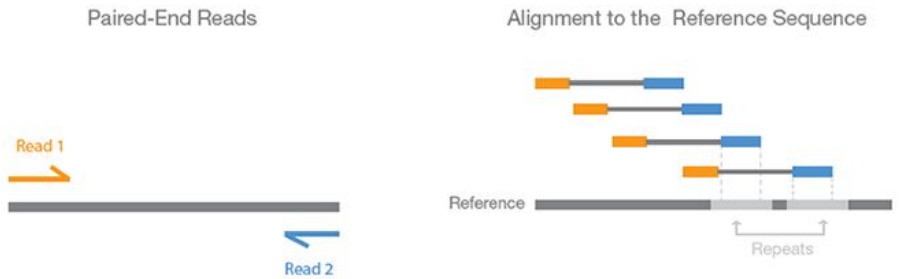
Genome Assembly

Team - II

Introduction

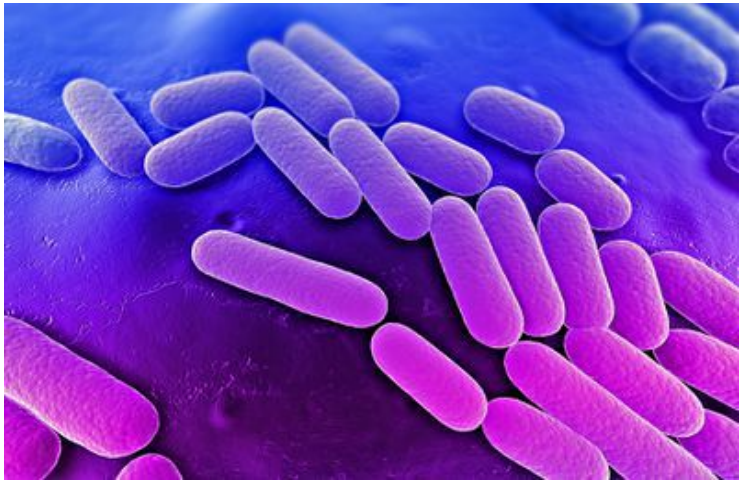
- Problem of antibiotic resistance
  - Heteroresistance in *Klebsiella spp.*
- The goals
  - To distinguish between susceptible and heteroresistant strains/species
  - To discover genomic determinants of antibiotic resistance
  - To develop a predictive web server
- Genome Assembly
- What we have
  - Reads from 260 *Klebsiella* genomes
  - Platform: Illumina MiSeq
  - Paired-End Sequencing



Illumina

Klebsiella spp.

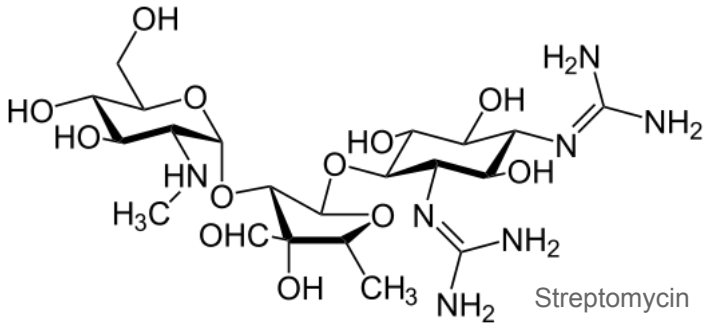
- Gram negative rod-shaped bacteria
- Genome size 5.3 - 5.9Mb
- *K. pneumoniae* & *K. oxytoca*: the most prevalent human pathogens
- *K. pneumoniae*: one of the leading causes of hospital acquired infections
- Disease states: pneumonia, urinary tract infections (contamination of urinary catheters), bacteremia and other systemic infections



Scanning electron microscope image of Klebsiella pneumoniae. From: Bioquell.com

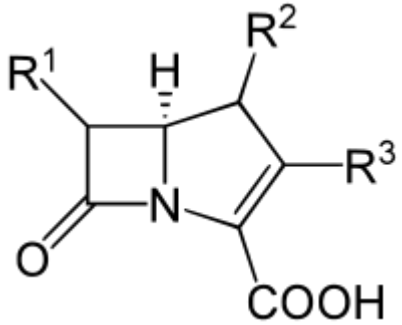
Antibiotics used for treatment and resistance

- Aminoglycoside resistance
  - Inhibit protein synthesis
  - Produce aminoglycoside-modifying enzymes



Streptomycin

- Beta-lactam resistance
  - Inhibit peptidoglycan synthesis, preventing cell wall formation
  - Produce Beta-lactamases or Carbamapenases

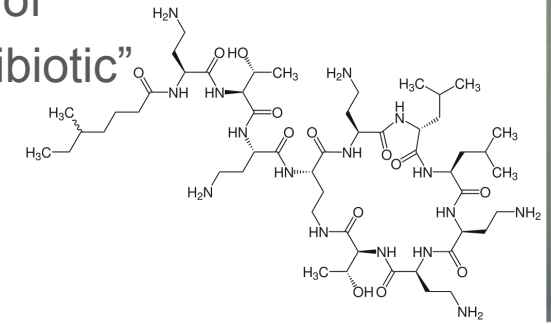


Core structure of carbapenem

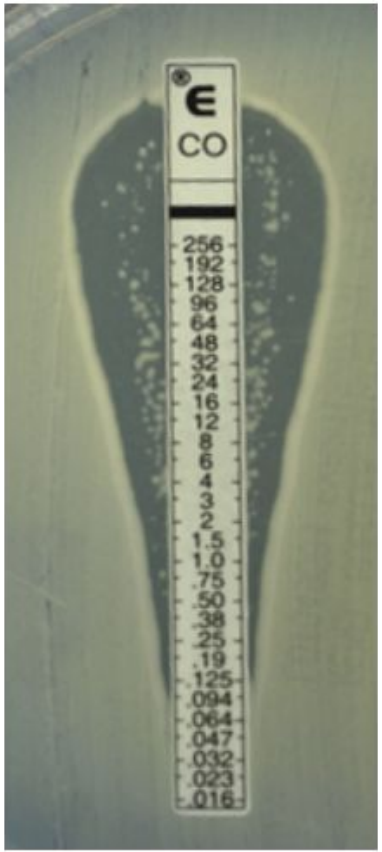
Wikipedia - Aminoglycoside, Carbapenem

Colistin & Heteroresistance

- Last-line antibiotics for multidrug-resistant Gram-negative bacteria
- Change the permeability of the cell wall
  - Cell leakage and death
- Heteroresistance: “subpopulations of seemingly isogenic bacteria exhibit a range of susceptibilities to a particular antibiotic”



