Topic: **Databases and Bioinformatics Tools**
PI: **DBT04**

TORNADO: AN AUTOMATED PIPELINE FOR DE NOVO HYBRID GENOME ASSEMBLY BASED ON FREE SOFTWARE PACKAGES FOR SANGER AND NEXT GENERATION SEQUENCING TECHNOLOGIES (NGS).

Herai R H[1,2], Costa G G L D[1], Júnior O R[1], Vidal R O[1,4], Nascimento L C[1], Parizzi L P[1], Pereira G G A[1,4], Carazzolle M F[1,3]

[1]*Genomics and Expression Laboratory (LGE), Genetics, Evolution and Bioagents Department, Biology Institute (IB), State University of Campinas (UNICAMP)*
[2]*Applied Bioinformatics Laboratory (LBA), Informatics and Agropecuary Subdivision (CNPTIA), Brazilian Agricultural Research Corporation (EMBRAPA)*
[3]*National Center for High Performance Computing (CENAPAD)*
[4]*National Laboratory of Biosciences (LNBio), Brazilian Association for Synchrotron Light Technology*

Next generation sequence technologies (NGS) made possible to sequence entirely genomes in a fast way and low cost, from unicellular to complex organisms, like plants and mammals. These sequences can be assembled (*i*) using a reference genome or by some *de novo* bioinformatics method, such as Velvet, SOAPDenovo, Edena, ABYSS, GS Assembler 454, Mira and ZORRO. They are mainly based on de Bruijin graphs or, in a few softwares, reads overlapping to form contigs and scaffolds. The involved filtering and assembly step are very sensitive for each type of tool, and can be a key factor to generate the best assembly results. This way, when a set of sequences from distinct technologies exists, from Sanger to NGS, it is necessary the use of distinct assembly strategies for each type of data.

Actually, at our knowledge, there is no automated hybrid strategies based on in-use of distinct assembly softwares that can be applied to assembly hybrid data generated by NGS or Sanger platforms.

This works presents TORNADO, and automated pipeline for hybrid genome assembly based on free software packages. TORNADO did not proposed new methods for genome assembly. It just uses the best described software strategies for each type of genomic data to perform the hybrid assembly. It was organized in two main modules that are configured by XML file. In the first module, input data are filtered for trimming and experimental artifacts clipping. In the second module, based on sequence type, TORNADO automatically performs the assembly task using Mira for 454 and Sanger reads, Velvet for Illumina/Solexa or Solid/Life Tech reads. Finally, each assembled data are merged in a single assembly using ZORRO. If there are paired-end (mate pairs data) reads, an additional step involves CloseGaps software, which closes the gaps between assembled scaffolds.

TORNADO was already applied to assembly hybrid genomic reads from *Moniliophthora perniciosa* fungi, Witche's broom causal in plant cacao. Results showed that our strategy can works like an useful method to automatically assembly hybrid genome data. TORNADO's was implemented using Java and PERL programming language technologies.