

# Systematic evaluation of horizontal gene transfer between eukaryotes and viruses

Nicholas A. T. Irwin<sup>1,2,3\*</sup>, Alexandros A. Pittis<sup>3</sup>, Thomas A. Richards<sup>2</sup>, Patrick J. Keeling<sup>3</sup>

<sup>1</sup> Merton College, University of Oxford, Oxford, OX1 4JD, United Kingdom.

<sup>2</sup> Department of Zoology, University of Oxford, Oxford, OX1 3SZ, United Kingdom.

<sup>3</sup> Department of Botany, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada.

\*Corresponding author. Email: nicholas.irwin@merton.ox.ac.uk.

**Gene exchange between viruses and their hosts acts as a key facilitator of horizontal gene transfer and is hypothesized to be a major driver of evolutionary change. Our understanding of this process comes primarily from bacteria and phage co-evolution, but the mode and functional significance of gene transfers between eukaryotes and their viruses remains anecdotal. Here we systematically characterized viral-eukaryotic gene exchange across eukaryotic and viral diversity, identifying thousands of transfers, and revealing their frequency, taxonomic distribution, and projected functions. Eukaryote-derived viral genes, abundant in the Nucleocytoviricota, highlighted common strategies for viral host-manipulation, including metabolic reprogramming, proteolytic degradation, and extracellular modification. Furthermore, viral-derived eukaryotic genes implicate genetic exchange in the early evolution and diversification of eukaryotes, particularly through viral-derived glycosyltransferases, which have impacted structures as diverse as algal cell walls, trypanosome mitochondria, and animal tissues. These findings illuminate the nature of viral-eukaryotic gene exchange and its impact on the evolution of viruses and their eukaryotic hosts.**

## Introduction

The exchange of genes between viruses and eukaryotes through horizontal gene transfer (HGT) is a key evolutionary driver capable of facilitating host manipulation and viral resistance<sup>1-4</sup>. Host-derived genes are known to be employed by viruses for replication and cellular control<sup>1,5</sup>. This is observed across a diversity of viral lineages which encode cellular-derived informational

genes like tRNA synthetases and polymerases, as well as operational genes, such as immune effectors and metabolic enzymes<sup>5–12</sup>. These genes counter host immunity, hijack cellular machinery, and circumvent nutritional bottlenecks, making them key resources for adaptation<sup>1,13</sup>.

Conversely, viral-derived genes in eukaryotic genomes have been frequently perceived as inconsequential remnants of viral interactions<sup>14</sup>. However, these genes can be co-opted, supplementing or supplanting existing cellular components or providing entirely novel functionality. For example, core proteins such as histones and E2F transcription factors have been replaced by viral proteins in dinoflagellates and fungi, respectively<sup>15,16</sup>, while viral structural proteins, fusogens, and proviruses are utilized for communication, cellular fusion, and antiviral defense, in mammals and other eukaryotes<sup>2,3,17–19</sup>. The co-option of such viral proteins has also been found to coincide with cellular innovation and the radiation of major eukaryotic lineages where these genes serve key functions<sup>20,21</sup>. Accordingly, these transfers have important evolutionary, ecological, and health implications, but we nonetheless lack a general understanding of the mode, tempo, and functional significance of viral-eukaryotic gene exchange, largely due to a lack of standardized analyses across diverse taxa.

## **Main text**

### **A systematic identification of gene transfer events reveals patterns of viral-eukaryotic gene exchange**

To reconcile this lack of a systematic survey, we comprehensively characterized viral-eukaryotic gene transfer in 201 eukaryotic and 108,842 viral taxa, covering the diversity of eukaryotic and viral species with genomic representation, by developing a phylogenetic pipeline capable of screening thousands of evolutionary trees for HGT-indicative topologies while accounting for phylogenetic statistics and contamination (Extended Data Fig. 1, 2). These analyses identified 1,333 candidate (i.e., both well- and weakly-supported) virus-to-eukaryote transfers, 4,807 eukaryote-to-virus transfers, and 600 transfers with unknown directionality, altogether affecting 2,841 distinct protein families (Fig. 1a, Supplementary Table 1), and including multiple previously characterized examples (e.g., transporters, signalling proteins, metabolic enzymes, and viral housekeeping genes<sup>5,6,9,22–24</sup>). To reduce false positives,

phylogenetically ambiguous ( $n = 607$ ) or long branching ( $n = 2,133$ ) HGTs were considered weakly-supported and were excluded in downstream analyses (Fig. 1a, included in Supplementary Table 1). Given our emphasis on specificity over sensitivity, along with limitations in taxon sampling and homology detection, these figures represent a conservative estimate of HGT events.

The resulting HGTs revealed trends regarding the nature of viral-eukaryotic gene exchange. Transfers from eukaryotes to viruses were observed approximately twice as frequently as transfers in the reverse direction (Fig. 1a). This imbalance is explained by the higher number of viral recipients compared to donors per eukaryotic taxa (Fig. 1b) and the greater number of genes transferred to each viral recipient relative to those received per viral donor (Fig. 1c, d). These data also demonstrate a correlation between the number of viral recipients and donors per eukaryote ( $r_{\text{Pearson}} = 0.49$ ,  $p < 1 \times 10^{-18}$ , Fig. 1b), suggesting that viral-eukaryotic gene transfer is reciprocal but biased towards viral acquisition. Although sampling bias could influence these numbers, taxon representation affects both recipient and donor frequencies, and bootstrap estimates based on random sampling of protein phylogenies corroborated the observed disparity (Fig. 1a). This bias may reflect the expanded repertoire of eukaryotic genes or differing recombination and fixation rates in eukaryotes and viruses<sup>25,26</sup>, all of which could generate greater opportunity for viral gene acquisition during host-pathogen interactions.

Identifying the taxonomy of donors and recipients revealed the propensity of certain lineages to participate in HGT. The vast majority of transfers involved double stranded DNA viruses (97.6%), particularly the nucleocytoplasmic large DNA viruses (NCLDV or Nucleocytoviricota, including groups such as the Phycodnaviridae, Mimiviridae, Iridoviridae, Pithoviridae, Asfarviridae, and Poxviridae) which were the main contributors to genetic exchange across eukaryotic diversity (82.5%,  $n = 3,109$ ), consistent with their large and flexible genomes, long and intimate co-evolution with eukaryotes, and wide host-breadth (Fig. 1e, f, Extended Data Figure 1c)<sup>8,27–30</sup>. However, transfers involving alternative lineages such as herpesviruses ( $n = 47$ ), baculoviruses (e.g., nucleopolyhedroviruses,  $n = 26$ ), RNA viruses (predominantly retroviruses,  $n = 93$ ) and bacteriophages (e.g., Caudovirales,  $n = 67$ ) were also documented, in accordance with previous observations (Extended Data Figure 3, 4, Supplementary Table 1)<sup>9,31–33</sup>. Amongst eukaryotes, gene exchange was more prevalent in unicellular compared to multicellular organisms (Fig. 1g), and particularly abundant in

unicellular opisthokonts (the protist relatives of animals and fungi), the diverse protist clade known as SAR (Stramenopila, Alveolata, and Rhizaria), and other ecologically important algal groups such as chlorophytes and haptophytes (Fig. 1e, f). This included numerous HGTs coinciding with the diversification of SAR and the largest influx of viral genes was detected around the origin of the dinoflagellates (Fig. 1e, f, Extended Data Fig. 3, 4). Elevated gene transfer amongst unicellular eukaryotes may result from more frequent encounters with NCLDV, which are hyper-diverse and abundant in aquatic environments<sup>8</sup>, as well as a lack of germline segregation, which likely contributes to the reduced frequency of HGTs observed in animals and plants (Fig. 1f, g)<sup>34</sup>. However, it is important to note that our methodology under-represents retroviral acquisitions, which are commonly observed throughout animal and plant lineages, but whose detection is limited in this analysis by the poor availability of host-free retroviral genome assemblies which are required for phylogenetic interpretation<sup>35</sup>.

We also noted eukaryotic species harboring particularly large numbers of viral genes (Fig. 1c, f). These included species previously described to contain substantial viral genomic insertions from phycodnaviruses (*Ectocarpus siliculosus* and *Tetrabaena socialis*), phycodnaviruses and asfarviruses (*Hyphochytrium catenoides*), or multiple poorly classified viruses (*Acanthamoeba castellanii*), indicating a single or few sources (Fig. 1c, f, Extended Data Fig. 4, Supplementary Table 1)<sup>24,36–38</sup>. Other species also exhibited elevated numbers of viral genes derived from multiple NCLDV viruses (Fig. 1c, f). Whether these genes retain functional roles, such as in anti-viral viroplasm production<sup>23,39</sup>, or reflect remnants of past infections<sup>40,41</sup>, is unclear. However, large multigene acquisitions (e.g., ten or more genes) were rarely observed at ancestral ( $n=1$ ) relative to terminal nodes ( $n=13$ ), with the exception of the dinoflagellate ancestor (Fig. 1f). This suggests that large-scale transfers, potentially resulting from viral integrations, have recurrently affected diverse eukaryotic lineages, but are generally only transiently retained, possibly providing an opportunity for the longer-term retention and co-option of individual viral genes given adaptive significance.