Colistin



Heteroresistance

Wikipedia - Colistin, Dr. David Weiss's Lecture Slide

WGS has helped solve the problem

Whole-Genome Comparative Analysis of Two

Comparative Genomics of *Klebsiella pneumoniae* Strains with Different Antibiotic Resistance

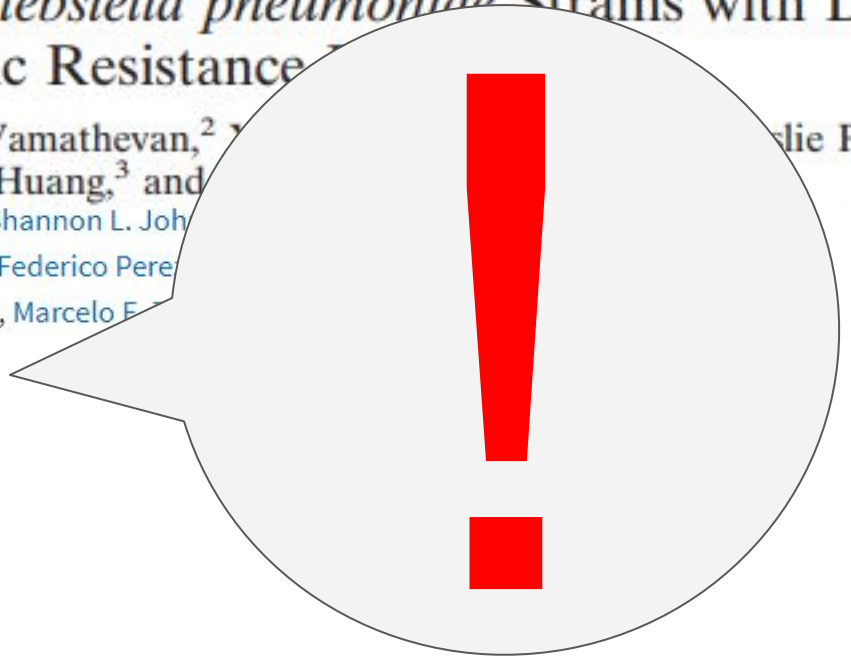
Vinod Kumar,[1] Peng Sun,[2‡] Jessica Vamathevan,[2] ... lie Palmer,[3] Jianzhong Huang,[3] and ...

María Soledad Ramirez, Gang Xie, German M. Traglia, Shannon L. Joh...

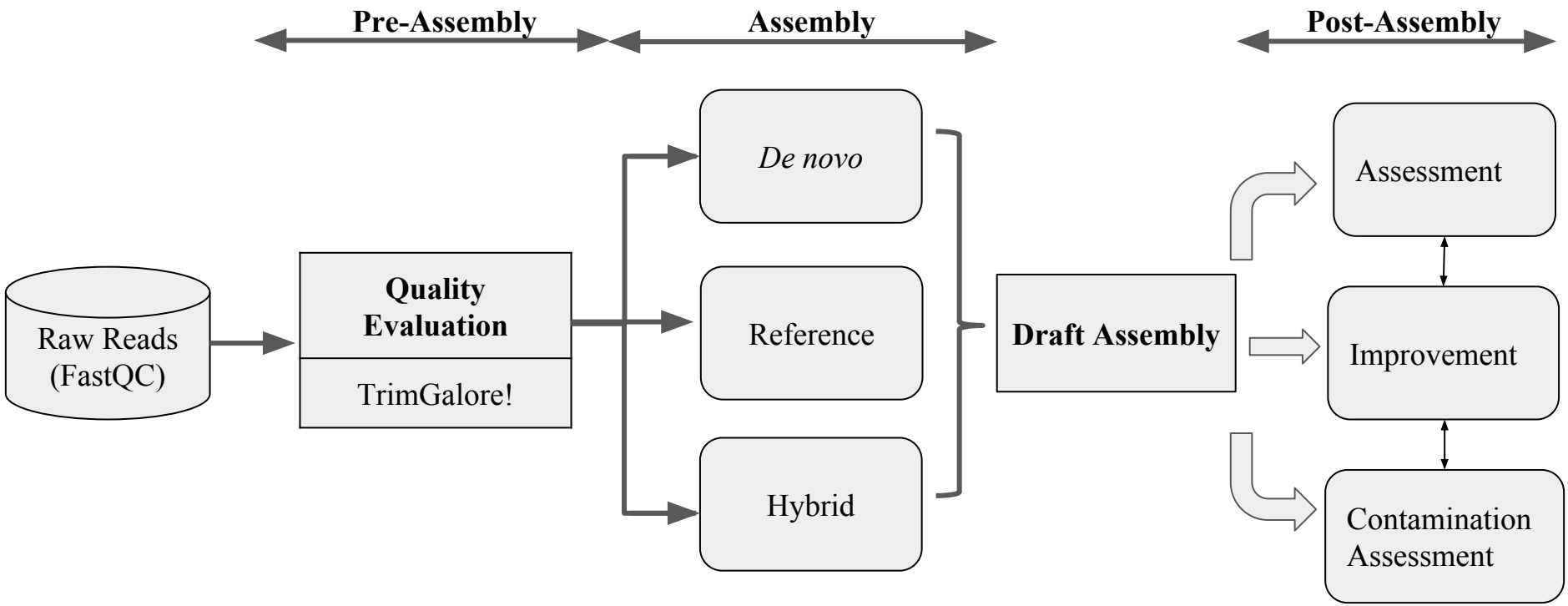
Karen W. Davenport, David van Duin, Azam Ramazani, Federico Pere...

David J. Sherratt Robert A. Bonomo, Patrick S.G. Chain, Marcelo F. ...

