PNAS Nexus, 2024, 3, 1–9
https://doi.org/10.1093/pnasnexus/pgad479
Advance access publication 18 January 2024
Research Report

Minor intron–containing genes as an ancient backbone for viral infection?

Stefan Wuchty[a,b,c,d,*], Alisa K. White[e], Anouk M. Olthof[e], Kyle Drake[e], Adam J. Hume[f,g,h], Judith Olejnik[f,g], Vanessa Aguiar-Pulido[a], Elke Mühlberger[f,g] and Rahul N. Kanadia[e,i,*]

[a]Department of Computer Science, University of Miami, Coral Gables, FL 33146, USA
[b]Department of Biology, University of Miami, Coral Gables, FL 33146, USA
[c]Institute of Data Science and Computing, University of Miami, Coral Gables, FL 33146, USA
[d]Sylvester Comprehensive Cancer Center, University of Miami, Miami, FL 33134, USA
[e]Physiology and Neurobiology Department, University of Connecticut, Storrs, CT 06269, USA
[f]Department of Virology, Immunology and Microbiology, Boston University Chobanian and Avedisian School of Medicine, Boston, MA 02118, USA
[g]National Emerging Infectious Diseases Laboratories, Boston University, Boston, MA 02118, USA
[h]Center for Emerging Infectious Diseases Policy and Research, Boston University, Boston, MA 02118, USA
[i]Institute for Systems Genomics, University of Connecticut, Storrs, CT 06269, USA
*To whom correspondence should be addressed: Email: wuchtys@cs.miami.edu (S.W.); Email: rahul.kanadia@uconn.edu (R.N.K.)

Edited By: Christopher Dupont

Abstract

Minor intron–containing genes (MIGs) account for <2% of all human protein–coding genes and are uniquely dependent on the minor spliceosome for proper excision. Despite their low numbers, we surprisingly found a significant enrichment of MIG-encoded proteins (MIG-Ps) in protein–protein interactomes and host factors of positive-sense RNA viruses, including SARS-CoV-1, SARS-CoV-2, MERS coronavirus, and Zika virus. Similarly, we observed a significant enrichment of MIG-Ps in the interactomes and sets of host factors of negative-sense RNA viruses such as Ebola virus, influenza A virus, and the retrovirus HIV-1. We also found an enrichment of MIG-Ps in double-stranded DNA viruses such as Epstein–Barr virus, human papillomavirus, and herpes simplex viruses. In general, MIG-Ps were highly connected and placed in central positions in a network of human–host protein interactions. Moreover, MIG-Ps that interact with viral proteins were enriched with essential genes. We also provide evidence that viral proteins interact with ancestral MIGs that date back to unicellular organisms and are mainly involved in basic cellular functions such as cell cycle, cell division, and signal transduction. Our results suggest that MIG-Ps form a stable, evolutionarily conserved backbone that viruses putatively tap to invade and propagate in human host cells.

Keywords: minor introns, viral infections

Significance Statement

Through a computational analysis of the role of minor intron–containing genes (MIGs), we found that the corresponding proteins were enriched in molecular interactomes and sets of host factors of a wide variety of viruses. Notably, MIGs are essential and highly ancestral (i.e. dating back to unicellular organisms), providing various basic cellular functions. Our observations suggest that MIG-encoded proteins may form a stable, evolutionarily conserved framework that viruses in general could tap into for their propagation.

Introduction

Based on the consensus sequences at the 5′ splice sites (SSs), branch point sequences, and the 3′ SSs, introns in human genes are classified as major (>99%) or minor (<1%) introns (1, 2). Major introns are spliced by the highly abundant and ubiquitous major spliceosome (3) that is composed of small nuclear RNAs (snRNAs: U1, U2, U4, U5, and U6) and over 150 associated proteins. In contrast, minor introns are spliced by the less abundant minor spliceosome (4) consisting of snRNAs (U11, U12, U4atac, U5, and U6atac) and many proteins that are shared with the major spliceosome, as well as 13 unique proteins (3, 5). Notably, phylogenetic reconstructions of the minor and major spliceosomal snRNAs and proteins indicated that both spliceosomes existed in the last eukaryotic common ancestor (5, 6). Furthermore, minor and major introns emerged roughly at the same time and have been conserved in eukaryotic genomes (7–10). While minor intron–containing genes (MIGs) execute disparate functions across various biological pathways (5), the reasons why a specific gene

Competing Interest: The authors declare no competing interest.
Received: March 27, 2023. **Accepted:** December 15, 2023

acquired a minor intron that persisted across evolution remain unclear. We have previously reported that MIGs are enriched in genes that are essential for survival (5), which ensured their maintenance across evolution.

The significance of roughly 700 MIGs in mammalian genomes and their role in biological functions has emerged through studies of human diseases that are linked to minor spliceosome components, or a targeted deletion of minor spliceosome components in mouse and other model systems (11). We and others have shown that inhibition of the minor spliceosome primarily results in either elevated retention of minor introns or alternative splicing around minor introns that is executed by the major spliceosome (11, 12). Unexcised minor introns in such transcripts are either removed by exosome degradation and nonsense-mediated decay or could potentially encode a truncated protein (11). Regardless, the original function of MIG-encoded proteins (MIG-Ps) including DNA transcription, replication and repair, RNA processing and translation, as well as cytoskeletal organization and vesicular transport, is lost upon inhibition of the minor spliceosome (2, 3, 5).

The role of the minor spliceosome and altered function of MIG-Ps has been discovered in autoimmune, congenital, and degenerative neurological diseases as well as cancer (13). However, little is known about the role of the minor spliceosome and MIG-Ps in viral infection. Since viruses are obligate intracellular parasites, and MIG-Ps appear in various biological functions including RNA and DNA processing, we hypothesized that viruses strongly rely on MIG-Ps to replicate in infected cells. Initial support for our hypothesis comes from the potential role of the minor spliceosome components, including *RNU4atac*, *snRNP25*, and *SNRPD2*, in reversing the block to HIV multiple splicing (14).

Through computational data analysis, we investigated the enrichment of MIG-Ps by considering large-scale screens of human proteins that interact with proteins of viral pathogens of major health concerns such as SARS-CoV-1, SARS-CoV-2, MERS coronavirus (MERS-CoV), Zika virus (ZIKV), hepatitis C virus (HCV), human immunodeficiency virus 1 (HIV-1), Ebola virus (EBOV), influenza A virus (IAV), Epstein–Barr virus (EBV), human papilloma virus 1 (HPV-1), and herpes simplex virus 1 (HSV-1). Despite the relatively small proportion of MIGs in the human genome, we show that MIG-Ps were significantly enriched among proteins that interact with viral proteins. Furthermore, MIG-Ps were also enriched among host factors that do not directly interact with viral proteins but are known to allow viruses to tap the cellular machinery of the host. Furthermore, we found that MIG-Ps that interact with viral proteins were mostly involved in elementary and essential cellular functions, dating back to unicellular organisms. In conclusion, our results suggest that viruses have generally evolved to access an evolutionary conserved backbone, presumably to tap elementary cellular host functions. Furthermore, our computational work sets the foundation for future experimental work that further explore the role of MIGs in viral infections.