Along with these transfers to eukaryotes, we identified a number of genes seemingly exchanged prior to the eukaryotic radiation. These transfers are inherently challenging to interpret given their antiquity, potential rooting uncertainty, and ambiguity resulting from intra-eukaryotic HGT<sup>42</sup>. Nonetheless, we observed multiple HGTs likely representing either ancient transfers from NCLDV viruses to early eukaryotic ancestors or recurrent viral acquisitions of

eukaryotic genes during eukaryogenesis. These included core informational genes originally derived from archaea such as RNA polymerases, DNA topoisomerase I, methionine aminotransferase 2, and replication factor C (RepC), the last of which involved transfers of three RepC subunits, similar to that observed for RNA polymerase (Extended Data Figure 5a-d) <sup>11</sup>. Furthermore, GDP-L-fucose synthase, which functions in fructose and mannose metabolism, was also involved in an ancient exchange (Extended Data Figure 5e) <sup>43</sup>. These data suggest that genetic exchange between ancestral eukaryotes and NCLDV viruses may have been important during eukaryogenesis, corroborating earlier observations and hypotheses <sup>11,21,44</sup>.

### **Direction and functional associations of gene transfers**

To further investigate the functional relevance of these HGTs, we examined the transfer direction and functional associations of exchanged protein families. Of the 1,859 families exhibiting HGT with known directionality, the majority (93%) underwent unidirectional transfer (i.e., eukaryote-to-virus or virus-to-eukaryote transfer; Fig. 2a). Dividing this dataset by direction, genes involved in eukaryote-to-virus exchanges were generally transferred unidirectionally (92% unidirectional), whereas a larger proportion of families undergoing virus-to-eukaryote transfer participated in bidirectional exchange (71% unidirectional and the remainder exhibiting both eukaryote-to-virus and virus-to-eukaryote transfers; Fig. 2a), suggesting that some of these exchanges may represent transduction (cell-virus-cell HGT). By moving across the phylogenies of all families exhibiting virus-to-eukaryote transfers, from viral donors towards the root, we estimated that 30.5% ( $n = 259$ ) of viral genes acquired by eukaryotes were originally eukaryotic, whereas fewer (8.2%,  $n = 70$ ) originated in prokaryotes, perhaps reflecting the differential utility or abundance of these genes in eukaryotic and viral systems (Extended Data Fig. 6, Supplementary Table 1). The remainder had unclear origins (24.2%,  $n = 205$ ) or were not attributable to a cellular lineage (37.1%,  $n = 315$ ), suggesting that these genes are either viral innovations, ancient viral acquisitions sharing deep cellular homology undetectable in our dataset, or are the result of taxon sampling deficiencies (Extended Data Fig. 6a). These data demonstrate that over evolutionary time, viruses have a capacity to mediate intra-eukaryotic and inter-domain HGT (i.e., transfers between eukaryotes and prokaryotes) through transduction, the relative frequencies of which will be important to assess comprehensively in the future. This provides further evidence that viruses act as a gene conduit between diverse eukaryotic lineages,

as suggested previously<sup>45,46</sup>, which is reminiscent of prokaryotes, where viral transduction is key in ecological adaptation and genome evolution<sup>47,48</sup>.

Direction of transfer was also associated with distinct functional biases. Relative to eukaryotic protein families as a whole, eukaryote-to-virus transfers were enriched in functions associated with cellular activity and house-keeping, such as metabolic proteins, E3-ligases, and tRNA synthetases (Fig. 2a, b, Supplementary Table 1, Supplementary Table 3). The enrichment of metabolic proteins implicates cellular-derived genes in reprogramming host metabolism during infection, which appears to be achieved through both *de novo* metabolite synthesis pathways and uptake (e.g., metabolic enzymes and/or nutrient transporters), as well as cellular recycling via proteolysis (e.g., proteasomal degradation and autophagy) (Fig. 2a, b, Supplementary Table 1, Supplementary Table 3)<sup>1</sup>. Additionally, signalling and stress response proteins were frequently acquired and likely also contribute to regulating host physiology, gene expression, immune responses, and viral processing<sup>9,49</sup>. The functions of viral-derived genes in eukaryotes were less obvious and have fewer functional associations, but are strongly enriched for proteins functioning in glycosylation ( $p < 10^{-6}$ ) and, to a lesser extent, nuclear proteins (Fig. 2a, c, Supplementary Table 1, Supplementary Table 3). Bidirectionally transferred genes are also enriched in metabolic processes, protein modification, and stress response proteins, which represent a subset of functions most often acquired by viruses (Fig. 2d, Supplementary Table 1, Supplementary Table 3). These data show that eukaryote-to-virus and virus-to-eukaryote HGTs both involve functional tendencies which are not equivalent, but may reflect the different adaptive contexts of viruses and eukaryotes<sup>50</sup>.

### **Eukaryote-derived viral genes are associated with distinct cellular processes and compartments**

To understand how these genes are used in viral and eukaryotic systems, we first examined the subcellular targets of eukaryote-derived viral proteins to understand where the proteins may operate in host cells. Cellular localizations were predicted using a neural network-based approach (DeepLoc)<sup>51</sup>, revealing that most eukaryote-to-virus HGTs likely function in the cytoplasm ( $n = 909$ ), nucleus ( $n = 482$ ), mitochondrion ( $n = 284$ ), and extracellular space ( $n = 214$ ) (Fig. 3a, Supplementary Table 1). However, relative to all eukaryotic protein families, viral-acquisitions were enriched in cytoplasmic, endoplasmic reticulum (ER), extracellular, and

180 peroxisomal proteins, the last of which suggests functions involving lipid catabolism and  
181 oxidation (Fig 3b). Moreover, predicted localizations were generally equivalent between donor  
182 and recipient proteins, with variation likely resulting from prediction inconsistencies, viral  
183 sequence divergence, or potentially neofunctionalization (Fig. 3c, 71% consistent)<sup>51,52</sup>. This  
184 indicates that eukaryote-derived gene products tend to function in the same subcellular context as  
185 the original host-encoded proteins.

186 To examine the processes that these genes impact in given cellular compartments, we  
187 conducted localization-based functional enrichments revealing the functional breadth and  
188 cellular processes associated with eukaryote-derived viral genes. Cytoplasmic proteins were  
189 largely involved in translation, metabolism, proteolysis, and signaling, whereas nuclear proteins  
190 mainly functioned in DNA processing, chromatin organization, cell cycle regulation, and protein  
191 modification (Fig 3d, e, Supplementary Table 1, Supplementary Table 4), in agreement with  
192 previous studies<sup>1,5,6,10,53–55</sup>. Endoplasmic reticulum proteins were predominantly associated with  
193 lipid metabolism and membrane remodeling (Fig. 3f, Supplementary Table 4). Proteins such as  
194 sphingolipid synthesis enzymes contributed to the localization bias, since many function in the  
195 ER, were frequently transferred (Supplementary Table 1), and are known to be used by diverse  
196 viruses for cellular regulation<sup>13,56,57</sup>. Additionally, ER remodeling is important for generating  
197 membrane-enclosed viral factories and for replication<sup>58</sup>. Extracellular proteins acquired by  
198 viruses were enriched for functions including carbohydrate metabolism and proteolysis,  
199 reflecting proteins such as glycosyl hydrolases, glycosyltransferases, and S1 peptidases, and  
200 implying a tendency for cell-surface modulation (Fig. 3g, Supplementary Table 4). This is  
201 consistent with repeated observations of viruses manipulating cell membranes and extracellular  
202 spaces through polysaccharide and protein modification<sup>13,59,60</sup>. These results therefore highlight  
203 the key cellular systems associated with eukaryote-derived viral genes which, given their known  
204 roles in host-manipulation<sup>1,6</sup>, may provide insights into common viral infection strategies.  
205 Indeed, many of these processes are also known to be manipulated by viruses that lack  
206 eukaryotic genes (e.g., many non-NCLDV viruses), which instead often rely on small effectors  
207 and host-encoded proteins<sup>61–63</sup>. This suggests that cellular manipulation strategies may be  
208 ubiquitous across viral lineages, but that the mechanism through which modification is  
209 accomplished may depend on viral coding capacity (e.g., large genome sizes and increased  
210 coding capacity in the NCLDV could permit the more flexible use of acquired eukaryotic genes).

Notably, The characterization of host-virus interactomes has been proposed as a promising avenue for host-targeting antiviral drug discovery<sup>64,65</sup>. Therefore, if host manipulation mechanisms are similar across viral lineages, we hypothesize that eukaryote-derived viral genes could facilitate the prediction of cellular components pertinent for infection by diverse viral lineages. Although indirect, this could provide an analytically simplistic (e.g., homology-based) approach for predicting therapeutic targets that could complement data from experimental host-virus model systems<sup>65</sup>.