- Colistin & Heteroresistance



Pipeline


Pre-Assembly
Assembly
Post-Assembly
Raw Reads (FastQC)
Quality Evaluation
TrimGalore!
De novo
Reference
Hybrid
Draft Assembly
Assessment
Improvement
Contamination Assessment

Pre Assembly

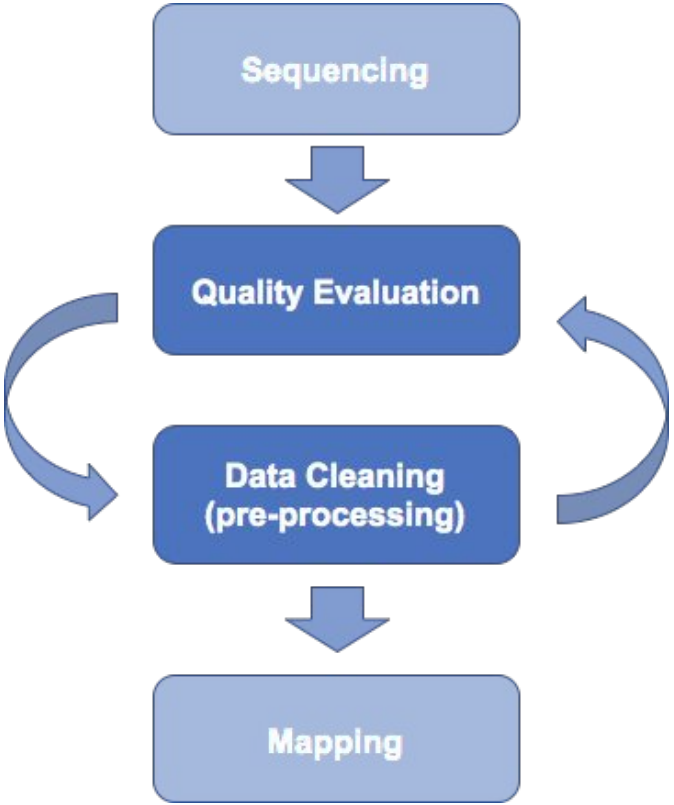
Quality Evaluation

It is important to check the quality of your sequenced reads!

Pre-processing

- Trim reads
- exclude low quality reads
- Contaminations

FastQC: reports quality profile of reads



Trim Galore!

WHAT IS Trim Galore!?

All in one pre-assembly tool

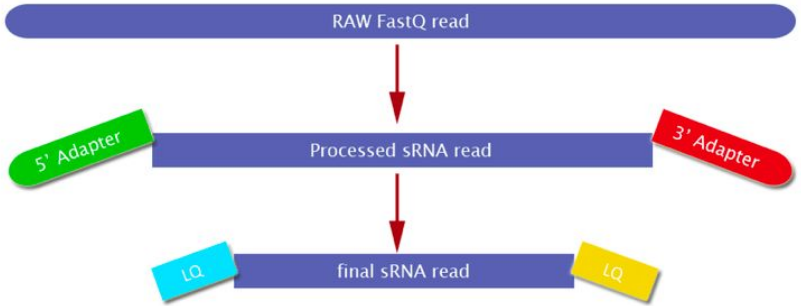
Trims

3’ adapter

5’ adapter

Low quality reads

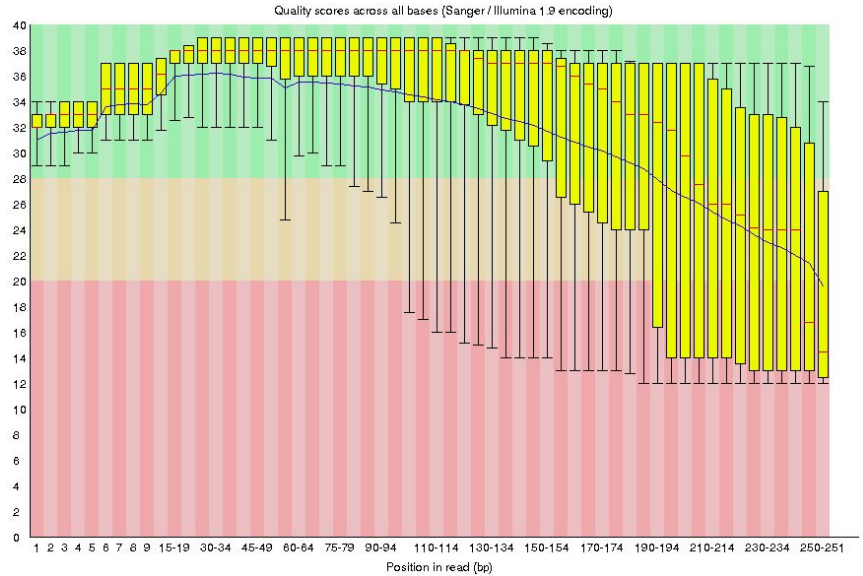
FastQC



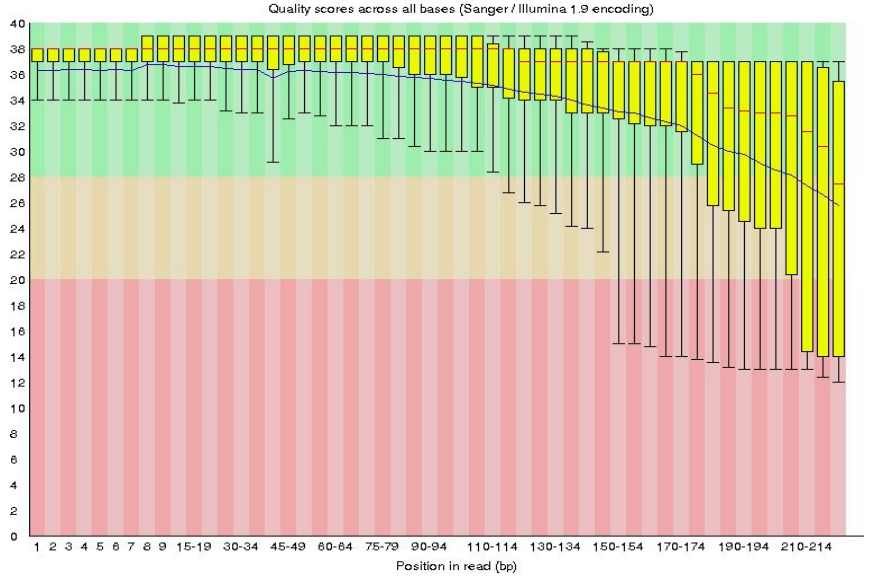
WHY DID WE CHOOSE IT?

