Last Metagenomic Assembler Standing

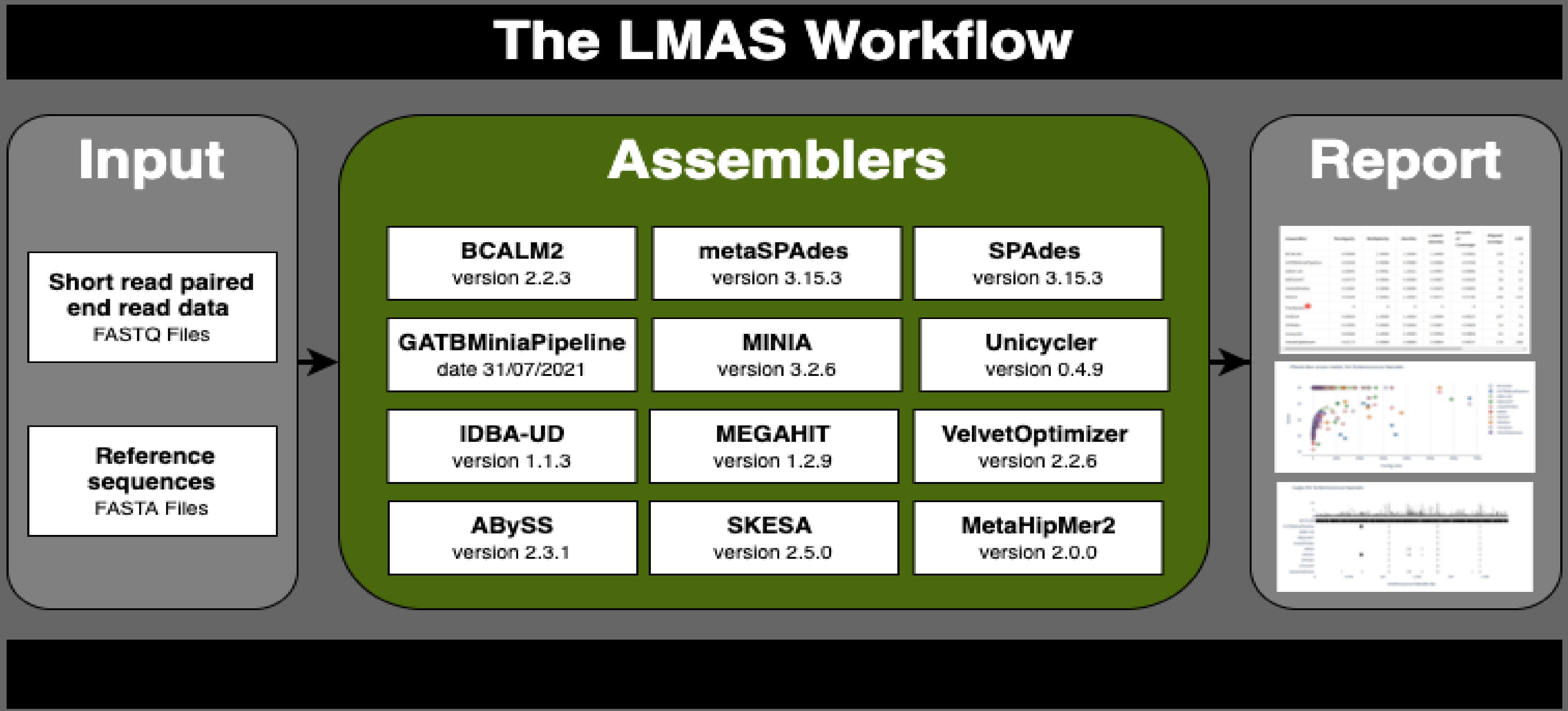
C I Mendes[1], P Vila-Cerqueira[1], Y Motro[2], J Moran-Gilad[2], J A Carriço[1], M Ramirez[1]

[1] Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal
[2] Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Email: cimendes@medicina.ulisboa.pt
Twitter: @ines_cim

ICCMg6

The LMAS Workflow



Input

Short read paired end read data
FASTQ Files

Reference sequences
FASTA Files

Assemblers

BCALM2 version 2.2.3	metaSPAdes version 3.15.3	SPAdes version 3.15.3
GATBMiniaPipeline date 31/07/2021	MINIA version 3.2.6	Unicycler version 0.4.9
IDBA-UD version 1.1.3	MEGAHIT version 1.2.9	VelvetOptimizer version 2.2.6
ABySS version 2.3.1	SKESA version 2.5.0	MetaHipMer2 version 2.0.0

Report

Abstract

The *de novo* assembly of raw sequence data is a **key process** when analysing data from shotgun metagenomic sequencing. It also represents one of the greatest bottlenecks when obtaining trustworthy, reproducible results.

