Development of NGSEP as an open-source comprehensive solution for analysis of high throughput sequencing data

Juan Fernando de la Hoz, Juan David Lobaton, Claudia Perea, Daniel Felipe Cruz, Juan Camilo Quintero, Paulo Izquierdo, Bodo Raatz and Jorge Duitama

[1] International Center for Tropical Agriculture (CIAT), Cali, Colombia. Email: j.duitama@cgiar.org

Project Website: https://sourceforge.net/p/ngsep/wiki/Home/
Source Code: https://sourceforge.net/projects/ngsep/files/SourceCode/
License: GNU General Public License, version 3 (GPL-3.0) (https://opensource.org/licenses/GPL-3.0)

The development and availability of high throughput sequencing (HTS) technologies revolutionized the research on genomics allowing to obtain genome-wide data on entire populations of nearly every form of life. A key step to extract relevant information from HTS data was the development of open-source software tools to perform different bioinformatic analyses. However, although even small labs are now able to efficiently produce large amounts of HTS data, comprehensive analysis of these data integrating different solutions remains a challenging task. We initially developed NGSEP as an open-source package that tightly integrates novel java implementations of algorithms for discovery of single nucleotide variants (SNVs), indels, and copy number variants (CNVs), called from a rich interface implemented in an Eclipse Plugin as well as basic command line usage. We built several functions to facilitate users processing HTS reads and genotype calls, including a one step wizard for parallel automated processing of entire populations, genotype filters and statistics, imputation for inbred populations and format conversion for integration with tools for assessment of population structure, GWAS, genomic prediction, among others. Benchmark using first Whole Genome Sequencing (WGS) data from human, yeast and rice samples and later Genotype by Sequencing (GBS) data from cassava and bean populations showed that NGSEP provides similar or better accuracy for SNP detection and genotyping compared to other tools such as GATK, Samtools and Tassel. NGSEP is now a useful software package for different research groups, as demonstrated by download statistics and recent scientific publications.

Continuing our efforts, we implemented in NGSEP three new algorithms to find structural variants (SVs) from WGS data: CNV-seq, for read-depth comparison between two samples; EWT to detect small CNVs; and a novel implementation of the read-pair and split-read strategies for detection of large indels with breakpoint resolution. Experiments with simulated data show that these functions provide similar accuracies compared to other tools such as Delly or Pindel. We now re-branded NGSEP as Next Generation Sequencing Experience Platform because we expanded the interfaces of NGSEP improving its documentation for command line usage and its integration with the Galaxy environment. We also made NGSEP available within the CyVerse (former iPlant) platform and we integrated the variants detector within the DNAnexus platform for SV discovery in the 3000 rice genomes project. We expect that all these development efforts facilitate the use of NGSEP for a growing number of researchers in different fields of basic and applied genomics.