Combining profile Hidden Markov Models and small RNA pattern based strategies to identify novel Endogenous Viral Elements (EVEs) and exogenous viruses

Eric R. G. R. Aguiar[1]*, Liliane S. Oliveira[2], João M. P. Alves[2], João Trindade Marques[1] and Arthur Gruber[2]*

[1]Department of Biochemistry and Immunology, Institute of Biological Sciences, UFMG, Belo Horizonte MG, Brazil; [2]Department of Parasitology, Institute of Biomedical Sciences, USP, São Paulo SP, Brazil.
**Correspondence: ericgdp@gmail.com and argruber@usp.br*

Endogenous Viral Elements (EVEs) are presumably derived from ancestral viruses that used to infect their hosts and had their sequences integrated into the host genomes. These elements show considerable sequence similarity to extant viral genomes. Thus, the accurate identification and characterization of novel EVEs are fundamental for the correct discrimination from exogenous viral sequences in metagenomic studies. In addition, the precise identification of EVEs enables a wide survey of ancestral viruses to which the host has been exposed (paleovirology), allowing a comparison to the viruses currently circulating in the host. Here we present a strategy to detect and discriminate EVEs and exogenous viral sequences using profile Hidden Markov Models (pHMMs) and an experimental validation using small RNA deep sequencing. We used virDB-Pack (see abstract by Oliveira *et al*. – X-Meeting 2016) to select a subset of 506 pHMMs from the vFam database according to virus-specific annotation terms. This set was used to screen a dataset composed of long RNAs sequenced from *Aedes aegypti*. Selected pHMMs showing the highest numbers of significant positive hits were employed as seeds for progressive assembly using GenSeed-HMM. The reconstructed sequences were submitted to similarity searches against the *nr* database, matching a wide variety of viral families. We have previously shown that EVEs present small RNA profiles with molecular characteristics distinct from exogenous viral sequences. Therefore, we mapped reads from small RNA libraries prepared from the same *A. aegypti* mosquitoes onto our candidate contigs to identify signatures that discriminate canonical EVEs from viruses. Notably, EVEs have small RNA profiles that show size in between 24-29 nt, U enrichment at the 1st nt of antisense reads, an enrichment at the 10th of sense reads, 10-nt overlap between 5' end of reads in opposite strands and asymmetrical small RNA density along sequence. Seventy-seven out of 381 contigs greater than 200 nt showed profiles consistent with EVEs, are not present on the current genome of *A. aegypti*, probably representing novel elements. We also found a 9-kb contig that showed a viral signature and represents a *Phasi Charoen Like-virus*, recently described by our group using small RNA-based strategy. This result confirms that both strategies, (1) pHMM screening followed by progressive assembly and (2) identification of viral signatures using small RNAs, can be jointly used to reliably identify and characterize new EVEs, even in the absence of a well-curated genome, and also to detect novel exogenous viruses. Support: CNPq and CAPES.