Results

MIGs are enriched in human proteins that interact with SARS-CoV-2 and other viruses

To investigate the role of MIG-Ps and the minor spliceosome during viral infections, we initially focused on SARS-CoV-2. Considering the first published SARS-CoV-2 host protein–protein interaction map (15), we surprisingly found that 20 out of 332 human host proteins that interacted with SARS-CoV-2 proteins were MIG-Ps (5). Notably, these 20 human MIG-Ps were involved in every stage of the viral replication cycle and executed disparate biological roles that SARS-CoV-2 uses to access various cellular pathways (Fig. 1A). Utilizing a hypergeometric test, we observed that the enrichment of the 20 MIG-Ps was statistically significant ($P = 5.1 \times 10^{-3}$), suggesting that the biological functions of MIG-Ps might be tapped by SARS-CoV-2. To corroborate our observation, we sampled 1,000 randomized lists of 332 protein-coding genes and found that a median of 10 MIG-Ps would occur by chance, indicating that the presence of 20 MIG-Ps in the SARS-CoV-2 interactome was well above background. Furthermore, we generated a randomized list of 332 protein-coding genes that initially included 0% MIGs, while MIGs were added progressively in increments of 10%. We found statistical significance at $\alpha < 0.05$ when 90% MIGs (i.e. 18 MIGs) were part of the list of 332 targeted genes ($P = 0.0210$), supporting our conclusion that the presence of 20 MIGs in the list of 332 human proteins that interacted with proteins of SARS-CoV-2 was statistically significant. However, the selection of a set of genes that possess minor intron(s) may introduce unexpected bias in our enrichment analysis. To explore if we would find similar enrichments in a list of genes with another intron feature, we selected a similarly sized set of 685 genes with introns that contain microRNAs (miRNAs). Our analysis showed that there was no significant enrichment of miRNA-containing genes among proteins that interacted with proteins of SARS-CoV-2 (Fig. 1B). Specifically, our results suggest that MIG-Ps did not occur randomly in the interactome of SARS-CoV-2 but point to host proteins that SARS-CoV-2 might tap into for viral replication.

Given this unique role that MIG-Ps play for SARS-CoV-2, a member of the large group of positive-sense RNA viruses, we hypothesized that MIG-Ps might show similar enrichment signals in the interactomes of other positive-sense RNA viruses, prompting us to extend our analysis to SARS-CoV-1, MERS-CoV, ZIKV, and HCV. Sampling 1,000 randomized lists for each of these viral protein–protein interactions (PPI), we found a distinct number of MIGs that would occur in the interactome of SARS-CoV-1 (12), MERS-CoV (9), ZIKV (16), and HCV (17) by chance (Fig. S1). While not significantly enriched for HCV ($P = 0.3138$), MIG-Ps were significantly enriched in the interactomes of SARS-CoV-1 ($P = 3.4 \times 10^{-3}$), MERS-CoV ($P = 2.0 \times 10^{-4}$), and ZIKV ($P = 0.0146$; Fig. 1C). Similar to SARS-CoV-2, we did not find any significant enrichment signals of genes with miRNA-containing introns in these viral interactomes (Fig. S2).

Next, we sought to investigate whether MIG-Ps were generally enriched in the interactomes of viruses beyond positive-sense RNA viruses, including HIV-1, a retrovirus, EBOV, and IAV as representatives of negative-sense RNA viruses, and EBV, HPV-1, and HSV-1 as representatives of double-stranded DNA viruses. Notably, we found a significant enrichment of MIG-Ps in the interactome of all viruses (HIV-1, EBOV, IAV [$P < 0.01$]; EBV [$P = 0.0604$]; HPV-1 [$P < 10^{-4}$]; HSV-1 [$P = 0.0053$]; Fig. 1C). With the exception of HIV-1 and IAV ($P < 0.01$), we did not observe significant enrichment for miRNA-containing genes within the interactome of the other analyzed viruses (Fig. S2).

MIG-Ps are enriched in host factors

The significantly increased presence of MIG-Ps in virus–host protein interactions prompted us to determine whether MIG-Ps were also enriched in sets of viral host factors. Here, we utilized data from genome-wide screens of viral host factors that are human proteins which are essential for virus replication, as demonstrated by genome-wide knockout screens (18). Applying our

