Gathering and processing Whole genome sequences

Date etc

NC STATE UNIVERSITY

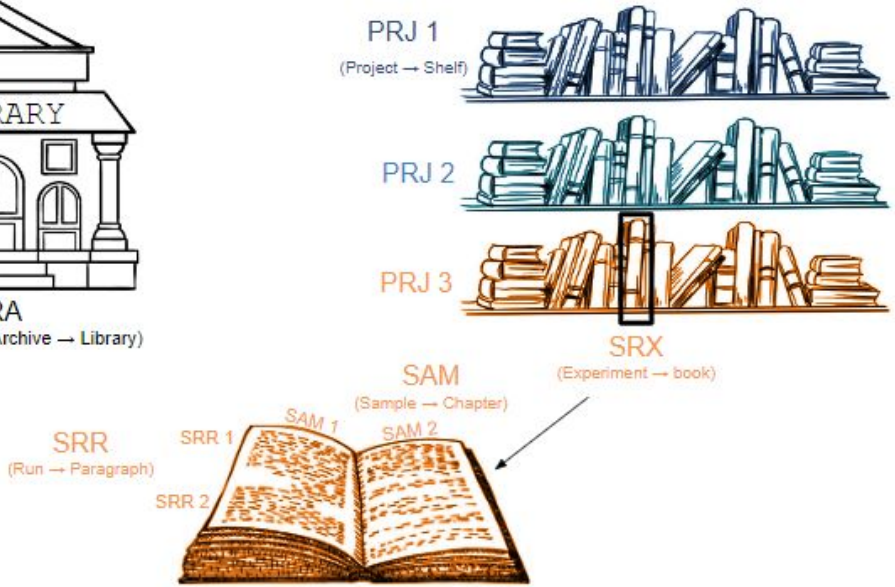
LANZAS

Infectious disease dynamics lab

Refresher

- Genomic data organization
- NCBI downloads
- Metadata
- Google colab

SRA
(Sequence Read Archive → Library)



Learning Outcomes - Day 2

- Use of the Bactopia pipeline to download sequence metadata, assess sequence QC and determine sequence completeness:
  - Obtain sequence metadata from NCBI IDs.
  - Application of FastQC to assess genome quality.
  - Produce genomic assemblies using SPAdes and Shovill.
  - Perform variant calling analysis using Snippy.
  - Produce genomic annotations using Prokka.
  - Assess genome completeness using BUSCO.

Bactopia

Open-source bioinformatic **pipeline** specifically designed for the complete analysis of bacterial genomes.

Analyze individual bacterial genomes or large datasets with thousands of genomes.

- Assemble and annotate bacterial genomes.
- Identify genes and their functions.
- Compare genomes to identify similarities and differences.
- Build phylogenetic trees to understand the evolutionary relationships between different bacteria.
- Identify potential virulence factors and antibiotic resistance genes.

| Research Article | 4 August 2020

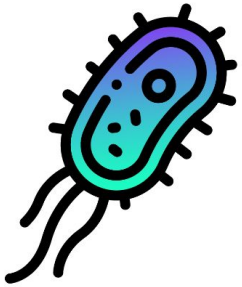
Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes

Authors: Robert A. Petit III, Timothy D. Read | AUTHORS INFO & AFFILIATIONS

DOI: https://doi.org/10.1128/msystems.00190-20 • Check for updates

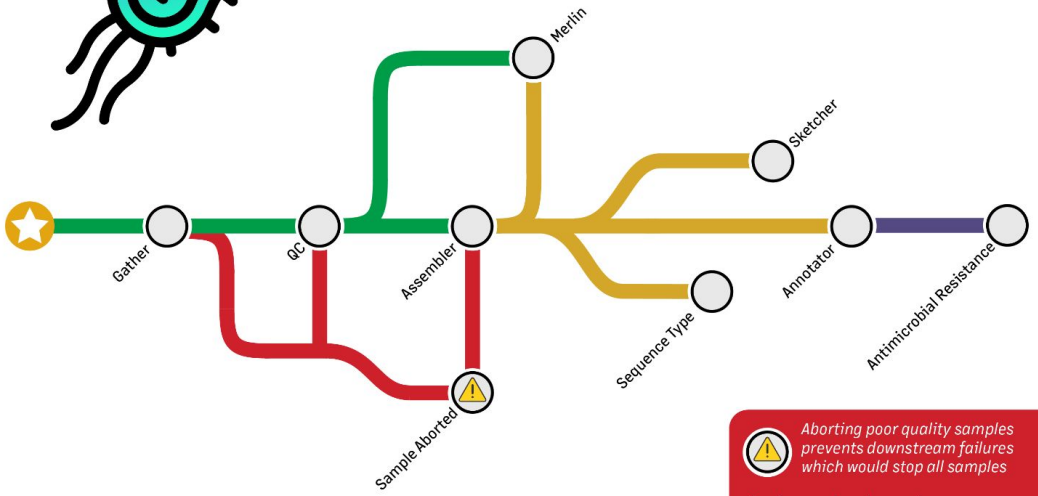
GitHub

Bactopia



BACTOPIA

COMPLETE ANALYSIS OF BACTERIAL GENOMES



Accepted Inputs

Illumina and/or Nanopore Reads
--R1/--R2, --SE, --ont. --hybrid, --sample
--samples with 'bactopia prepare' file-of-filenames

DDBJ/ENA/SRA Accessions
--accession 'Experiment Accession'
--accessions with 'bactopia search' results

Assemblies
--assembly, --sample
--samples with 'bactopia prepare' file-of-filenames

NCBI Assembly Accessions
--accession 'Assembly Accession'
--accessions 'file with accessions'

Abort Reasons

Aborting poor quality samples prevents downstream failures which would stop all samples

- Too few reads or basepairs
- Coverage below minimum
- Paired-end with different read counts, mismatched IDs, or skewed proportions
- Genome size outside expectation
- 0 assembled contigs
- Assembled size below minimum

Legend

- Process uses FASTQs
- Process uses Contigs
- Process uses Contigs and Proteins
- Minimum QC not met, sample aborted

Bactopia Processes

Gather
Collect local files and/or download sequences from SRA/ENA/DDBJ or NCBI Assembly accessions

QC
Trim and filter low quality reads, subsample to specified coverage, and generate summary metrics

Assembler
Create a de novo assembly (standard, hydbrid, or short read polished) and summary metrics

Merlin
Use Mash distances to automatically execute species specific tools, requires --ask_merlin

Sketcher
Create minmer sketches and query them against GTDB and RefSeq

Annotator
Predict genes and proteins from the assembled contigs

Antimicrobial Resistance
Identify presence of AMR and/or virulence genes

Sequence Type
Determine sequence type base on PubMLST profiles

Quality Control

Controlling the quality of raw data helps to quickly identify poor-quality samples in addition to flagging data issues.

This often means saving a great amount of time in later analysis.

FastQC: Quality control tool for high throughput sequence data.

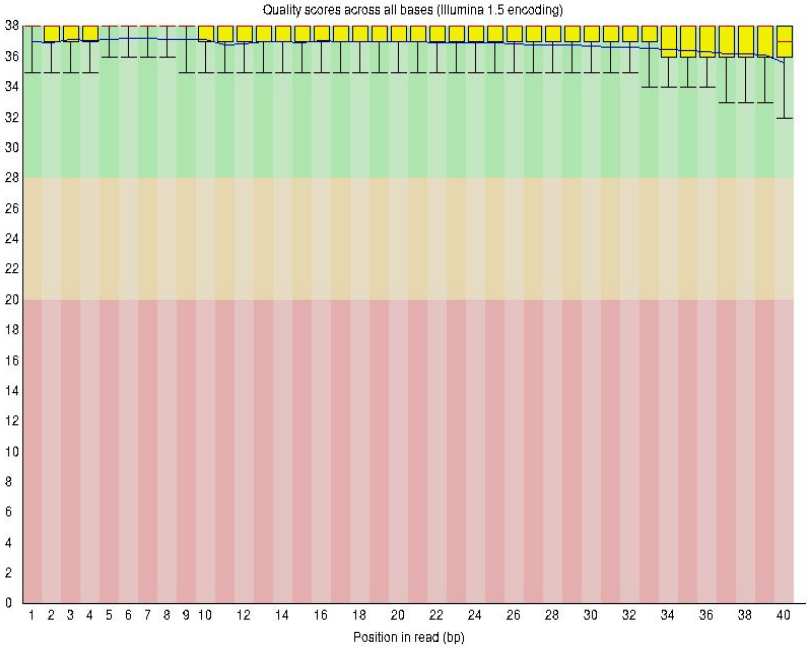
1. Import of data from BAM, SAM or FastQ files (any variant)
2. Providing a quick overview to tell you in which areas there may be problems
3. Summary graphs and tables to quickly assess your data
4. Export of results to an HTML based permanent report
5. Offline operation to allow automated generation of reports without running the interactive application