### **The acquisition of viral-derived glycosyltransferases correlates with eukaryotic morphological transitions**

Lastly, to gain insights into the role viral genes play in eukaryotic systems, we inspected the distributions and functions of viral-derived glycosyltransferases, which were strongly enriched in the identified virus-to-eukaryote HGTs (Fig. 2c). We identified 63 instances of eukaryotes acquiring viral glycosyltransferases, of which 13 mapped to ancestral nodes, implying functional relevance under long term selection (Supplementary Table 5). Plotting transfer events and annotations over a eukaryotic phylogeny revealed the functional diversity and recurrent acquisitions of these enzymes across eukaryotic lineages (Fig. 4a, Extended Data Fig. 7). These HGTs were often correlated with morphological and structural synapomorphies including algal cell wall elaboration (e.g., lipopolysaccharide (LPS) and cellulose synthesis enzymes)<sup>66</sup>, long-chain polyamine-containing scale formation in haptophytes (spermidine synthase)<sup>67</sup>, cellular aggregation in the opisthokonts and dictyostelid slime molds (hyaluronan synthase and GlcNAc transferase), and mitochondrial structural divergence in the kinetoplastids (fucosyltransferase), a group primarily comprised of animal parasites such as trypanosomes (Fig 4a). Experimental data supported a number of these correlations, including the unusual identification of LPS in the cell walls of *Chlorella*<sup>68</sup>, the importance of hyaluronan in vertebrate tissues<sup>69</sup>, and the role of the dictyostelid N-acetylglucosamine transferase, Gnt2, in calcium-independent cellular aggregation<sup>70</sup>, indicating that virally sourced genes are co-opted during the evolution of novel cellular traits (Fig. 4a). We further examined two glycosyltransferase acquisitions in kinetoplastids, hypothesizing that, given the correlation between the HGT acquisitions and the origin of the highly derived kinetoplastid mitochondria (called kinetoplasts), they should function in that compartment. Phylogenetic analyses revealed that both genes were derived from the NCLDV, highlighted the prokaryotic origin of the fucosyltransferase



(C000231), and confirmed that both genes are conserved throughout kinetoplasts (Fig. 4b, c). Moreover, both proteins localized to the kinetoplast in *Trypanosoma brucei* (identifiable as a non-nuclear DNA-stained foci) both when tagged with mNeonGreen (Fig. 4d) and by organellar proteomics (Fig. 4e). A recent report also suggested an essential role for the fucosyltransferase in mitochondrial function in *T. brucei*<sup>71</sup>, altogether indicating that these viral-derived glycosyltransferases were co-opted for use in the kinetoplast at the same time as it underwent massive evolutionary change. These data, in combination with the capacity for viruses to modify cell surfaces and induce morphological alterations in their hosts (e.g., cytopathic effects)<sup>72</sup>, suggests that viral-derived genes may have played various roles in the evolution of cellular morphology across the eukaryotic tree of life.

## Discussion

Horizontal gene transfer between viruses and eukaryotes has been observed and assumed to impact the evolution of both participants, but until now we lacked the systematic characterization necessary to generalize the mode and functional significance of these transfers in both viral and eukaryotic contexts<sup>2,29,37</sup>. As with all computational surveys, our dataset is limited by specificity and sensitivity, but nonetheless it provides an extensive resource from which phylogenetic patterns can be observed and their genomic and functional importance may be predicted. From a viral perspective, the preponderance of host-derived genes in the NCLDV reiterates the importance of gene exchange in the evolution of these viruses<sup>27</sup>, and underscores the ubiquity of certain viral host-manipulation strategies. Indeed, many important emerging human pathogens, such as Zika and coronaviruses, depend on the manipulation of similar eukaryotic systems, such as autophagy, proteolysis, ER modification, and sphingolipid metabolism<sup>57,73,74</sup>. Similarly, functional investigations of eukaryote-derived viral genes, particularly using heterologous expression<sup>6</sup>, may also provide insights into how viruses manipulate these cellular pathways while circumventing the need for tractable host-virus model systems. From a eukaryotic perspective, our analyses provide further evidence that viruses participate in eukaryotic transduction and implicate viral-eukaryotic gene exchange in eukaryogenesis and the evolution of eukaryotic morphology. In particular, horizontally-acquired glycosyltransferases have recurrently impacted transitions as fundamental as the evolution of tissues and divergence of mitochondria, reminiscent of how retroviral genes, such as fusogens,

have repeatedly driven placental evolution in animals<sup>19,75</sup>. Our survey also identified protein candidates for which experimental characterizations could help reveal the impact of these genes on cellular systems and their roles in driving the evolution of eukaryotic complexity.

## Methods

### Dataset assembly

To systematically and conservatively identify instances of viral-eukaryotic gene exchange, groups of homologous eukaryotic, viral, and prokaryotic proteins were clustered into protein families and phylogenetic analyses were performed (Extended Data Fig. 1a). To do this, a eukaryotic dataset was generated from 196 genome-predicted eukaryotic proteomes, primarily from UniProt (release 2018\_11), derived from diverse species from all available major eukaryotic supergroups. These proteomes were individually clustered at 99% percent-identity with Cd-hit v4.8.1<sup>76</sup> to reduce redundancy resulting from recent paralogs and isoforms, and combined. The eukaryotic dataset was further supplemented with five complete transcriptomes (four dinoflagellates (MMETSP0224, MMETSP0227, MMETSP0228, MMETSP0790) and a cercozoan (SRR3221671) with over 90% BUSCOs (Benchmarking Universal Single Copy Orthologs) present using the alveolata\_odb10 (for the dinoflagellates) or eukaryota\_odb10 (for *Paulinella*) databases, assessed using BUSCO v4.1.4) to fill taxonomic gaps in lineages with poor genomic sampling<sup>77,78</sup>. Viral proteins predicted from the genomes of diverse viral taxa including DNA and RNA viruses, were obtained from UniProt and filtered to exclude those derived from Human Immunodeficiency Virus-1, which were over-represented (Extended Data Fig. 1b, c). Additional viral proteins were acquired from nucleocytoplasmic large DNA virus (NCLDV) metagenomes previously assembled from diverse environments and assessed as having low contamination based on gene content (see **Contamination Scoring**)<sup>8</sup>. Viral taxonomic annotations were assigned to metagenomes based on previously conducted phylogenomic analyses<sup>8</sup>.

Eukaryotic and viral proteins were then clustered into protein families using a similarity-based approach and the Markov clustering (MCL) algorithm (inflation = 2) after comparing sequences to one another using Diamond v2.0.2 BLASTp (sensitive mode, E-value < 10<sup>-5</sup>, query coverage > 50%) (Extended Data Figure 1a, step i)<sup>79,80</sup>. Protein families containing both viral and eukaryotic representatives were retained, aligned with MAFFT v.7.397<sup>81</sup>, and used to generate profile hidden Markov models (HMMs) which were used to search 9,035 prokaryotic

proteomes from UniProt with HMMER v.3.1b2 ( $E\text{-value} < 10^{-5}$ ,  $\text{incE} < 10^{-5}$ ,  $\text{domE} < 10^{-5}$ )<sup>82</sup> (Extended Data Fig. 1a, step i). In this case, HMMs were used to improve the detection of distant prokaryotic homologs. Due to the large number of prokaryotic sequences, the resulting hits were reduced by taking the most significant hit (based on E-value) per genus or per strain, to a maximum of 150 sequences, this allowed for diverse taxon sampling while avoiding an overabundance of prokaryotic proteins (Extended Data Figure 1a, step i). Sequences assigned to viral-eukaryotic protein families were then combined with the prokaryotic proteins and re-clustered, as above (Extended Data Fig. 1a, step ii).

### Phylogenetic analysis

Phylogenetic trees were generated from clustered protein families to infer the evolutionary relationships between viral and eukaryotic homologs. Protein families were filtered to retain only those with viruses and eukaryotes, aligned with MAFFT (--auto), trimmed using a gap-threshold of 20% in trimAl v1.2, and sequences with less than 50 amino acid positions were removed (Extended Data Figure 1a, d, e)<sup>83</sup>. Maximum likelihood phylogenies were conducted in IQ-Tree v1.6 using the robust and generic LG+F+R5 substitution model, and statistical support was calculated using SH-aLRT (Shimodaira-Hasegawa approximate likelihood ratio test,  $n = 1,000$ ), which was chosen due to its speed, insensitivity to model violations and taxon sampling, and its comparable conservativeness to standard bootstrapping<sup>84–86</sup>. Phylogenies for large protein families with over 1,500 sequences ( $n = 103$ ) were generated using the fast search mode in IQ-Tree. Phylogenetic rooting was done using minimal ancestral deviation (MAD), which is a rooting method that is more robust to heterotachy than midpoint rooting<sup>87</sup>.

For individual phylogenies of particular interest, such as those shown in Fig. 4, and Extended Data Figs. 5, 6 and 7, analyses were repeated as above but after alignment with the more accurate L-INS-i algorithm in MAFFT (or --auto for C000038) and limited curation (e.g., the removal of long-branching taxa as defined below, see **Horizontal gene transfer detection**). Additionally, substitution models were selected using ModelFinder in IQ-Tree<sup>86,88</sup> and phylogenies were visualized and annotated using iTOL v4<sup>89</sup>. Notably, the topologies of these trees were consistent with their initial iterations and ModelFinder consistently selected the LG substitution model similar to that used in the other phylogenies, corroborating the use of the aforementioned methods (see Extended Data Fig. 6, 7). Prior to phylogenetic inference, the trimmed alignments for protein families inferred to exhibit ancient gene transfers (e.g., Extended

Data Fig. 5) were recoded using 4-bin Dayhoff recoding to reduce the effects of saturation and compositional heterogeneity, as done previously<sup>11,42,90</sup>.

### **Horizontal gene transfer detection**

To detect instances of HGT, we developed a conservative algorithmic approach that emphasized specificity over sensitivity, given the potential for contamination in the underlying dataset and the risk of phylogenetic artifacts. We developed an automated pipeline using the python package, ETE 3<sup>91</sup>, to identify HGT-indicative topologies in the phylogenetic trees generated from each protein family. Specifically, we aimed to identify eukaryotic species nested within viral clades (viral-to-eukaryote HGT) or viral taxa within eukaryotic clades (eukaryote-to-virus HGT) (Extended Data Fig. 2a). To this end, phylogenies were initially processed to account for statistical support and directionality (i.e., rooting), and to assign taxonomic annotations. Firstly, phylogenetic nodes with SH-aLRT values below 0.8 were collapsed, a threshold with a false positive rate similar to a standard bootstrap support of 60%<sup>84</sup> (Extended Data Fig. 2a, step i). Collapsed phylogenies were then rooted using MAD rooting<sup>87</sup> and taxa were annotated as eukaryotic, viral, or prokaryotic using the National Centre for Biotechnology Information (NCBI) Taxonomy database (Extended Data Fig. 2a, step i)<sup>92</sup>. Viral taxonomic annotations in the NCBI Taxonomy database are in part based on ICTV (International Committee of the Taxonomy of Viruses) classifications<sup>93</sup>.