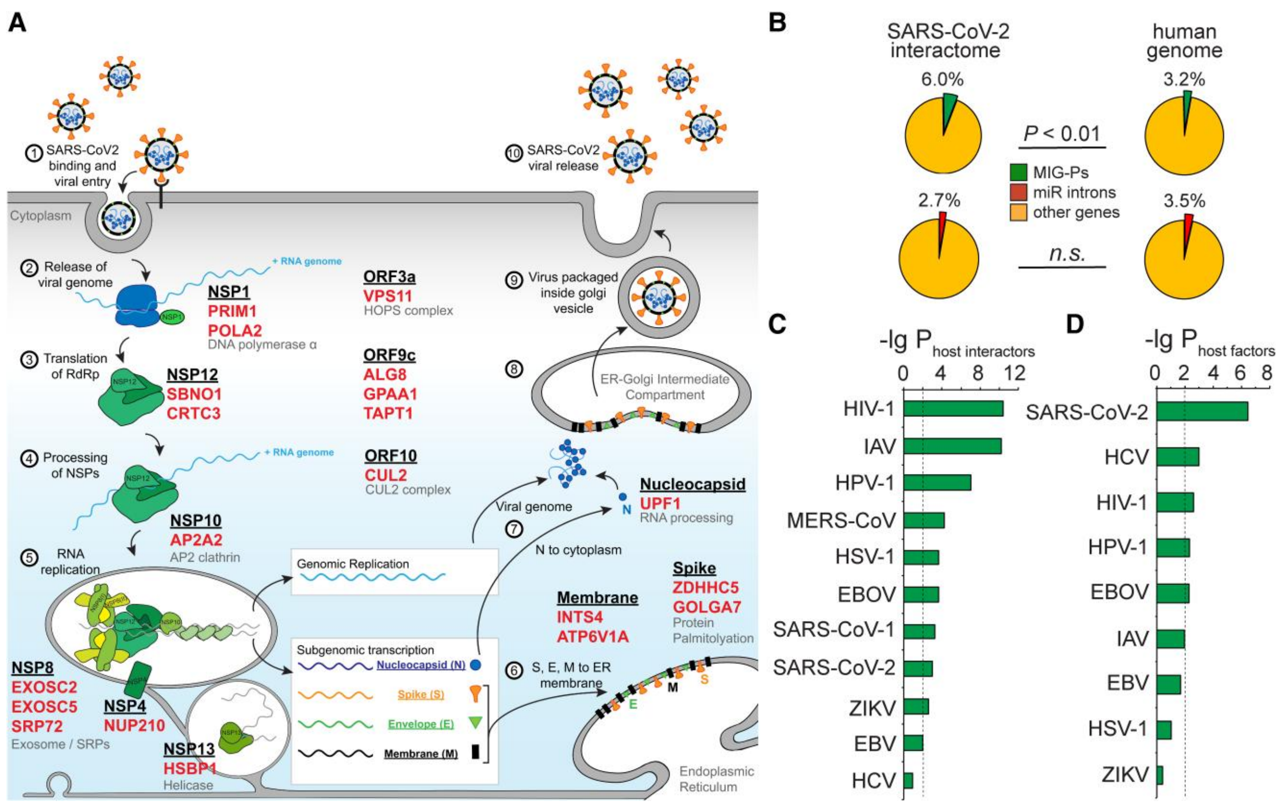

Fig. 1. Enrichment of MIG-Ps among human host proteins that interact with viral targets and host factors. A) A scheme of the SARS-CoV-2 replication cycle indicating viral proteins (black, underlined) that target MIG-Ps (red) as initially reported in the study by Gordon et al. (15). B) Comparison of the proportion of MIG-Ps and genes with miRNA-containing introns in the SARS-CoV-2 interactome with their proportion in the human genome employing hypergeometric tests. C) Significant enrichment of MIG-Ps in multiple sets of human proteins that are targeted by viral proteins. D) Significant enrichment of MIG-Ps in sets of host factor genes that are required by different viruses to infect their host cells. The statistical significance of enrichment in C and D) is established by performing hypergeometric tests. The dashed lines indicate a $P < 0.01$ threshold.

quantitative enrichment framework, we found a significant enrichment of MIG-Ps in host factors of SARS-CoV-2, HIV-1, HPV-1, IAV, EBOV, and HCV, while there was no significant enrichment of host factors when we considered EBV, ZIKV, or HSV-1 (Fig. 1D).

MIG-Ps in the vicinity of viral targets

Given that only a handful of MIG-Ps interact with viral proteins or serve as viral host factors, we hypothesized that the remaining MIG-Ps could potentially appear downstream of the viral interactome and host factors. To investigate this conjecture, we utilized a network of human PPI (19) and calculated the shortest distance of each human protein to the nearest human protein that interacted with a viral protein or served as a host factor. As a result, proteins that directly interacted with a viral protein or host factors occurred at a distance of $d = 0$, while a protein that interacted through a single intermediary protein appeared at a distance of $d = 1$. To determine whether MIG-Ps were enriched within groups of proteins that were at a given distance away from the viral interactors or host factors in the underlying human protein interaction network, we performed a permutation analysis combining all viral protein targets and host factors, respectively. Specifically, we randomly sampled 10^5 sets of MIG-Ps from all human proteins that appeared in the underlying interaction network. Calculating the lg_2 of the ratio of the observed and expected numbers of MIG-Ps as a function of distance d, we found that MIG-Ps were significantly enriched at a distance of $d = 1$ compared with non-MIG-Ps. Vice versa, MIG-Ps were found depleted at further distances ($d \geq 2$; Fig. S3), suggesting that MIG-Ps also appeared immediately downstream of proteins that interacted with viral proteins or served as host factors.

Network characteristics of MIG-Ps

To further characterize MIG-Ps topologically, we investigated their placement in the underlying human protein interaction network, given that viral proteins tend to interact with highly connected host proteins (20–22). We binned proteins according to their number of interactions (23), randomly sampled 10^5 sets of MIG-Ps out of all proteins in the underlying interaction network, and determined the enrichment of MIG-Ps in the different bins of connectivity. As shown in Fig. 2A, we observed that MIG-Ps preferably occurred in bins of highly interacting proteins. As a different measure that captures the level of a node's global centrality in the underlying human protein interaction network, we considered a protein's betweenness centrality, reflecting a protein's propensity to increasingly appear in shortest paths between all protein pairs (24). Sorting all proteins according to their betweenness centrality, we defined the top 20% of the proteins with the highest betweenness centrality as a set of bottlenecks (23, 25) and hypothesized that MIG-Ps were preferentially bottleneck proteins. Randomly sampling sets of MIG-Ps out of all proteins in the human interaction network, we calculated the lg_2 of the ratio of the observed and expected numbers of MIG-Ps in the set of bottleneck nodes and, indeed, found that MIG-Ps were significantly enriched as bottlenecks (Fig. 2B).

As another indicator of the central topological role of proteins, we determined control nodes (26, 27) in a network of directed

