Title: Origins and Challenges of Viral Dark Matter

Authors: Siddharth R. Krishnamurthy, David Wang

PII: S0168-1702(16)30801-2
DOI: http://dx.doi.org/doi:10.1016/j.virusres.2017.02.002
Reference: VIRUS 97072

To appear in: *Virus Research*

Received date: 5-12-2016
Revised date: 31-1-2017
Accepted date: 6-2-2017

Please cite this article as: Krishnamurthy, Siddharth R., Wang, David, Origins and Challenges of Viral Dark Matter.Virus Research http://dx.doi.org/10.1016/j.virusres.2017.02.002

Title: Origins and Challenges of Viral Dark Matter

Authors: Siddharth R Krishnamurthy[1], David Wang*[1]

Affiliations: [1]Departments of Molecular Microbiology and Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA

Corresponding author: David Wang (davewang@wustl.edu)

Highlights:

- Primary sequence alignment is the most common method to identify viral sequences
- Alignment methods are limited by low identity query sequences
- Low diversity in viral databases limit alignment based methods
- Experimental methods can specifically enrich viral sequences in sequencing datasets
- Computational methods use virus specific attributes to identify viral sequences

Abstract: The accurate classification of viral dark matter –metagenomic sequences that originate from viruses but do not align to any reference virus sequences– is one of the major obstacles in comprehensively defining the virome. Depending on the sample, viral dark matter can make up from anywhere between 40-90% of sequences. This review focuses on the specific nature of dark matter as it relates to viral sequences. We identify three factors that contribute to the existence of viral dark matter: the divergence and length of virus sequences, the limitations of alignment based classification, and limited representation of viruses in reference sequence databases. We then discuss current methods that have been developed to at least partially circumvent these limitations and thereby reduce the extent of viral dark matter.

Abbreviations[1]

---

[1]Abbreviations:

Contiguous sequence – contig

Double stranded RNA – dsRNA

Internal transcribed spacer – ITS

Naïve Bayes Classifier – NBC