- Specifically designed for Illumina data
- Has an Illumina adapter library
- Fast

FastQC



Before Trim Galore!

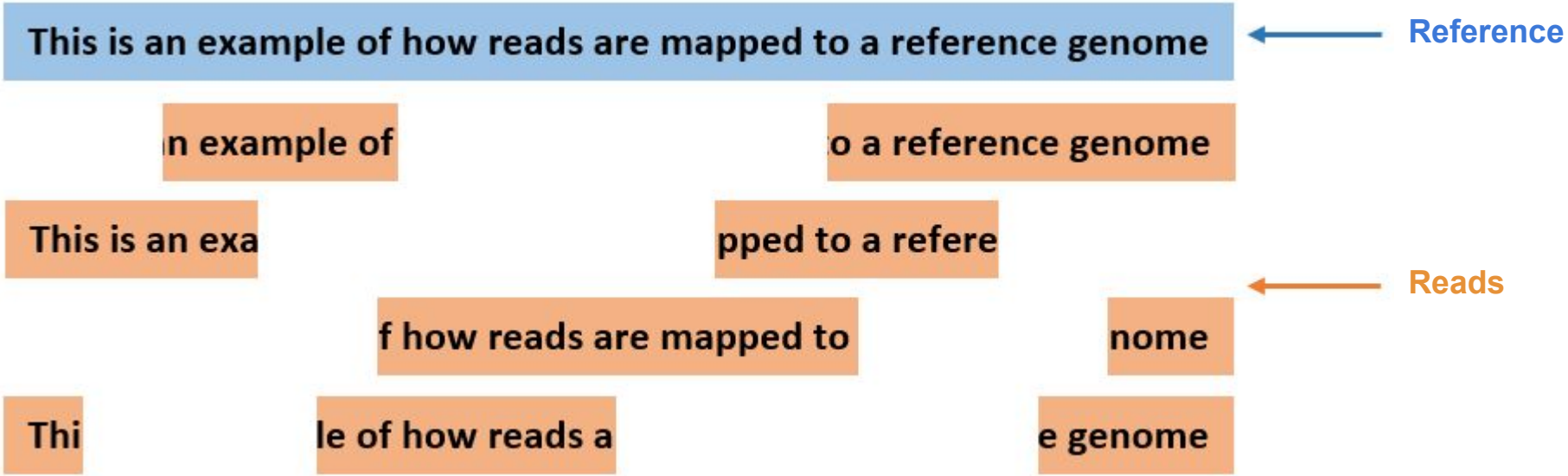


After Trim Galore!

Reference Assembly

Reference Assembly: Introduction

A type of genome assembly where the reads are mapped (or compared) to a known version of the organism's genome.



Reference Assembly: Introduction

- A type of genome assembly where the reads are mapped (or compared) to a known version of the organism's genome.

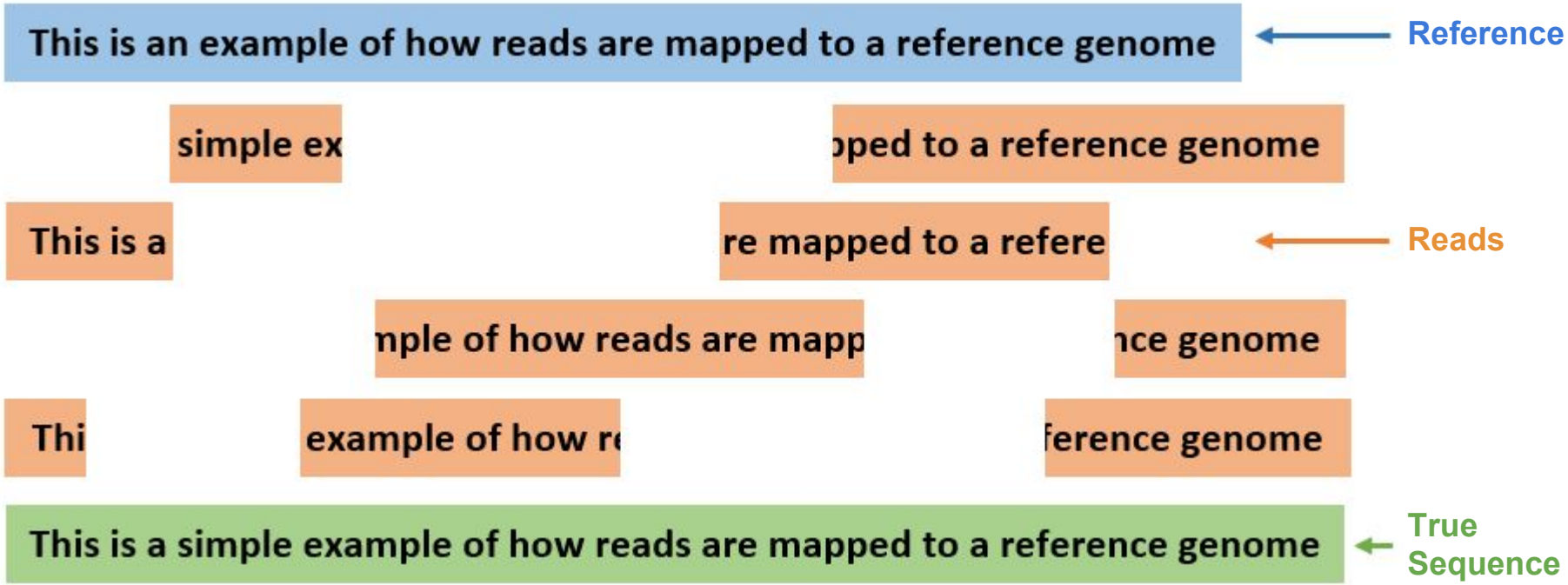
Reference Assembly: Advantages

- Saves time
- Is significantly more accurate
- You obtain less contigs

- SNPs and very small variations are more easily positioned and compared among groups