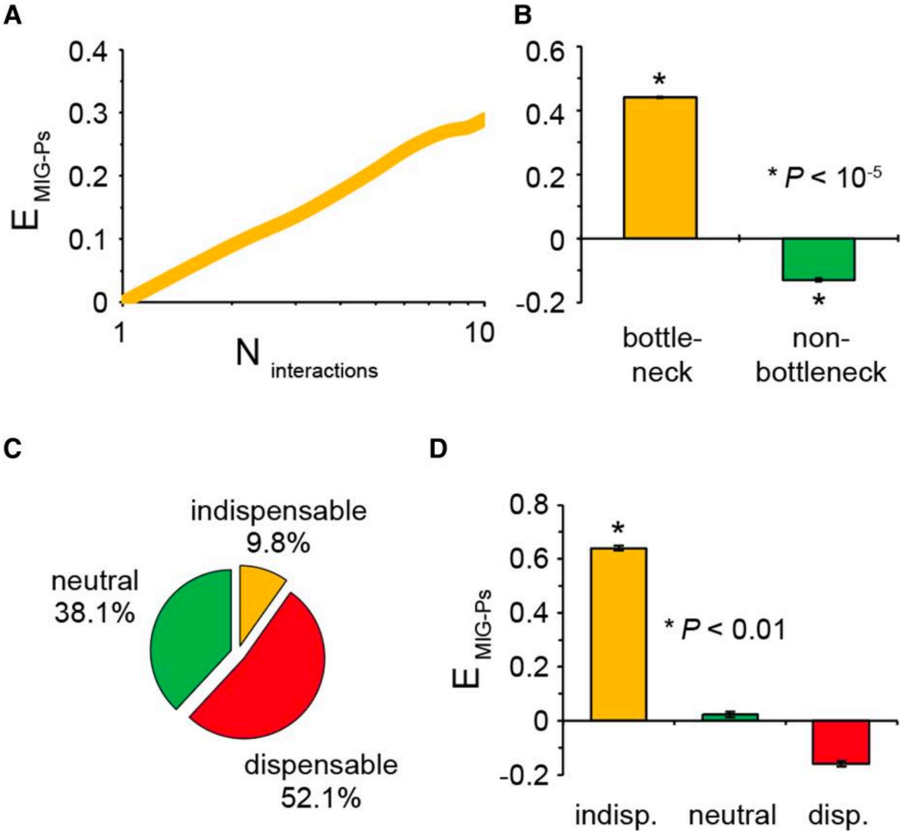

Fig. 2. MIG-Ps are topologically central in networks of human protein interactions. A) Sets of MIG-Ps were randomly sampled out of all proteins in a human protein–protein interaction network, and the enrichment of MIG-Ps in the different bins of connectivity was determined. MIG-Ps were increasingly enriched in bins of highly interacting proteins. B) As a measure of global connectedness in a network, we considered a protein's betweenness centrality in the underlying protein interaction network. The top 20% of the proteins with the highest betweenness centrality were defined as bottlenecks. These proteins were significantly enriched with MIG-Ps. C) The pie chart suggests that <10% of the proteins were indispensable, while the remaining 90% nodes were neutral or dispensable for the control of the underlying interaction network. D) Randomly sampling such control-specific labels revealed that MIG-Ps were significantly enriched with indispensable control proteins, while the opposite was observed when neutral and dispensable control nodes were considered.

protein interactions that we obtained from parsing pathways from the KEGG database (16), capturing protein and regulatory interactions. Specifically, we determined driver nodes that were sufficient for the structural controllability of linear dynamics of the underlying network (see Materials and methods). To establish their importance for the controllability of the underlying network, we deleted each node and determined the number of driver nodes in the perturbed network separately. If the number of driver nodes increased, the underlying protein was considered indispensable for network control. Conversely, nodes were considered neutral or dispensable for network control if the number of driver nodes remained the same or decreased in the network upon deletion of the nodes in question. Our analysis revealed that roughly 10% of the proteins in the underlying network were indispensable (Fig. 2C), suggesting that MIG-Ps may be enriched with control relevant proteins. Randomly sampling 10^5 sets of MIG-Ps out of all proteins in the underlying directed interaction network, we observed that MIG-Ps in general were significantly enriched with indispensable nodes, while we found the opposite when we considered neutral or dispensable proteins (Fig. 2D).

Central placement in interaction networks also comes with an increased propensity that these proteins are essential. Indeed, previous analyses of human–viral protein interactomes indicated that viral proteins preferentially interact with essential human host proteins (20). Since our analyses showed that MIG-Ps were significantly enriched in sets of host interactors and host factors of individual viruses, we next explored whether each of the analyzed viruses targets host genes that are involved in essential cellular functions. To test this hypothesis, we used a collection of genes from 342 cell lines that were identified as essential for cell survival (28) when knocked out through a genome-wide CRISPR/Cas9 screen. Our enrichment analysis, indeed, revealed that enrichment levels were higher among viral host interactors compared with the enrichment of all MIG-Ps (Fig. 3A). We obtained similar results when we analyzed the enrichment of MIG-P host factors in this set of essential genes (Fig. 3B).

Expression of MIGs remains unchanged upon SARS-CoV-2 infection

Assuming that MIG-Ps might play a vital role in virus replication, we hypothesized that the expression levels of MIGs would not change upon virus infection to facilitate a stable environment for viral replication. Examining the transcriptomic profiles in SARS-CoV-2, MERS-CoV, and SARS-CoV-1 infections (29), we indeed found that MIG-Ps that interacted with viral proteins remained mostly unchanged (i.e. nondifferentially expressed) over the course of infection (Fig. S4). Notably, this is markedly different compared with the backdrop of all genes in cells infected with SARS-CoV-2 that showed a considerable spread of expression change during viral infection.

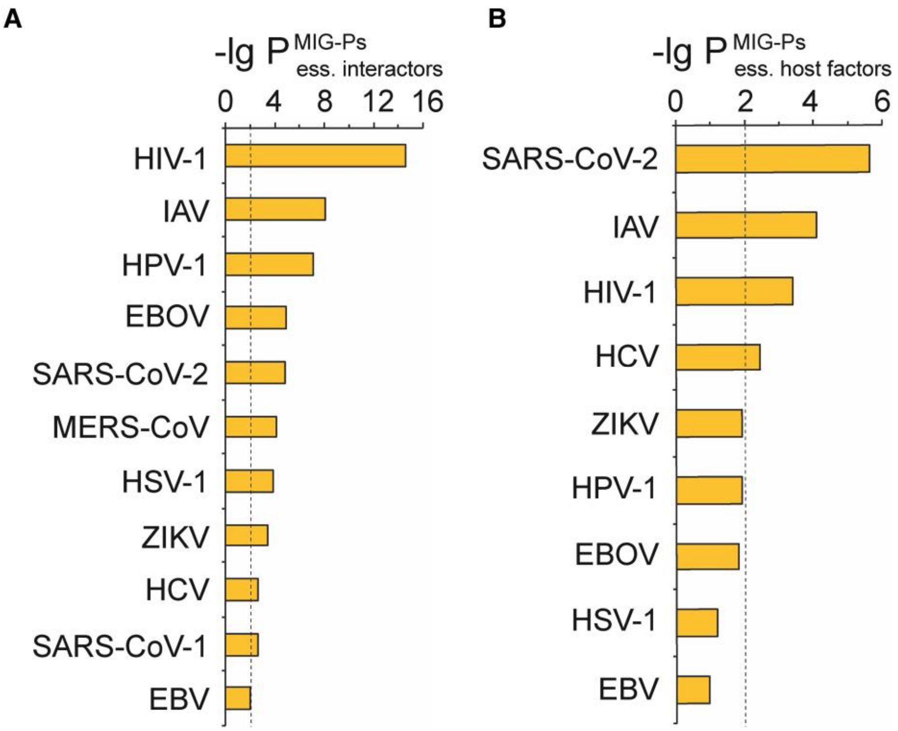

Fig. 3. Essential MIG-Ps are viral targets. A) Essential MIG-Ps were significantly enriched in sets of human proteins that interacted with proteins of a variety of viruses. B) Similarly, yet less frequently, essential MIG-Ps were enriched in sets of host factor genes that viruses require to infect their host cells. The statistical significance of enrichment in A and B) is established by performing hypergeometric tests. The dashed lines indicate a $P < 0.01$ threshold.

