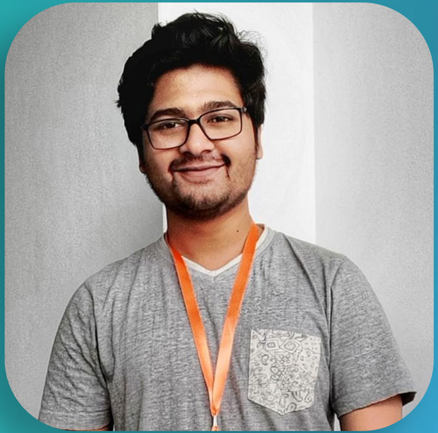


Atiq Israk Niloy
Researcher
Metagenomic Assembly
Southeast University
atiqisrak@gmail.com

Metavelvet Overview

A Presentation on the Metagenomic Assembly Assembler

Abstract and Introduction

1. A limitation of a single-genome assembler for de novo metagenome assembly is that sequences of highly abundant species are likely misidentified as repeats in a single genome, resulting in a number of small fragmented scaffolds.

2. The de Bruijn graph-based assembly program identifies the overlaps between reads using a de Bruijn graph and merges the reads to reconstruct longer sequences.

3. Our fundamental strategy for metagenome assembly was to consider that a de Bruijn graph constructed from mixed sequence reads of multiple species is equivalent to the mixture of multiple de Bruijn subgraphs, each of which is constructed from sequence reads of individual species and to decompose the mixed de Bruijn graph into individual subgraphs and build scaffolds based on each decomposed subgraph

4. We made use of two features, the coverage (abundance) difference and graph connectivity, for the decomposition of the de Bruijn graph.

5. For simulated datasets, MetaVelvet succeeded in generating significantly higher N 50 scores than any single-genome assemblers.

Abstract and Introduction

1. If use Single genome assembler > de novo > misidentified species > many small fragmented scaffolds > limitation
2. Overlaps between reads > De Bruijn > Reconstruct longer sequences.
3. Mixed species > a de Bruijn graph = Multiple de Bruijn sub graphs
4. Each subgraph = sequence reads of individual species
5. Decompose mixed de Bruijn > Individual Sub graphs > Build scaffolds
6. To decompose > Used Coverage difference & Graph Connectivity
7. Generates N 50 scores => any single-genome assemblers

Materials & Methods

Need to Discuss

1. In Velvet, the de Bruijn graph is implemented slightly differently, such that each node represents a series of overlapping k-mers where adjacent k-mers overlap by k 1 nucleotides.

2. The ordered set is cut whenever an overlap with another read begins or ends. a node is created. Two nodes can be connected by a directed edge. If two nodes are connected, the last k-mer of an origin node overlaps by k 1 nucleotides with the first of its destination node. New directed edges are created by tracing the read through the constructed graph.

3. Second, Velvet executes three functions, ‘simplification’ for node merging, and ‘removing tips’ and ‘removing bubbles’ for error removal.

