The Polyploid Gene Assembler (PGA)

Leandro Costa do Nascimento[1,2], Gonçalo Amarante Guimarães Pereira[1], Marcelo Falsarella Carazzolle[1]

1- Laboratório de Genômica e Expressão (LGE) – Deparatamento de Genética, Evolução e Bioagentes – Instituto de Biologia – Universidade Estadual de Campinas, 2 - Laboratório Central de Tecnologias de Alto Desempenho (LaCTAD) – Universidade Estadual de Campinas

In the last years, hundreds of genomes were sequenced, including species with complex genomes like papaya, panda, orange, etc. The increase of the number of sequenced genomes is directly related with advances in sequencing technologies, which nowadays allows the generation of millions of reads with low costs. In the case of complex genomes, the methodologies for *de novo* assembly remain a bottleneck due to a variety of biological and computational problems. Plant genomes, in particular, have genome size larger than mammals, requiring more elaborated and expansive methodologies for sequencing and high-performance computing to perform the analysis. Moreover, about 80% of the plants have high levels of polyploidy and heterozygosity, which complicate the assembly process for generating drastic variation in the genome sequencing coverage (intronic and intergenic regions have lower coverage than exonic regions). Considering that all genome assembly, including Velvet, SOAPdenovo and Abyss, work with the concept of uniform coverage, i.e., the sequencing coverage remains uniform throughout all regions of the genome, varying only in repetitive regions, *de novo* assembly of plant genomes have resulted in highly fragmented assemblies (millions of contigs lower than 1,000 bp) complicating further analysis, such as gene prediction and annotation. In this context, we present the **P**olyploid **G**ene **A**ssembler, a new methodology for reference-assisted sequence assembly focused in genic regions (including UTRs, exons, introns and, in some cases, promoter sequences) using low DNA sequencing coverage (around 3-10x). The pipeline was developed in PERL scripts for running in Linux system that integrates various software for read mapping, de novo assembling and scaffolding. In order to solve the assembly problems related to sequencing coverage variations, a *de novo* transcriptome assembly was used because it allows coverage oscillations during De Bruijn graph exploration. Although, PGA was developed for gene assembly from plant genomes, it can be used to any organism that has a closely related species with sequenced genome. PGA has been successfully applied in two complex and very well studied plant genomes: soybean (*Glycine max*) and wheat (*Triticum aestivum*), identifying a total of 99 and 90% of the known genes, respectively. PGA was also used to generate a gene catalogue from *Saccharum officinarum,* an important plant for production of sugar and ethanol, composed by 29,828 transcripts (27,768 genes; mean size of 936 bp), being 27,124 (90.9%) with similarity against the NCBI protein database.