PLOS PATHOGENS

RESEARCH ARTICLE

Cellular receptors for mammalian viruses

Ana Valero-Rello, Carlos Baeza-Delgado, Iván Andreu-Moreno, Rafael Sanjuán*

Institute for Integrative Systems Biology (I2SysBio), Consejo Superior de Investigaciones Científicas-Universitat de València, Paterna, València, Spain

* rafael.sanjuan@uv.es

Abstract

The interaction of viral surface components with cellular receptors and other entry factors determines key features of viral infection such as host range, tropism and virulence. Despite intensive research, our understanding of these interactions remains limited. Here, we report a systematic analysis of published work on mammalian virus receptors and attachment factors. We build a dataset twice the size of those available to date and specify the role of each factor in virus entry. We identify cellular proteins that are preferentially used as virus receptors, which tend to be plasma membrane proteins with a high propensity to interact with other proteins. Using machine learning, we assign cell surface proteins a score that predicts their ability to function as virus receptors. Our results also reveal common patterns of receptor usage among viruses and suggest that enveloped viruses tend to use a broader repertoire of alternative receptors than non-enveloped viruses, a feature that might confer them with higher interspecies transmissibility.

Author summary

Specific interactions between viruses and cellular entry factors are a critical initial step in viral infection. The identification of virus receptors and other key entry factors is important for understanding viral tropism, pathogenesis and cross-species transmissibility, as well as for the design of antiviral drugs. Here, we provide a comprehensive meta-analysis of our current knowledge of virus receptors and attachment factors. We use the newly assembled dataset to reveal general patterns of receptor use across viruses and to derive predictions about the propensity of each cell surface protein to serve as a virus receptor. This work may assist future research on receptor discoveries, and suggests new implications of virus-receptor interactions, including viral emergence.

OPEN ACCESS

Citation: Valero-Rello A, Baeza-Delgado C, Andreu-Moreno I, Sanjuán R (2024) Cellular receptors for mammalian viruses. PLoS Pathog 20(2): e1012021. https://doi.org/10.1371/journal.ppat.1012021

Editor: Welkin E. Johnson, Boston College, UNITED STATES

Received: October 5, 2023

Accepted: February 2, 2024

Published: February 20, 2024

Data Availability Statement: Data are available in the supplementary tables. Code is available at https://github.com/cbaezadelgado/Viral_receptors.

Funding: This work was financially supported by Advanced Grant 101019724—EVADER from the European Research Council, grant PID2020-118602RB-I00—ZooVir from the Spanish Ministerio de Ciencia e Innovación, and grant 202130-31 from Fundació La Marató de TV3 to R. S. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. R.S. and A.V-R.

Introduction

Mammals can be infected by thousands of viruses belonging to tens of different families. Cellular receptors and other entry factors critically determine infectivity and play a major role in viral cross-species transmission [1–4]. There are numerous examples showing the evolution of key mutations in viral receptor-binding proteins that promote transmissibility. For instance, certain changes in the influenza virus hemagglutinin determine sialic acid preferences and the

received a salary from the European Research Council. C.B-D. received a salary from the Fundació La Marató de TV3.

Competing interests: The authors have declared that no competing interests exist.

ability of avian strains to infect humans [5,6]. Similarly, changes in the affinity of the viral spike protein for human ACE2 have been instrumental in the emergence and evolution of SARS-CoV-2 [7,8]. Conversely, several receptor-coding genes have evolved under virus-driven selection, such NPC1 in bats [9] and TFRC in rodents [10], among others. Therefore, the identification of cellular factors involved in viral entry is a cornerstone in our understanding of viral tissue tropism and pathogenesis. The discovery and characterization of virus receptors also facilitate the development of entry inhibitors and allow the targeting of therapeutic viruses to specific cells [11].

However, virus receptor studies can be technically challenging, particularly due to the complex nature of viral entry, which often involves redundant receptors, co-receptors, and accessory receptors, as well as different attachment factors. The use of multiple functional receptors by viruses has been extensively documented, two examples being SARS-related [12,13] and Zika [14] viruses. Strategies for the identification of cellular factors determining viral entry include systematic perturbation methods such as RNAi, CRISPR-Cas, and overexpression of candidate genes, as well as biochemical and biophysical methods used to demonstrate and quantify virus-receptor binding, including protein microarrays, affinity-purification mass spectrometry, biolayer interferometry, and plasmon resonance [15,16]. Virus receptor inference can involve full viruses or use pseudotypes, an approach that focuses precisely on viral entry and allows handling non-culturable viruses [17].

Currently, virus receptors are collected in a few databases, such as ViralZone (viralzone.expasy.org), KEGG (genome.jp/kegg), and VTHunter (db.cngb.org/VThunter). Moreover, previous articles have relied on this information in combination with different search strategies to investigate viral entry factors [18–21]. Overall, these databases and previous works report about 100 cellular receptors for a similar number of viruses. However, they may not provide a comprehensive view of the literature, and their content is probably biased towards human and economically relevant viruses. Here, we aim to provide a more thorough analysis of the actual diversity of known receptors and attachment factors used by mammalian viruses. To achieve this goal, we implemented systematic and semi-automated search strategies that allowed us to double the amount of information extracted from the literature compared to previous work, and to pinpoint the role of each cellular factor in viral entry. Using machine learning, we identify cell surface proteins that are more likely to function as virus receptors, as well as common features of these proteins. We also explore how the repertoire of cellular receptors varies according to viral species, families and other viral features, and show that this repertoire correlates with viral cross-species transmissibility.

Results

Dataset of virus receptors

We defined receptors as host factors promoting viral entry through direct interactions with the surface of viral particles. This includes receptors believed to be both necessary and sufficient for viral entry, but also alternative receptors (sufficient but not necessary for viral entry), co-receptors (necessary but not sufficient), and accessory receptors (promoting entry but not necessary or sufficient). We also included attachment factors, defined as moieties that promote initial virus binding. A PubMed search of mammalian virus names associated with the keywords “receptor”, “entry”, “binding”, or “attach” yielded 67,492 results, which were filtered and analyzed using a combination of automated text mining and manual revision of abstracts or full-length articles when needed (**S1 Fig**). This allowed us to retrieve 705 distinct virus-host interactions involving 233 viral species and 204 cellular factors. Of these, 61.3% were identified by the manual approach, 7.7% by the automated method, and 31.0% by both methods.

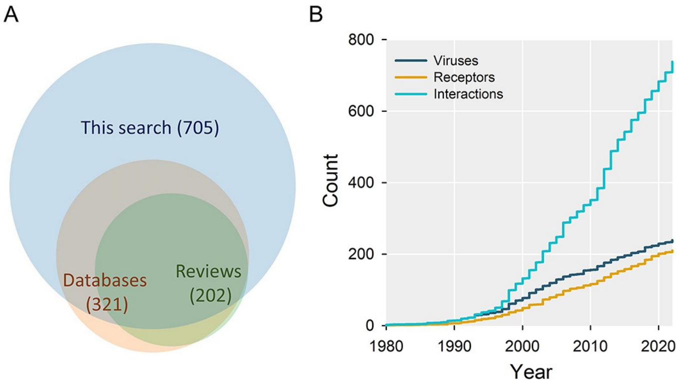

Fig 1. Virus receptor discovery trends. A. Venn diagram showing the number of virus-receptor interactions in three databases combined (ViralZone, KEGG, VTHunter), two previous reviews or meta-analyses (see text), and the search performed in this study. Areas are proportional to the number of interactions described, and numbers indicate the total number of pairs from each source. See Tables 1 and S1 for details. **B.** Cumulated numbers of distinct virus-receptor interactions, viruses, and host factors discovered per year.

https://doi.org/10.1371/journal.ppat.1012021.g001

Compared to three previous databases (ViralZone, KEGG, VTHunter) and two previous meta-analyses [18,20], our search increased by 2.2-fold the number of interactions, while capturing 95.5% of those reported in these other sources (**Fig 1A and Table 1**). After pooling all data, we obtained a final dataset consisting of 210 cellular factors, 239 viral species, and 738 interactions. For each of these, we provide the name of the virus, the name, chemical nature (protein, carbohydrate, lipid) and functional role of the cellular factor (main receptor, alternative receptor, co-receptor, accessory receptor, or attachment factor), the gene symbol for protein-coding genes, the PMID of the original publication, year of discovery, and whether the interaction was reported in previous reviews and databases (**S1 Table**). Analysis of publication years of the original articles showed that virus receptor discovery rates have been approximately constant since 2000, with roughly 7.6 new cellular factors, 7.7 new viruses, and 28.2 new interactions per year, the latter rate showing a slight acceleration in recent years (**Fig 1B**).

