Science Advances

AAAS

advances.sciencemag.org/cgi/content/full/1/8/e1500527/DC1

Supplementary Materials for

A phylogenomic data-driven exploration of viral origins and evolution

Arshan Nasir and Gustavo Caetano-Anollés

Published 25 September 2015, *Sci. Adv.* **1**, e1500527 (2015)
DOI: 10.1126/sciadv.1500527

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at advances.sciencemag.org/cgi/content/full/1/8/e1500527/DC1)

Supplementary Materials

Text S1. Phylogenetic assumptions and models.

We focused on the abundance and occurrence of FSF domains in proteomes. It can be argued that abundance of some folds could be increased by non-vertical evolution such as HGT or decreased artificially due to incomplete or biased sampling or simply due to evolutionary bottlenecks (e.g. loss of an ancient fold from the ancestor of a superkingdom). However, the abundance-based approach is relatively more robust against non-vertical evolutionary forces, mainly HGT. The effect of HGT-related artificial increases in genomic abundance for ancient FSFs would be almost negligible (as those already have high abundance count in genomes). In turn, HGT gain of some of the recently evolved FSFs that are present in genomes with low count (e.g. 1-2 per genome) could be significant but would only affect the very derived branches of the ToD. We also note that the ancestry of FSFs in ToDs depends upon the 'profile' distribution of FSFs in proteomes. For example, some immunoglobulin superfamily domains are very abundant in some eukaryotes. Despite their very high abundance in some organisms, they are not the most ancient FSFs in our ToDs. This implies that both abundance and spread of FSFs determine the position of FSFs in timelines derived from phylogenetic trees. Still, comparing phylogenies obtained from occurrence and abundance counts of FSFs can experimentally validate polarization [e.g. *(132)*]. For example, distance-based phylogenies yield topologies that are congruent with ToPs (Figs. 7 and 8), and similar conclusions were strongly supported by comparative genomic experiments. Thus, the ancient history of the viral supergroup should be considered reliable unless strong evidence is presented to suggest otherwise.

In terms of character polarization, it could be argued that viruses with very small proteomes can be artificially attracted to basal branches of ToPs making the construction of a *u*ToL problematic. This interpretation however is erroneous since polarization also involves spread in the nested lineages of the *u*ToL and is only applied *a posteriori*, allowing gains and losses throughout branches of the tree *(132)*. We note that assumptions of character polarization comply with Weston's generality criterion of phylogenetic rooting *(133, 134)* and are consistent with the proposal of a simpler progenote organism (community) at the beginning of evolution. A number of theoretical arguments and experimental evidence support the assumption that ancient genes have more time to accumulate and spread in diversifying lineages. For example, the 'P-loop containing NTP hydrolases' FSF (c.37.1) includes ubiquitous and highly abundant proteins that are involved in membrane transport and metabolic processes. There is general agreement that these proteins evolved first in evolution. FSF c.37.1 was also the first to appear in our ToD and this result was consistently recovered in many previous phylogenomic reconstructions [e.g. *(135)*].

Finally, it is noteworthy that the new evoPCO methodology is free from typical artifacts that complicate ToL reconstruction, most importantly character independence *(16, 136)*. ToLs are generally built from nucleotide or amino acid site information in nucleic acid or protein sequences, which are generally not independent from each other because of the mere existence of molecular structure *(137)*. This violates the phylogenetic requirement of character independence, unless suitable representations of structure-based dependencies are incorporated into the evolutionary tree-building model. In contrast, FSF ages used in evoPCO were calculated from a ToD, a tree derived from domain abundance counts in proteomes, which are used as characters. Since proteomes generally evolve independently from each other (except for symbiotic or trophic interactions) and any possible interaction between them occurs at levels of organization that are much higher than molecular structure, ToDs (and temporal information they provide) are

therefore impervious to the need to budget molecular structural dependencies of characters in evolutionary models.