MIG-Ps date back to unicellular organisms

The observation that viruses specifically target essential human host proteins that carry out fundamental cellular functions (20) led us to hypothesize that this might be a consequence of evolutionary adaptation, as viruses coevolve with their hosts. To evolutionarily classify human genes, we obtained the point of their emergence in evolutionary history by phylostratigraphy (30). In this analysis, human genes were classified into 16 phylostrata, representing their major evolutionary innovations based on the most distant species with a clear ortholog. According to Trigos et al. (30), human genes assigned to phylostrata 1–3 date back to unicellular ancestors (UC genes), while genes assigned to later phylostrata (4–15, 18) emerged in multicellular ancestors (MC genes). Out of 17,318 genes, 6,684 were classified as UC genes (38.6%), while 10,374 (61.4%) were considered MC genes (inset, Fig. 4A) (30). Intriguingly, the ratio between UC and MC genes changed when we focused on MIGs. Our analysis shows that 70.9% of all MIGs emerged from UC genes, suggesting that MIGs were enriched among UC genes ($P < 10^{-20}$, Fisher's exact test). When we randomly sampled UC and MC genes out of all phylostratified genes, we observed a significant enrichment of MIGs among UC genes (Fig. 4A), while MIGs were diluted among MC genes. Together, our data suggest that MIGs that were classified as UC genes might correspond to elementary, evolutionarily conserved cellular functions. To obtain a high-level summary of biological functions, we resorted to Gene Ontology (GO) Slim terms, which are cut-down versions of the original GO ontologies and provide a broad overview of the ontology content without the details of the specific, fine-grained terms. Functionally analyzing GO Slim terms of UC MIGs using GOATools (31), we found that these genes were significantly involved in cell cycle and division as well as signal transduction functions (hypergeometric tests, false discovery rate (FDR) < 0.05; Fig. 4B). Based on these results, we hypothesized that host interactors and host factors might be potentially enriched with UC genes and determined the enrichment of ancestral UC and MC MIG-Ps among host interactors and host factors. In support of our hypothesis, our results indicate that UC MIG-Ps were more frequently enriched significantly with proteins that interacted with viral proteins and were host factors compared with MC MIG-Ps (Fig. 4C and D).

Discussion

Our results demonstrate that MIG-Ps were significantly enriched in host virus interactomes of various viruses and were also enriched among host factors that play a role in virus replication but do not necessarily interact with viral proteins (20). Although our analyses were restricted to statistical analysis of data sets of viruses for which large-scale data exist, our observations allow us to formulate hypotheses of the putative roles of MIG-Ps in viral infections that can be experimentally probed. Notably, we performed our analyses with a range of different viruses, including positive-sense and negative-sense RNA viruses, retroviruses, and DNA viruses, and obtained the same results across all virus groups.

Our results suggest that MIG-Ps might play a universal role in viral infections, allowing viruses to tap cellular host functions for their propagation. Given that MIG-Ps were enriched among cellular proteins that interact with viral proteins, we surmise that viruses might leverage MIG-Ps as a consequence of their unusually high degree of conservation (7, 32) as the origin of MIG-Ps dates back to unicellular organisms. Furthermore, MIG-Ps have previously been categorized as "information-processing genes" (9), covering common RNA, DNA, vesicular transport, and signaling functions (33). Specifically, MIG-Ps that date back to unicellular organisms capture such functions. Since viruses and their hosts

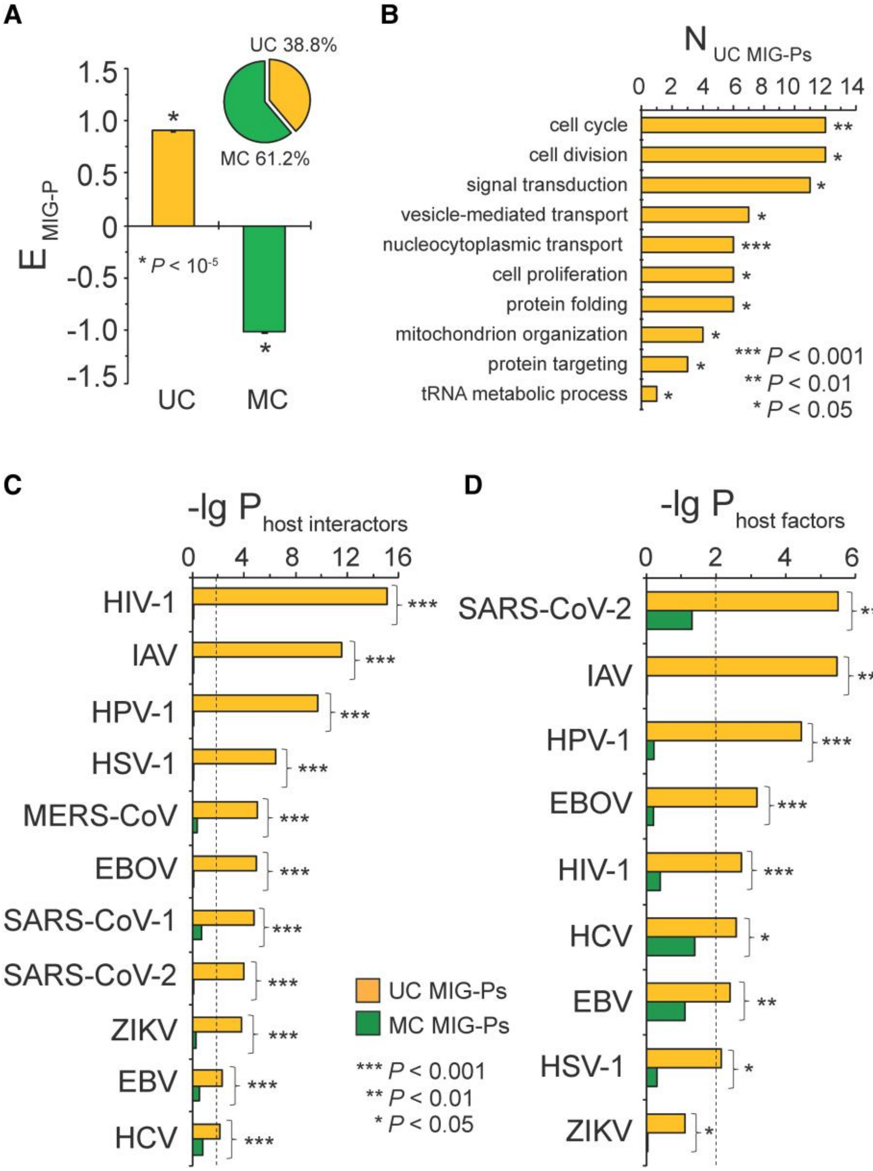

Fig. 4. MIG-Ps that interact with viral proteins date back to unicellular organisms. A) The inset shows the percentage of human genes that date back to MC or UC origin. MIG-Ps were enriched among genes of UC origin and diluted among gens that first appeared in MC organisms. B) Regarding biological processes, MIG-Ps were mostly enriched in pathways that regulate the elementary functions of unicellular cells. MIG-Ps of UC origin were more frequently enriched among proteins that interacted with viral proteins C) and host factors D) compared with MIG-Ps that date back to MC organisms. The statistical significance of enrichment in C and D) is established by performing hypergeometric tests. The dashed lines indicate a $P < 0.01$ threshold. The significance of enrichment differences is established by proportion tests.