Table 1. Numbers of distinct host factors, viruses, and virus-host interactions collected in this study as well as previous publications and databases.

	Host factors	Viruses[g]	Interactions[g]
KEGG[a]	75	89	174
ViralZone[b]	76	105	215
VTHunter[c]	87	105	215
Zhang et al.[20][d]	69	94	176
Wang et al.[18][d]	81	91	190
Combined[e]	112	140	324
New search	204	233	705
Fold increase[f]	1.8	1.7	2.2
Total	210	239	738

[a] Extracted from genome.jp/kegg

[b] From viralzone.expasy.org/5356

[c] Extracted from db.cngb.org/VThunter

[d] See text for reference.

[e] Combination of all the above sources.

[f] This study compared to the five other sources of information combined.

[g] Five viral species were split into 11 subgroups as the pattern of their receptor usage was markedly different.

https://doi.org/10.1371/journal.ppat.1012021.t001

Table 2. Virus-receptor interactions classified according to the nature and role of the host factors involved, showing differences between enveloped and non-enveloped viruses.

Nature	Role	Total	Enveloped	Non-enveloped
Proteins	Main receptor	164 (22.2%)	127 (21.8%)	37 (23.7%)
	Alternative receptor	166 (22.5%)	146 (25.1%)	20 (12.8%)
	Co-receptor	79 (10.7%)	67 (11.5%)	12 (7.7%)
	Accessory receptor	207 (28.0%)	174 (29.9%)	33 (21.2%)
Moieties	Attachment	122 (16.5%)	68 (11.7%)	54 (34.6%)
Total		738 (100%)	582 (100%)	156 (100%)

https://doi.org/10.1371/journal.ppat.1012021.t002

Of the 738 virus-host interactions identified, 616 involved 201 receptors constituted by host proteins or protein complexes. Overall, 22.2% corresponded to main receptors, 22.5% to alternative receptors, 10.7% to co-receptors, and 28.0% to accessory receptors (Table 2). However, these assigned roles can be uncertain, since they depend on how data were interpreted in the original publications. For instance, a receptor might be misclassified as sufficient for entry if experimental evidence was obtained in cells that expressed an unknown co-receptor. Also, the main receptor could be functionally equivalent to alternative receptors, the only difference being the time of discovery. Additionally, a given cellular protein could play different roles depending on the virus. The remaining 122 interactions (16.5%) corresponded to 9 different moieties linked to unspecified proteins. In most cases, these moieties were sialic acids or other glycans such as heparan sulfate. For some viruses such as influenza virus, the molecular details of the interaction between carbohydrates and viral surface proteins have been characterized extensively, but in most cases such details are unknown. Moreover, carbohydrate moieties typically function as attachment factors, such that viral particles bind to the carbohydrate molecules but may not enter the cell unless these moieties are found in a cell surface protein capable of mediating viral internalization.

Overview of the patterns of receptor use across viruses

Among the total 201 individual proteins or protein complexes identified, some were used by many viruses, the most frequent being different integrin subunits, followed by CD209 (DC-SIGN) and CLEC4M (C-type lectins), HAVCR1 (TIM-1), TFRC and AXL, each associated to >10 different viruses belonging to more than five families (Fig 2). As previously noted [3,22], this underscores that viral entry frequently exploits cellular functions related to cell-cell adhesion (e.g. integrins), carbohydrate-mediated signalling (lectins), and autophagy (HAVCR1, AXL), and shows that non-proteinaceous components of the virion surface can play a central role in this process. For instance, carbohydrates in viral surface glycoproteins can bind lectins [23–25], and HAVCR1 can interact with the lipid membrane of many enveloped viruses to promote viral endocytosis in a process known as apoptotic mimicry [26,27].

A hierarchical cluster analysis in which entry factors were grouped according to virus sharing suggested some general patterns (Fig 2). A large group (C1) was formed by host proteins with heterogeneous functions used predominantly by plus-strand RNA viruses, particularly flaviviruses, coronaviruses, and togaviruses. Another cluster (C2) included AXL, TYRO3, HAVCR1, and NPC1, which are frequent receptors for flaviviruses, togaviruses, arenaviruses, and filoviruses. These viruses are often internalized nonspecifically in cells by apoptotic mimicry and use downstream specific receptors that mediate membrane fusion in the endosome, such as NPC1. Other receptor clusters were associated with a given viral family, such as solute carriers and immune signalling proteins for retroviruses (C3).