LMAS is an automated workflow enabling the benchmarking of traditional and metagenomic prokaryotic *de novo* assembly software using **defined mock communities**. Several steps were implemented to ensure the transparency and reproducibility of the results. The mock communities can be provided by the user to better reflect the samples of interest. New assemblers can be added with minimal changes to the pipeline, so that LMAS can be expanded as novel algorithms are developed.

Usage

LMAS accepts as input **raw short-read paired-end sequencing data** and **reference sequences in a single file**. The resulting assembled sequences are processed and the **global and per reference quality assessment is performed**. The results are presented in an **interactive HTML report** where **selected global and reference specific performance metrics can be explored.**

Implementation

LMAS is implemented in Nextflow using Docker containers to provide flexibility. The use of **Docker containers** for each assembler allows versions to be tracked, and the use of **Nextflow**, a workflow management software, allows the effortless deployment of LMAS in any UNIX-based system, **from local machines to high-performance computing clusters** with a container engine installation.

ZymoBIOMICS Community Standards

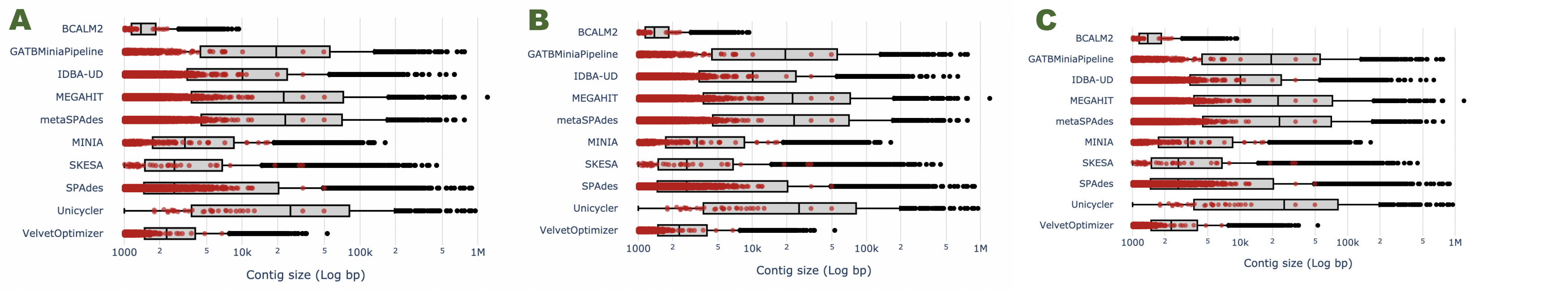
The **eight bacterial genomes and four plasmids of the ZymoBIOMICS Microbial Community Standards** were used as reference, and raw sequence data of the mock communities, with an even and logarithmic distribution of species, and a simulated sample of the evenly distributed reads generated from the reference genomes were used as **input for LMAS**.

Our results show that the choice of a *de novo* assembler depends greatly on the computational resources available and the species of interest, with the **performance of each assembler varying greatly with the abundance in the sample.** Overall, **multiple k-mer De Bruijn graph assemblers** outperform the alternatives but **no major performance gains were obtained when using dedicated metagenomic assemblers.** No single assembler emerged as an undisputed ideal choice.

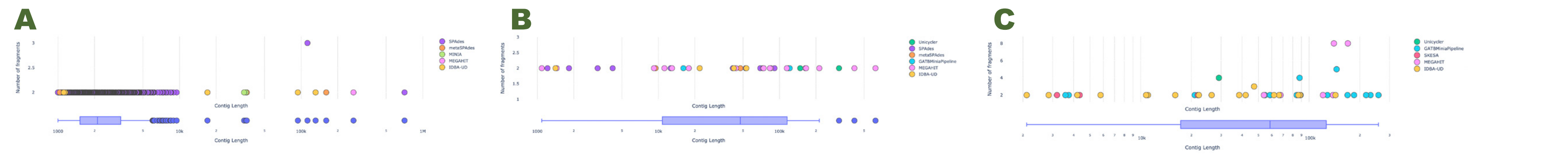
LMAS is open-source and freely available at https://github.com/cimendes/LMAS.

A demo LMAS report is available at https://lmas-demo.herokuapp.com.

Contig size distribution for log (A), even (B) and mock (C) ZymoBIOMICS samples



Misassemblies for log (A), even (B) and mock (C) ZymoBIOMICS samples



PLS Metrics for log (A), even (B) and mock (C) ZymoBIOMICS samples