have coevolved in a constant biological arms race (17), our data support the idea that viruses may preferentially leverage evolutionarily conserved genes that are required for basic cellular functions as these genes are not given much leeway to escape viral hijacking through mutation. This conclusion is based on a significant enrichment of genes of unicellular origin among human proteins that interact with viral proteins or were host factors, while genes of multicellular origin were significantly diluted in these groups. Notably, the observed enrichment was enforced when we focused on MIG-Ps, presumably pointing to a pan-viral backbone closely, which is tied to the evolutionary history of genes and provides pan-viral entry points to tap elementary and fundamental cellular processes.

It is also conceivable that MIG-Ps serve as a cellular defense mechanism that viruses need to interrupt. In particular, apoptosis of an infected host cell has been identified as a powerful mechanism to curtail viral spread, prompting viruses to evolve strategies to subvert host cell apoptotic defenses (34, 35). When we focused on 1,037 genes that are involved in the apoptotic process as of the GO database (36), we found that MIG-Ps that interact with viral proteins were indeed enriched with genes that are involved in cell death functions ($P = 8.2 \times 10^{-5}$, Fisher's exact test), lending credence to the idea that viruses may target MIG-Ps to evade shutdown of their host cells.

Independently of the functional roles MIG-Ps might play in the infection process, we also showed that the expression levels of MIG-Ps during viral infections hardly changed, putatively indicating an essential role of these proteins in elementary cellular functions. A stable expression of essential cellular proteins might help to provide a supportive environment for viruses to mediate the replication and production of progeny viruses. Recent evidence points to the role of the minor spliceosome in posttranscriptional regulation, as intron retention leads to the degradation of MIG transcripts by nonsense-mediated decay, alternative splicing of the transcript, or translation into a truncated protein (11). According to this mechanistic concept, the overarching function of minor introns in MIGs could be to act as "molecular switches" that regulate protein expression (37). Notably, we found that MIG-Ps were enriched in sets of virus-specific host factors that are essential for virus replication as shown by knockout studies. As a reverse conclusion, it is tempting to speculate that "turning off" MIG-Ps through inhibition of the minor spliceosome might lead to a pan-viral therapeutic strategy to combat viruses effectively.

While such interpretations of our results appear plausible, we stress that our observations were exclusively based on computational data analyses and do not provide experimental validation. While more experimental work needs to be done to determine the role of minor intron retention in the regulation of MIG expression and its relevance for viral infections, our work should be considered as a first step toward understanding the role of MIG-Ps, building the foundation for future in vitro and in vivo studies.

Materials and methods

Determination of the essentialome

Meyers et al. (28) identified genes that were essential for the survival of 342 rapidly dividing cancer cell lines through a genome-wide CRISPR/Cas9 screen. In accordance with recommendations from the Broad Institute, we removed the PK59_PANCREAS cell line from our analysis, since it failed subsequent quality controls (38). Classification of the remaining 341 cell lines by cancer origin/type was performed based on the cell line data provided in Supplementary Table 1 by Meyers et al. (28). Based on their thresholding, we identified 4,360 genes in this essentialome (5).

Phylostratigraphy of human genes

Trigos et al. (30) mapped a total of 17,318 human genes to a phylogenetic tree that captured 16 clades (i.e. phylostrata), ranging from all cellular organisms (phylostratum 1) to homo sapiens (phylostratum 16). To classify human genes, they assigned the most ancient phylostratum to a gene if its orthologs appeared in the most distant species using the OrthoMCL database version 5 (39) and considered this the point of emergence of the human protein (40). Following their guidelines, we defined that genes of unicellular origin capture phylostrata 1–3, while we considered genes of multicellular origin in the remaining phylostrata.

MIG and miRNA-containing gene enrichment analysis

The proportion of MIGs and genes that harbor miRNAs in their introns was compared with that one found in the human genome through a hypergeometric test. Additionally, a pollution analysis was performed for SARS-CoV-2, where 332 non-MIGs were selected randomly, and MIGs were added increasingly, in increments of 10%. Hypergeometric tests were utilized to determine the minimum number of MIGs necessary to reach significance at $\alpha < 0.05$. Finally, random sampling was performed to retrieve subsets of randomly selected genes with the size of the different viral interactomes. This procedure was repeated 1,000 times, allowing us to obtain the median number of MIGs for each interactome.

Network enrichment analysis

To consider the enrichment of MIG-Ps in a given set (e.g. bottleneck proteins), we first determined their frequency in the underlying set of interest (i.e. observed frequency f). As a null model, we randomly sampled proteins of the same set size 10^5 times and calculated the corresponding random expected frequency, f_r, in each sample. Finally, we defined the enrichment/depletion of proteins as the average over all $E = \lg_2 (f/f_r)$.

As for determining the enrichment of MIG-Ps in bins of proteins with a certain number of interaction partners, k, we similarly calculated their observed frequency up to a given k, $f_{\geq k}$. We determined the expected frequencies $f^r_{\geq k}$ of randomly sampled MIG-Ps out of all proteins in the underlying human interaction network 10^5 times and defined the enrichment/depletion of MIG-Ps in given bins of connectivity as the average over all $E_{\geq k} = \lg_2 (f_{\geq k}/f^r_{\geq k})$.

Pathway and GO Slim enrichment analysis

As for GO Slim analysis, we utilized GOATools (31). In particular, we applied hypergeometric tests and obtained FDR-corrected P-values using the Benjamini–Hochberg correction (41).

Human–viral protein interactions

We collected 1,659 human proteins that were experimentally found to interact with proteins of SARS-CoV-2 from the study by Gordon et al. (15). Furthermore, we used 365 targets of SARS-CoV-1 and 292 targets of MERS-CoV from the study by Gordon et al. (42). As for other viruses, we collected human proteins that are linked to proteins of ZIKV (823), HIV-1 (1,709), HPV (2,126), IAV (2,797), EBV (1,141), EBOV (332), HSV-1 (2,197), and HCV (1,033) using the HVIDB database (43).