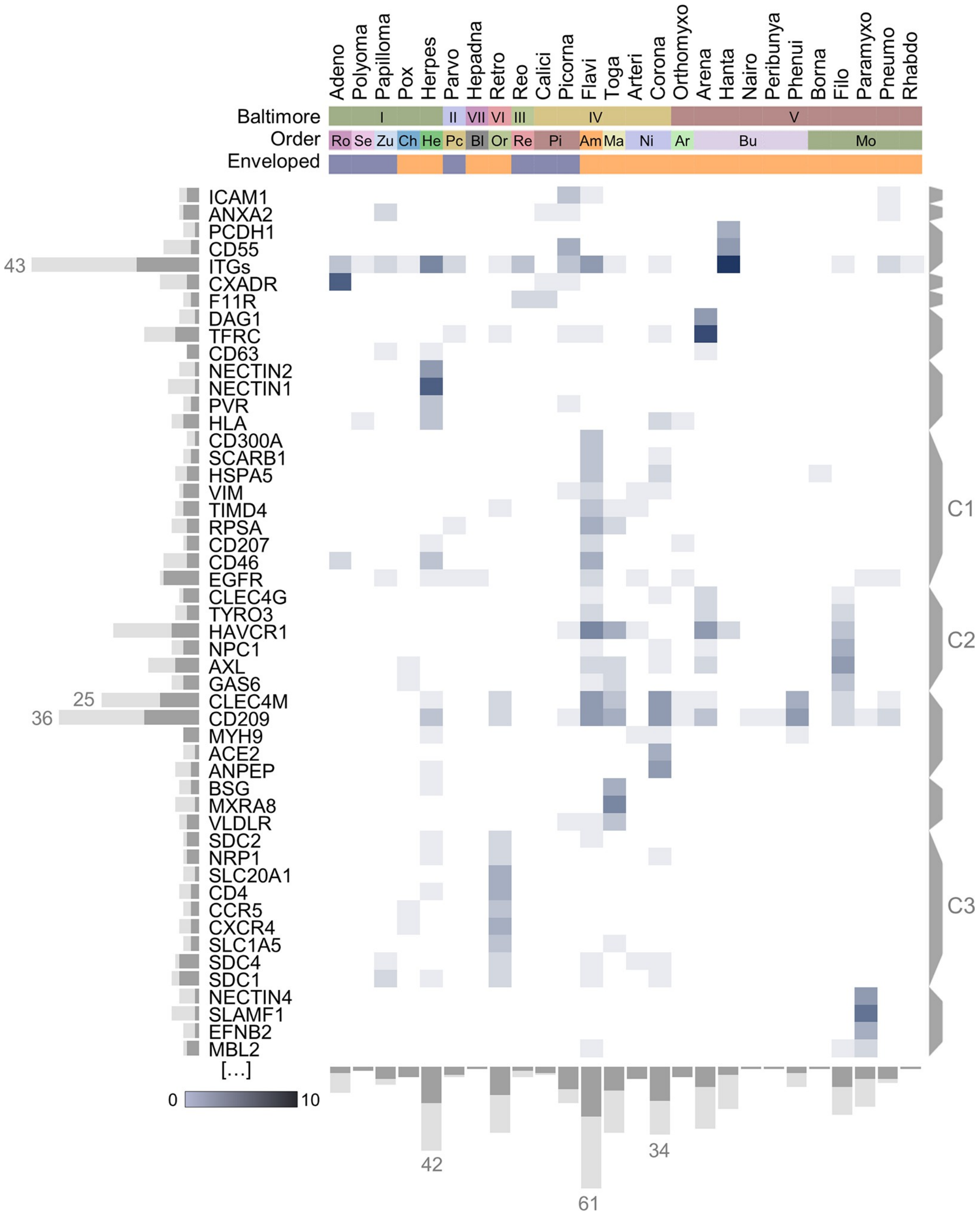

Fig 2. Heat map of the 50 cellular proteins most frequently used as receptors by mammalian viruses. HUGO gene names are shown, except for the HLA, VGGC, and integrin (ITGs) complexes. Shades of grey indicate the number of viruses known to use each protein as a receptor. Bars on the left show the total number of viruses (light grey) and viral families (dark grey) using each protein (top three values indicated). Brackets on the right indicate protein clusters obtained in a hierarchical cluster analysis, using the cosine similarity metric to group proteins according to levels of virus sharing. Three such clusters (C1-C3) are highlighted. In columns, viruses are aggregated by viral families. Baltimore group, taxonomical order, and whether the virus is enveloped are indicated. For each family, grey bars at the bottom show the number of known virus-receptor pairs (light grey; top values indicated), and the number of distinct receptors used (dark grey). Ro: Rowavirales; Se: Sepolyvirales; Zu: Zurhausenvirales; Ch:Chitovirales; He: Herpesvirales; Pc: Piccovirales; Bl: Blubervirales; Or: Ortervirales; Re: Reovirales; Pi: Picornavirales; Am: Amarillovirales; Ma:Martellivirales; Ni: Nidovirales; Ar: Articulavirales;Bu: Bunyavirales; Mo: Mononegavirales.

https://doi.org/10.1371/journal.ppat.1012021.g002

Predictability of virus receptors

We set out to explore features that could predict whether a protein may serve as a virus receptor. For this, we compared 175 known receptors with 2668 plasma membrane proteins located at the cell surface (surfaceome) that have not been previously involved in viral entry. Using a generalized boosted model (GBM), we analyzed a large number of features including functional domains (PFAMs), expression profiles in 54 healthy human tissues, protein size, the number of human protein interactors, post-translational modifications (glycosylation, lipidation, disulfide bonds), sequence distance and synonymous to nonsynonymous evolution rate ratios (dN/dS) between human proteins and their orthologs in four mammal species, and >13,000 Gene Ontology annotation terms, excluding virus-related functions. Our best model showed an AUC of 84.0%, improving the performance of previous models [28] (Fig 3A). Of the 12 proteins with a score >0.90, 11 were known receptors and, overall, the distributions of scores assigned to known receptors and the rest of surfaceome proteins were well differentiated (Fig 3B). Using a threshold score of 0.5, the model detected 72.6% of the known receptors, but also predicted 570 that have not been so far described (S2 Table). These could be false positives, but also undiscovered receptors. The 20 proteins with the highest scores among those not known to be receptors are shown in Table 3.

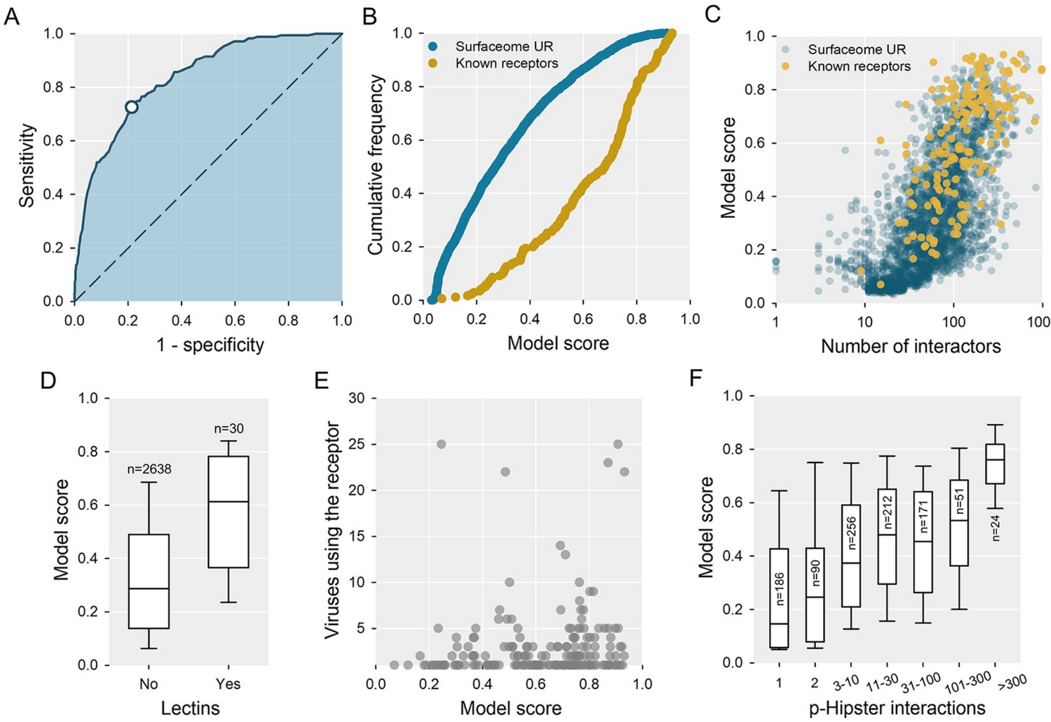

Fig 3. Predictability of virus receptors. A. AUC plot of the selected model. The white dot shows the sensitivity and specificity achieved when a threshold score of 0.5 was used to classify cell surface proteins as virus receptors. **B.** Cumulative frequency plots of the scores obtained for 175 known receptors included in the model and 2668 surfaceome proteins not known to be virus receptors (surfaceome UR). **C.** Model scores against the number of protein interactors, measured as the degree parameter in the STRING database. **D.** Model scores for the subset of surfaceome proteins classified as lectins versus non-lectins. This subset was obtained from unilectin.expasy.org/humanLectome, selecting curated lectins only. The number of proteins in each group is indicated. **E.** For receptors known to be used by at least one virus, relationship between the model score and the actual number of viruses using this receptor. **F.** Relationship between the model score and the number of interactions with viruses predicted by p-Hipster (phipster.org). This information was available for 990 surfaceome proteins. For visualization, the number of p-Hipster interactions was categorized into the indicated groups. Group sizes are shown.

https://doi.org/10.1371/journal.ppat.1012021.g003

Table 3. List of the proteins with the highest GBM scores among surfaceome proteins currently unknown to function as virus receptors.

