MINTIA : Metagenomic INserT BIoinformatic Annotation in Galaxy

Sandrine Laguerre, Sarah Maman, Sabrina Legoueix, Elisabeth Laville, Gabrielle Veronese, Christophe C. Klopp

MINTIA

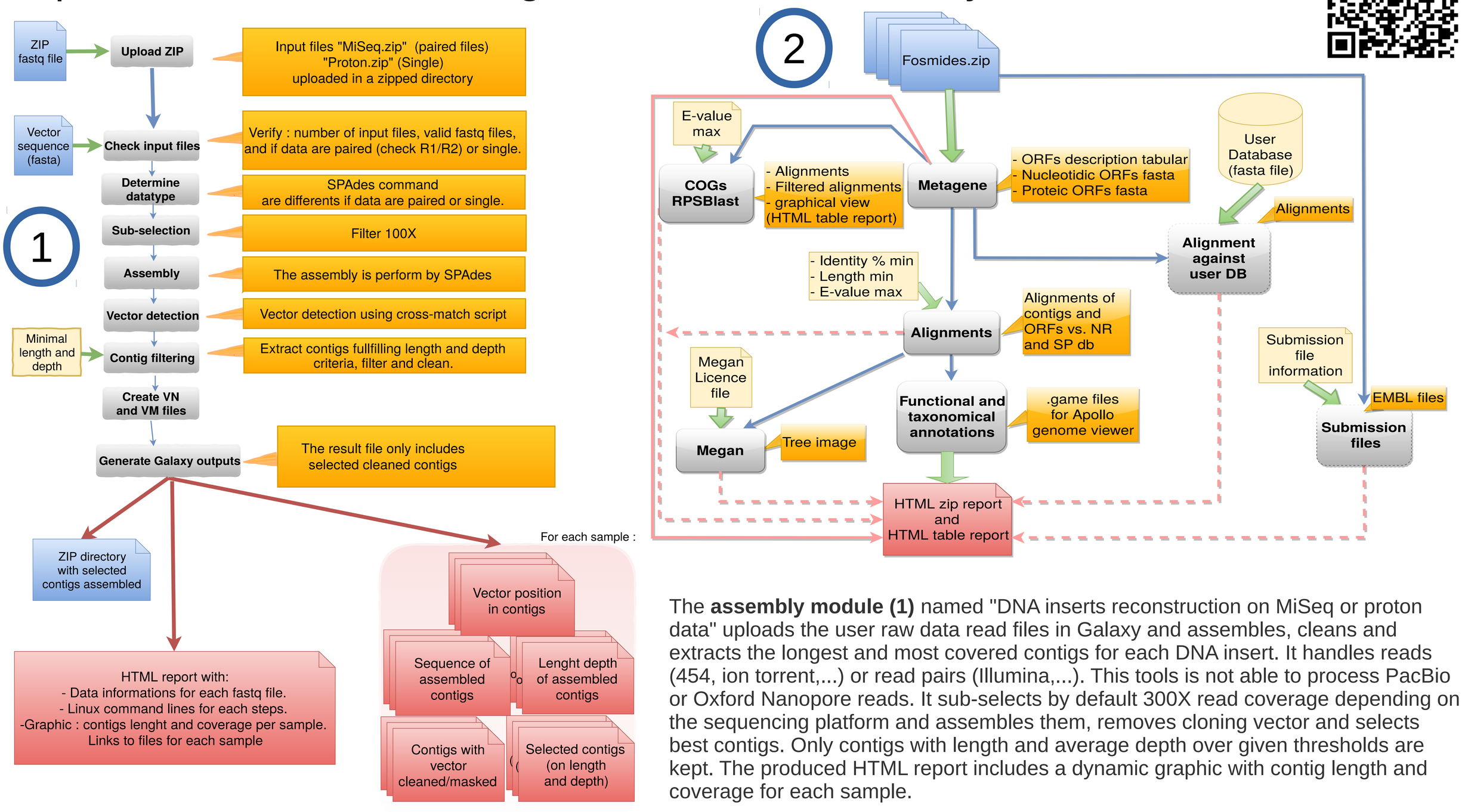
Metagenomic INsertT BIoinformatic Annotation in Galaxy

Sandrine Laguerre[1,2,3], Sarah Maman[4], Sabrina Legoueix[5], Élisabeth Laville[1,2,3], Gabrielle Potocki-Véronèse[1,2,3] & Christophe Klopp[4,6]

[1] Université de Toulouse; INSA, UPS, INP, 135 Avenue de Rangueil, F-31077, Toulouse, France, [2] INRA, UMR792 Ingénierie des Systèmes Biologiques et des Procédés, LISBP, F-31400, Toulouse, France, [3] CNRS, UMR5504, F-31400, Toulouse, France, [4] SIGENAE, GenPhySE, Université de Toulouse, INRA, INPT, ENV, Castanet Tolosan, France, [5] TWB, Université de Toulouse, INRA, INSA, CNRS, Ramonville Saint-Agne, France, [6] Plate-forme bio-informatique Genotoul, Mathématiques et Informatique Appliquées de Toulouse, INRA, Castanet Tolosan, France

Functional metagenomics is used to understand who is doing what in microbial ecosystems. DNA sequencing can be prioritized by activity-based screening of libraries obtained by cloning and expressing metagenomic DNA fragments in an heterologous host. When large insert libraries are used, allowing a direct access to the functions encoded by entire metagenomic loci sizing several dozens of kbp, NGS is required to identify the genes that are responsible for the screened function. MINTIA is an easy to use pipeline assembling and annotating metagenomic inserts.

Pipeline modules : 1/ metagenomic insert assembly & 2/ annotation.

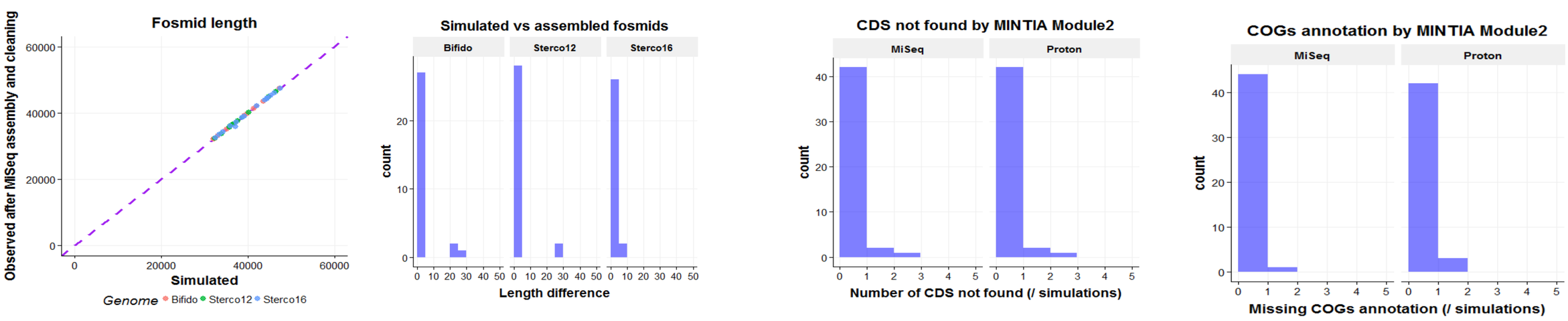


The **assembly module (1)** named "DNA inserts reconstruction on MiSeq or proton data" uploads the user raw data read files in Galaxy and assembles, cleans and extracts the longest and most covered contigs for each DNA insert. It handles reads (454, ion torrent,...) or read pairs (Illumina,...). This tools is not able to process PacBio or Oxford Nanopore reads. It sub-selects by default 300X read coverage depending on the sequencing platform and assembles them, removes cloning vector and selects best contigs. Only contigs with length and average depth over given thresholds are kept. The produced HTML report includes a dynamic graphic with contig length and coverage for each sample.

The **annotation module (2)** aims at obtaining main gene functions and a functional classification. The pipeline launches at least metagene ORF detection, generating fasta files for genes and proteins as well as a tabular file containing ORFs description.
Depending on additional options selected, contigs and ORFs are aligned against NCBI NR (Non Redundant) as well as SP (SwissProt) and COGss databases. These blast or RPSBlast results used for functional and taxonomical annotations produce .game formatted files up-loadable in Apollo for visualization and Megan functional classification tree to find the corresponding organisms. This module can also generate a GeneBank submission file.

First results on simulated datasets

15 sets of reads were simulated for paired end MiSeq and single end Proton from Bifidobacterium adolescentis ATCC15703 whole genome and Bacteroides stercoris ATCC43183 scaffold_02_12 and scaffold_02_16. Those sets were analyzed using MINTIA module 1 and 2.



Length of simulated and assembled fosmid sequences are very close. For all simulations, the assembled and cleaned sequence covers more than 99% of the simulated sequence with a percentage of similarity of 100% showing the high reliability of MINTIA module 1. The graph hereover on the left side compares the length between simulated and assembled sequence. The one on the right shows the length of the non corresponding part in base pairs.

Most of the times MINTIA module2 finds all the CDS of the simulated sequences. The graph on the left shows the number of missing CDS in the assembled sequences.
Close to all the CDS annotated with COGs on the simulated sequences are also annotated with MINTIA module 2. The graph on the right shows the CDS having different annotation between simulated and assembled sequences.

You can contact sandrine.laguerre@insa-toulouse.fr to get access to the modules in their actual state. They will be deposited on the Toolshed later this year.