Analysis of hypervariable DNA sequences by NGS technologies: QuasiFlow

Díaz-Martínez L[a,b], Seoane P[c], Grande-Pérez A[a,b], Claros MG[c] and Viguera E[a*].

a, Departamento de Genética, Universidad de Málaga, 29071, Malaga, Spain. eviguera@uma.es
b, Instituto de Hortofruticultura Subtropical y Mediterránea "La Mayora" (IHSM-UMA-CSIC), 29071, Malaga, Spain
c, Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, 29071, Malaga, Spain.

The development of Next Generation Sequencing (NGS) technologies has allowed deep characterization of highly variable sequences such as viral or mitochondrial genomes. With respect to RNA and ssDNA viruses, their low replication fidelity generates viral populations consisting of complex mutant spectra termed viral quasispecies. Their study is of special interest as they can be considered a phenotypic reservoir[1]. Similarly, heteroplasmy of human mitochondrial genomes, in which different sequences are found within a single individual, might have important clinical consequences.

For the analysis of the mutant spectrum of such hypervariable sequences from NGS data, we have developed QuasiFlow, a workflow designed in AutoFlow[2] that uses Illumina reads. QuasiFlow provides information about DNA variability, such as SNPs, indels and recombination events (Figure 1). Furthermore, it allows haplotype reconstruction of viral quasispecies and characterization of its diversity through normalized Shannon index, nucleotide diversity and mutation networks. Quasiflow performs also a comparative study among samples, based on correlation, ANOVA and PCA analysis, in order to determine which parameters are affected by the experiment and how the samples behave according to their biological origin.

In this work, we have applied QuasiFlow to analyze the population structure of the begomovirus *Tomato yellow leaf curl virus* (TYLCV) infectious clone inoculated in *Arabidopsis thaliana* plants, using HiSeq or MiSeq reads. Their analysis allowed detection of minor quasispecies variants with a frequency of 10^{-4} to 10^{-5} and reconstructed the haplotypes present in the sample. In addition, QuasiFlow was used to discover variants and recombinants in mixed infections of tomato plants. These results show the fast generation of recombinant genomes in geminivirus mixed infections and demonstrate the potential of QuasiFlow for the analysis of mutant spectra using Illumina MiSeq sequencing data.

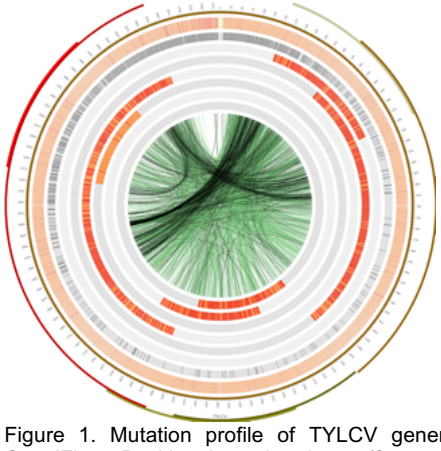

Figure 1. Mutation profile of TYLCV generated by QuasiFlow. Positional results shows (from outside to inside): (1) Mutation frequency/nt; (2) Mutations with highest mutation frequency; (3-8) Mutation frequency/aminoacid ORF1-ORF6 (9) Recombination events.

We have extended the use of QuasiFlow to the analysis of highly variable sequences such as the mitochondrial DNA. For that, we have analyzed DNA Illumina Miseq reads from 47 human mitochondrial samples from different cell lines obtained from the NCBI SRA database. Quasiflow generated automatically SNPs, SNP frequencies, indels and analyzed up to 23 variables using PCA analysis and performed an hierarchical clustering of the samples. Our analysis was able to detect pathological variants presented in a frequency lower than 1%.

This research was funded by Junta de Andalucía and EU through the ERDF 2014-2020, Projects P10-CVI-6075 to M. G.C. and P10-CVI-6561 to A.G-P.

References

1. Josep Gregori, Celia Perales, Francisco Rodriguez-Frias, Juan I. Esteban, Josep Quer, and Esteban Domingo. *Virology. Viral quasispecies complexity measures,* 2016, 493, 227-237.
2. Pedro Seoane, Sara Ocana, Rosario Carmona, Rocío Bautista, Eva Madrid, Ana M. Torres and M Gonzalo Claros. Current Bioinformatics. *AutoFlow, a Versatile Workflow Engine Illustrated by Assembling an Optimised de novo Transcriptome for a Non-Model Species, such as Faba Bean (Vicia faba)*, 2016, Volume 11 *in press*.