Gene symbol	Description	GBM score
APP	Amyloid beta precursor protein	0.914
ENG	Endoglin	0.897
CSF3R	Colony Stimulating Factor 3 Receptor	0.891
THBS1	Thrombospon-din 1	0.890
CD44	Hyaluronate Receptor	0.886
CXCR3	C-X-C Motif Chemokine Receptor 3	0.882
CD33	Sialic Acid-Binding Ig-Like Lectin 3	0.881
CD59	CD59 Molecule, Complement Regulatory Protein	0.880
ALCAM	Activated Leukocyte Cell Adhesion Molecule	0.879
CD226	Platelet And T Cell Activation Antigen 1	0.873
RHOB	Ras Homolog Gene Family, Member B	0.870
AOC3	Amine Oxidase, Copper Containing 3 (Vascular Adhesion Protein 1)	0.868
ITGB4	Integrin Subunit Beta 4	0.867
MCAM	Melanoma Cell Adhesion Molecule	0.863
CD22	Sialic Acid-Binding Ig-Like Lectin 2	0.846
ATP1B1	ATPase Na+/K+ Transporting Subunit Beta 1	0.843
HLA-E	Major Histocompatibility Complex, Class I, E	0.843
SELL	Selectin L	0.842
STAB1	Stabilin 1	0.840
MME	Membrane Metalloendopeptidase	0.840

https://doi.org/10.1371/journal.ppat.1012021.t003

Protein features were ranked according to their relative importance in the model (S3 Table), the top single variable being the number of protein interactors (16.0% gain; Fig 3C), followed by the GO term "protein binding" (GO.0005515; 8.3% gain). This underscores that the more exposed location of highly-interacting proteins to be accessible to their ligands also makes them more accessible to viruses [20]. Indeed, many proteins that bind or transport other ligands are virus receptors, and it is known that cell adhesion proteins are preferred receptors for viruses [29]. Consistently, the "cell adhesion" GO term (GO.0007155) was the third most important variable in the model (gain 6.6%). Collectively, the gene expression levels in 54 human tissues were a highly relevant feature (46.1% gain), although none of them reached a high value individually (<6.3% gain), meaning that proteins expressed at high levels in at least some tissues are more likely to be receptors. Other features were frequently used for classification (high cover) although they were not among the most decisive features. This included having a high density of disulfide bonds, high glycosylation, containing Immunoglobulin domains (PFAM term PF07686), or participating in cell chemotaxis (S3 Table). The role of protein glycosylation was expected considering that many viruses are known to use carbohydrates as attachment factors [25,30]. Conversely, although GO terms associated with lectins did not appear among the most important variables in the model, lectins tended to show higher scores than other proteins (Fig 3D). This was also expected because the binding of cellular lectins to carbohydrates present in viral particles can promote viral entry [31]. Finally, despite virus receptors being frequently under positive selection [18], dN/dS ratios and divergence values among mammals did not feature among the top variables of the model (<1% gain).

The quality of the predictions was supported by two additional observations from information not used for training the model. First, the model used binary information about whether or not each protein is a known virus receptor but did not consider how many different viruses have been shown to use each receptor. Nevertheless, we found a weak but significant correlation between the model score of known receptors and the log number of viruses reported to use each receptor (Pearson r = 0.186, P = 0.009; **Fig 3E**). As a second external check of the model, for each surfaceome protein we obtained the number of interactions with viruses predicted in p-Hipster (phipster.org) [32]. The algorithms used for p-Hipster predictions are structured-based and hence do not use the same type of information as our model. Despite this, we found a significant correlation between the log number of predicted interactions in p-Hipster and our model score (Pearson r = 0.370, P < 0.001; **Fig 3F**).

Variations in the type and number of receptors used by different viruses

Our tentative functional classification of receptors suggested that non-enveloped viruses tend to rely on a single receptor more often than enveloped viruses, which on the contrary are more likely to use alternative receptors, each sufficient for entry (**Table 2** and **S2 Fig**). We also found that non-enveloped viruses are three times more often reported to use carbohydrate moieties associated with undefined proteins than enveloped viruses (34.6% versus 11.7% of all interactions, respectively). Dependence on such moieties is seemingly strongest for caliciviruses and polyomaviruses (60.0% and 80.0% of the total interactions, respectively) and weakest for retroviruses (1.3%).

We then examined in more detail how many different cellular proteins are used as receptors by each virus. We found ample variation across viral families, with over 50 described for flaviviruses, retroviruses, and herpesviruses, whereas other families such as *Circoviridae*, *Nairoviridae*, *Bornaviridae*, and *Polyomaviridae* showed five or fewer (**Fig 2**). However, these observations can be strongly biased by the research effort dedicated to each virus. To address this, we used the number of publications in PubMed as a proxy of research effort, and estimated the effect of this variable on the known number of proteins used as receptors by each virus, using a generalized linear model (GLM; **Fig 4A**). This allowed us to identify viruses that

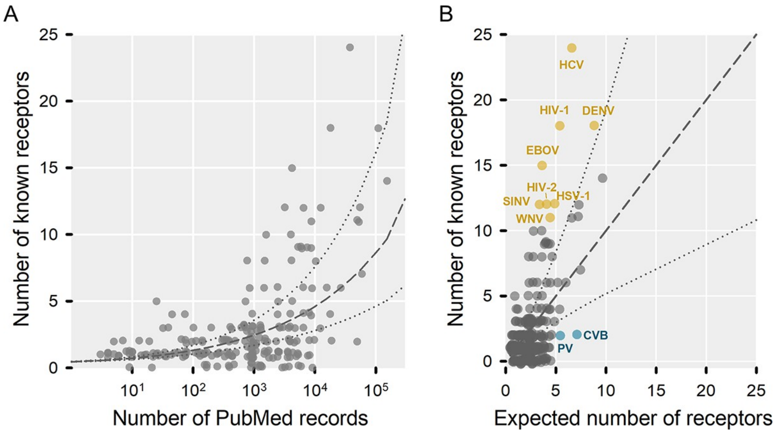

Fig 4. Variation in the number of known receptors across viruses. A. Relationship between research effort, measured the number of PubMed records and the number of different proteins used as receptors for each virus. The dashed line shows the expected number of receptors obtained from a GLM (null model), and dotted lines correspond to 95% confidence intervals. **B.** Known receptors versus the expected number under the null model shown in panel A. The dashed line indicates no deviation from the model, and the dotted lines the 95% confidence interval. Viruses shown in yellow exceed the upper limit of the confidence interval and have more than 10 known receptors. Viruses shown in blue fall below the lower limit of the confidence interval, and their expected number of receptors according to the model is higher than five. CVB: Coxsackievirus B (enterovirus B); DENV: dengue virus; EBOV: Zaire ebolavirus; HCV: hepatitis C virus (hepacivirus C); HSV-1: herpes simplex virus 1 (human alphaherpesvirus 1); PV: poliovirus (enterovirus C); SINV: Sindbis virus; WNV: West Nile virus.

https://doi.org/10.1371/journal.ppat.1012021.g004

use more receptors than expected from research effort alone, such as hepatitis C virus, dengue virus, HIV-1 and HIV-2, and Ebola virus (Fig 4B). In some cases, such as HIV-2, the excess of known receptors could be explained by knowledge acquired from a related virus, whereas other viruses may be subject to research biases not accounted for here, or might be truly promiscuous in terms of receptor usage. Some highly studied viruses did not exhibit particularly high numbers of host proteins used as receptors, such as influenza viruses, and some enteric viruses. Specifically, according to our search, no cellular proteins have been found to serve as receptors for influenza B or Norwalk viruses despite these being well-studied viruses research and the fact that several receptors have been described for the related influenza A virus and murine norovirus, respectively.