Process

DNA Sequence input >
Merge Overlaps > Repeat

Read

Sequence fragment

Contig

Loger Sequence

Tip

Chain of nodes > One end disconnected

Bubble

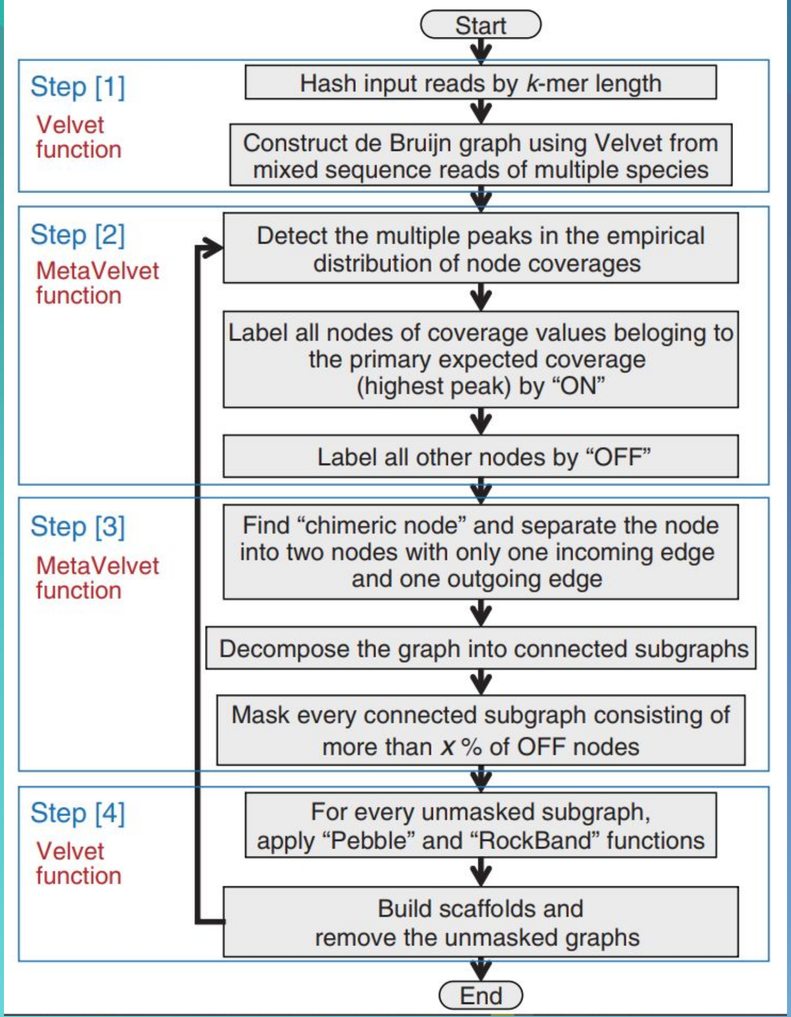
2 redundant paths > start+end point > Contain similar sequences

Tips & Bubble

Tips and bubbles are created by sequencing errors or biological variants, such as single nucleotide polymorphisms (SNPs).

Pebble & Rock Band

Called for constructing the scaffold and for repeat resolution using paired-end information and long read information



Start

Step [1]

Velvet function

Hash input reads by *k*-mer length

Construct de Bruijn graph using Velvet from mixed sequence reads of multiple species

Step [2]

MetaVelvet function

Detect the multiple peaks in the empirical distribution of node coverages

Label all nodes of coverage values beloging to the primary expected coverage (highest peak) by "ON"

Label all other nodes by "OFF"

Step [3]

MetaVelvet function

Find "chimeric node" and separate the node into two nodes with only one incoming edge and one outgoing edge

Decompose the graph into connected subgraphs

Mask every connected subgraph consisting of more than *x* % of OFF nodes

Step [4]

Velvet function

For every unmasked subgraph, apply "Pebble" and "RockBand" functions

Build scaffolds and remove the unmasked graphs

End

Process de Velvet

01

Cutting

The ordered set is cut whenever an overlap with another read begins or ends.

02

Uninterrupted

Create node of the original uninterrupted subset.

03

Connect Nodes

Connects 2 nodes by joining K of the first node with K-1 of the previous one.

3 Functions

04

Simplification

Join 1 ending and 1 beginning edge

05

Tip removal

Chain of nodes disconnected on one end is removed.

06

Bubble

Loops are merged

Metavelvet

01

Construction of a de Bruijn graph from the input reads.

02

Detection of multiple peaks on k-mer frequency distribution.

03

Decomposition of the constructed de Bruijn graph into individual subgraphs.

04

Assembly of contigs and scaffolds based on the decomposed subgraphs.