Viral host factors

We collected 799 human host factors of SARS-CoV-2 from (44–49) as well as 917 in HIV-1 (18, 50), 790 in ZIKV (51, 52), 315 in HPV (53), 1,251 in IAV (54–57), 144 in EBV (58), 458 in EBOV (59), 358 in HSV-1 (60), and 262 in HCV (61).

Gene expression

Data from samples that were infected with three different coronaviruses, including SARS-CoV-2, MERS-CoV, and SARS-CoV (29), were analyzed for changes in gene expression. DESeq2-processed data were downloaded for all three coronaviruses. Data pertaining to MRC5 cell lines 24 h postinfection at a multiplicity of infection of 3 were included in the current analysis.

Controllability analysis

To determine nodes that are important for the control of a directed, unweighted molecular interaction network, we first determined driver nodes by following Liu et al. (26). In particular, we map such a structural controllability problem to a maximum matching problem, given that a network of direct interactions is a graph-based proxy of the underlying dynamical system. We solve the maximum matching problem by the Hopcroft–Karp algorithm (62), mapping a directed to a bipartite network. Specifically, we mapped directed links to edges between partitions of nodes that start and end edges. In the matching, a subset of edges M is a matching of maximum cardinality in a directed network if no two edges in M share a common starting and ending vertex. Vertices that do not appear in M are unmatched and have been shown to be nodes that structurally control the underlying network (26). As a corollary, a maximum matching implies the presence of a minimum set of such driver nodes of size N_D. To assess the influence of network nodes on the controllability of the underlying directed network, we followed the following heuristic (27, 63, 64): when a node is removed from the underlying network, we determined the size N'_D of driver nodes in the changed network. If $N'_D > N_D$, the node is classified as indispensable (i.e. a control node) as the number of driver nodes increased. In other words, the deletion of a node increased the number of nodes that allowed the control of the underlying network. In turn, if $N'_D \leq N_D$, the node is classified as noncontrolling as the number of driver nodes remained unchanged (neutral node) or decreased (dispensable node).

Bottleneck proteins

As a measure of global topological centrality, we defined the betweenness centrality of a node v as $c_B(v) = \sum_{s \neq t \neq v \in V} (\sigma_{st}(v)/\sigma_{st})$. In particular, σ_{st} is the number of shortest paths between node pairs s and t out of all nodes in the underlying network (65), while $\sigma_{st}(v)$ is the number of shortest paths running through protein v. As a representative set of bottleneck proteins, we selected the top 20% of the most central proteins (23, 25).

Acknowledgments

The authors thank M. Wolfinger, I. Hofacker, A. Emili, and A. Yousef for fruitful discussions.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

R.K. acknowledges support from a grant from SPARK technology Commercialization Fund of the University of Connecticut. E.M. acknowledges suppport from a grant from Fast Grants.

Author Contributions

R.N.K., S.W., and V.A.-P. conceived the project. S.W., J.O., A.J.H., A.K.W., K.D., A.M.O., and V.A.-P. analyzed the data. S.W., R.N.K., V.A.-P., and E.M. supervised the research. S.W. wrote the first draft of the manuscript. S.W., R.N.K., V.A.-P., E.M., J.O., and A.J.H. edited the manuscript.

Preprints

A version of this manuscript was posted on a preprint: https://www.biorxiv.org/content/10.1101/2022.09.30.510319v1.

Data Availability

All data that support the findings of this study are available in this manuscript and the Supplementary Material.

References

1 Akinyi MV, Frilander MJ. 2021. At the intersection of major and minor spliceosomes: crosstalk mechanisms and their impact on gene expression. *Front Genet*. 12:700744.

2 Olthof AM, Hyatt KC, Kanadia RN. 2019. Minor intron splicing revisited: identification of new minor intron-containing genes and tissue-dependent retention and alternative splicing of minor introns. *BMC Genomics*. 20:686.

3 El Marabti E, Malek J, Younis I. 2021. Minor intron splicing from basic science to disease. *Int J Mol Sci*. 22:6062.

4 Montzka KA, Steitz JA. 1988. Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc. *Proc Natl Acad Sci U S A*. 85:8885–8889.

5 Baumgartner M, Drake K, Kanadia RN. 2019. An integrated model of minor intron emergence and conservation. *Front Genet*. 10: 1113.

6 Rogozin IB, Carmel L, Csuros M, Koonin EV. 2012. Origin and evolution of spliceosomal introns. *Biol Direct*. 7:11.

7 Basu MK, Makalowski W, Rogozin IB, Koonin EV. 2008. U12 intron positions are more strongly conserved between animals and plants than U2 intron positions. *Biol Direct*. 3:19.

8 Moyer DC, Larue GE, Hershberger CE, Roy SW, Padgett RA. 2020. Comprehensive database and evolutionary dynamics of U12-type introns. *Nucleic Acids Res*. 48:7066–7078.

9 Burge CB, Padgett RA, Sharp PA. 1998. Evolutionary fates and origins of U12-type introns. *Mol Cell*. 2:773–785.

10 Larue GE, Roy SW. 2023. Where the minor things are: a pan-eukaryotic survey suggests neutral processes may explain much of minor intron evolution. *Nucleic Acids Res*. 51: 10884–10908.

11 Baumgartner M, *et al*. 2018. Minor spliceosome inactivation causes microcephaly, owing to cell cycle defects and death of self-amplifying radial glial cells. *Development*. 145:dev166322.

12 Olthof AM, *et al*. 2021. Disruption of exon-bridging interactions between the minor and major spliceosomes results in alternative splicing around minor introns. *Nucleic Acids Res*. 49:3524–3545.

13 Verma B, Akinyi MV, Norppa AJ, Frilander MJ. 2018. Minor spliceosome and disease. *Semin Cell Dev Biol*. 79:103–112.

14 Moron-Lopez S, *et al*. 2020. Human splice factors contribute to latent HIV infection in primary cell models and blood CD4+ T cells from ART-treated individuals. *PLoS Pathog*. 16:e1009060.

15 Gordon DE, *et al*. 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*. 583:459–468.

16 Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. 2023. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*. 51:D587–D592.

17 Tenthorey JL, Emerman M, Malik HS. 2022. Evolutionary landscapes of host-virus arms races. *Annu Rev Immunol*. 40:271–294.

18 Brass AL, *et al*. 2008. Identification of host proteins required for HIV infection through a functional genomic screen. *Science*. 319: 921–926.