More information:

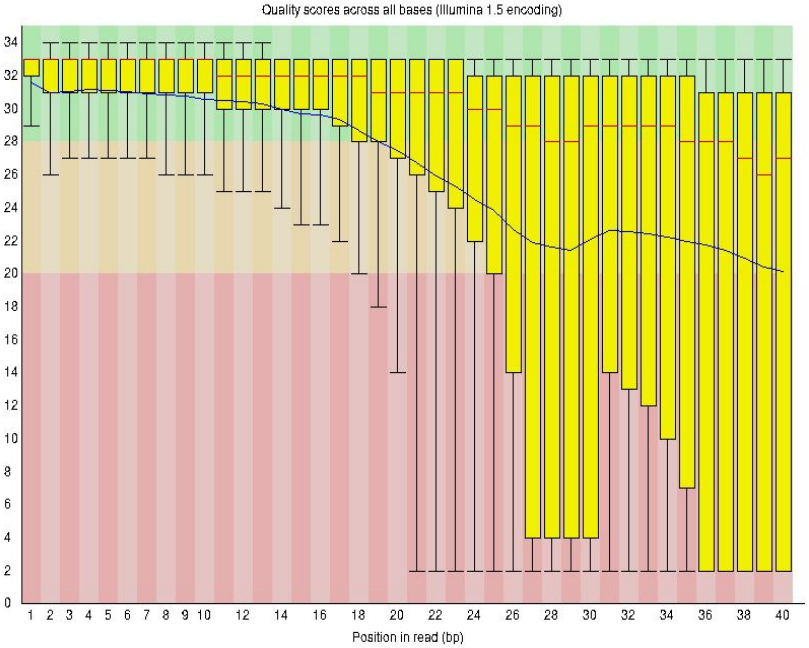
Babraham Institute

Quality Control

Good sequence data



Bad sequence data



More information:

Genome Assembly

Taking each piece and placing them back together in order to reconstruct the original message.

Short-read sequencing produces numerous smaller fragments, while long-read sequencing generates longer fragments.

Genome Assembly

Genome assembly is the process of reconstructing a genome from short sequencing reads generated by WGS.

To do this, Bactopia uses **SPAdes** and **Shovill.**

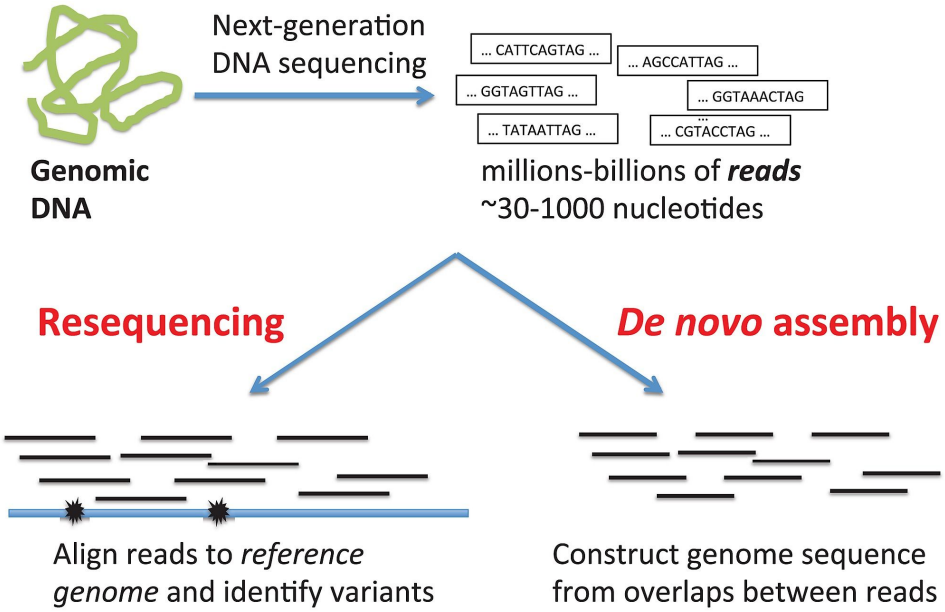


Figure from: Raphael, B. J. (2012). Chapter 6: Structural variation and medical genomics. PLoS computational biology, 8(12), e1002821.