To test for more general patterns, we added to our GLM two more factors: the viral family, and whether or not the virus is enveloped. This showed that, overall, enveloped viruses use a wider variety of host proteins as receptors than non-enveloped viruses, and also more receptors that are sufficient for entry according to the literature (main or alternative receptors; $P < 0.001$). We inferred from this model that, after controlling for research effort, enveloped viruses use on average 2.4 times more cellular proteins as receptors than non-enveloped viruses. This excess was more marked for phenuiviruses, togaviruses, and filoviruses, whereas polyomaviruses and caliciviruses showed particularly low numbers of such receptors (S3 Fig).

The repertoire of alternative receptors correlates with the viral host range

In a previous publication, we used known virus-host associations to show that enveloped viruses infect on average a larger number of different host species than non-enveloped viruses [33]. Since we have shown here that enveloped viruses also tend to display a broader repertoire of receptors, we reasoned that the ability to use alternative receptors could allow viruses to infect more host species. Interspecies variability in receptor genes can prevent cross-species transmission, but this barrier might be less stringent if a virus can use other receptors that are also sufficient for entry. We found that the average number of host species was over twofold higher for viruses known to use alternative receptors (26.7 ± 4.2) than for viruses using only one receptor (11.6 ± 1.4; Fig 5). However, this might be again due to differences in research effort. To address this, we implemented a GLM in which, as above, the number of PubMed records per virus was used to quantify research effort. This confirmed that the use of alternative receptors is associated with the ability to infect a larger number of host species ($P < 0.001$), with an estimated 1.6-fold increase in the number of host species per virus. Therefore, the broader host range shown by enveloped viruses might be in part attributed to their greater tendency to use alternative receptors, compared to non-enveloped viruses.

Discussion

By performing a systematic search of virus receptors, we have provided a general overview of the state of the art in this area of research and identified knowledge gaps. Our dataset may assist future research on receptor discovery, antiviral therapeutics, and virus-host interactions. The search was restricted to mammalian viruses due to their relevance, but also due to current limitations in our ability to perform more generalized analyses that would also include other vertebrate, invertebrate, plant, fungal, and prokaryotic viruses. Despite being as exhaustive as possible, our search strategy was mostly based on reviewing article titles, abstracts, and MESH terms, whereas full-length documents were analyzed only in cases of uncertainty. The automated text mining approach allowed us to ensure objectivity and to expand our search capacities, but was also based on abstracts and required that the virus and host gene terms be in the same sentence, limiting its power. Further optimizations of this automated pipeline could

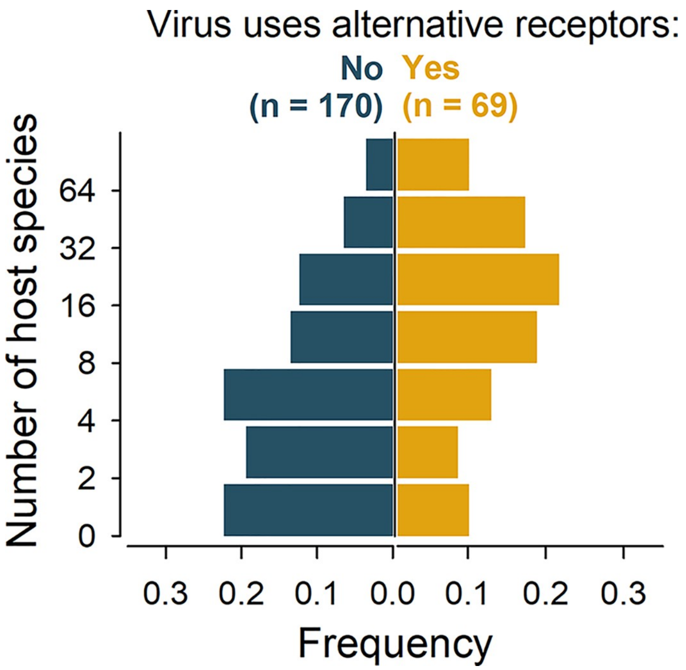

Fig 5. Relationship between viral host range and use of alternative receptors for viral entry. The distribution of the number of host species per virus is shown for viruses for which there is only one known protein receptor functionally sufficient for entry, versus those with alternative receptors. Virus-host associations were obtained from a previous publication (see text).

https://doi.org/10.1371/journal.ppat.1012021.g005

include full text, but this would strongly increase computing demands. In addition, we provide information about the role of each cellular protein in viral entry, which can vary from strictly necessary and sufficient to an accessory role. However, this classification is challenging, and in several instances, a given protein was classified as an alternative receptor by some authors and as an accessory receptor by others. We attempted to solve these incongruences by using clear definitions and checking raw results in the articles when necessary, but current knowledge gaps inherently limit our ability to robustly assign a functional role to each receptor. Finally, we classified all moieties such as sialic acids and other glycans as attachment factors, since their role in viral entry is probably dependent on the protein to which these moieties are linked. For instance, it has been shown that influenza virus infection depends on attachment to sialic acids, but that entry is triggered by host proteins such as EFGR [34], CACNAC1 [35], or CEACAM6 [36].

High throughput methods such as CRISPR-Cas knockout libraries are often used to select candidate receptor genes, but these require subsequent experimental testing, which can be a complex task. Our machine-learning model could help researchers prioritize candidates at this stage of receptor discovery. According to this model, approximately 700 proteins located at the plasma membrane may function as virus receptors for mammalian viruses, of which most remain unreported. The latter could be real but undiscovered receptors, host proteins that are currently not used by any virus but could potentially serve as receptors, or incorrect predictions of the model. In agreement with previous reports, plasma membrane proteins with functions related to ligand-binding, cell-adhesion, and a high interaction degree are more prone to be used as receptors by mammalian viruses [20,29]. Regardless of the accuracy of these predictions, our data show that the yearly discovery rate of virus receptors has not reached a plateau, indicating that there is still room for a substantial expansion of the virus receptorome. It has been estimated that there are approximately 40,000 viruses in nature capable of infecting mammals [37] but, for several viral families, only a few receptors are known. The VirHostNet database, which reports experimentally validated virus-host interactions, lists >10,000 human proteins that physically interact with at least one viral protein [38], and the p-Hipster database predicted approximately 282,000 such protein-protein interactions [32]. We have shown that top-scoring proteins in our model tend to have large numbers of predicted interactions in p-Hipster. Examples include THBS1 (thrombospondin 1) with the norovirus major capsid protein, APP (amyloid beta precursor protein) with influenza A virus hemagglutinin and the envelope protein of several flaviviruses, and NTRK1 with betaherpesvirus membrane glycoproteins.

Overall, we have found that enveloped viruses display a broader repertoire of known receptors than non-enveloped viruses. Some properties of viral envelopes could explain such differences. Envelope proteins might be structurally less constrained than capsid proteins, which form highly rigid and complex structures with potentially lower ability to accommodate different types of interactions with cellular proteins. Relatedly, this might also allow envelope proteins to accumulate increased genetic diversity, facilitating their evolution toward new receptor usages. We also found that viruses that use alternative receptors tend to show broader host ranges. These findings are consistent with our previous observation that enveloped viruses exhibit a greater propensity to cross-species transmission and zoonosis than non-enveloped viruses [33]. In part, the larger number of virus-receptor interactions found in enveloped viruses could in principle be a result of differences in research effort, but this confounder was accounted for in our analyses.

