Published on Apr 11, 2023 by Arcadia Science **DOI**: 10.57844/arcadia-75v2-nj84

Streamlining genome assembly and QC with the reads2genome workflow

We want to swiftly generate genome assemblies and produce quality control statistics to gauge the need for more curation. We built a Nextflow pipeline that assembles Illumina, Nanopore, or PacBio sequencing reads for a single organism and runs QC checks on the resulting assembly.

Contributors (A-Z)

Feridun Mert Celebi, Megan L. Hochstrasser, Elizabeth A. McDaniel, Taylor Reiter, Peter S. Thuy-Boun

Version 1 · Mar 31, 2025

Purpose

We want to ensure that we assemble high-quality genomes in a reproducible manner. We built a Nextflow workflow, hifi2genome, to assemble PacBio HiFi reads from a single organism and produce quality control statistics for the resulting assembly. The product of this pipeline is an assembly, mapped reads, and interactive visualizations reported with MultiQC. The final HTML report addresses assembly quality, lineage-specific checks, and mapping statistics that will help us make more informed decisions about downstream curation and functional annotation efforts.

We built this pipeline using open-source software and tools, and we hope others will shape and extend this resource to fit their needs.

- This pub is part of the **project**, "Useful computing at Arcadia." Visit the project narrative for more background and context.
- The **hifi2genome Nextflow pipeline** is available at this GitHub repository.
- We've included a **sample report** of assemblies that we generated from 24 microorganisms in the PacBio HiFi "Food safety and infectious microbes" dataset.

The resource

The problem

Running the commands for assembly and quality control (QC) checks from sequencing efforts of single organisms can be fairly straightforward but repetitive, depending on the desired outcomes. We want to quickly generate assemblies and resulting statistics to decide if further curation is needed before moving forward with downstream steps.

Our solution

We've developed a computational resource that automates genome assembly and quality control checks from HiFi reads, called hifi2genome.

> The **hifi2genome Nextflow workflow** is available at this GitHub repository (DOI: 10.5281/zenodo.7706592).

An overview of the hifi2genome workflow

The hifi2genome pipeline injects a sample sheet that includes the sample name and the local path, URL, or URI of the HiFi reads in FASTQ format.

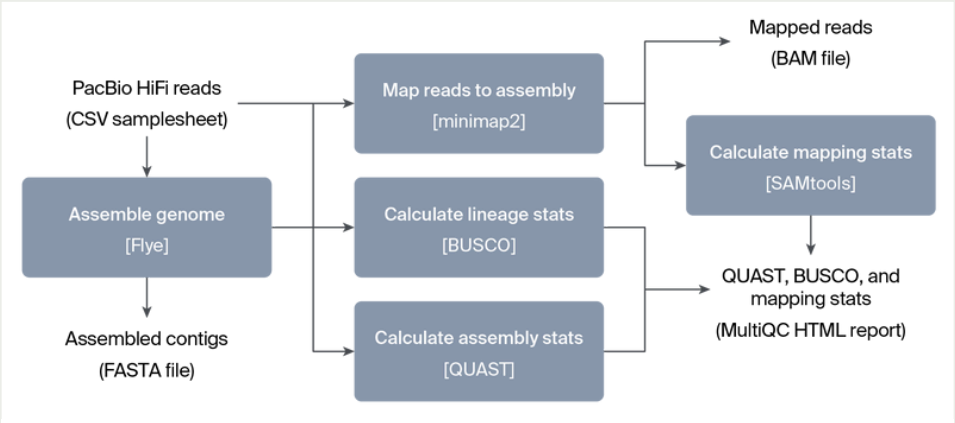

Figure 1

Computational steps in the hifi2genome workflow.

Each box represents a process step and the other text describes the input and outputs of the pipeline. File types are indicated in parentheses and software in brackets.

The first step in the pipeline runs Flye **[1]** to assemble the PacBio HiFi reads into contigs (Figure 1). Subsequent steps then run in parallel on the assembly for generating QC statistics. Currently, the checks we include are 1) assembly QC statistics with QUAST **[2]**, 2) lineage-specific QC statistics with BUSCO **[3]**, and 3) mapping stats generated with SAMtools **[4]** from mapping the reads back to the assembly with minimap2 **[5]**.

In addition to the sample sheet containing the path to the reads, the only other input the user must provide is the closest BUSCO lineage of the target organism for calculating lineage-specific completeness and redundancy statistics.

The final step of the pipeline aggregates the results from QUAST, BUSCO, `samtools stats`, and the information about the pipeline run and software versions into an HTML report with MultiQC **[6]** (Figure 1). MultiQC can generate an HTML report from the log files of numerous bioinformatics programs, and you can use it with or without running a Nextflow pipeline. The MultiQC report currently outputs general information about the

assemblies and mapping statistics, as well as more detailed information about each assembly from QUAST, including the distribution of sizes of contigs that were assembled, BUSCO lineage assessment results, and outputs from `samtools stats`, including percentages of the reads that mapped to each corresponding assembly and alignment metrics.

View an example of the MultiQC HTML report from the pipeline below from a run on a publicly available dataset of PacBio HiFi sequencing of 24 microorganisms from the "Food safety and infectious microbes" dataset with `nextflow run main.nf -input samplesheet.csv -outdir microbial_hifi_assemblies -profile docker -lineage bacteria`:

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

> This report has been generated by the Arcadia-Science/hifi2genome analysis pipeline. The purpose of this pipeline is to assemble PacBio Hifi reads and produce QC stats about the assembly. This pipeline performs assembly with Flye, assembly statistics with QUAST lineage specific QC statistics with BUSCO, mapping of reads back to the assembly with minimap2 and subsequent mapping statistics with samtools.

Report generated on 2023-02-24, 20:27 UTC

Download the sample report:

html `hifi2genome_multiqc_report.html` Download

Deployment

We deploy the pipeline with continuous integration testing using subsampled PacBio HiFi reads of two strains of *E. coli* from the PacBio HiFi "Food safety and infectious microbes" dataset for ensuring proper execution of the workflow as new features are added.

We are currently deploying all of our Nextflow workflows, including hifi2genome, through Nextflow Tower using our AWS Batch setup **[7]**. The pipeline is still fully executable locally via the command line and works on diverse compute infrastructure setups.

We found that for organisms with small genomes, such as bacteria and archaea, hifi2genome assembles the reads fairly quickly, and can run these jobs on interruptible AWS EC2 spot instances and complete successfully. However, for higher-order eukaryotes with larger genomes, like humans and ticks **[8]**, which might take multiple days to assemble, we needed to reconfigure the Nextflow Tower queue directive settings so that assemblies running via on-demand instances would not be interrupted.

> The **hifi2genome Nextflow workflow** is available at this GitHub repository (DOI: 10.5281/zenodo.7706592).

Next steps

This first version of the hifi2genome pipeline is a simple way to assemble PacBio HiFi reads and QC the resulting assembly. In the future, we would like to:

- Provide the user with the option to use other assembly algorithms (such as Hifiasm) in place of Flye or concurrently to compare assembly outputs.
- Add an optional endosymbiont detection subworkflow for pulling out contigs that do not belong to the host genome and are likely symbiont(s) sequences.
- Extend the workflow or apply its methods to Nanopore- and Illumina-based single-organism assembly workflows

For these efforts, we have created GitHub issues in the hifi2genome GitHub repository and welcome outside suggestions and contributions through pull requests!

Contributors

(A–Z)

- Feridun Mert Celebi
  - Critical Feedback, Validation
- Megan L. Hochstrasser
  - Editing, Visualization
- Elizabeth McDaniel
  - Conceptualization, Software, Visualization, Writing
- Taylor Reiter
  - Critical Feedback, Validation
- Peter S. Thuy-Boun
  - Critical Feedback

References

1. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. (2019). Assembly of long, error-prone reads using repeat graphs. https://doi.org/10.1038/s41587-019-0072-8
2. Gurevich A, Saveliev V, Vyahhi N, Tesler G. (2013). QUAST: quality assessment tool for genome assemblies. https://doi.org/10.1093/bioinformatics/btt086
3. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. https://doi.org/10.1093/bioinformatics/btv351

4 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. (2009). The Sequence Alignment/Map format and SAMtools. https://doi.org/10.1093/bioinformatics/btp352

5 Li H. (2018). Minimap2: pairwise alignment for nucleotide sequences. https://doi.org/10.1093/bioinformatics/bty191

6 Ewels P, Magnusson M, Lundin S, Käller M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. https://doi.org/10.1093/bioinformatics/btw354

7 Celebi FM, McDaniel EA, Reiter T. (2024). Creating reproducible workflows for complex computational pipelines. https://doi.org/10.57844/ARCADIA-CC5J-A519

8 Chou S, Poskanzer KE, Rollins M, Thuy-Boun PS. (2024). De novo assembly of a long-read Amblyomma americanum tick genome. https://doi.org/10.57844/ARCADIA-9B6J-Q683