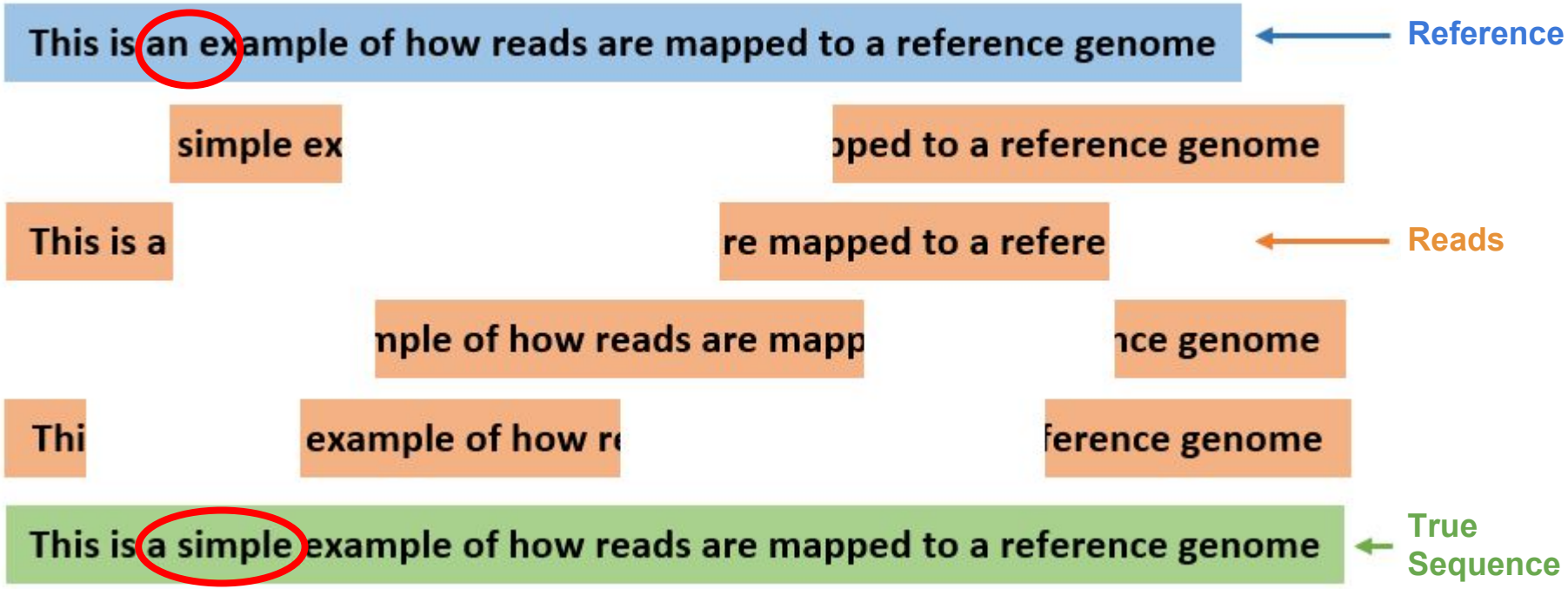
Reference Assembly: Disadvantages

New (completely different) sequences are lost



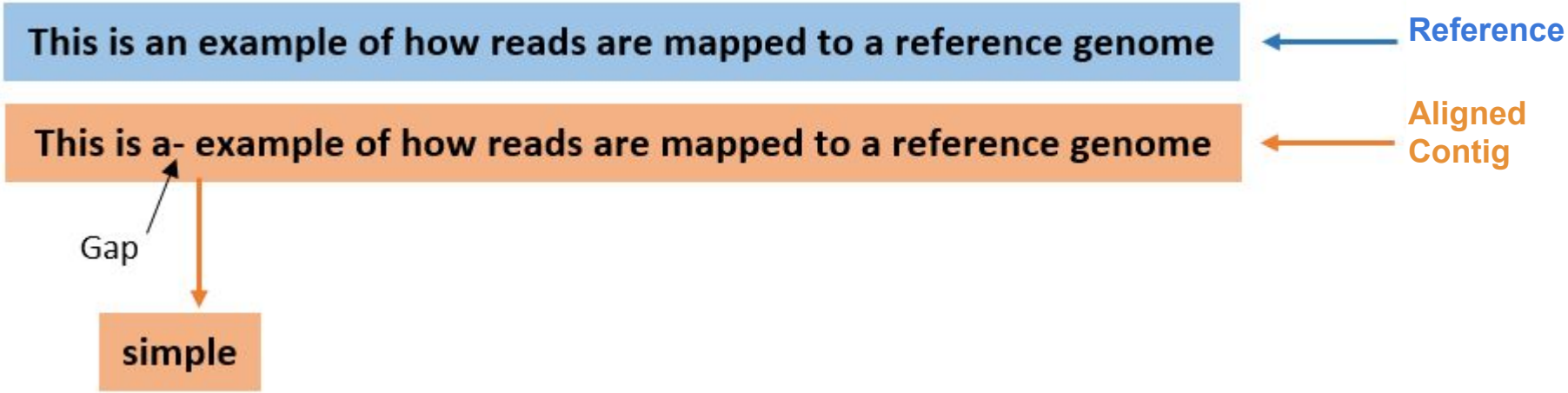
Reference Assembly: Disadvantages

New (completely different) sequences are lost



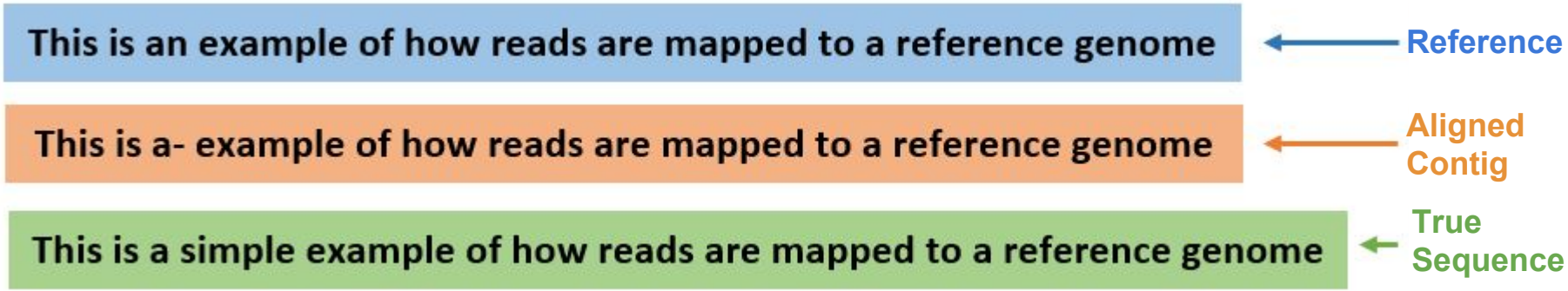
Reference Assembly: Disadvantages

New (completely different) sequences are lost



Reference Assembly: Disadvantages

New (completely different) sequences are lost



Consequence: Finalized alignment does not necessarily correspond entirely to the original sequence

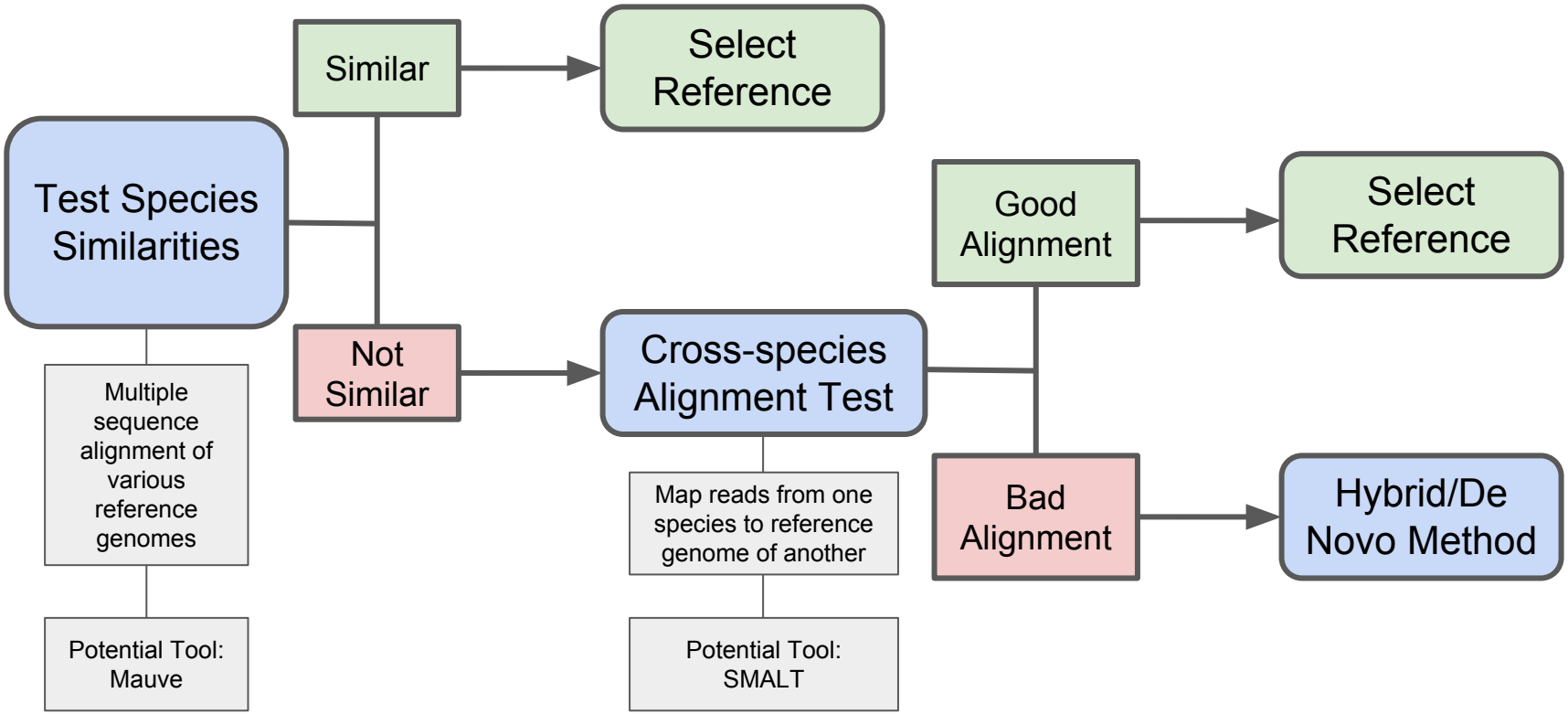
Reference Assembly: Disadvantages

- If multiple positions on the reference genome are equally likely for a read, then:
  - Reads are ignored
  - Reads are placed at multiple locations