In conclusion, this work may promote future research in the field by providing a comprehensive dataset of known virus-host interactions involved in viral entry, which could be probed in related viruses, and by suggesting candidate cell surface proteins for virus receptor discovery. It also reveals previously unrecognized differences in the usage patterns of cellular entry factors among viruses, viral families, and major viral groups.

Methods

Manual search strategy

A list of 6034 viruses known to infect mammals, available from previous work [33], was used as a query. NCBI Entrez was interrogated for publications meeting the following criteria (**S1 Fig**): (i) to contain the name of the query virus, or any of their aliases obtained from NCBI Taxonomy, coded as "organism"; (ii) to contain the keywords "receptor", "entry", "binding", or "attach" in the title, abstract or MESH terms; (iii) publication date later than 1999; (iv) not including the terms "clinical", "trial", "therapy", "therapeutics", "cohort", "biomarker", "RNA-binding", or "DNA-binding". The resulting dataset comprised 67,492 articles encompassing 503 viral species. For 423 viruses that had less than 100 articles each, we performed a manual review (6257 total articles), whereas for the 80 viruses with more than 100 articles (61,235 total articles) we carried out a search strategy based on analysis of selected reviews.

Automated text mining strategy

Our pipeline was inspired by pubmedKB [39]. The above list of 67,492 abstracts was preprocessed by replacing acronyms with their original references in the text, and abstracts were divided into sentences. All resulting sentences were analyzed using the entity recognition and normalization tool BERN2 [40]. Based on the entities labelled by BERN2 as species or diseases, a search was performed in the NCBI Taxonomy database [41] to identify those corresponding to viruses and normalize their names. Sentences containing both virus and gene entities were then analyzed using the two-relations extraction models OpenIE (stanfordnlp.github.io/CoreNLP) and Spacy (spacy.io). The relationships obtained from both models were manually curated to determine which ones correctly identify actual cellular receptors used for viral entry. The complete pipeline is available at github.com/cbaezadelgado/Viral_receptors.

Additional virus receptor datasets

Our database was completed by extracting receptors from ViralZone (viralzone.expasy.org), KEGG (www.genome.jp/kegg/annotation/br03220.html), QuickGO (ebi.ac.uk/QuickGO/annotations), and three previous meta-analyses [18–20].

Database curation

The database was manually curated as follows. First, virus taxonomy was normalized according to NCBI Taxonomy, and viruses were grouped into species representatives, which agglutinates all receptors used by its members. For groups of viruses using very different receptor profiles within a species, the groups were kept separated. Endogenous viruses were removed. Second, ambiguous protein names, mammalian receptors without human orthologs, and moieties that are already known to be part of glycoprotein receptors were also removed. Third, only host proteins located at the cell surface were included, with the exception of a few intracellular receptors such as NPC1. The final list of host factors contained 214 protein-coding genes. Of these, 17 encoded integrin subunits, which were pooled into a single receptor class since integrins need to be combined to form functional complexes. Three other complexes with unspecified subunits were included (laminins, HLA, and VGCC), resulting in 201 proteins or protein complexes. In addition, the database included 9 carbohydrate or lipid moieties, thus making 210 total host factors. Virus metadata included taxonomy according to NCBI or ICTV and host data from a curated dataset originally extracted from the VIRION database [33,42].

Metadata used in the gradient boosting model

The goal of the GBM was to identify plasma membrane-associated proteins located at the cell surface (surfaceome) that are more likely to be used as virus receptors. To define the list of candidate proteins, we first selected human plasma membrane proteins using Gene Ontology (GO) annotation terms GO.0005886 and GO.0009897, which yielded 5112 proteins. We then discarded proteins with the annotation term GO.0009898 (cytoplasmic side of plasma membrane). We also filtered out all pseudogenes listed in GeneCards. Next, among the remaining proteins, we selected only those included in at least one of the following curated databases: (i) the mass spectrometric-derived cell surface protein atlas (wlab.ethz.ch/cspa), the in silico human surfaceome (wlab.ethz.ch/surfaceome), or the cancer surfaceome atlas (fcgportal.org/TCSA). For the latter database, we selected only proteins with a GESP score >4, as recommended by the authors [43]. The curated dataset contained 2843 high-confidence surfaceome proteins, including 175 of the 214 individual proteins known to be virus receptors. For each, we obtained mRNA expression levels in 54 healthy human tissues obtained from the Human Protein Atlas (proteinatlas.org;rna_tissue_consensus.tsv.zip), >13,000 GO terms from Geneontology.org (excluding virus entry-related terms), post-translational modifications (lipidation, glycosylation, disulfide bonds), length, protein families (PFAM) from UniProt (uniprot.org), and the number of human protein interactors in the STRING database (string-db.org). In addition, we calculated the normalized amino-acidic distance between 444 human proteins and their orthologs in four mammal species (*Bos taurus*, *Canis lupus familiaris*, *Mus musculus*, and *Myotis lucifugus*) using the R package ape [44], as well as site (MEME model)

[45], and branch-site (aBSREL model) [46] dN/dS ratios of non-synonymous to synonymous evolution rates using Hyphy [47].

Implementation of the gradient boosting model

An XGBoost classifier was run using the xgboost R package [48]. To address class imbalance, we weighted the positive class (i.e. known virus receptors) by the ratio of the number of negative to positive instances. Then, ten-fold stratified cross-validation was performed to estimate the model predictive ability using a maximum of 10,000 boosting iterations with 50 rounds as an early stopping criterion (xgb.cv function). The area under the ROC curve (AUC) was used as the preferred evaluation metric to determine optimal model complexity. A suitable combination of model hyperparameters was found by Bayesian optimization using the R package ParBayesianOptimization (CRAN.R-project.org/package=ParBayesianOptimization). We optimized the maximum depth of boosted trees (max_depth), the fraction of training samples used to construct each tree (subsample), the fraction of predictors used to construct each tree (colsample_bytree), the learning rate (eta), the L1 regularization term (alpha), the L2 regularization term (lambda) and the lagrangian control for tree split (gamma). Because instances labeled as negatives in our dataset might represent undiscovered virus receptors, false positive predictions may not be so. Therefore, higher recall was preferable to higher precision in selecting the optimal model. To account for this in our hyperparameter optimization, we maximized the combination (or average) of precision and recall. We initialized the Bayesian optimization with 100 random parameter combinations and sampled another 10 parameter sets during each refinement epoch, for a total of 50 epochs. The 50 top-ranked models were run again 100 times to take into account the randomness in model training and data subsets generated for cross-validation. Finally, the model with a significantly larger combined precision and recall was selected, which was determined by performing one-tailed one-sample t-test. Finally, we used the average number of boosting iterations to train a final model with full data (xgboost function), which was then used to analyze the relative importance of the predictor variables (xgb.importance function). The complete pipeline and gene database used for the predictive model are available at github.com/cbaezadelgado/Viral_receptors.

Generalized linear models

GLMs were used to test for differences in the number of known receptors among viruses. The data were assumed to follow a Tweedie distribution, and a log link function was used. The log number of PubMed records available for each virus was used to control for research effort. To test for differences across viral families, the viral family was added to the GLM, and data corresponding to families with less than 5 viral species were removed. The viral family factor was nested within a binary variable indicating whether the virus is enveloped. A GLM with underlying Tweedie distribution and log link function was also used to test for differences in the number of host species known for each virus. This analysis also controlled for research effort using the log number of PubMed records and included a binary variable indicating whether the virus is known to use multiple receptors each sufficient for viral entry (main and alternative receptors). As an alternative metric of research effort, we used the log number of sequences of each virus deposited in Genbank. All the factors shown to be significant using PubMed records were also significant using Genbank sequences.

Supporting information

S1 Fig. Workflow for the manual and automatic text-mining searches. A starting list of 6034 mammal viruses was used to obtain, which were then reviewed using both manual and

PubmedKB-based automated strategies. The resulting virus-receptor pairs were combined with known databases and manually curated (see text for full description). M, manual strategy; PKB, PubmedKB strategy.
(TIF)

S2 Fig. Roles of known receptors according to viral family. Only families with at least 10 known virus-host interactions are represented. Families of enveloped and non-enveloped viruses are shown, and within each group, families are sorted by the fraction of known receptors that are sufficient for viral entry (main plus alternative receptors).
(TIF)

S3 Fig. Research effort versus the known number of host proteins used as receptors for different viral families. Data points correspond to individual viruses. Those corresponding to the indicated family are shown in color (blue for non-enveloped viruses; yellow for enveloped viruses), and grey points correspond to all other viruses. The colored and grey dashed lines show the GLM prediction obtained specifically for the family and all viruses, respectively. Only families with at least 5 viral species in the dataset were considered.
(TIF)

S1 Table. Database of virus receptors generated in this study. The following information is provided: the viral species, number of PubMed records and Genbank sequences available for each virus, viral family, presence of an envelope, number of host species, receptor symbol, receptor nature, functional role of the receptor, corresponding gene symbol, original publication PMID, year of discovery, and whether the virus-receptor interaction was reported in previous reviews and databases.
(XLSX)

S2 Table. Scores obtained from the GBM. The gene symbol, assigned score (probability of being a receptor), and whether the corresponding protein is a known receptor are indicated.
(XLSX)

S3 Table. Relevant features identified by the GBM. Gain represents the relative contribution of each variable to the model prediction. Cover indicates the relative number of observations that are related to a given variable.
(XLSX)

Acknowledgments

We thank José M. Cuevas Torrijos and Jaime Buigues for their support during the first stages of the study. Computer-intensive analyses were performed in the HPC cluster Garnatxa at I2SysBio.

Author Contributions

Conceptualization: Rafael Sanjuán.

Data curation: Ana Valero-Rello, Carlos Baeza-Delgado.

Formal analysis: Ana Valero-Rello, Carlos Baeza-Delgado, Iván Andreu-Moreno, Rafael Sanjuán.

Funding acquisition: Rafael Sanjuán.

Investigation: Ana Valero-Rello, Carlos Baeza-Delgado, Iván Andreu-Moreno.

Supervision: Rafael Sanjuán.

Visualization: Carlos Baeza-Delgado, Rafael Sanjuán.

Writing – original draft: Ana Valero-Rello, Iván Andreu-Moreno, Rafael Sanjuán.

Writing – review & editing: Rafael Sanjuán.

References

1. Escudero-Pérez B, Lalande A, Mathieu C, Lawrence P. Host-Pathogen Interactions Influencing Zoonotic Spillover Potential and Transmission in Humans. Viruses. 2023; 15: 599. https://doi.org/10.3390/v15030599 PMID: 36992308
2. Warren CJ, Sawyer SL. How host genetics dictates successful viral zoonosis. PLoS Biol. 2019; 17: e3000217. https://doi.org/10.1371/journal.pbio.3000217 PMID: 31002666
3. Maginnis MS. Virus–Receptor Interactions: The Key to Cellular Invasion. J Mol Biol. 2018; 430: 2590–2611. https://doi.org/10.1016/j.jmb.2018.06.024 PMID: 29924965
4. Warren CJ, Sawyer SL. Identifying animal viruses in humans. Science. 2023; 379: 982–983. https://doi.org/10.1126/science.ade6985 PMID: 36893227
5. Taubenberger JK, Kash JC. Influenza virus evolution, host adaptation, and pandemic formation. Cell Host Microbe. 2010; 7: 440–451. https://doi.org/10.1016/j.chom.2010.05.009 PMID: 20542248
6. AbuBakar U, Amrani L, Kamarulzaman FA, Karsani SA, Hassandarvish P, Khairat JE. Avian Influenza Virus Tropism in Humans. Viruses. 2023; 15: 833. https://doi.org/10.3390/v15040833 PMID: 37112812
7. Kang L, He G, Sharp AK, Wang X, Brown AM, Michalak P, et al. A selective sweep in the Spike gene has driven SARS-CoV-2 human adaptation. Cell. 2021; 184: 4392–4400.e4. https://doi.org/10.1016/j.cell.2021.07.007 PMID: 34289344
8. Wrobel AG. Mechanism and evolution of human ACE2 binding by SARS-CoV-2 spike. Curr Opin Struct Biol. 2023; 81: 102619. https://doi.org/10.1016/j.sbi.2023.102619 PMID: 37285618
9. Ng M, Ndungo E, Kaczmarek ME, Herbert AS, Binger T, Kuehne AI, et al. Filovirus receptor NPC1 contributes to species-specific patterns of ebolavirus susceptibility in bats. eLife. 2015; 4: e11785. https://doi.org/10.7554/eLife.11785 PMID: 26698106
10. Kerr SA, Jackson EL, Lungu OI, Meyer AG, Demogines A, Ellington AD, et al. Computational and functional analysis of the virus-receptor interface reveals host range trade-offs in New World arenaviruses. J Virol. 2015; 89: 11643–11653. https://doi.org/10.1128/JVI.01408-15 PMID: 26355089
11. Zhao Z, Ukidve A, Kim J, Mitragotri S. Targeting Strategies for Tissue-Specific Drug Delivery. Cell. 2020; 181: 151–167. https://doi.org/10.1016/j.cell.2020.02.001 PMID: 32243788
12. Jeffers SA, Tusell SM, Gillim-Ross L, Hemmila EM, Achenbach JE, Babcock GJ, et al. CD209L (L-SIGN) is a receptor for severe acute respiratory syndrome coronavirus. Proc Natl Acad Sci U S A. 2004; 101: 15748–15753. https://doi.org/10.1073/pnas.0403812101 PMID: 15496474
13. Jeffers SA, Hemmila EM, Holmes KV. Human coronavirus 229E can use CD209L (L-SIGN) to enter cells. Adv Exp Med Biol. 2006; 581: 265–269. https://doi.org/10.1007/978-0-387-33012-9_44 PMID: 17037540
14. Lee I, Bos S, Li G, Wang S, Gadea G, Desprès P, et al. Probing Molecular Insights into Zika Virus−Host Interactions. Viruses. 2018; 10. https://doi.org/10.3390/v10050233 PMID: 29724036
15. Barrass SV, Butcher SJ. Advances in high-throughput methods for the identification of virus receptors. Med Microbiol Immunol (Berl). 2020; 209: 309–323. https://doi.org/10.1007/s00430-019-00653-2 PMID: 31865406
16. Murali S, Rustandi RR, Zheng X, Payne A, Shang L. Applications of Surface Plasmon Resonance and Biolayer Interferometry for Virus-Ligand Binding. Viruses. 2022; 14: 717. https://doi.org/10.3390/v14040717 PMID: 35458446
17. Li Q, Liu Q, Huang W, Li X, Wang Y. Current status on the development of pseudoviruses for enveloped viruses. Rev Med Virol. 2018; 28. https://doi.org/10.1002/rmv.1963 PMID: 29218769
18. Wang W, Zhao H, Han G-Z. Host-Virus Arms Races Drive Elevated Adaptive Evolution in Viral Receptors. J Virol. 2020; 94: e00684–20. https://doi.org/10.1128/JVI.00684-20 PMID: 32493827
19. Chen D, Tan C, Ding P, Luo L, Zhu J, Jiang X, et al. VThunter: a database for single-cell screening of virus target cells in the animal kingdom. Nucleic Acids Res. 2022; 50: D934–D942. https://doi.org/10.1093/nar/gkab894 PMID: 34634807
20. Zhang Z, Zhu Z, Chen W, Cai Z, Xu B, Tan Z, et al. Cell membrane proteins with high N-glycosylation, high expression and multiple interaction partners are preferred by mammalian viruses as receptors.