Following tree processing and annotation, but before identifying HGT events, the phylogenies were analyzed to assess rooting ambiguity (Extended Data Fig. 2a, step ii). In particular, we checked whether viral and eukaryotic sequences could be separated into two monophyletic groups using alternative root placements. In this case rooting becomes unclear unless the phylogeny is strongly biased toward viral or eukaryotic species representation (e.g., it is unlikely that a gene conserved throughout a eukaryotic supergroup was derived from a single virus). To evaluate this, if a phylogeny could be split into two discrete taxonomic clades, the ratio of eukaryotic to viral species was determined. If the ratio was heavily skewed towards eukaryotes or viruses (eukaryote:viral species ratio > 49 or < 0.15, reflecting the top and bottom 20% of all protein families), the tree was rooted normally. Otherwise, the topology would be classified as an HGT with unknown directionality. Lastly, single prokaryotic sequences and HGTs between prokaryotes and viruses or eukaryotes (identified as below) were removed to

reduce topology complexity, but this did not increase the false positive rate amongst viral-eukaryotic HGTs (Extended Data Fig. 2a, step ii).

After processing, phylogenies were screened for HGT topologies (Extended Data Fig. 2a, step iii). To achieve this, viral and eukaryotic clades were identified and the taxonomy of their sister group (i.e., the most closely related phylogenetic group) and ‘cousin’ group (i.e., the second most closely related phylogenetic group) were determined. A eukaryote-to-virus HGT topology was defined as a viral clade with a eukaryotic sister and cousin whereas a virus-to-eukaryote HGT required a eukaryotic clade with a viral sister and cousin (Extended Data Fig. 2a, step ii). Initially, viral and eukaryotic clades were identified and the taxonomy of their sister and cousin groups were assessed. To classify the taxonomy of these groups, the number of viral, eukaryotic, and prokaryotic sequences in each group was counted. Sister and cousin groups were then classified as viral, eukaryotic, or prokaryotic if the taxonomies were consistent across the members of the group. If the taxonomies of a group were mixed (e.g., if both viral and eukaryotic sequences were present), but viral or eukaryotic taxa dominated at least 80% of the sequences, the group was described as ‘probably’ viral or eukaryotic, or else the group received an ambiguous designation. In the event of a polytomy, multiple sister and cousin groups could be present. To account for this, the taxonomy of the polytomy-wide group would be summarized by determining the taxonomy of each group within the polytomy (as above). If all candidate sisters or cousins within the polytomy were classified consistently, the group would be identified as viral, eukaryotic, or prokaryotic. Likewise, if a majority (more than two-thirds) of the groups were consistently classified, the sister or cousin would be denoted as ‘probably’ viral or eukaryotic, otherwise it would be labeled as ambiguous. After classifying both sister and cousin groups, if the topology was consistent with one of the aforementioned scenarios, an HGT event would be noted (Extended Data Fig. 2a, step iii). Each phylogeny was screened for eukaryote-to-virus and virus-to-eukaryote HGTs three times iteratively, given that once a viral or eukaryotic clade had been classified as an HGT, it would be interpreted as eukaryotic or viral, respectively, in subsequent iterations. Finally, after three cycles of HGT identification, if there were remaining viral and eukaryotic clades sister to one another with ambiguous or prokaryotic cousins, they were labeled as HGTs with unknown transfer directionality.

Once an HGT was identified, characteristics including the recipient, donor, phylogenetic statistics, and topology notes were recorded (Extended Data Fig. 2a, step iv; see Supplementary

Table 1). Recipient and donor taxa were assessed by determining the last common ancestor of the recipient and donor (sister), respectively, based on the NCBI Taxonomy database <sup>92</sup>. Moreover, node support values were recorded along with the branch length of the recipient (the distance from recipient node or tip to the transfer node). If the branch length of the recipient or donor represented an extreme outlier (defined as the median branch length of the phylogeny after removing identical branch lengths, plus three times the interquartile range), the HGT was highlighted as a potential long branch attraction (LBA) artifact (Extended Data Fig. 2a, step iv). Additionally, if the donor only had a ‘probable’ taxonomic classification, ambiguity would also be noted. In both of these cases, HGTs were labeled as weakly supported and excluded in downstream analyses. Lastly, the approximate origin of viral-derived eukaryotic genes was determined by moving up through phylogenetic nodes from the donor clade towards the root until a cellular lineage was reached, if possible (Extended Data Fig. 3a). The effect of gene sampling on the number of HGTs was assessed using 95% bootstrap confidence intervals, calculated by randomly sampling gene families with replacement to an equal number of families as the original dataset ( $n = 1,000$ ) (Fig. 1a).

### **Contamination scoring**

After identifying the HGTs, individual transfer events were assessed for possible alternative sources of phylogenetic incongruence, specifically contamination. This is important given that eukaryotic and viral genes can be artifactually present in viral and eukaryotic genomes, respectively, which may give the impression that HGT has occurred. To address this, only viral reference proteomes and metagenomes with low contamination scores (estimated previously based on the representation of core NCLDV genes in a given metagenome relative to those observed in viruses from a related superclade, for more information see <sup>8</sup>) were included in the analysis and individual viral-derived eukaryotic genes were assessed based on a series of criteria and a contamination scoring scheme (Extended Data Fig. 2b-e). Contamination was assessed based on two main attributes: 1) the presence of related taxa in the HGT, and 2) the characteristics of the genomic contig upon which the gene was encoded. Firstly, the taxonomic composition of the HGT-recipients was assessed based on the assumption that the same contamination is unlikely to occur in multiple independently sampled genomic datasets, particularly if the species from which they are derived are closely related. Therefore, points were

given if the HGT-recipients included multiple members of the same (+3) or different (+1) phyla as the species encoding the gene of interest, based on NCBI Taxonomy (Extended Data Fig. 2b, step i).

Secondly, the characteristics of the genomic contigs encoding each viral-derived gene were inspected based upon the notion that they should share attributes with the host genome, such as consistent GC-content, reasonable contig size, and that the gene should be flanked by eukaryotic regions. Accordingly, contigs were identified by mapping proteins to the genome using tBLASTn (e-value < 10<sup>-5</sup>) and points were given if the contig was within one standard deviation of the median genomic GC-content (+1) and if the contig was a reasonable size (greater than half of the L50) (+1) (Extended Data Fig. 2b (step ii-iv), c). Lastly, the genomic context was inspected by extracting DNA regions (5 kbp) upstream (-2.5 kbp) and downstream (+2.5 kbp) of each gene. Extracted regions were then taxonomically classified by comparing them to the SWISS-PROT database (release 2019\_11) using BLASTx and assessing the taxonomy of the resulting hits. A maximum of 20 hits (e-value < 10<sup>-3</sup>) were evaluated and normalized scores for viral, eukaryotic, and prokaryotic classifications were calculated to account for database bias. These scores were calculated for each taxonomic group as:

$$score_t = -\log_{10}(t_{prop}) \times S_t \times n_t$$

where  $t_{prop}$  is the proportion of a taxonomic group in the database, and  $n_t$  and  $S_t$  are the number and median bit-score of hits to that taxonomic group, respectively. The taxonomy was assigned to the group with the maximum score and points were given for eukaryotic classifications (+1 per region) and subtracted for prokaryotic ones (-1 per region), which are more indicative of contamination. Viral and uncertain classifications were neutral (+0) to account for the possibility of large viral insertions (Extended Data Fig. 2b, step v-vi).

After assigning contamination scores to each putative eukaryotic sequence involved in a virus-to-eukaryote HGT or an HGT with unknown directionality, sequences with scores less than two were excluded and HGTs and recipient taxonomies were reassessed (Extended Data Fig. 2b, d, e). Due to the strict criteria applied during filtering and HGT identification, false positive rates should be low, however, transfers associated exclusively with transcriptomic data should be interpreted with care as these sequences are more challenging to assess (e.g., transfers to *Alexandrium*, *Protoceratium*, *Togulla*, *Polarella*, and *Paulinella*).

## Functional analyses

To examine HGT function, eukaryotic and viral proteins were annotated with eggNOG, PANTHER (Protein analysis through evolutionary relationships), and Pfam using a combination of eggNOG-Mapper v2, InterProScan v.5.48, and HMMER v3.1b2 (e-value  $< 10^{-3}$ ) with the default parameters<sup>82,94,95</sup>. For clarity, the resulting gene ontology (GO) terms were simplified by mapping the terms to the yeast GO-slim subset using Map2Slim (see <https://github.com/owlcollab/owltools/wiki/Map2Slim>). Protein families were given functional annotations based on a majority rule (Supplementary Table 1) and labeled with gene ontology (GO) terms if a given term was assigned to at least 20% of annotated proteins within a family. Similarly, Pfam domains were assigned to a given family if they were detected in 20% of annotated proteins. Ultimately, of the 2,841 protein families exhibiting HGT, 2,747 (98.3%) received an annotation, while those that did not tended to be small and divergent (median number of sequences = 9, range = 3-60). To conduct GO-enrichment analyses, we tested the null hypothesis that GO-terms associated with the HGTs reflect a random sampling of eukaryotic protein families. To this end, protein families exhibiting HGT were compared against a eukaryotic background comprising all eukaryotic protein families containing either a virus or at least ten eukaryotic species, generated during both MCL clustering steps. The frequencies of individual GO-terms in the HGT families were compared to the eukaryotic background using permutation tests which involved randomly sampling equally sized sets of annotated protein families without replacement ( $n = 10^7$ ). Significantly enriched GO terms ( $p < 0.01$ ) were summarized and visualized using REVIGO<sup>96</sup>.