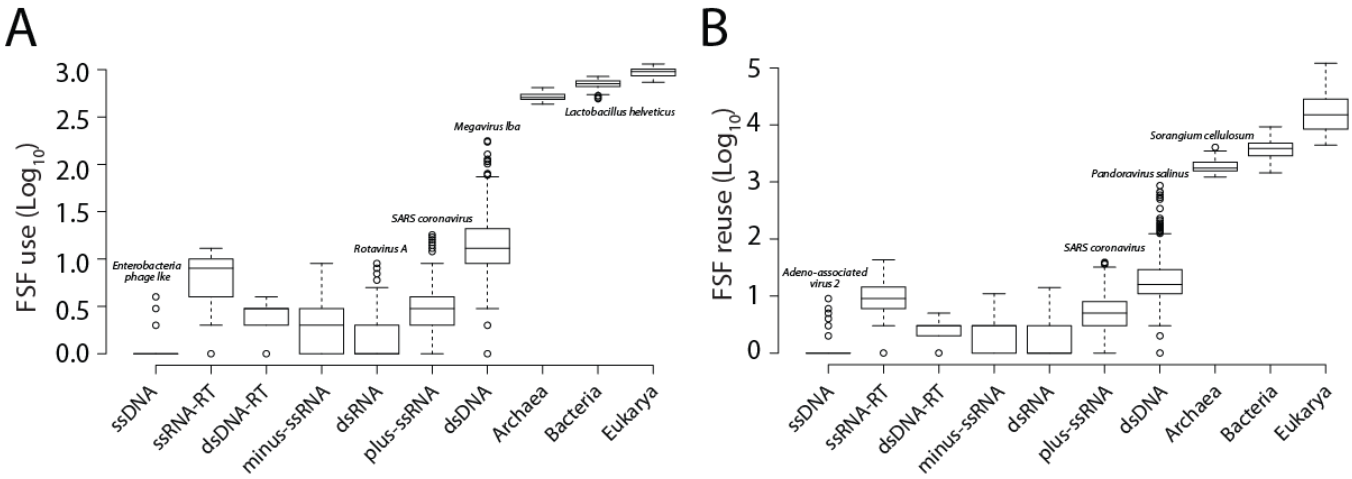

Fig. S1. FSF use and reuse for proteomes in each viral subgroup and for free-living cellular organisms. Both values are given in logarithmic scale.

