RESUMO - VIROLOGIA VEGETAL

SEQUENCE CURATION AS A CRITICAL STEP IN BIOINFORMATIC ANALYSES: INSIGHTS FROM AN UNEXPECTED RESULT INVESTIGATING THE VIROME IN PAPAYA

Yam De Sousa Santos (yamssantos@gmail.com)

Dennis Knierim (dek13@dsmz.de)

Eliza Bellard Do Nascimento (LIZABELLARD@GMAIL.COM)

Caroline Bezerra (bezerracarol@gmail.com)

Viviana Ramirez Rios (vivianaramirez0830@gmail.com)

Rosana Blawid (rosana.blawid@ufrpe.br)

Emanuel Felipe Medeiros Abreu (emanuel.abreu@embrapa.br)

Simone Ribeiro (simone.ribeiro@embrapa.br)

Paolo Margaria (paolo.margaria@dsmz.de)

High-throughput sequencing (HTS) combined with bioinformatic pipelines has become essential in modern virology. As part of a virome survey of papaya conducted in plants from Northeastern Brazil, total RNA extracts obtained from leaf samples were combined into four pools, each one composed of six samples

collected from different regions, and were subsequently subjected to Illumina HTS. Contigs were assembled using Geneious Prime software after filtering out host-derived reads. The analysis enabled the assembly of contigs assigned to papaya ringspot virus (PRSV, Potyvirus papayanuli; Potyviridae) in two pools. However, reconstruction of the full-length viral genome sequence was hindered by insufficient coverage in the coat protein (CP) gene region. A detailed evaluation of the assembly workflow and input datasets revealed that the papaya reference sequences (NCBI Bioproject: PRJNA264084) used for host-read subtraction contained a PRSV CP-derived sequence. This sequence was likely present in the dataset as part of a transgenic construct associated with CP-mediated resistance strategies in papaya. As a consequence, reads covering the CP region were inadvertently removed during host-derived sequences subtraction, preventing the complete reconstruction of PRSV genome sequence. Nevertheless, the full-length viral sequence was successfully recovered through iterative assembly when using the complete set of trimmed reads. These findings demonstrate that the composition of reference datasets can directly influence the outcome of bioinformatic analyses and the recovery of viral sequences from HTS data. Furthermore, this study highlights the importance of careful curation of databases for host-read subtraction to improve analytical accuracy and ensure reliable reconstruction of viral genomes sequences.

Palavras-chave: papaya virome; bioinformatic pipeline; viral genome reconstruction.