Combined genome guided and long reads assembly of the *Coffea arabica* transcriptome

Rezende P. M., Ribeiro T. H. C., Fernandes-Brum C. N., Schumacher P. V., Ferrara-Barbosa B. C., Chalfun-Junior A.

Laboratory of Plant Molecular Physiology, Federal University of Lavras

Coffee represents a great source of income for several countries and Brazil poses itself as the world's biggest producer and exporter of this commodity. The genus *Coffea* has more than 124 species, however, only two are economically relevant: *Coffea arabica* and *Coffea canephora*. *Coffea arabica*, the only tetraploid species, originated from a crossing event between *Coffea canephora* and *Coffea eugenioides*. The elevation of global temperatures caused by climate change is a threat to coffee production and is estimated that it may cause losses of up to $2,9 bi by 2020. Given this scenario, it becomes necessary to take measures to assure the global production of coffee. The study of the coffee genome and transcriptome can provide the basis for the improvement of coffee production and quality, through the development of new cultivars using techniques such as genetic engineering. However, only the genome of *C. canephora* has been sequenced so far. Thus, in this study, we used paired-end RNAseq libraries of two *C. arabica* cultivars ("Acauã" and "Catuaí Vermelho"), grown under to two different temperature ranges (19/23 °C, 26/30 °C), along with the *C. canephora* genome and EST sequences, from the CAFEST database, to reassemble the *C. arabica* transcriptome. The RNAseq libraries were aligned to the *C. canephora* genome, using the aligner STAR v2.4.2, and approximately 85% of coverage with this genome was obtained. The transcriptome assembly was performed with Trinity v2.2.0 tool, using as a reference genome the libraries assembly, and to enrich the assembly we also added as long reads ESTs from a public database of coffee ESTs (CAFEST). As results, 108940 putative genes and 144480 putative transcripts were identified. The N50 length, the statistics that define assembly quality, was 1781bp for the transcripts and 1296bp for the biggest transcripts of each gene. Hence, from these results we demonstrate that the strategy used in this study for the transcriptome assembling was wide and robust, and it can be used as a genomic resource for future investigation on *C. arabica*.