Phage Isolation, Genome Analysis and Comparative Genomics Workshop

Bioinformatics Analysis

By Fredrick Kebaso

Outline

- Pre-processing raw reads
  - ✓ Trimming
  - ✓ post-trim QC
- Genome assembly
  - ✓ Long and short read sequences
- Post-assembly QC

Pre-processing Raw Reads

Includes initial QC checks

Tools: **FastQC** or **MultiQC** to identify read quality issues:

- ✓ Per-base quality
- ✓ Adapter contamination
- ✓ Overrepresented sequences

Common issues in raw data

- ✓ Low base quality in 3’ ends
- ✓ Presence of adapter/primer sequences
- ✓ Contaminating host reads (e.g., bacterial host, not the phage)

Read Trimming

Tools: Trimmomatic, Cutadapt, Trimgalore etc

Why Trimming

- ✓ Trim low-quality bases at read ends
- ✓ Remove adapter sequences
- ✓ Isolate contaminating sequences (if known)

Parameters to consider

- ✓ Minimum Phred quality score cutoff (e.g., Q20 or Q30)
- ✓ Minimum read length after trimming (avoid too short reads)
- ✓ Adapter sequence trimming

Impact on assembly

- ✓ Improves accuracy
- ✓ Reduces errors in overlap/extension

Post-trim QC

Tools: FastQC or **MultiQC** again

QC checks after trimming

- ✓ Evaluate if the quality improved
- ✓ Confirm adapter sequences are removed

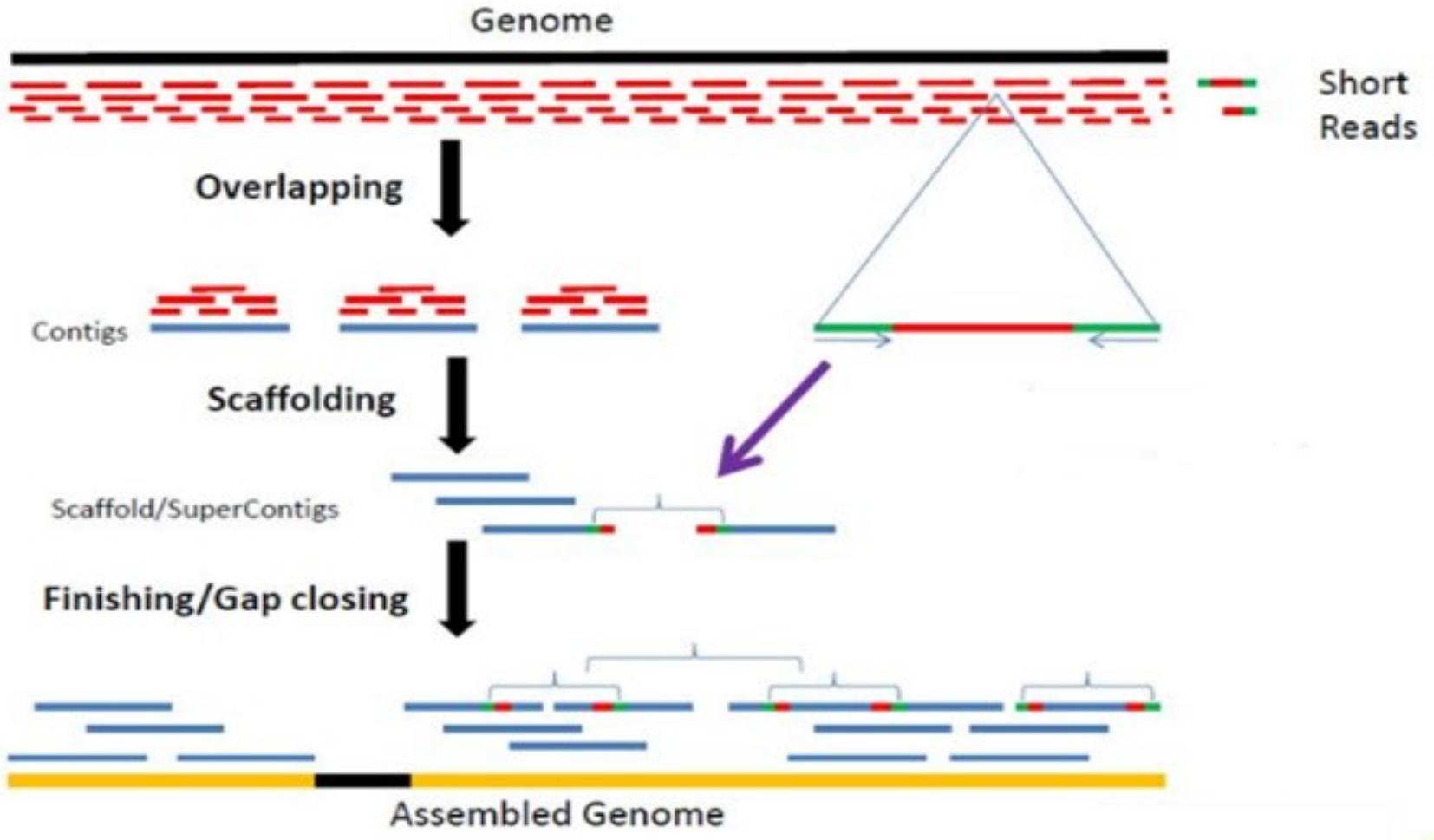
Coverage estimation

- ✓ Roughly estimate read coverage for the phage genome (if reference or approximate genome size is known)

Potential issues

- ✓ Over-trimming leading to short reads
- ✓ Persisting low-quality regions if trimming settings were too lenient

Genome Assembly Overview



Adapted from; https://www.arraygen.com/De-novo-Assembly.php

Short vs. Long Read Genome assemblies

Short reads (Illumina)

- ✓ Pros: high accuracy, high throughput
- ✓ Cons: difficulty spanning large repeats, potential fragmentation of the assembly

Long reads (Nanopore/PacBio)

- ✓ Pros: resolves repetitive regions, produces more contiguous assemblies
- ✓ Cons: higher error rates (though improving), requires careful error correction