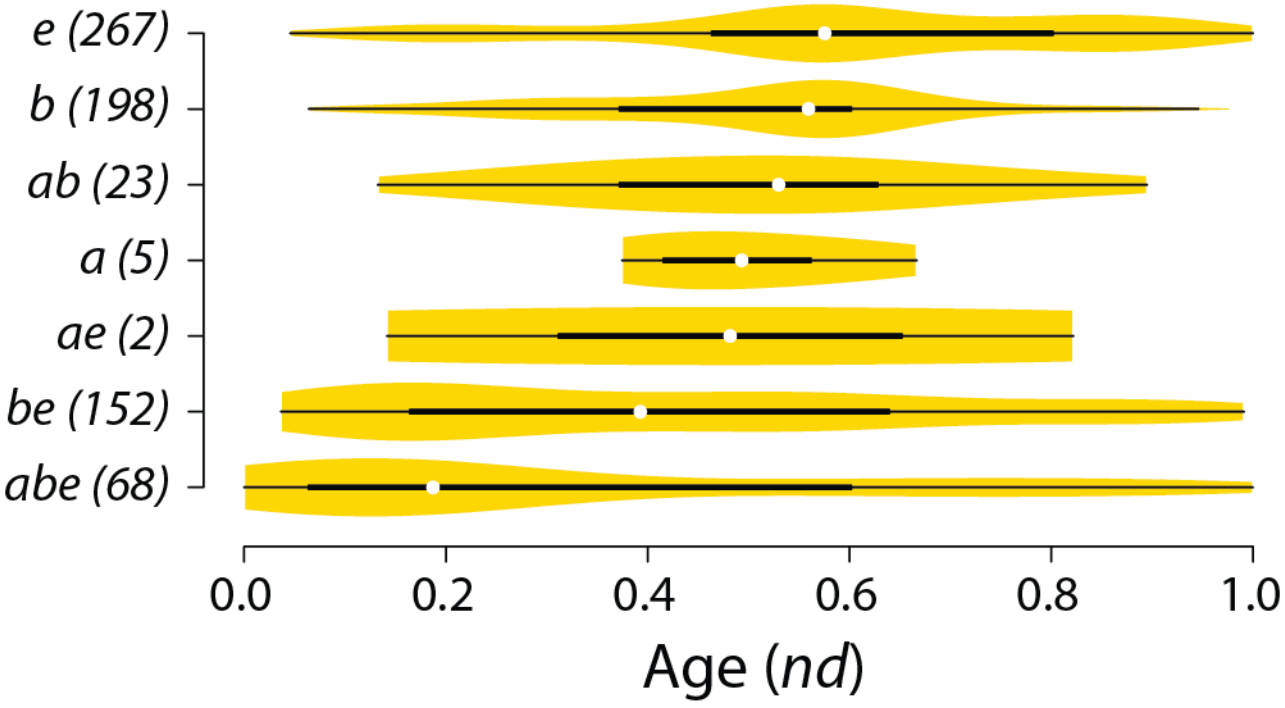

Fig. S2. Distribution of FSFs in each of the seven Venn groups defined in Fig. 3B along the evolutionary timeline (nd). The nd represents the node distance and was calculated from a phylogenetic tree of domains (ToD) describing the evolution of FSF domains (see text). It ranges from 0 (most ancient FSF) to 1 (most recent). Numbers in parenthesis indicate total number of FSFs in each Venn group.

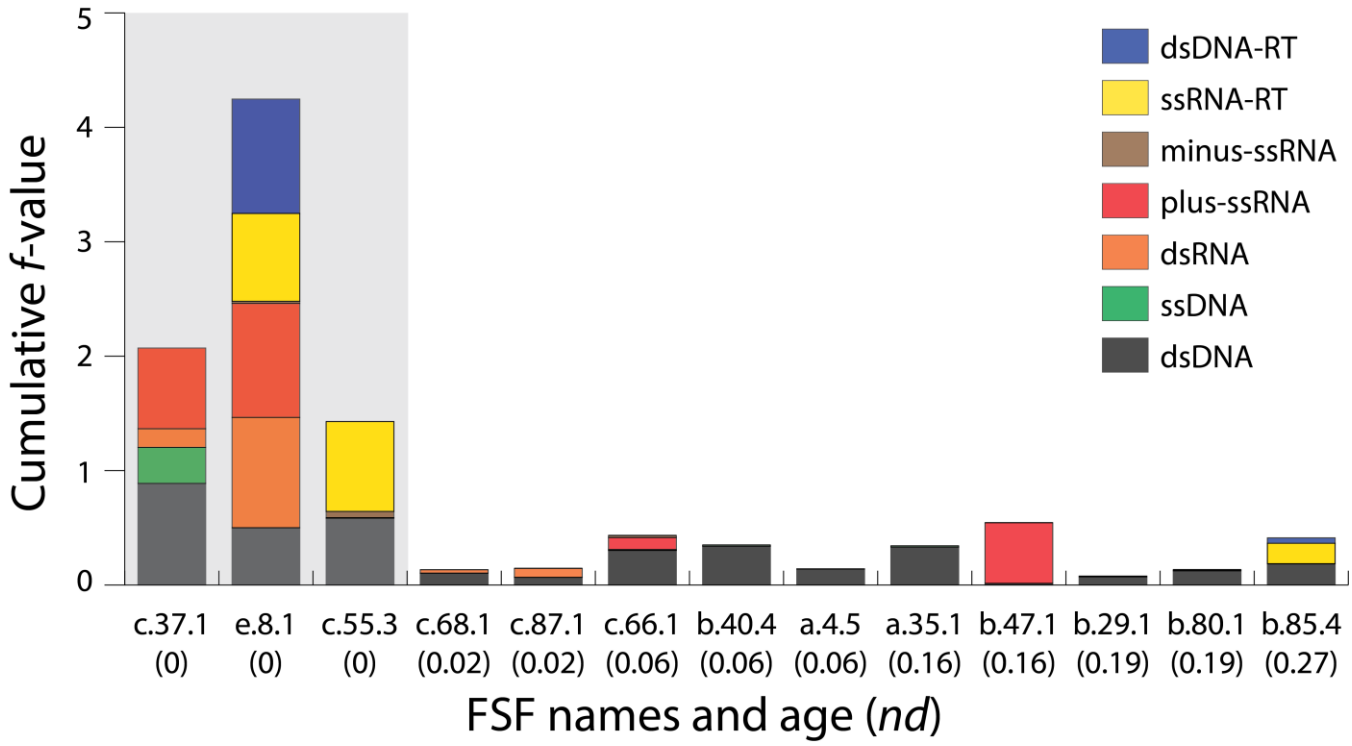

Fig. S3. Spread of *abe* core FSFs in viral subgroups. Out of the 68 *abe* FSFs, 49 were unique to dsDNA viruses. From the remaining that was shared by at least more than one viral subgroup, 13 were of ancient origin (*nd* < 0.3). *nd* values for individual FSFs are shown in parenthesis.

