A DATA-DRIVEN APPROACH TO EXPLAIN VIRAL ORIGINS AND EVOLUTION A. Nasir[1,2*], F.J. Sun[3], K.M. Kim[4], G. Caetano-Anollés[1], [1]Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, University of Illinois, Urbana, IL 61801, USA; [2]Department of Bioscience, COMSATS Institute of Information Technology, Islamabad 44000, Pakistan; [3]School of Science and Technology, Georgia Gwinnett College, Lawrenceville, GA 30043, USA; and [4]Microbial Resource Center, KRIBB, Daejeon 305-806, Korea. *E-mail: anasir@illinois.edu

Background: The origin and evolution of viruses remains difficult to explain. This in part owes to the great genetic and morphotype diversity that is present in viruses [1,2] and their atypical mode of survival. Viruses are holoparasites of cells of organisms in the three domains of life, Archaea, Bacteria, and Eukarya, that utilize cellular machinery for synthesis of viral proteins. Viruses also possess numerous forms of nucleic acid genetic material (DNA, RNA, and retrotranscribing), which is sometimes single or double stranded and even segmented. The unique virion morphotypes exhibited by many groups of viruses [1,2] adds another layer of complexity in explaining viral origins. Different viral groups are also very dissimilar at gene level. These observations suggest viruses likely evolved multiple times in evolution and perhaps via different mechanisms. However, a recent focus on molecular structure and gene function, features that are more conserved in evolution than the sequences of genes, has shed new light into viral evolution. The existence of a 'viral supergroup' is slowly gaining popularity [3]. Moreover, viruses infecting distant hosts, sometimes even in different domains of life, encode similar type of capsids and harbor homologous packaging signals. This suggests that viruses are very ancient and perhaps evolved even prior to the origin of modern cells. Recently the discovery of giant viruses, which resemble cellular parasites in lifestyle and genome and particle size, and their virophages, has further challenged the century old views regarding viral evolution (e.g. [4]).

Results: Here we mine thousands of completely sequenced viral and cellular proteomes with hidden Markov models of structural recognition, using established techniques [3]. Both, comparative and evolutionary genomic approaches, were applied to the global analysis of protein structural domains defined at the fold superfamily (FSF) level of the Structural Classification of Proteins. Our study uncovered remarkable evolutionary patterns. Viruses harboring different replicon types and infecting distantly related hosts shared many metabolic and informational FSFs that were also widespread in cellular proteomes and were of ancient origin (Fig. 1). This conserved structural core suggests an ancient origin for the viral supergroup. Importantly, a substantial number of viral proteins lacked cellular homologs, strongly negating the idea that viruses merely evolve by capturing cellular genes. Phylogenetic trees describing the evolution of protein structural domains and tRNA structures suggest that viruses evolved via reductive evolution from ancient cells (an idea previously put forward independently by Claverie and Bandea [5,6]). Reconstruction of a truly universal tree of life defined viruses as a distinct and ancient supergroup, comparable in significance to known cellular superkingdoms.

Conclusions: The study challenges the way we have historically perceived viruses and identifies them as crucial biological agents that impact the evolution of cellular organisms. Phylogenomics revealed that modern viruses originated from ancient cells that harbored segmented RNA genomes. These ancient cells reduced into modern viruses once diversified life took over the planet. The phylogenomic trees showed viruses belonged to a 'fourth supergroup' along with superkingdoms Archaea, Bacteria, and Eukarya. The new model for the origin and evolution of viruses and cells is backed by strong genomic and structural evidence and can be reconciled with existing models of viral evolution, if one considers viruses originated from ancient cells and not from modern cellular organisms.

References: [1] Nasir A. et al. (2014) *Front. Microbiol.*, *5*, 194. [2] Pietilä M.K. et al. (2014) *Trends Microbiol., 22*, 334-344. [3] Nasir A. et al. (2012) *BMC Evol. Biol., 12,* 146. [4] Philippe N. et al. (2013) *Science*, 341, 281-286. [5] Bandea C.I. (2009) *Nature Prec.*, http://hdl.handle.net/10101/ npre.2009.3886.1. [6] Claverie J.M. and Abergel C. (2013) *Adv. Virus Res, 85*, 25-56.

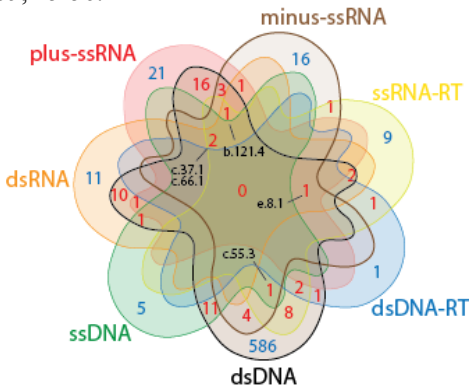

Fig. 1. The seven-set Venn diagram describes FSFs shared or uniquely present in the viral supergroup.