Assembly Approaches

De novo assembly (most common for unknown or highly variable phage genomes)

Reference-guided assembly (less common, used if a closely related reference is available)

Tools for short-read assembly:

- ✓ **SPAdes** (commonly used for phage/bacterial genomes)
- ✓ **Velvet**, **Unicycler** (also used for microbial assemblies)

Tools for long-read assembly

- ✓ **Flye**, **Canu**, **Minimap2/Minasm** (for error correction and assembly)

Hybrid assembly:

- ✓ Combining short and long reads for better contiguity and accuracy

Short vs. Long Read Assembly

Short Reads (Illumina)

Pros:

- ✓ High accuracy (low error rates ~0.1–1%), High throughput, cost-effective

Cons:

- ✓ Struggles with large repeats → fragmented assemblies, Contigs may remain short/disjointed

Long Reads (Nanopore/PacBio)

Pros:

- ✓ Spans repetitive regions → more contiguous assemblies
- ✓ Better detection of large structural variants

Cons:

- ✓ Higher error rates (needs polishing)
- ✓ More expensive, specialized equipment

Post-assembly QC

Tools: Quast, Busco, Compleasm

Assembly quality metrics

- ✓ Assembly statistics (N50, total length, GC content)
- ✓ Check for completeness

Contiguity checks

- ✓ Look for large contigs vs. many small contigs
- ✓ Evaluate if assembly represents the expected genome size

Confirmation

- ✓ Map reads back to the assembled genome to check coverage uniformity

Common Challenges & Best Practices

Challenges

- ✓ Host DNA contamination
- ✓ Small genome size leading to coverage misinterpretation
- ✓ Repetitive elements or terminal redundancies

Best Practices

- ✓ Thorough QC at every step
- ✓ Use multiple tools for assembly and compare results
- ✓ Manual inspection of genome ends (especially for phages)
- ✓ Keep metadata organized (host strain, isolation details, library prep methods)

Lets get our hands dirty

Terminal is your Magic Wand

Ctrl+Alt+T

Time for Practicals