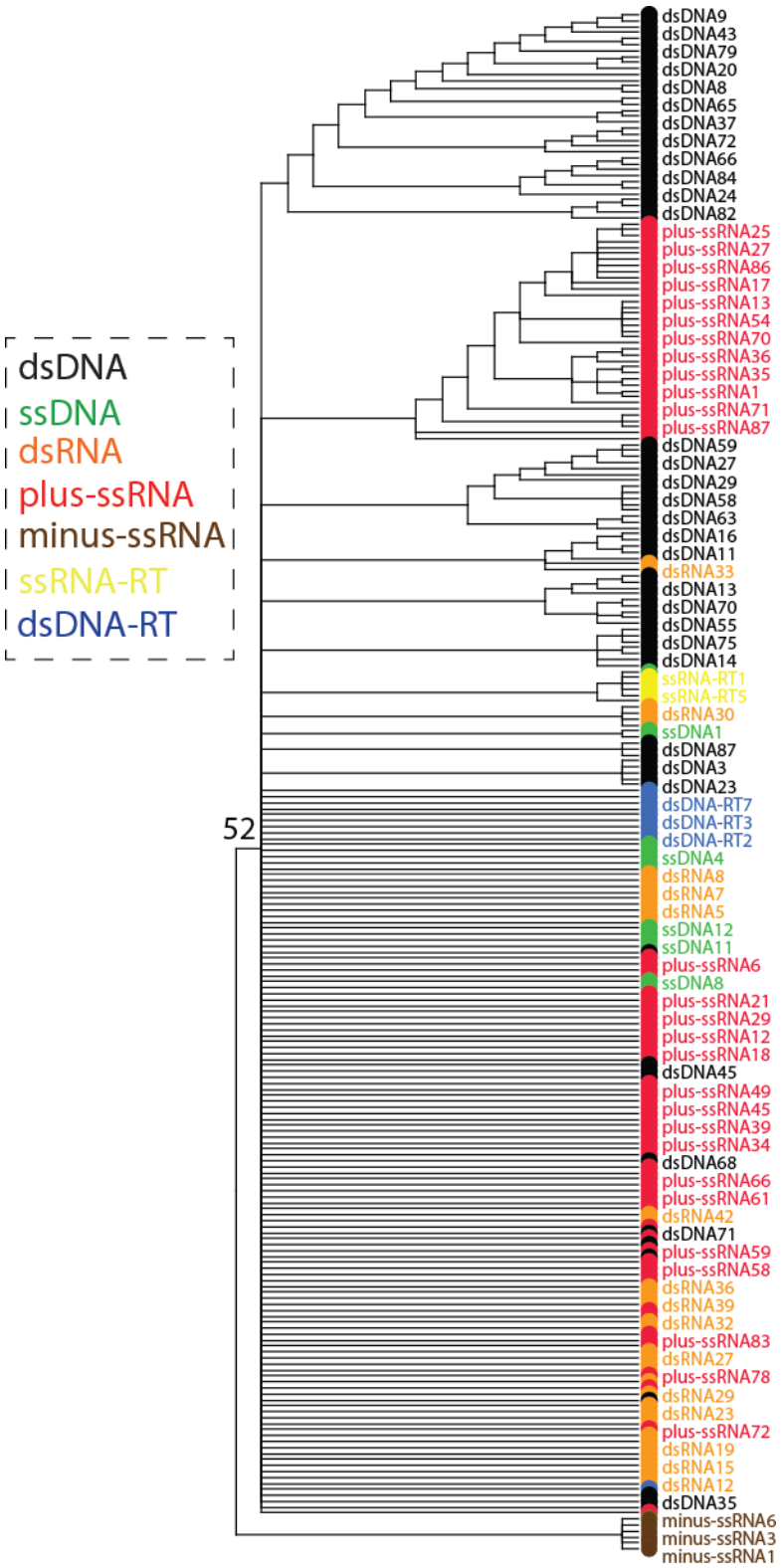

Fig. S4. Evolutionary relationships within the viral subgroup. A ToV highlights the evolutionary relationships between different viral subgroups. A total of 258 viral proteomes (taxa) were randomly sampled from viruses and were distinguished by the abundance of 68 *abe*

core FSFs (characters). A strict consensus of two most parsimonious trees is shown (Tree length = 3,644; Retention Index = 0.80; g_1 = -0.37; 66 parsimony informative characters). Each taxon was given a unique tree Id (see table S1). Not all taxa were labeled, as they would not be legible.

