Development of a new structural variant detection software based on graph clustering algorithms from long reads

***Nicolás Gustavo Gaitán Gómez, *Jorge Duitama**

***Systems and Computing Engineering Department, Universidad de los Andes, Bogotá, Colombia**

**Presenter´s E-mail: ng.gaitan@uniandes.edu.co*

Structural variants (SV) are a type of genetic polymorphisms, which are usually defined by their length (>50 bp) and alter the structure of chromosomes. There are many types of SVs such as deletions, insertions, translocations, inversions, and copy number variants. Recently, the relevance of SV analysis has increased because SVs can be related to phenotypic variation (including human diseases), and to evolutionary events such as speciation. This has led to the development of computational methods to identify and genotype SVs from high throughput sequencing data. Given the limitations of short reads, new algorithms aim to increase the accuracy of SV detection using long reads.

In this work, we present an accurate and computationally efficient software to predict structural variation events from long-read sequencing data from an organism compared to a reference genome of the same species. This tool is implemented as a new functionality of the Next generation Sequencing Experience Platform (NGSEP). This facilitates the integration with other functionalities implemented in NGSEP for the analysis of genomics data. The software execution process can be divided into three main phases. First, collect the signatures or evidence of SVs from read alignments. Then, cluster the signatures as SV hypotheses and, finally, transform the clusters into genotyped SVs, based on the evidence that supports each hypothesis. For the clustering phase, a graph is created based on the length and genomic position distances between different signatures, and two clustering algorithms were implemented based on this graph. First, a maximal clique detection clustering procedure was programmed with a Bron Kerbosch algorithm implementation (MCC). The other approach was to implement the DBSCAN algorithm which allowed for a more specific delimitation of clusters. In the benchmark phase, both algorithms and SVIM were tested on PacBio HiFi data from an individual of *Oryza sativa Minghui* and then compared to the structural variant caller SVIMasm as reference, which detects SVs based on *de novo* genome assembly alignments, which was reported to be more accurate. For indel calls, the MCC algorithm achieved a

precision and recall of 60% and 86%, while the DBSCAN algorithm provided values of 58% and 89% respectively. Although the SVIM algorithm had better results in terms of recall with 91%, its precision is lower (49%) because it reported many low-quality variants. Additionally, the DBSCAN algorithm presented a better differentiation of close events. Further work is needed to perform better benchmarking, given that the reference call-set is not a gold standard.