Genome Assembly

SPAdes is the *de facto* standard for genome assembly for Illumina WGS, however, its components can be slow and it traditionally did not handle overlapping paired-end reads well.

Shovill uses SPAdes at its core, but alters the steps before and after the primary assembly step to get similar results in less time.

Journal of Computational Biology > Vol. 19, No. 5 > Original Articles

SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing

Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner

Published Online: 7 May 2012 | https://doi.org/10.1089/cmb.2012.0021

More information

Genome Assembly QC

To assess the quality of the assemblies, Bactopia uses **MASH** and **QUAST**.

More information

Method | Open access | Published: 05 November 2019

Mash Screen: high-throughput sequence containment estimation for genome discovery

Brian D. Ondov, Gabriel J. Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B. Buck & Adam M. Phillippy

Genome Biology **20**, Article number: 232 (2019) | Cite this article

13k Accesses | **117** Citations | **64** Altmetric | Metrics

JOURNAL ARTICLE

QUAST: quality assessment tool for genome assemblies FREE

Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, Glenn Tesler Author Notes

Bioinformatics, Volume 29, Issue 8, April 2013, Pages 1072–1075,
https://doi.org/10.1093/bioinformatics/btt086

Published: 19 February 2013 **Article history**

Variant calling

To identify genetic variants by comparing sequencing reads to a reference genome.

Types of variants

Indels

- Insertion (ins) **A ➡ AC**
- Deletion (del) **ACCG ➡ AG**

Substitutions

- Single nucleotide polymorphism (SNP) **A ➡ C**
- Multiple nucleotide polymorphism (MNP) **AG ➡ TC**

Complex

- Compound events **AC ➡ T**

To do this, Bactopia uses **Snippy**.

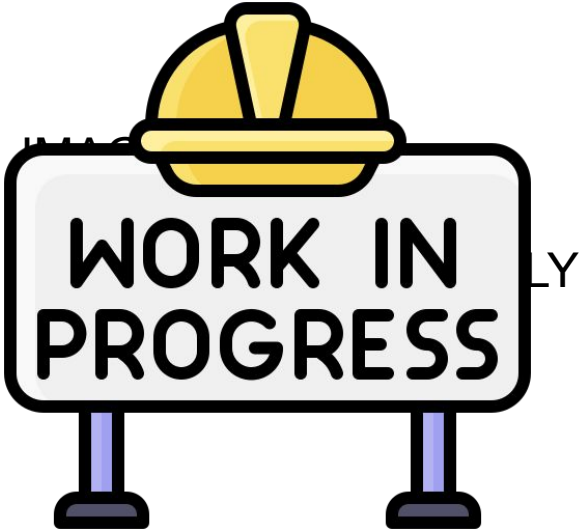
More information

Genome Annotation

Once you have the reconstructed book in your own language, genome annotation is like reading and interpreting the text.

It involves identifying and understanding the functional elements within the genome sequence:

- Predicting genes, their functions, regulatory elements…
- Repetitive regions, non-coding RNAs, transposable elements…



Genome Annotation

Genome annotation involves identifying genes and other genomic features and assigning functional annotations.

To do this, Bactopia uses **Prokka**

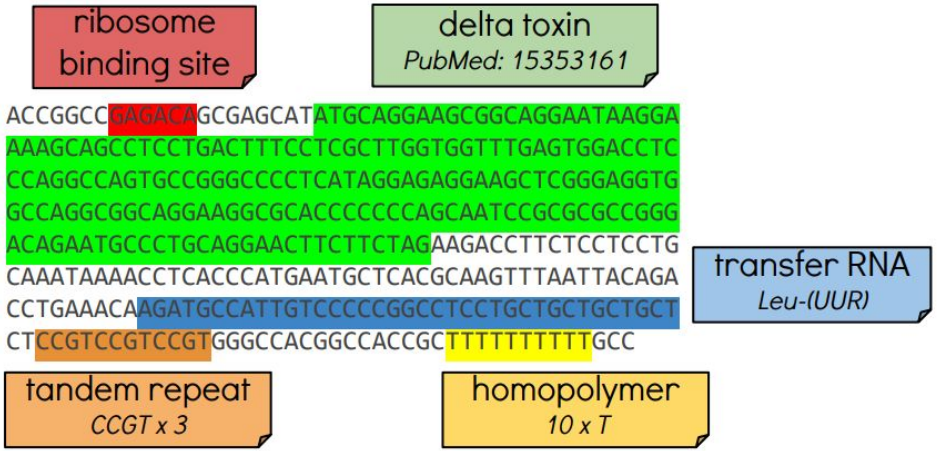
JOURNAL ARTICLE

Prokka: rapid prokaryotic genome annotation FREE

Torsten Seemann Author Notes

Bioinformatics, Volume 30, Issue 14, July 2014, Pages 2068–2069,
https://doi.org/10.1093/bioinformatics/btu153