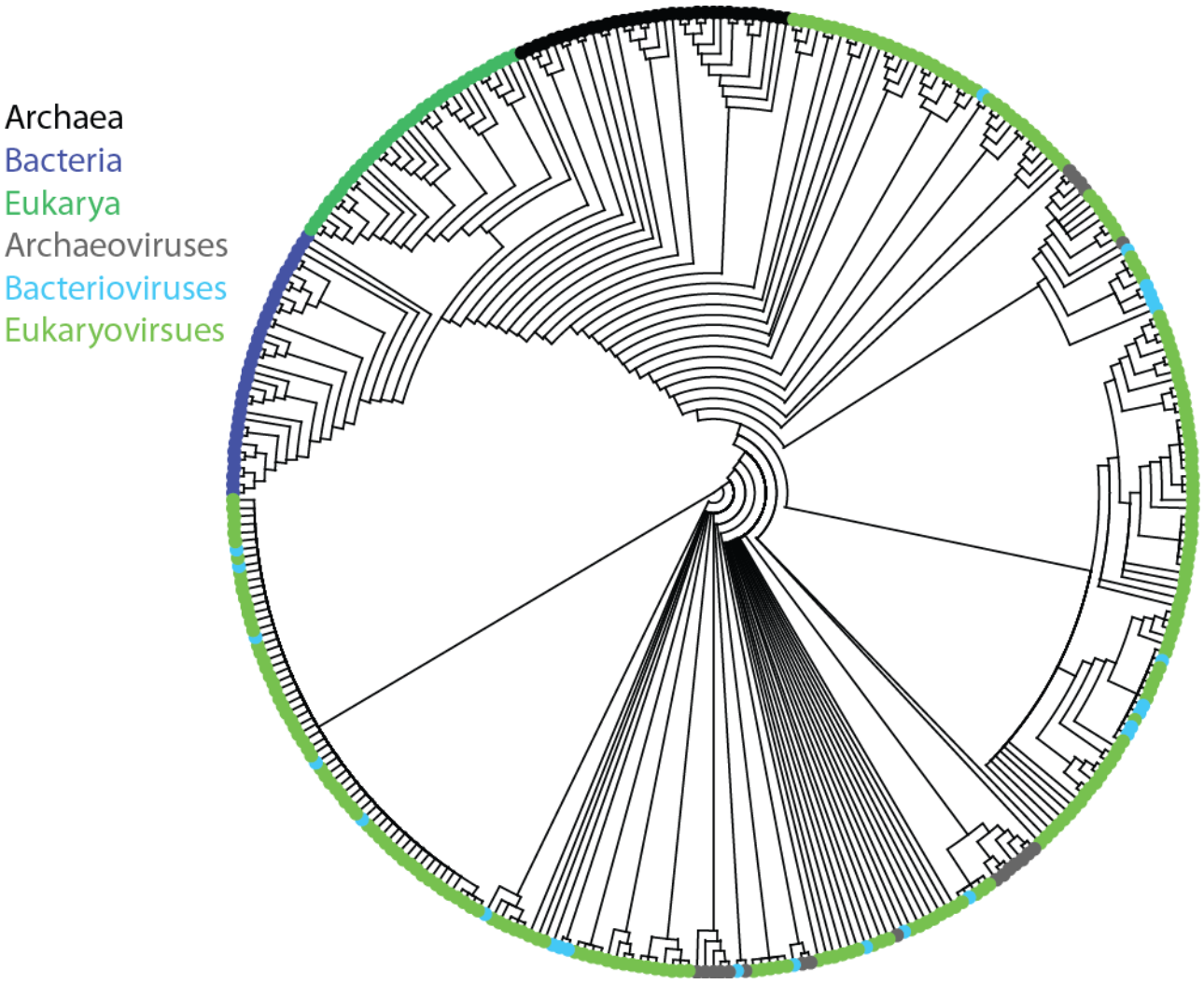

Fig. S5. Evolutionary relationships between cells and viruses. A ToP highlights the evolutionary relationship between viral and cellular proteomes. A total of 368 proteomes (taxa) were randomly sampled from viruses and cells and were distinguished by the abundance of 442 ABEV FSFs (characters) (Tree length = 45,935; Retention Index = 0.83; g_1 = -0.31). All characters were parsimony informative. Viruses were classified according to their host type. Taxa were colored for better visualization.

Table S1. List of viruses sampled in this study. Tree Id is a unique id assigned to each viral taxon that was sampled in the phylogenetic analysis. Viruses included in evolutionary analysis are highlighted in yellow.

Table S2. List of cellular organisms sampled in this study. Tree Id is a unique id assigned to each cellular taxon that was sampled in the phylogenetic analysis. 'Free-living' organisms included in evolutionary analysis are highlighted in yellow.

Table S3. VSFs and their spread in cellular (X) proteomes.

Table S4. **FSF use and reuse values for all proteomes.** 'Free-living' proteomes that were included in evolutionary analysis are highlighted in yellow.

Table S5. List of FSFs corresponding to each of the seven Venn groups defined in Fig. 3B. The *f*-value for each FSF in the three superkingdoms and viral subgroups is also given.

Table S6. **FSFs mapped to structure-based viral lineages.** Family names in red were not mapped in our HMM-based search while names in blue were proposed novel additions. FSFs highlighted in yellow participate in capsid assembly but were not signature FSFs.

Table S7. Significantly enriched "biological process" GO terms in EV FSFs (FDR < 0.01). No term was enriched in AV or BV FSFs.