19 Das J, Yu H. 2012. HINT: high-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol*. 6:92.

20 Mariano R, Khuri S, Uetz P, Wuchty S. 2016. Local action with global impact: highly similar infection patterns of human viruses and bacteriophages. *mSystems*. 1:e00030-15.

21 Wuchty S, Siwo G, Ferdig MT. 2010. Viral organization of human proteins. *PLoS One*. 5:e11796.

22 Blasche S, *et al*. 2014. The EHEC-host interactome reveals novel targets for the translocated intimin receptor. *Sci Rep*. 4:7531.

23 Mariano R, Wuchty S, Vizoso-Pinto MG, Häuser R, Uetz P. 2016. The interactome of *Streptococcus pneumoniae* and its bacteriophages show highly specific patterns of interactions among bacteria and their phages. *Sci Rep*. 6:24597.

24 Tyagi M, Hashimoto K, Shoemaker BA, Wuchty S, Panchenko AR. 2012. Large-scale mapping of human protein interactome using structural complexes. *EMBO Rep*. 13:266–271.

25 Wuchty S, Boltz T, Küçük-McGinty H. 2017. Links between critical proteins drive the controllability of protein interaction networks. *Proteomics*. 17:e1700056.

26 Liu YY, Slotine JJ, Barabási AL. 2011. Controllability of complex networks. *Nature*. 473:167–173.

27 Devkota P, Wuchty S. 2020. Controllability analysis of molecular pathways points to proteins that control the entire interaction network. *Sci Rep*. 10:2943.

28 Meyers RM, *et al.* 2017. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet*. 49:1779–1784.

29 Blanco-Melo D, *et al.* 2020. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell*. 181: 1036–1045.e1039.

30 Trigos AS, Pearson RB, Papenfuss AT, Goode DL. 2017. Altered interactions between unicellular and multicellular genes drive hallmarks of transformation in a diverse range of solid tumors. *Proc Natl Acad Sci U S A*. 114:6406–6411.

31 Klopfenstein DV, *et al.* 2018. GOATOOLS: a Python library for gene ontology analyses. *Sci Rep*. 8:10872.

32 Patel AA, Steitz JA. 2003. Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol*. 4:960–970.

33 Yeo GW, Van Nostrand EL, Liang TY. 2007. Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet*. 3:e85.

34 Orzalli MH, Kagan JC. 2017. Apoptosis and necroptosis as host defense strategies to prevent viral infection. *Trends Cell Biol*. 27: 800–809.

35 Liu Z, *et al.* 2021. SARS-CoV-2 encoded microRNAs are involved in the process of virus infection and host immune response. *J Biomed Res*. 35:216–227.

36 Gene Ontology Consortium. 2023. The gene ontology knowledgebase in 2023. *Genetics*. 224:iyad031.

37 Younis I, *et al.* 2013. Minor introns are embedded molecular switches regulated by highly unstable U6atac snRNA. *eLife*. 2: e00780.

38 Viswanathan SR, *et al.* 2018. Genome-scale analysis identifies paralog lethality as a vulnerability of chromosome 1p loss in cancer. *Nat Genet*. 50:937–943.

39 Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13:2178–2189.

40 Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet*. 23:533–539.

41 Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 57:289–300.

42 Gordon DE, *et al.* 2020. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science*. 370:eabe9403.

43 Yang X, *et al.* 2021. HVIDB: a comprehensive database for human–virus protein–protein interactions. *Brief Bioinformatics*. 22: 832–844.

44 Wang R, *et al.* 2021. Genetic screens identify host factors for SARS-CoV-2 and common cold coronaviruses. *Cell*. 184: 106–119.e114.

45 Baggen J, Vanstreels E, Jansen S, Daelemans D. 2021. Cellular host factors for SARS-CoV-2 infection. *Nat Microbiol*. 6:1219–1232.

46 Baggen J, *et al.* 2021. Genome-wide CRISPR screening identifies TMEM106B as a proviral host factor for SARS-CoV-2. *Nat Genet*. 53:435–444.

47 Daniloski Z, *et al.* 2021. Identification of required host factors for SARS-CoV-2 infection in human cells. *Cell*. 184:92–105.e116.

48 Wei J, *et al.* 2021. Genome-wide CRISPR screens reveal host factors critical for SARS-CoV-2 infection. *Cell*. 184:76–91.e13.

49 Schneider WM, *et al.* 2021. Genome-scale identification of SARS-CoV-2 and pan-coronavirus host factor networks. *Cell*. 184:120–132.e114.

50 Park RJ, *et al.* 2017. A genome-wide CRISPR screen identifies a restricted set of HIV host dependency factors. *Nat Genet*. 49: 193–203.

51 Li Y, *et al.* 2017. Induction of expansion and folding in human cerebral organoids. *Cell Stem Cell*. 20:385–396.e383.

52 Savidis G, *et al.* 2016. Identification of Zika virus and dengue virus dependency factors using functional genomics. *Cell Rep*. 16: 232–246.

53 Aydin I, *et al.* 2014. Large scale RNAi reveals the requirement of nuclear envelope breakdown for nuclear import of human papillomaviruses. *PLoS Pathog*. 10:e1004162.

54 Brass AL, *et al.* 2009. The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. *Cell*. 139:1243–1254.

55 Karlas A, *et al.* 2010. Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature*. 463: 818–822.

56 König R, *et al.* 2010. Human host factors required for influenza virus replication. *Nature*. 463:813–817.

57 Shapira SD, *et al.* 2009. A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell*. 139:1255–1267.

58 Ma Y, *et al.* 2017. CRISPR/Cas9 screens reveal Epstein-Barr virus-transformed B cell host dependency factors. *Cell Host Microbe*. 21:580–591.e587.

59 Martin S, *et al.* 2018. A genome-wide siRNA screen identifies a druggable host pathway essential for the Ebola virus life cycle. *Genome Med*. 10:58.

60 Griffiths SJ, *et al.* 2013. A systematic analysis of host factors reveals a Med23-interferon-lambda regulatory axis against herpes simplex virus type 1 replication. *PLoS Pathog*. 9:e1003514.

61 Li Q, *et al.* 2009. A genome-wide genetic screen for host factors required for hepatitis C virus propagation. *Proc Natl Acad Sci U S A*. 106:16410–16415.

62 Hopcroft JE, Karp RM. 1973. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM J Comput*. 2:225–231.

63 Vinayagam A, *et al.* 2016. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc Natl Acad Sci U S A*. 113:4976–4981.

64 Paul I, *et al.* 2023. Parallelized multidimensional analytic framework applied to mammary epithelial cells uncovers regulatory principles in EMT. *Nat Commun*. 14:688.

65 Brandes U. 2001. A faster algorithm for betweenness centrality. *J Math Sociol*. 25:163–177.