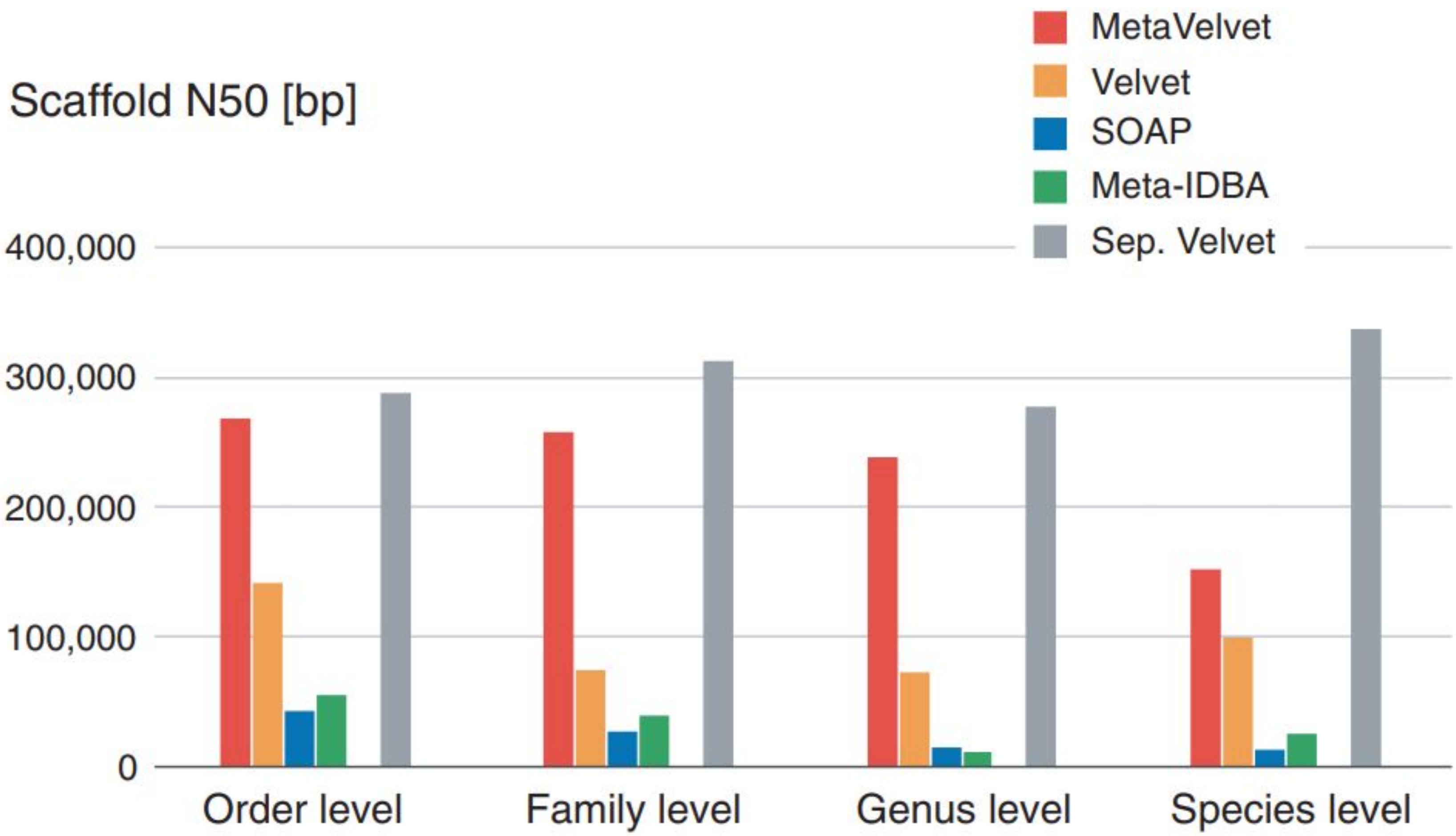
Chimeric Nodes

Shared Nodes

Compared with

1. Velvet
2. SOAP
3. Meta-IDBA

Meta Velvet Performed the best


Scaffold N50 [bp]
🟥 MetaVelvet
🟧 Velvet
🟦 SOAP
🟩 Meta-IDBA
⬜ Sep. Velvet
400,000
300,000
200,000
100,000
0
Order level
Family level
Genus level
Species level

