A survey of bacterial and archaeal genomes reveals novel casposon elements and hosts

Liliane Santana Oliveira Kashiwabara, Bas Dutilh, Arthur Gruber, Táriky Meirelles Rocha

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ, UNIVERSIDADE DE SÃO PAULO

Abstract

Casposons comprise a superfamily of self-synthesizing DNA transposons containing cas1, a gene for endonuclease Cas1, a key enzyme of the adaptive immunity CRISPR-Cas system. All casposons also harbor polb, a type B DNA polymerase gene. Some other genes are also found but are not shared by all elements. Casposons are typically flanked by terminal inverted repeats (TIRs) and target site duplications (TSDs) and these features can be used to predict their boundaries within the host's genome. Four casposon families have been characterized so far. Family 1 is composed of elements of archaeal hosts, presenting a protein-primed PolB, an enzyme closely related to polymerases of archaeal viruses. Elements of the other families were found in Archaea (Families 2 and 4) and Bacteria (Family 3). Phylogenetic analyses suggest that casposons originated CRISPR-Cas systems. To better understand the evolutive role of casposons in the emergence of adaptive immunity in prokaryotes, we decided to perform a survey on PATRIC, a public repository of assembled genomes. First, we used TABAJARA, a program developed by our group, to construct sets of profile HMMs derived from Cas1 and PolB sequences. All models were validated against bona fide datasets of Cas1 and PolB sequences derived from casposons or other sources. Profile HMMs specific to casposon elements were used together with e-Finder, a generic tool that use the models to detect and extract multigene elements from assembled genomes. All elements were annotated with EGene2 and the functional annotation was curated on Artemis. TIRs and TSDs were detected with UGENE program on regions flanking the predicted casposon elements. The survey revealed 136 elements, 90 in archaea and 46 in bacteria. From this set, 44 elements did not present the flanking repeats, with 39 of them showing truncated sequences in at least one end. In total, 92 full-length casposons were found with the expected flanking repeats. A phylogenetic analysis of the elements confirmed their monophyletic character in regard to CRISPRs. Based on the phylogenies of both Cas1 and PolB sequences, we found evidence that some casposons of Family 2 may in fact constitute a novel family. More detailed genome structure information and higher taxa sampling with be necessary to confirm this result. Another interesting finding was the presence of a full Type II CRISPR embedded within casposons of Hyphomonadaceae bacteria. Finally, new hosts were also identified, expanding the current knowledge on the occurrence of casposon elements.

Funding:

Link to Video: