Sequence-independent metagenomic analysis of animal viromes based on molecular characteristics of small RNAs

Queiroz, L.R., Olmo, R.P., Marques, J.T. and Aguiar, E.R.G.R.

Department of Biochemistry and Immunology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Viruses are obligatory intracellular parasites that require the host machinery to replicate. Since viral RNAs normally present different molecular characteristics from those present in cellular RNAs, viral RNA intermediates can be recognized and degraded by different antiviral mechanisms that include RNA interference pathways, miRNA, siRNA and piRNA. These virus-derived small RNAs are imprinted with molecular characteristics that reflect their origin. Thus, since molecular features of small RNAs such as size, polarity and base preferences depend on the type of viral substrate and host mechanism of degradation, the pattern of small RNAs generated in infected cells can be used as a molecular footprint to identify and characterize viruses independent on sequence homology searches against known references. In this work, we analyzed 27 small RNA libraries from a broad range of organisms infected with viruses, including plants, arthropods and mammals, to determine how molecular characteristics of small RNAs could be applied to identify viral and non-viral sequences. We observed that production of small RNAs ranging from ~20 to 23 nt symmetrically from both strands, typically associated with the activation of siRNA pathway, was a conserved and specific response to viral infection in different organisms. Additionally, in arthropods, we noticed that other small RNA sources such as Endogenous Viral Elements (EVEs) and Transposable Elements (TEs), showed profiles more consistent with the activation of the piRNA pathway (e.g. small RNAs ranging from 24 to 30 nt, asymmetrical in polarity, and a strong 5′ U preference). These molecular characteristics of small RNAs allowed us to use Hierarchical Clustering based on Pearson correlation to classify sequences independently of homology searches against reference databases. Our results indicated that small RNA patterns are able to separate clusters of sequences containing viruses, EVEs and transposable elements. Using pattern-based analysis we successfully differentiate 10 viral sequences from 76 EVEs and 8311 TEs in six different organisms. This strategy can help overcome a great limitation of virus discovery by metagenomic strategies, since it does not require sequence similarity searches against known references. We are currently expanding our strategy to include the analysis of the ORF profile along assembled contigs and the use of di- and tri-nucleotide frequencies to identify and classify viral sequences.

Financial Support: CNPq, CAPES and FAPEMIG