To investigate the predicted subcellular localizations of eukaryote-derived viral genes, all eukaryotic proteins were annotated using DeepLoc v1.0 and the BLOSSUM62 matrix<sup>51</sup>. Localization predictions with likelihoods less than 0.5 were re-classified as unknown and cellular targets were assigned to individual eukaryote-to-virus HGTs based on the majority localization of the donor (i.e., eukaryotic) sequences. Eukaryotic donor sequences were used for localization characterizations given that DeepLoc is trained and optimized using eukaryote encoded proteins<sup>51</sup>. When shown, predictions of eukaryote-derived viral proteins are displayed in the context of their donor sequence localizations. Enrichments were assessed by comparing the frequency of individual localizations in the HGTs to an equally sized random sampling of annotated eukaryotic proteins ( $p < 0.05$ ,  $n = 10^6$ ). The null hypothesis was that viruses randomly acquire



eukaryotic genes irrespective of their predicted subcellular localizations. Bootstrap confidence intervals were calculated based on bootstrap resampling of HGT donors ( $n = 1,000$ ). Subcellular localizations for *Trypanosoma brucei* proteins were assessed using fluorescent localization and organellar proteomic data obtained from TrypTag and TriTrypDB, respectively<sup>97,98</sup>.

## Data Availability

All data, including proteomes, protein families, annotations, alignments, phylogenies, and summaries of detected HGTs (both before and after contamination filtering) are available from Dryad (<https://datadryad.org/stash/dataset/doi:10.5061/dryad.z08kprrc9>).

## Code Availability

All the code used for phylogenetic interpretation, contamination scoring, and functional enrichments and analyses are available from Dryad (<https://datadryad.org/stash/dataset/doi:10.5061/dryad.z08kprrc9>).

## Acknowledgements

We thank Richard Wheeler for providing fluorescent micrographs of *Trypanosoma brucei*, as part of TrypTag. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-2014-03994) and from the Gordon and Betty Moore Foundation (<https://doi.org/10.37807/GBMF9201>) to P.J.K. N.A.T.I. was supported by a Junior Research Fellowship from Merton College, Oxford and an NSERC Canadian Graduate Scholarship. A.A.P. was supported by European Molecular Biology Organization (EMBO) Long-term Fellowship (ALTF 118-2017). T.A.R. was supported by a Royal Society University Research Fellowship (UF130382).

## Author Contributions

Conceptualization, N.A.T.I. and A.A.P.; Funding acquisition, P.J.K. and T.A.R.; Investigation N.A.T.I. and A.A.P.; Resources, P.J.K. and T.A.R.; Supervision, P.J.K. and T.A.R.; Writing N.A.T.I. with input from all authors.

## Competing Interests

The authors declare no competing interests.

## Materials and Correspondence

Material requests and correspondence should be addressed to N.A.T.I.

## References

1. Zimmerman, A. E. *et al.* Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat. Rev. Microbiol.* **18**, (2019).
2. Koonin, E. V. & Krupovic, M. The depths of virus exaptation. *Curr. Opin. Virol.* **31**, 1–8 (2018).
3. Frank, J. A. & Feschotte, C. Co-option of endogenous viral sequences for host cell function. *Curr. Opin. Virol.* **25**, 81–89 (2017).
4. Moreira, D. & Brochier-Armanet, C. Giant viruses, giant chimeras: The multiple evolutionary histories of Mimivirus genes. *BMC Evol. Biol.* **8**, 12 (2008).
5. Filée, J., Pouget, N. & Chandler, M. Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evol. Biol.* **8**, 320 (2008).
6. Monier, A. *et al.* Host-derived viral transporter protein for nitrogen uptake in infected marine phytoplankton. *Proc. Natl. Acad. Sci.* **114**, E7489–E7498 (2017).
7. Monier, A. *et al.* Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome Res.* 1441–1449 (2009) doi:10.1101/gr.091686.109.
8. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
9. Aswad, A. & Katzourakis, A. Cell-derived viral genes evolve under stronger purifying selection in rhadinoviruses. *J. Virol.* **92**, e00539-18 (2018).
10. Schulz, F. *et al.* Giant viruses with an expanded complement of translation system components. *Science* **356**, 82–85 (2017).
11. Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of

- giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19585–19592 (2019).
12. Enav, H., Mandel-Gutfreund, Y. & Béjà, O. Comparative metagenomic analyses reveal viral-induced shifts of host metabolism towards nucleotide biosynthesis. *Microbiome* **2**, 9 (2014).
  13. Vardi, A. *et al.* Host-virus dynamics and subcellular controls of cell fate in a natural coccolithophore population. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 19327–19332 (2012).
  14. Biémont, C. A brief history of the status of transposable elements: From junk DNA to major players in evolution. *Genetics* **186**, 1085–1093 (2010).
  15. Gornik, S. G. *et al.* Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. *Curr. Biol.* **22**, 2303–2312 (2012).
  16. Medina, E. M., Turner, J. J., Gordân, R., Skotheim, J. M. & Buchler, N. E. Punctuated evolution and transitional hybrid network in an ancestral cell cycle of fungi. *Elife* **5**, e09492 (2016).
  17. Pastuzyn, E. D. *et al.* The neuronal gene Arc encodes a repurposed retrotransposon Gag protein that mediates intercellular RNA transfer. *Cell* **172**, 275–288 (2018).
  18. Mi, S. *et al.* Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**, 785–789 (2002).
  19. Cornelis, G. *et al.* An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental *Mabuya* lizard. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E10991–E11000 (2017).
  20. Irwin, N. A. T. *et al.* Viral proteins as a potential driver of histone depletion in dinoflagellates. *Nat. Commun.* **9**, 1535 (2018).
  21. Forterre, P. & Prangishvili, D. The major role of viruses in cellular evolution: Facts and hypotheses. *Curr. Opin. Virol.* **3**, 558–565 (2013).
  22. Delaroque, N., Maier, L., Knippers, R. & Müller, D. G. Persistent virus integration into the genome of its algal host, *Ectocarpus siliculosus* (Phaeophyceae). *J. Gen. Virol.* **80**, 1367–1370 (1999).
  23. Blanc, G., Gallot-Lavallée, L. & Maumus, F. Provirophages in the *Bigelowiella* genome bear testimony to past encounters with giant viruses. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5318–E5326 (2015).

24. Leonard, G. *et al.* Comparative genomic analysis of the ‘pseudofungus’ *Hyphochytrium catenoides*. *Open Biol.* **8**, (2018).
25. McCrone, J. T. & Luring, A. S. Genetic bottlenecks in intraspecies virus transmission. *Curr. Opin. Virol.* **28**, 20–25 (2018).
26. Awadalla, P. The evolutionary genomics of pathogen recombination. *Nat. Rev. Genet.* **4**, 50–60 (2003).
27. Koonin, E. V. & Yutin, N. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. *Intervirology* **53**, 284–292 (2010).
28. Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479–480**, 2–25 (2015).
29. Filée, J. & Chandler, M. Gene exchange and the origin of giant viruses. *Intervirology* **53**, 354–361 (2010).
30. Sun, T.-W. *et al.* Host range and coding potential of eukaryotic giant viruses. *Viruses* **12**, 1337 (2020).
31. Fraser, M. J., Smith, G. E. & Summers, M. D. Acquisition of host cell DNA sequences by Baculoviruses: Relationship between host DNA insertions and FP mutants of *Autographa californica* and *Galleria mellonella* nuclear polyhedrosis viruses. *J. Virol.* **47**, 287–300 (1983).
32. Liu, H. *et al.* Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. *J. Virol.* **84**, 11876–11887 (2010).
33. Bordenstein, S. R. & Bordenstein, S. R. Eukaryotic association module in phage WO genomes from *Wolbachia*. *Nat. Commun.* **7**, 13155 (2016).
34. Richards, T. A., Hirt, R. P., Williams, B. A. P. & Embley, T. M. Horizontal gene transfer and the evolution of parasitic protozoa. *Protist* **154**, 17–32 (2003).
35. Hayward, A., Cornwallis, C. K. & Jern, P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 464–469 (2015).
36. Cock, J. M. *et al.* The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* **465**, 617–621 (2010).
37. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, (2020).

