Characterization of the virome in mosquitoes using a small RNA-based approach

João Paulo Pereira de Almeida, Eric Roberto Guimarães Rocha Aguiar, Roenick Proveti Olmo, Yaovi Mathias Honore Todjro, Jean-Luc Imler, João Trindade Marques

Universidade Federal da Bahia

Abstract

High-throughput sequencing techniques and bioinformatics have guided to the elucidation of viral diversity (virome) in different organisms. Since all viruses produce RNA molecules at some point in their replication cycle, RNA-seq is an efficient and unbiased approach to assess viral diversity. In insects, RNA interference is a major antiviral response and generates small viral RNAs of specific sizes. Virus-derived small RNAs (vsRNAs) produced by the small interfering RNA (siRNA) pathway are phased duplex small RNAs of 20–23nt while products of the piwi-interacting RNA (piRNA) pathway are 24–30nt long and show specific patterns of base enrichment. Different from standard RNA-seq approaches that focus on long RNA fragments, sequencing of small RNAs allows the assembly of viral contigs, but also the inference of antiviral mechanisms and sequence-independent identification of viral contigs based on the pattern of vsRNAs. This strategy overcomes the limitation of sequence similarity searches of highly divergent viral sequences. In this work, we improved our previously developed method to identify viral sequences using small RNA libraries. Our pipeline is integrated into a Perl script requiring a fasta||fastq file as input. After pre-processing, reads are aligned against the host genome, bacterial and transposons reference sequences. Unmapped reads are kept and used to assemble contigs. Viral contigs are then identified by sequence similarity using Blastn or Blastx followed by ORF prediction and protein domain search. Filtered reads are aligned to the assembled contigs to generate a small RNA pattern that is used for hierarchical grouping using UPGMA and Pearson correlation coefficient. The script is run with a single command line in a terminal Linux. Parsed files with viral taxonomy information and plots with read coverage and small RNA profiles per contig are generated as output and are ready to be used in further analysis or publications. We applied our strategy to 56 unpublished small RNA libraries from wild mosquitoes (Aedes, Hemagogus, and Sabethes) collected around the globe. The summary of assembly metrics for viral contigs identified by sequence similarity in all libraries: an average of 43 contigs per library; average N50 of 746nt; contigs size average of 376nt; largest contigs average size of 1968nt. In total, we identified 17 viruses, 9 of which are possibly new viral species. The summary of assembly metrics for contigs without a similar correspondent in the databases: an average of 1866 contigs per library; average N50 of 66; contigs size average of 69nt; largest contigs average size of 465nt. Interestingly, in these last group of contigs, we identified 197 with only siRNA pattern, 62 with only piRNA pattern and 61 with both siRNA and piRNA patterns, all evidence to suggest the origin from unknown viruses of these contigs. Our improved approach allowed us to have an overview of viral diversity and provided information about antiviral defenses of these mosquitoes, knowledge that can be applied in the control and monitoring of virus circulation in natural conditions.

Funding: CAPES, CNPq, FAPEMIG, ZikAlliance, ANR, Labex NET-RNA