  - Read is placed at one location, which is selected at random
  - Read is placed at the first likely position

- Requires a reference that is very similar to the sequenced data

- Limited by read length for feature detection

Reference Assembly: Our Plan



De Novo Assembly

Naïve Solution: Shortest Common Superstring

- Shortest common superstring: shortest string containing all of the reads
- Advantage: Well defined computer science problem
- Disadvantages:
  - NP-hard -- Fast approximation only guarantees max ~2.5x longer sequence
  - Repeats collapsing:

Consider reads of length 3

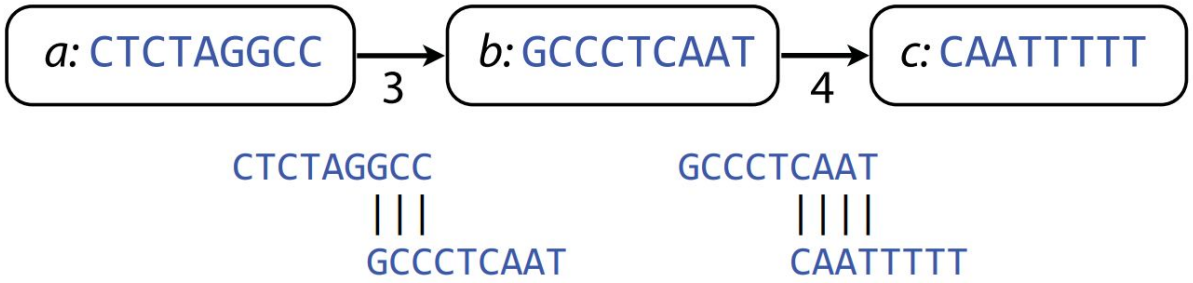
Genome: AAT**TCC**AGCTGAT**TCC**AGT

Assembly: AATTCCAGCTGA**TA**GT

Every 3-mer is in the assembly, yet 2-mer **TA** is not present in the original genome