38. Maumus, F. & Blanc, G. Study of gene trafficking between *Acanthamoeba* and giant viruses suggests an undiscovered family of amoeba-infecting viruses. *Genome Biol. Evol.* **8**, 3351–3363 (2016).
39. Fischer, M. G. & Hackl, T. Host genome integration and giant virus-induced reactivation of the virophage mavirus. *Nature* **540**, 288–291 (2016).
40. Maumus, F., Epert, A., Nogu  , F. & Blanc, G. Plant genomes enclose footprints of past infections by giant virus relatives. *Nat. Commun.* **5**, 4268 (2014).
41. Patel, M. R., Emerman, M. & Malik, H. S. Paleovirology - ghosts and gifts of viruses past. *Curr. Opin. Virol.* **1**, 304–309 (2011).
42. Williams, T. A., Embley, T. M. & Heinz, E. Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large DNA viruses. *PLoS One* **6**, e21080 (2011).
43. Tonetti, M., Sturla, L., Bisso, A., Benatti, U. & De Flora, A. Synthesis of GDP-L-fucose by the human FX protein. *J. Biol. Chem.* **271**, 27274–27279 (1996).
44. Forterre, P., Grihaldo, S., Gadelle, D. & Serre, M. C. Origin and evolution of DNA topoisomerases. *Biochimie* **89**, 427–446 (2007).
45. Malik, S. S., Azem-e-Zahra, S., Kim, K. M., Caetano-Anoll  s, G. & Nasir, A. Do viruses exchange genes across superkingdoms of life? *Front. Microbiol.* **8**, (2017).
46. Gilbert, C. & Cordaux, R. Viruses as vectors of horizontal transfer of genetic material in eukaryotes. *Curr. Opin. Virol.* **25**, 16–22 (2017).
47. Chen, J. *et al.* Genome hypermobility by lateral transduction. *Science* **362**, 207–212 (2018).
48. Touchon, M., Moura de Sousa, J. A. & Rocha, E. P. Embracing the enemy: The diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Curr. Opin. Microbiol.* **38**, 66–73 (2017).
49. Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E2401–E2410 (2017).
50. Moreira, D. & L  pez-Garc  a, P. Ten reasons to exclude viruses from the tree of life. *Nat. Rev. Microbiol.* **7**, 306–311 (2009).
51. Almagro Armenteros, J. J., S  nderby, C. K., S  nderby, S. K., Nielsen, H. & Winther, O. DeepLoc: Prediction of protein subcellular localization using deep learning.

- Bioinformatics* **33**, 3387–3395 (2017).
52. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: Patterns and determinants. *Nat. Rev. Genet.* **9**, 267–276 (2008).
  53. Jung, J. U., Stäger, M. & Desrosiers, R. C. Virus-encoded cyclin. *Mol. Cell. Biol.* **14**, 7235–7244 (1994).
  54. Erives, A. J. Phylogenetic analysis of the core histone doublet and DNA topo II genes of *Marseilleviridae*: Evidence of proto-eukaryotic provenance. *Epigenetics and Chromatin* **10**, 55 (2017).
  55. Schvarcz, C. R. & Steward, G. F. A giant virus infecting green algae encodes key fermentation genes. *Virology* **518**, 423–433 (2018).
  56. Pagarete, A., Allen, M. J., Wilson, W. H., Kimmance, S. A. & De Vargas, C. Host-virus shift of the sphingolipid pathway along an *Emiliania huxleyi* bloom: Survival of the fattest. *Environ. Microbiol.* **11**, 2840–2848 (2009).
  57. Schneider-Schaulies, J. & Schneider-Schaulies, S. Sphingolipids in viral infection. *Biol. Chem.* **396**, 585–595 (2015).
  58. Fernández De Castro, I., Tenorio, R. & Risco, C. Virus assembly factories in a lipid world. *Curr. Opin. Virol.* **18**, 20–26 (2016).
  59. Hiramatsu, S., Ishihara, M., Fujie, M. & Usami, S. Expression of a chitinase gene and lysis of the host cell wall during *Chlorella* virus CVK2 infection. *Virology* **260**, 308–315 (1999).
  60. Klenk, H. D. & Garten, W. Host cell proteases controlling virus pathogenicity. *Trends Microbiol.* **2**, 39–43 (1994).
  61. Thaker, S. K., Ch'ng, J. & Christofk, H. R. Viral hijacking of cellular metabolism. *BMC Biol.* **17**, 59 (2019).
  62. Mahalingam, S., Meanger, J., Foster, P. S. & Lidbury, B. A. The viral manipulation of the host cellular and immune environments to enhance propagation and survival: A focus on RNA viruses. *J. Leukoc. Biol.* **72**, 429–439 (2002).
  63. Ravindran, M. S., Bagchi, P., Cunningham, C. N. & Tsai, B. Opportunistic intruders: How viruses orchestrate ER functions to infect cells. *Nat. Rev. Microbiol.* **14**, 407–420 (2016).
  64. de Chassey, B., Meyniel-Schicklin, L., Vonderscher, J., André, P. & Lotteau, V. Virus-host interactomics: New insights and opportunities for antiviral drug discovery. *Genome*

*Med.* **6**, 115 (2014).

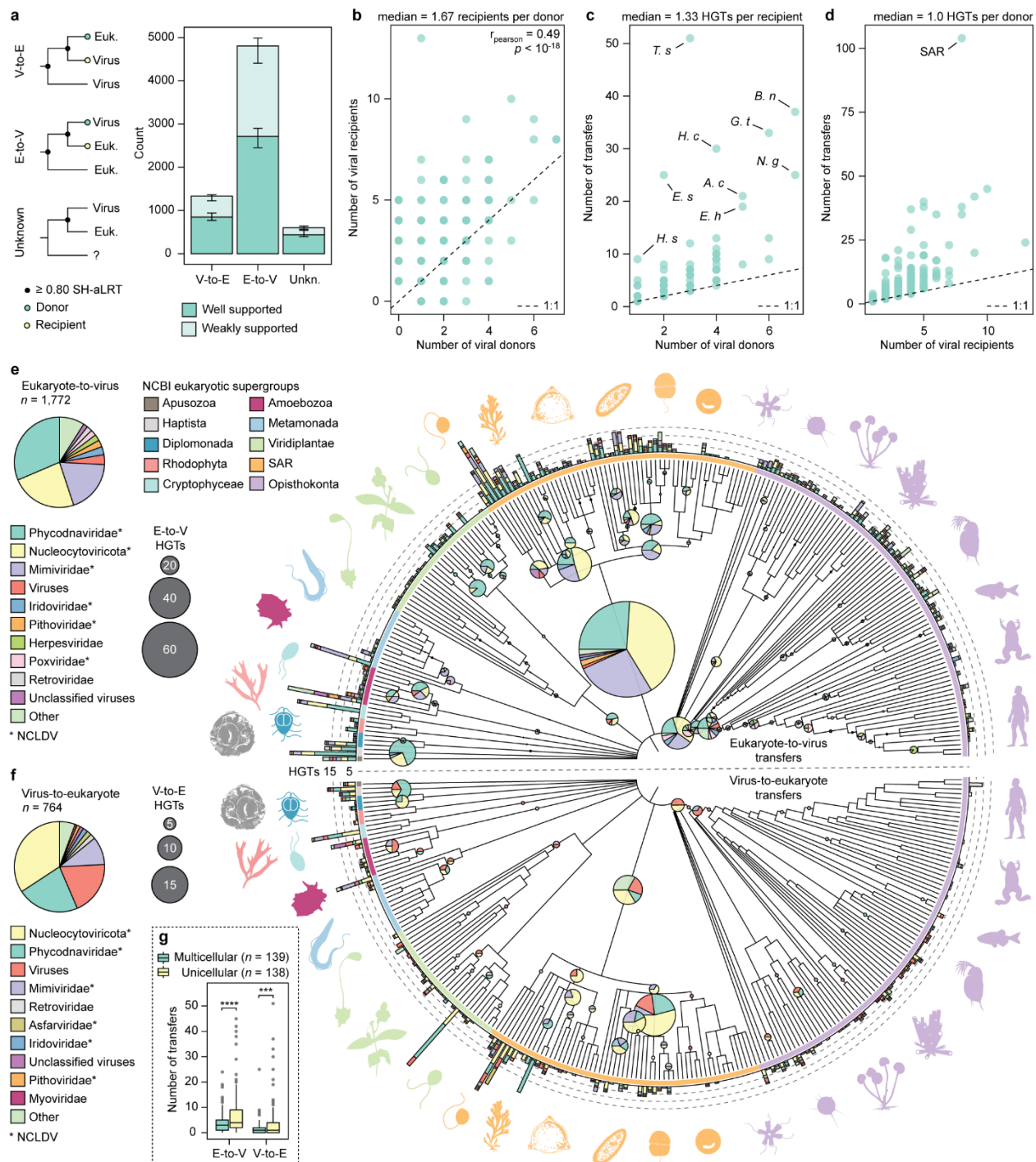
65. Puschnik, A. S., Majzoub, K., Ooi, Y. S. & Carette, J. E. A CRISPR toolbox to study virus-host interactions. *Nat. Rev. Microbiol.* **15**, 351–364 (2017).
66. Michel, G., Tonon, T., Scornet, D., Cock, J. M. & Kloareg, B. Central and storage carbon metabolism of the brown alga *Ectocarpus siliculosus*: Insights into the origin and evolution of storage carbohydrates in eukaryotes. *New Phytol.* **188**, 67–81 (2010).
67. Durak, G. M. *et al.* A role for diatom-like silicon transporters in calcifying coccolithophores. *Nat. Commun.* **7**, 10543 (2016).
68. Armstrong, P. B., Armstrong, M. T., Pardy, R. L., Child, A. & Wainwright, N. Immunohistochemical demonstration of a lipopolysaccharide in the cell wall of a eukaryote, the green alga, *Chlorella*. *Biol. Bull.* **203**, 203–204 (2002).
69. Laurent, T. C. & Fraser, J. R. E. Hyaluronan. *FASEB* **6**, 2397–2404 (1992).
70. Loomis, W. F., Wheeler, S. A., Springer, W. R. & Barondes, S. H. Adhesion mutants of *Dictyostelium discoideum* lacking the saccharide determinant recognized by two adhesion-blocking monoclonal antibodies. *Dev. Biol.* **109**, 111–117 (1985).
71. Bandini, G. *et al.* An essential, kinetoplastid-specific GDP-Fuc:  $\beta$ -D-Gal  $\alpha$ -1,2-fucosyltransferase is located in the mitochondrion of *Trypanosoma brucei*. *Elife* **10**, e70272 (2021).
72. Schrom, M. & Bablanian, R. Altered cellular morphology resulting from cytotoxic virus infection. *Arch. Virol.* **70**, 173–187 (1981).
73. Raaben, M. *et al.* The ubiquitin-proteasome system plays an important role during various stages of the coronavirus infection cycle. *J. Virol.* **84**, 7869–7879 (2010).
74. Leier, H. C. *et al.* A global lipid map defines a network essential for Zika virus replication. *Nat. Commun.* **11**, 3652 (2020).
75. Chuong, E. B. The placenta goes viral: Retroviruses control gene expression in pregnancy. *PLoS Biol.* **16**, e3000028 (2018).
76. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
77. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* **12**, e1001889 (2014).

78. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
79. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
80. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2014).
81. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
82. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
83. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
84. Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C. & Gascuel, O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol.* **60**, 685–699 (2011).
85. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116 (1999).
86. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
87. Tria, F. D. K., Landan, G. & Dagan, T. Phylogenetic rooting using minimal ancestor deviation. *Nat. Ecol. Evol.* **1**, 0193 (2017).
88. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
89. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**, 256–259 (2019).
90. Susko, E. & Roger, A. J. On reduced amino acid alphabets for phylogenetic inference.



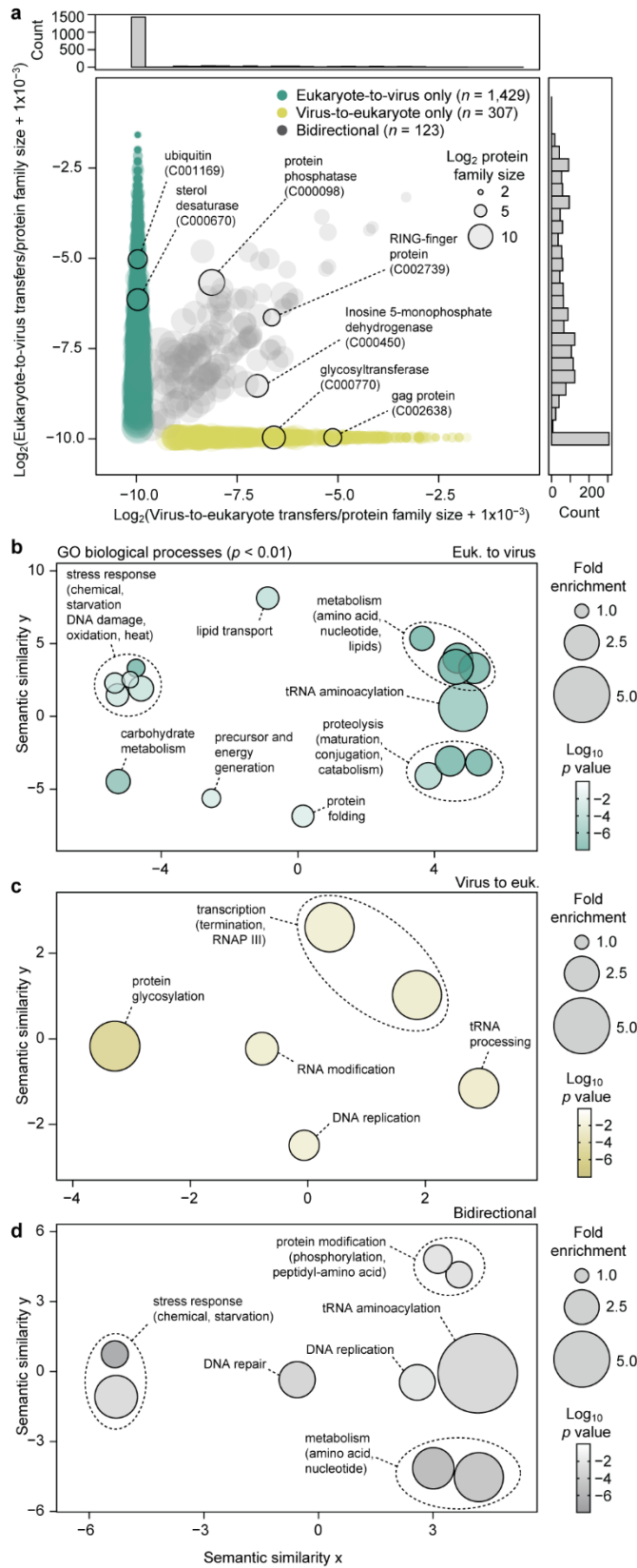
*Mol. Biol. Evol.* **24**, 2139–2150 (2007).

91. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
92. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, 136–143 (2012).
93. Walker, P. J. *et al.* Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Arch. Virol.* **166**, 2633–2648 (2021).
94. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
95. Jones, P. *et al.* InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
96. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).
97. Dean, S., Sunter, J. D. & Wheeler, R. J. TrypTag.org: A trypanosome genome-wide protein localisation resource. *Trends Parasitol.* **33**, 80–82 (2017).
98. Aslett, M. *et al.* TriTrypDB: A functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res.* **38**, 457–462 (2009).
99. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The new tree of eukaryotes. *Trends Ecol. Evol.* **35**, 43–55 (2020).



**Fig. 1. The mode and taxonomic distribution of viral-eukaryotic gene exchange.** **a.** Number of transfers from eukaryotes-to-viruses (E-to-V), viruses-to-eukaryotes (V-to-E), and those with unknown directionality (Unkn.). Recipients and donors were based on the last common ancestor of the recipients and their sister clades. Weakly-supported transfers had long branching or ambiguous participants and were excluded from subsequent analyses (see Methods). Error bars

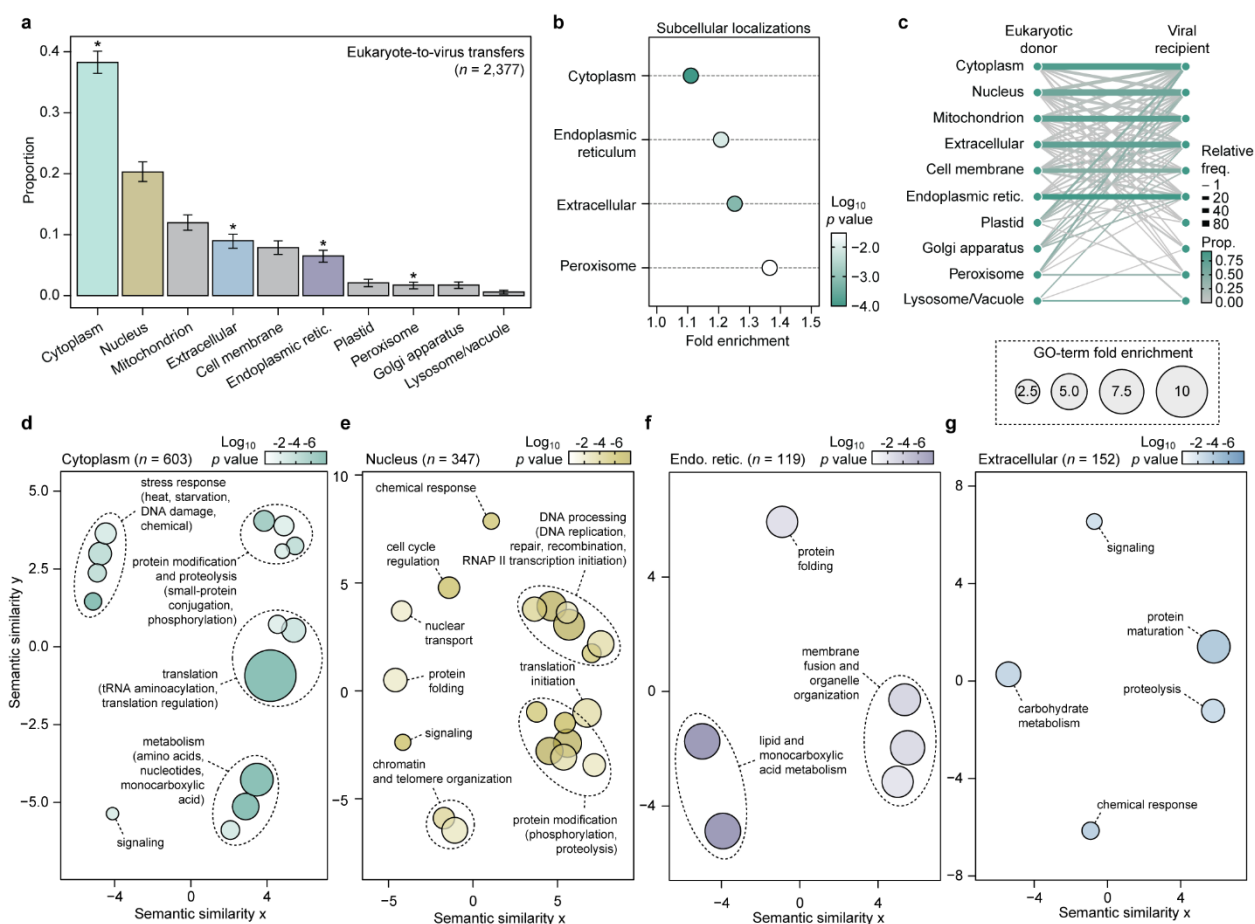
represent 95% confidence intervals estimated from bootstrap pseudo-replicates ( $n = 1,000$ ; random sampling of protein families with replacement). **b-d.** Scatter plots comparing gene exchange statistics. Points represent eukaryotic taxa and dashed lines represent lines of equality. **e, f.** Gene transfers from E-to-V (**e**) and V-to-E (**f**) across a eukaryotic phylogeny. Bar charts represent HGTs present in an individual genome, whereas pie charts present inferred ancestral HGTs. Bar height and pie diameter reflect transfer frequency and colours denote viral taxonomy. For clarity, viral taxa were mapped to their nearest family, phylum, or higher-level classification. Because of this, multiple families from the same phylum are shown, such as the NCLDV lineages which are denoted with an asterisk (note that some unclassified viruses include candidate NCLDV lineages). Eukaryotic supergroups are labeled with a coloured ring and exemplary taxa are represented using Phylopic images (see <http://phylopic.org/>). Taxonomic information and phylogenies are based on the NCBI (National Center for Biotechnology Information) Taxonomy database<sup>92</sup>. Transfers assigned to the last eukaryotic common ancestor are excluded but are listed in Supplementary Table 1. **g.** Boxplots comparing the number of HGTs observed in multicellular and unicellular taxa participating in HGT. The boxes span from the first to third quartile with whiskers extending 1.5 times the interquartile range. Black bars represent the median and  $p$  values were calculated using two-sided Welch's  $t$ -tests (\*\*\*\*:  $p = 5.61 \times 10^{-5}$ , \*\*\*:  $p = 4.36 \times 10^{-4}$ ). Higher level taxa encompassing both uni- and multicellular organisms were omitted. Abbreviations: SAR, Stramenopila-Alveolata-Rhizaria; *H. s.*, *Homo sapiens*; *E. s.*, *Ectocarpus siliculosus*; *T. s.*, *Tetrabaena socialis*; *H. c.*, *Hyphochytrium catenoides*; *A. c.*, *Acanthamoeba castellanii*; *E. h.*, *Emiliania huxleyi*; *G. t.*, *Guillardia theta*; *B. n.*, *Bigelowiella natans*; *N. g.*, *Naegleria gruberi*.



784

785

**Fig. 2. Gene function is related to transfer direction.** **a.** A scatter plot relating the frequency of transfer events to protein families, normalized for family size (number of sequences). Individual points represent protein families and functional annotations for exemplary families are highlighted. Histograms denoting point density along the x- and y-axis are displayed above and to the right of the scatter plot. **b-d.** Scatter plots displaying enriched gene ontology (GO) biological process terms from protein families participating in unidirectional (**b, c**) and bidirectional transfers (**d**), relative to all eukaryotic protein families. Labeling has been summarized for clarity, but complete terms are available in Supplementary Table 3. Semantic similarity was determined using REVIGO<sup>96</sup> and statistical significance was assessed using two-sided permutation tests ( $p < 0.01$ ,  $n = 10^7$ ).



**Fig. 3. Predicted subcellular localizations and functions of eukaryote-derived viral genes. a.**

Proportions of sub-cellular localizations for eukaryote-to-virus HGTs based on the predicted targeting of eukaryotic donor sequences. Asterisks denote statistically significant enrichments ( $p < 0.05$ , see **b**). Error bars represent 95% confidence intervals determined from bootstrap pseudo-replicates ( $n = 1,000$ ; random sampling of HGTs with replacement). **b.** The enrichment of subcellular compartments relative to total eukaryotic proteomes. Significance was assessed using two-sided permutation tests ( $n = 10^6$ ). **c.** A comparison between the predicted localization of proteins from eukaryotic donors and their viral recipients. The relative frequencies and proportions are indicated by edge thickness and colour, respectively. **d-g.** Scatter plots displaying enriched GO biological process terms for families with a given donor localization relative to all eukaryotic protein families for localizations to (from left to right, colour coded as in **a**) the cytoplasm, nucleus, endoplasmic reticulum, and extracellular space. Labeling has been summarized for clarity but complete terms are available in Supplementary Table 4. Semantic

815 similarity was determined using REVIGO and statistical significance was assessed using two-  
816 sided permutation tests ( $p < 0.01$ ,  $n = 10^7$ ).

817

818

819

820

821

822

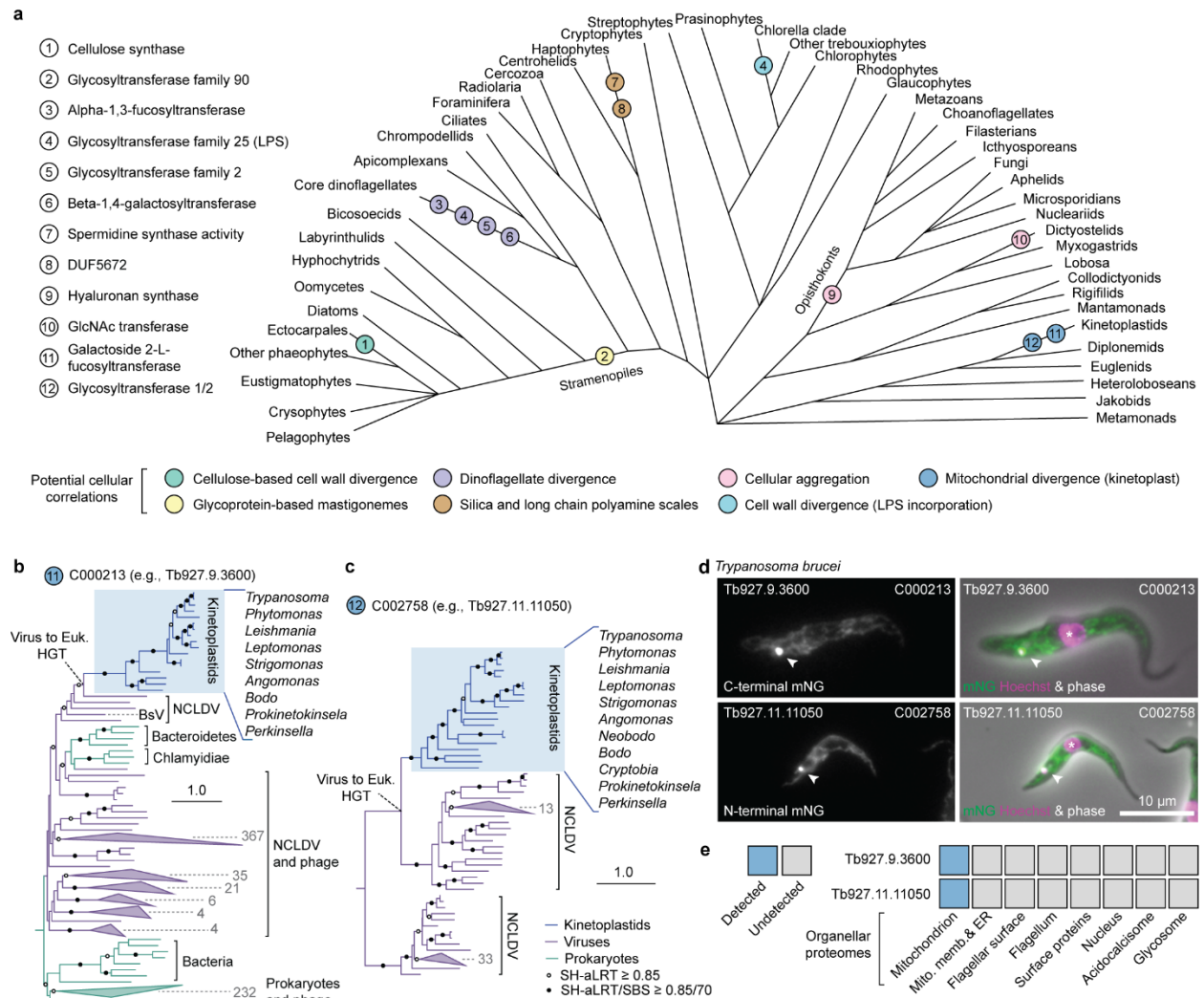
823

824

825

826

827



**Fig. 4. Recurrent acquisition of viral glycosyltransferases across the eukaryotic tree of life.**

**a.** A schematic eukaryotic phylogeny, based on previous phylogenomic studies<sup>99</sup>, with glycosyltransferase acquisitions plotted at ancestral nodes. Protein families and annotations are denoted with numbers and phenotypic correlations are noted with colours. **b, c.** The phylogenetic origin of kinetoplastid glycosyltransferases from NCLDV by HGT. Maximum likelihood phylogenies were generated in IQ-Tree using the LG+F+R10 (**b**) and LG+F+R6 (**c**) substitution models as selected using ModelFinder, and statistical support was assessed using SH-aLRT and standard bootstraps ( $n = 1,000$ )<sup>86,88</sup>. Kinetoplastid genera and the gene identifiers for *Trypanosoma brucei* orthologs are noted. **d.** Viral-derived glycosyltransferases function in highly-derived kinetoplastid mitochondria (kinetoplasts). Representative TrypTag fluorescent micrographs, based on observations of at least twenty individual cells, depicting the localization of both glycosyltransferases after labeling with mNeonGreen (left panel) and Hoechst DNA stain



841 (right panel). White arrowheads denote the kinetoplast. **e.** Organellar proteomic data displaying  
842 the presence or absence of both glycosyltransferases. Proteomic data was assessed using  
843 TriTrypDB.

844