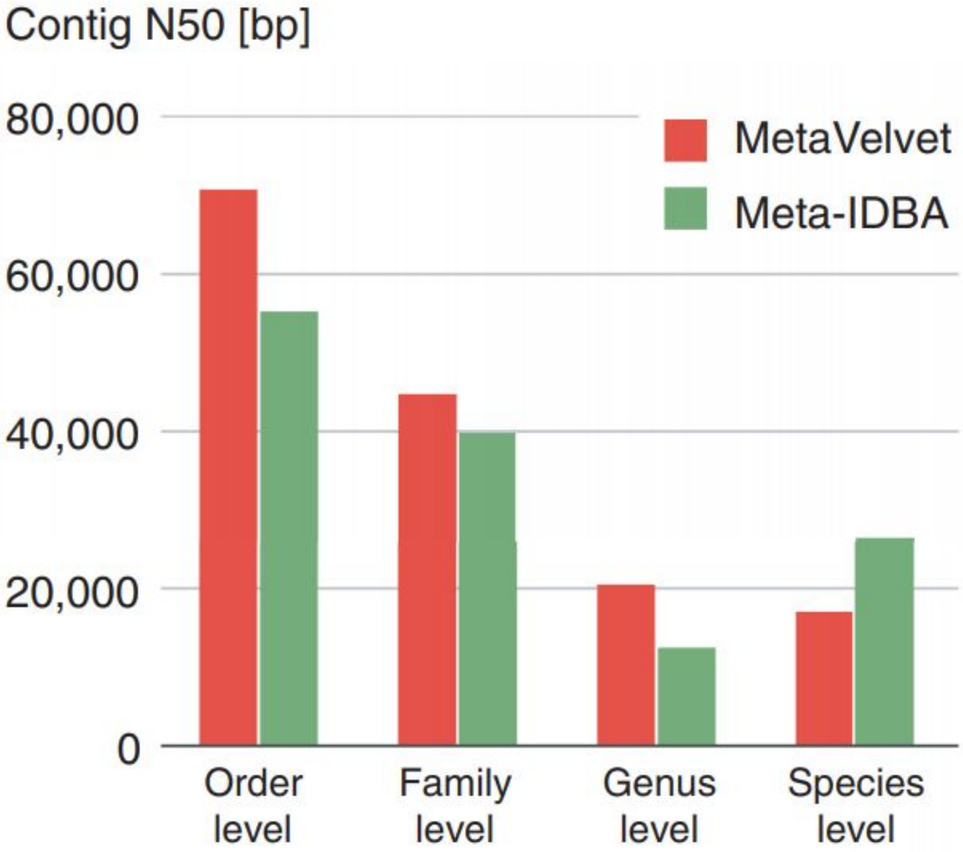
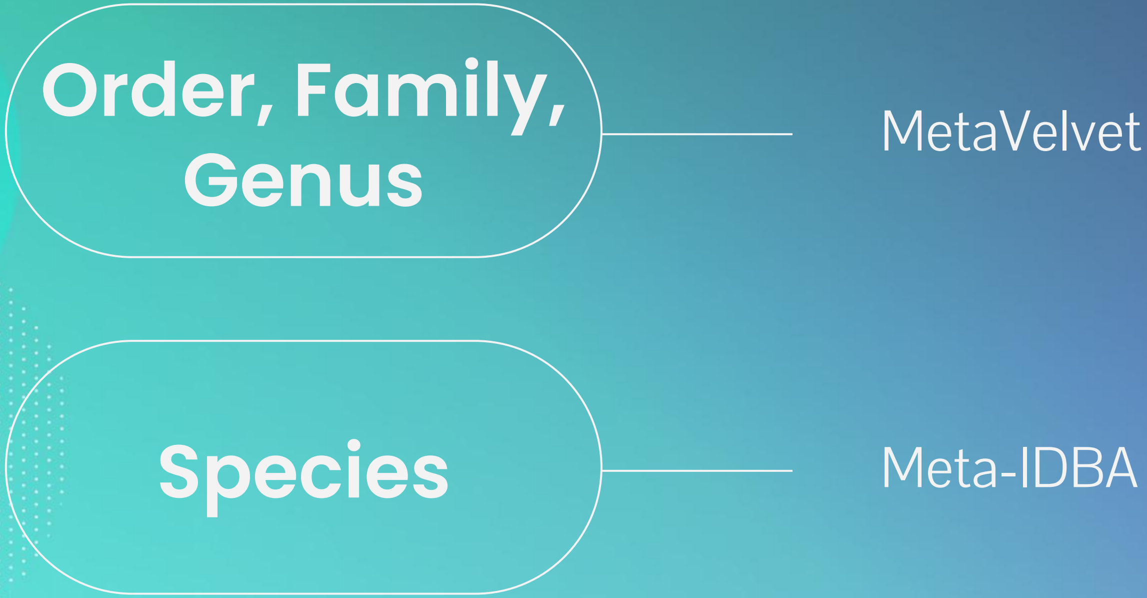


Figure 13. N50 scores for contigs generated by MetaVelvet and Meta-IDBA.

Best Performer



Order, Family, Genus — MetaVelvet

Species — Meta-IDBA

Meta-IDBA is designed to solve the metagenome assembly problem caused by polymorphisms in similar species in metagenomic environments.

In this aspect, Meta-IDBA might be more useful for analyzing slight variants in the genomes of subspecies within a same species.

The End
Thank You