Overlap Layout Assembly

- Build a string graph and deduct assembly from paths in the graph
- Nodes: Reads, Edges: Overlap
- Complexity: ***O(n^2)*** -- every pair of reads has to be compared to determine overlap lengths
- Not applicable for a large amount of short reads

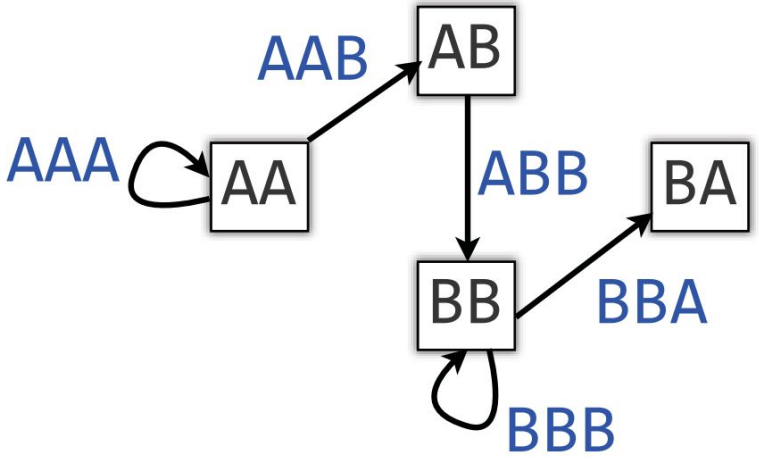


Overlap Layout Assembly

- Result?
- - Mess
- Many techniques to remove suboptimal edges
- Unambiguous paths which arise after edge removal are reported as contigs

De Bruijn Graph Assembly

- All reads are split into k-mers
- Form left and right k-1 mers: nodes in the graphs
- Nodes are connected by the original k-mers -- edges
- Eulerian walk (visits each edge exactly once) is the assembly

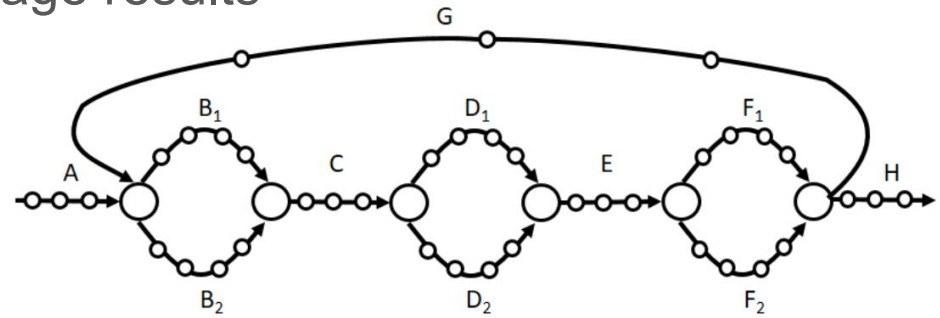


De Bruijn Graph Assembly: Speed

- Complexity analysis:
  - Add an edge: *O(1)* -- nodes are stored in a hashmap
  - Add all edges: *O(n*l)* where *l* is the average length of a read
  - Find Eulerian walk: *O(n*l)* -- linear with respect to the number of edges
  - Overall: ***O(n*l)***
- Speed is the key advantage of De Bruijn graph assembly approach

De Bruijn Graph Assembly: Issues

- The speed comes at a cost
- All reads split into independent k-mers -> even if an ideal graph is constructed, some reads may not appear in the final assembly
- Incomplete and erroneous coverage results in disconnected graphs
- Multiple optimal Eulerian walks may exist due to repeats
- Solution to all of the above: report only **unitigs** -- safe unambiguous segments



De Bruijn Graph Assembly: Tools

- Commonly used tools:
  - SPAdes
  - Unicycler
  - Skesa
  - MaSuRCA
- The tools are all based on De Bruijn graphs, difference lies in:
  - Determining optimal k-mer size
  - Edge pruning in the graph (removal of low quality edges)
  - Final contig extraction
  - Combination with overlap layout graphs

Hybrid Assembly

What is hybrid about it?

- Combines 2nd gen sequencing with 3rd gen sequencing
  - Short Reads (Illumina)
  - Long Reads (PacBio)
- Differ in accuracy
- Longer reads essentially used as reference
  - Important with long repeats, similar sequences

Challenges for Our Project

- **NEWS FLASH**: We don’t possess PacBio or other 3rd gen sequencing reads for these single-cell sequences

- Could simulate long reads from a reference
  - Same hangups as reference assembly with that method

Tools to be Tested

- SPAdes (hybrid options)
  - Pros
    - Easy to install
    - Relatively fast
  - Cons
    - Would have to simulate long reads

- Unicycler
  - Pros
    - Tested in previous years
  - Cons
    - Would have to simulate long reads