Bioinforma Oxf Engl. 2019; 35: 723–728. https://doi.org/10.1093/bioinformatics/bty694 PMID: 30102334

21. Jolly CL, Sattentau QJ. Attachment factors. Adv Exp Med Biol. 2013; 790: 1–23. https://doi.org/10.1007/978-1-4614-7651-1_1 PMID: 23884583

22. Bhella D. The role of cellular adhesion molecules in virus attachment and entry. Philos Trans R Soc Lond B Biol Sci. 2015; 370: 20140035. https://doi.org/10.1098/rstb.2014.0035 PMID: 25533093

23. Feng T, Zhang J, Chen Z, Pan W, Chen Z, Yan Y, et al. Glycosylation of viral proteins: Implication in virus-host interaction and virulence. Virulence. 2022; 13: 670–683. https://doi.org/10.1080/21505594.2022.2060464 PMID: 35436420

24. Li Y, Liu D, Wang Y, Su W, Liu G, Dong W. The Importance of Glycans of Viral and Host Proteins in Enveloped Virus Infection. Front Immunol. 2021; 12: 638573. https://doi.org/10.3389/fimmu.2021.638573 PMID: 33995356

25. Koehler M, Delguste M, Sieben C, Gillet L, Alsteens D. Initial Step of Virus Entry: Virion Binding to Cell-Surface Glycans. Annu Rev Virol. 2020; 7: 143–165. https://doi.org/10.1146/annurev-virology-122019-070025 PMID: 32396772

26. Amara A, Mercer J. Viral apoptotic mimicry. Nat Rev Microbiol. 2015; 13: 461–469. https://doi.org/10.1038/nrmicro3469 PMID: 26052667

27. Bohan D, Maury W. Enveloped RNA virus utilization of phosphatidylserine receptors: Advantages of exploiting a conserved, widely available mechanism of entry. PLoS Pathog. 2021; 17: e1009899. https://doi.org/10.1371/journal.ppat.1009899 PMID: 34555126

28. Zhang Z, Ye S, Wu A, Jiang T, Peng Y. Prediction of the Receptorome for the Human-Infecting Virome. Virol Sin. 2021; 36: 133–140. https://doi.org/10.1007/s12250-020-00259-6 PMID: 32725480

29. Wang J huai. Protein recognition by cell surface receptors: physiological receptors versus virus interactions. Trends Biochem Sci. 2002; 27: 122–126. https://doi.org/10.1016/s0968-0004(01)02038-2 PMID: 11893508

30. Mathez G, Cagno V. Viruses Like Sugars: How to Assess Glycan Involvement in Viral Attachment. Microorganisms. 2021; 9: 1238. https://doi.org/10.3390/microorganisms9061238 PMID: 34200288

31. Dugan AE, Peiffer AL, Kiessling LL. Advances in glycoscience to understand viral infection and colonization. Nat Methods. 2022; 19: 384–387. https://doi.org/10.1038/s41592-022-01451-0 PMID: 35396476

32. Lasso G, Mayer SV, Winkelmann ER, Chu T, Elliot O, Patino-Galindo JA, et al. A Structure-Informed Atlas of Human-Virus Interactions. Cell. 2019; 178: 1526–1541.e16. https://doi.org/10.1016/j.cell.2019.08.005 PMID: 31474372

33. Valero-Rello A, Sanjuán R. Enveloped viruses show increased propensity to cross-species transmission and zoonosis. Proc Natl Acad Sci U S A. 2022; 119: e2215600119. https://doi.org/10.1073/pnas.2215600119 PMID: 36472956

34. Sieben C, Sezgin E, Eggeling C, Manley S. Influenza A viruses use multivalent sialic acid clusters for cell binding and receptor activation. PLoS Pathog. 2020; 16: e1008656. https://doi.org/10.1371/journal.ppat.1008656 PMID: 32639985

35. Fujioka Y, Nishide S, Ose T, Suzuki T, Kato I, Fukuhara H, et al. A Sialylated Voltage-Dependent Ca2+ Channel Binds Hemagglutinin and Mediates Influenza A Virus Entry into Mammalian Cells. Cell Host Microbe. 2018; 23: 809–818.e5. https://doi.org/10.1016/j.chom.2018.04.015 PMID: 29779930

36. Rahman SK, Ansari MA, Gaur P, Ahmad I, Chakravarty C, Verma DK, et al. The Immunomodulatory CEA Cell Adhesion Molecule 6 (CEACAM6/CD66c) Is a Protein Receptor for the Influenza a Virus. Viruses. 2021; 13: 726. https://doi.org/10.3390/v13050726 PMID: 33919410

37. Carlson CJ, Zipfel CM, Garnier R, Bansal S. Global estimates of mammalian viral diversity accounting for host sharing. Nat Ecol Evol. 2019; 3: 1070–1075. https://doi.org/10.1038/s41559-019-0910-6 PMID: 31182813

38. Guirimand T, Delmotte S, Navratil V. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic Acids Res. 2015; 43: D583–587. https://doi.org/10.1093/nar/gku1121 PMID: 25392406

39. Li P-H, Chen T-F, Yu J-Y, Shih S-H, Su C-H, Lin Y-H, et al. pubmedKB: an interactive web server for exploring biomedical entity relations in the biomedical literature. Nucleic Acids Res. 2022; 50: W616–W622. https://doi.org/10.1093/nar/gkac310 PMID: 35536289

40. Sung M, Jeong M, Choi Y, Kim D, Lee J, Kang J. BERN2: an advanced neural biomedical named entity recognition and normalization tool. Bioinforma Oxf Engl. 2022; 38: 4837–4839. https://doi.org/10.1093/bioinformatics/btac598 PMID: 36053172

41. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database J Biol Databases Curation. 2020; 2020: baaa062. https://doi.org/10.1093/database/baaa062 PMID: 32761142

42. Carlson CJ, Gibb RJ, Albery GF, Brierley L, Connor RP, Dallas TA, et al. The Global Virome in One Network (VIRION): an Atlas of Vertebrate-Virus Associations. mBio. 2022; 13: e0298521. https://doi.org/10.1128/mbio.02985-21 PMID: 35229639

43. Hu Z, Yuan J, Long M, Jiang J, Zhang Y, Zhang T, et al. The Cancer Surfaceome Atlas integrates genomic, functional and drug response data to identify actionable targets. Nat Cancer. 2021; 2: 1406–1422. https://doi.org/10.1038/s43018-021-00282-w PMID: 35121907

44. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinforma Oxf Engl. 2019; 35: 526–528. https://doi.org/10.1093/bioinformatics/bty633 PMID: 30016406

45. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012; 8: e1002764. https://doi.org/10.1371/journal.pgen.1002764 PMID: 22807683

46. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol. 2015; 32: 1342–1353. https://doi.org/10.1093/molbev/msv022 PMID: 25697341

47. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005; 21: 676–679. https://doi.org/10.1093/bioinformatics/bti079 PMID: 15509596

48. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2016. pp. 785–794. https://doi.org/10.1145/2939672.2939785