A reliable pipeline for a transcriptome reference in non-model species

Hicham Benzekri[1], Rocío Bautista[1], Darío Guerrero-Fernández[1], Noé Fernández-Pozo[2], M. Gonzalo Claros[1] ✉

[1]University of Málaga, Spain
[2]The Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, United States

Motivation and objectives

Next-generation sequencing (NGS) platforms can sequence a particular transcriptome in a fast and cost-effective way. However, most bioinformatics tools are focused in model species where a reference sequence is available. But *de novo* transcriptome assemblies occur commonly when working with non-model species.

Methods

Here it is presented a pipeline for obtaining a reliable transcriptome in a plant non-model species, such as pine and sole, using NGS reads. It should be noted that the non-model species selected do not have a homogeneous genome, since individuals sampled were highly heterozygotic from natural populations. That means that single-nucleotide variations in reads can be due to SNPs or sequencing errors, and there is no way to discern both possibilities. The pipeline is outlined in figure 1.

The pipeline starts with the pre-processing software SeqTrimNext, developed by the authors, that extracts the reliable reads and removes from low-quality ends to contaminant fragments. It can work both on short and long reads. The assembling strategy is based on well-know, dedicated software for transcriptomics. Several assemblers were tested (SOAP, Trinity, ABySS, CABOG, Newbler, etc.) and the ones that better behaved were Oases, MIRA3 and EULER-SR. CD-HIT has used for selection of longest contigs derived from short reads. Since there is no reference to compare the reliability of assemblies, several confirmations were included: (1) original reads were mapped on contigs using Bowtie2 in order to discard artefacts; (2) FullLengtherNext (developed by the authors) was used to discard contigs that do not seem to be coding. Finally, all (long) contigs were reconciled using CAP3 (Minimus can also be used) to obtain the final candidate unigenes. FullLengtherNext can also be used to

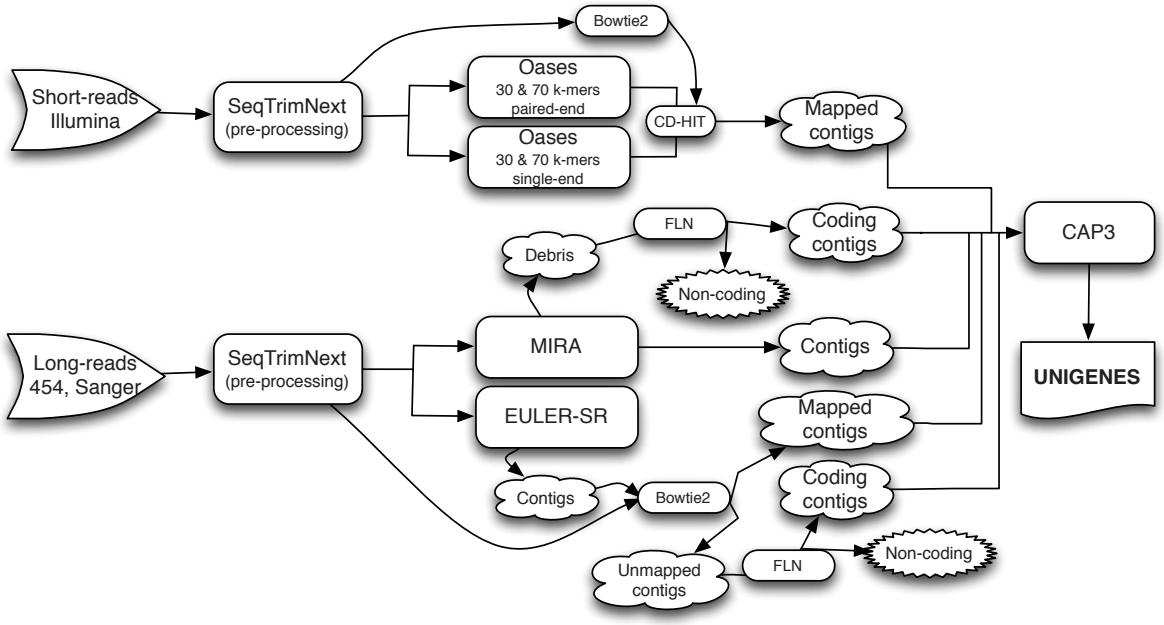

Figure 1. Pipeline for assembling non-model transcriptomes using well-known software as well as new algorithms developed by the authors. It can also handle both long-read and short-reads, including paired-ends.

test which assemblies are the better ones when several strategies are conducted. Unigenes were finally annotated using Sma3Annot (developed in collaboration with O. Trelles) and AutoFact.

Results and discussion

Reference transcriptomes obtained with this approach have been used for printing microarrays (whose hybridisation provided significant and useful results), and perform RNA-Seq analyses that were confirmed by RT-PCR, suggesting that the pipeline is adequate for transcriptome assembling of non-model organisms.