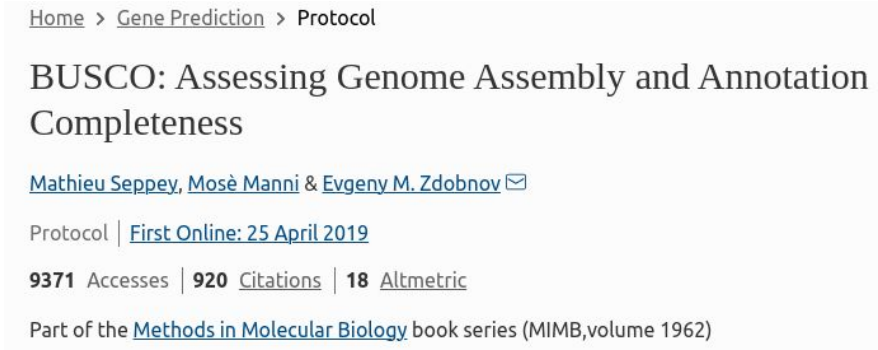
Published: 18 March 2014 **Article history** ▾



Genome Completeness

BUSCO assesses the completeness of genome assemblies by searching for a set of conserved single-copy orthologous genes.

A high BUSCO completeness score indicates a well-assembled genome with minimal fragmentation or gene loss.


Home > Gene Prediction > Protocol

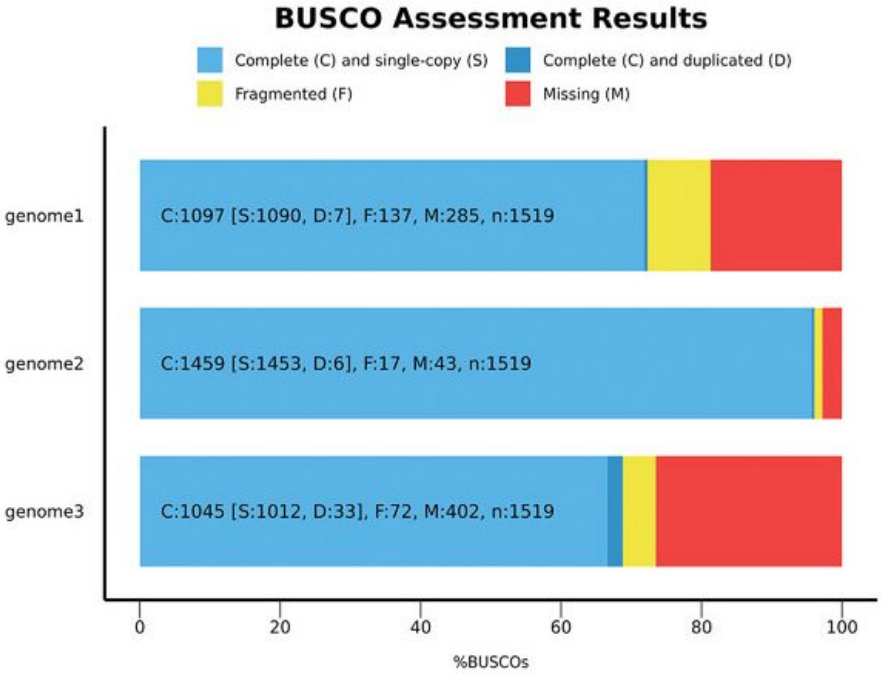
BUSCO: Assessing Genome Assembly and Annotation Completeness

Mathieu Seppey, Mosè Manni & Evgeny M. Zdobnov

Protocol | First Online: 25 April 2019

9371 Accesses | 920 Citations | 18 Altmetric

Part of the Methods in Molecular Biology book series (MIMB,volume 1962)



More information:

BUSCO UNIVERSITÉ DE GENÈVE FACULTÉ DE MÉDECINE SIB Swiss Institute of Bioinformatics

Questions