PhageTerm: a tool for fast and accurate determination of phage termini and packaging mechanism using next-generation sequencing data

Garneau Julian[1], Depardieu Florence[2], Louis-Charles Fortier[3], David Bikard*[2], and Marc Monot†‡[1,3,4]

[1]Laboratoire Pathogenèse des Bactéries Anaérobies (LPBA) – Institut Pasteur de Paris – France
[2]Groupe de Biologie de Synthèse – Institut Pasteur de Paris – France
[3]Département de microbiologie et infectiologie, Université de Sherbrooke, Canada – Canada
[4]Université Paris Diderot, Sorbonne Paris Cité – Université Paris Diderot - Paris 7 – France

Résumé

The worrying rise of antibiotic resistance in pathogenic bacteria is leading to a renewed interest in bacteriophages as a treatment option. Novel sequencing technologies enable description of an increasing number of phage genomes, a critical piece of information to understand their life cycle, phage-host interactions, and evolution. In this work, we demonstrate how it is possible to recover more information from sequencing data than just the phage genome. We developed a theoretical and statistical framework to determine DNA termini and phage packaging mechanisms using NGS data. Our method relies on the detection of biases in the number of reads, which are observable at natural DNA termini compared with the rest of the phage genome. We implemented our method with the creation of the software PhageTerm and validated it using a set of phages with well-established packaging mechanisms representative of the termini diversity, i.e. 5'cos (Lambda), 3'cos (HK97), pac (P1), headful without a pac site (T4), DTR (T7) and host fragment (Mu). In addition, we determined the termini of nine Clostridium difficile phages and six phages whose sequences were retrieved from the Sequence Read Archive.

PhageTerm could also be used to detect whether a contigs reconstructed from virome data has a phage origin or not. An issue with contigs from virome data is to sort phage sequences from background (host sequence, assembly artefact...). For this purpose, current available methods (Phaster, MetaPhinder, Virsorter...) are all based on sequence homology. PhageTerm is the only one to our knowledge that is not based on homology as it detects termini using biases in the number of sequencing reads. Thus, we tested PhageTerm on contigs obtained with virome data and it was able to detect termini on severals contigs. PhageTerm found termini on contigs detected by other software but also on contigs that are never defined as phage sequence by other software.

PhageTerm is freely available (https://sourceforge.net/projects/phageterm), as a Galaxy ToolShed and on a Galaxy-based server (https://galaxy.pasteur.fr).

*Auteur correspondant: david.bikard@pasteur.fr

†Intervenant

‡Auteur correspondant: mmonot@pasteur.fr