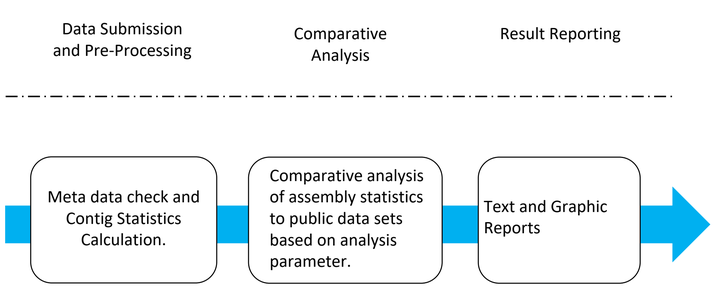

Figure 1: Flow diagram of ARGOS-QC.

The workflow of ARGOS-QC is shown in Figure 1. ARGOS-QC is composed of three components: (1) assembly metric calculation, (2) comparison to curated reference assemblies and (3) report generation. Input for the QC pipeline includes both the assembly files and accompanying metadata. The default analysis compares each input assembly to curated NCBI GenBank assemblies of the same species and the user can define additional restraints for the analysis. ARGOS-QC produces a tab-delimited file capturing the genome assessment result for each individual assembly and a summary graphic.

Assessment of the NCTC 3000 Project

The NCTC 3000 project is a collaborative project between Public Health England, Pacific Biosciences and the Wellcome Sanger Institute to sequence 3000 bacterial species using Pacific Biosciences' Single Molecule, Real-Time (SMRT) sequencing technology [2]. The ARGOS-QC pipeline was used to assess the quality of the publicly available NCTC 3000 assemblies. The NCTC assemblies were compared to a curated set of NCBI GenBank assemblies, TCC assemblies [3] and publicly available ARGOS assemblies [4]. The TCC assemblies from the Los Alamos laboratory were chosen as examples of high quality 454-Illumina hybrid assemblies and ARGOS assemblies were chosen as a collection of PacBio-Illumina hybrid assemblies. A summary comparison of NCTC, GenBank, TCC and ARGOS assemblies is shown in Table 1.

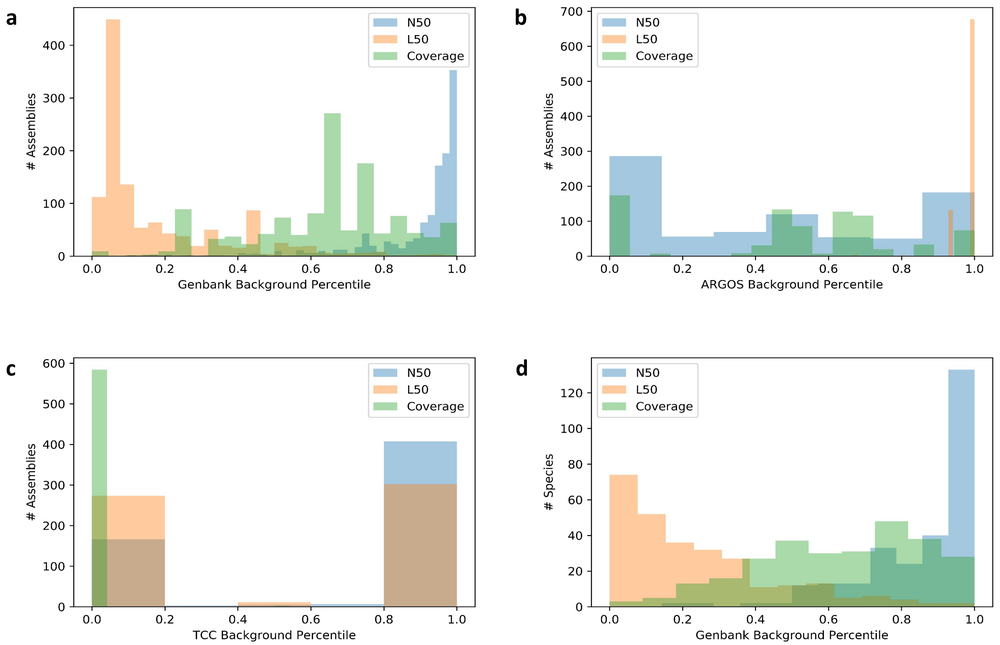

Figure 2: (a) Histogram of the percentile metric of N50, L50 and Coverage within the NCTC assemblies in comparison to GenBank assemblies. (b) Comparison to ARGOS assemblies. (c) Comparison to TCC assemblies. (d) Histogram of the percentile metric of N50, L50 and Coverage within the NCTC species in comparison to GenBank assemblies

Using ARGOS-QC, each NCTC assembly is characterized by the percentage of assemblies of the same species within the comparison group (NCBI Genbank, TCC, ARGOS) with lower N50, L50, and Coverage. A high percentage score for N50 and a low percentage score for L50 would mean that the specific assembly has higher continuity than the existing assemblies of the same species. Additionally, the GC% and assembly size of each NCTC assemblies were also compared to the distribution of assembly size and GC% of assemblies from the same species. The histograms of these three metrics are shown in Figure 2.

Discussion

	NCTC	GenBank	TCC	ARGOS
# Assemblies	1199	120011	195	332
# Species	276	1480	50	123
# Median / Species	2	12	1	1

Table 1: Summary comparison of NCTC, GenBank, TCC and ARGOS. Total number of assemblies, number of species and median number of assemblies per species metrics are provided.

The ARGOS-QC analysis demonstrated that the majority of NCTC assemblies showed better L50 and N50 values than curated NCBI GenBank assemblies of the same species (Fig 2a). However, the coverage of the NCTC assembly did not display the same consistency. In comparison to ARGOS assemblies, the L50 value of assemblies from both projects had an L50 of 1. Additionally, N50 and Coverage of NCTC assemblies were also similar to that of ARGOS (Fig 2b). In contrast, comparison to the TCC project showed a distinct pattern: the coverage of NCTC assemblies was typically less than the coverage of TCC assemblies and the N50 and L50 comparison showed a bi-modal distribution (Fig 2c). The species-level comparison to Genbank assemblies highlighted the species-dependent variation of assembly quality (Fig 2d).

Conclusion

- ARGOS-QC can be used to quickly assess the quality of genome assemblies leveraging user-defined references.
- NCTC assemblies demonstrated higher continuity and average coverage. The coverage is similar to other PacBio assemblies.
- Additional work will be needed to correlate these ARGOS-QC results to specific algorithm performance.

References

1. Yang, Li-An, et al. "SQUAT: a Sequencing Quality Assessment Tool for data quality assessments of genome assemblies." BMC Genomics 19.9 (2019): 238.
2. https://www.phe-culturecollections.org.uk/collections/nctc-3000-project.aspx
3. Chain, Patrick, et.al. “Genome Project Standards in a New Era of Sequencing.” Science 10.1126 (2009): 326.
4. Sichtig, Heike, et al. "FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science." Nature Communications 10.1 (2019): 3313.