Post Assembly

Post Assembly Assessment

Contig Weighted Score

$$\frac{\log_{10}(\mathrm{N50} \cdot \mathrm{Length})}{\#\mathrm{contigs}}$$

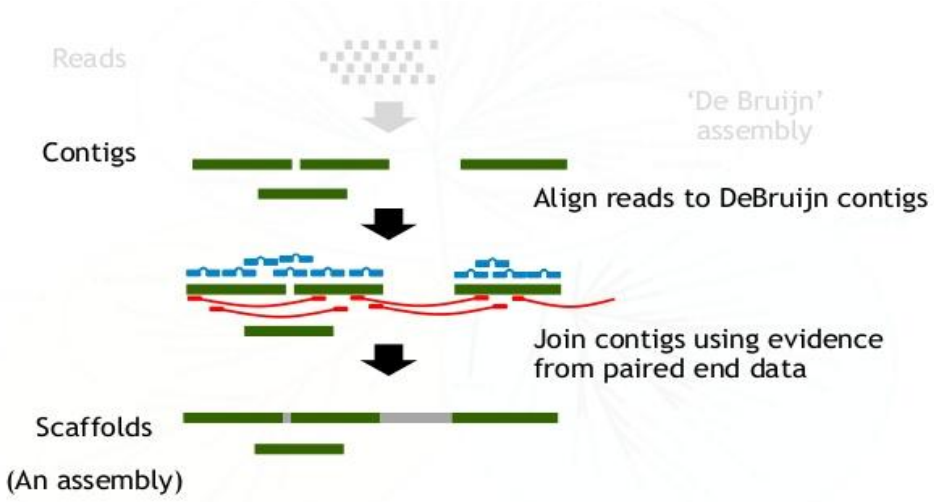
L50 Weighted Score

$$\frac{\log_{10}\left(\frac{\mathrm{N50}}{\#\mathrm{contigs}}\right)}{\left(\frac{\mathrm{AssemblyLength}}{\mathrm{ExpectedLength}}\right)^2}$$

* ↓ # contigs , ↑ length = better score

Post Assembly Improvement: Scaffolding

- Contigs are unordered mass of stretches of DNA
- Scaffolding tries to bring order and direction to these stretches of DNA
- Algorithms attempt to join multiple contigs using insert info and paired end data of reads.



Post Assembly Improvement: Scaffolding

Tools used:

- Bambus, Bambus2, SSPACE (standalone scaffolding tools)
- SOAP, SOAPdenovo2, SOPRA, SGA, velvet (Integrated into tools)
- CLA includes error checking

Post Assembly Improvement: Closing Gaps

- Gaps of undetermined bases (N) occur after scaffolding between super-contigs
- Tools:
  - Sealer: Local reassembly of gap regions. Useful in regions of repetitive sequences.
  - GapFiller: uses aligning paired end reads
  - GapCloser
  - GFinisher: In addition to gap closing, identifies errors as well.

Contamination Evaluation

- How Does Contamination Arise
  - Sample contamination
  - In-silico
- How Is Contamination Measured
  - Bacterial single-copy core genes

Contamination Evaluation

- Estimate Completeness
  - Used to easily filter out poor assemblies
  - Variance of single-copy genes

- Useful When No Direct Reference Present

- Further Down Pipeline
  - Weighting Certain Assemblies

Contamination Evaluation: Tools

CheckM

Pros

- Industry Standard
- Robust and speedy (~10 mins/bacterial genome)

Cons

- Computationally Expensive (16GB RAM minimum)
- Lot of Dependencies

Anvi’o

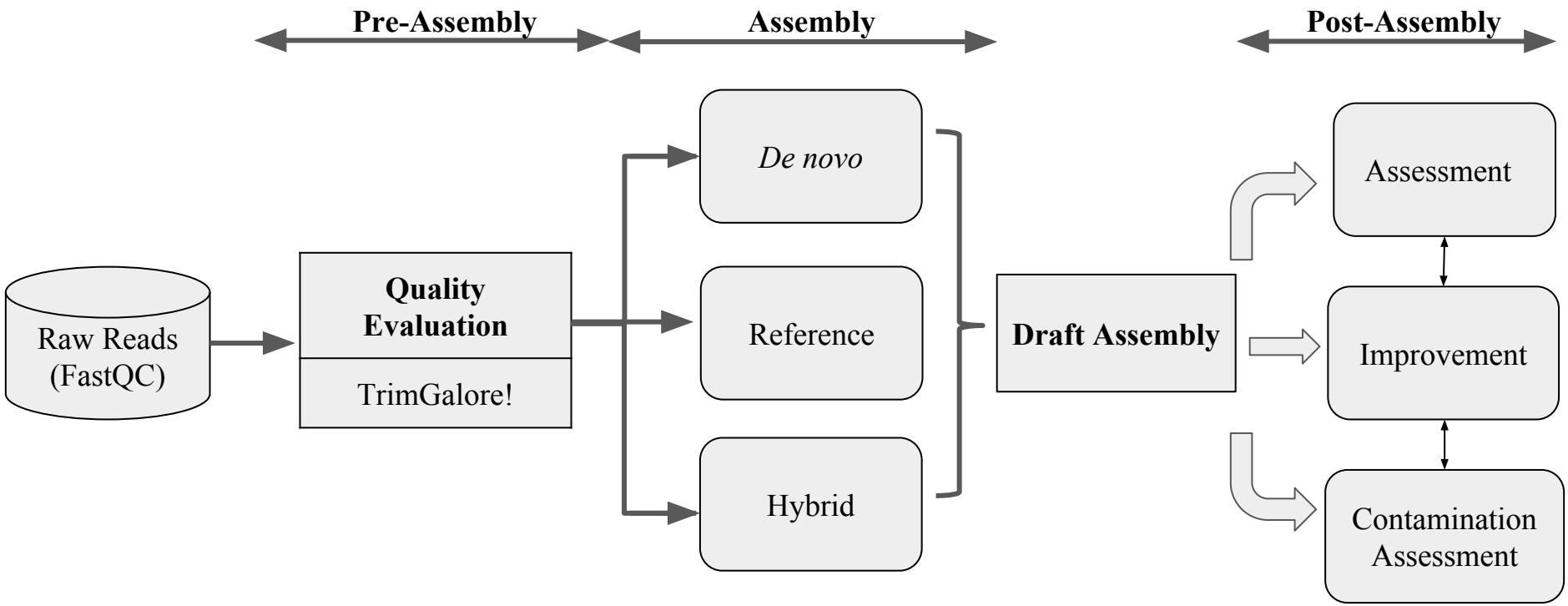
Pros

- Less computationally intensive
- Visualization Module
- Easy Install

Cons

- Not completely command line centric

Conclusion


Pre-Assembly
Assembly
Post-Assembly
Raw Reads (FastQC)
Quality Evaluation
TrimGalore!
De novo
Reference
Hybrid
Draft Assembly
Assessment
Improvement
Contamination Assessment

Questions?