

---

The following resources related to this article are available online at <http://stke.sciencemag.org>.  
This information is current as of 30 September 2010.

---

- Article Tools** Visit the online version of this article to access the personalization and article tools:  
<http://stke.sciencemag.org/cgi/content/full/sigtrans;3/137/rs2>
- Supplemental Materials** "*Supplementary Materials*"  
<http://stke.sciencemag.org/cgi/content/full/sigtrans;3/137/rs2/DC1>
- Related Content** The editors suggest related resources on *Science's* sites:  
<http://stke.sciencemag.org/cgi/content/abstract/sigtrans;3/137/eg6>
- References** This article cites 73 articles, 48 of which can be accessed for free:  
<http://stke.sciencemag.org/cgi/content/full/sigtrans;3/137/rs2#otherarticles>
- Glossary** Look up definitions for abbreviations and terms found in this article:  
<http://stke.sciencemag.org/glossary/>
- Permissions** Obtain information about reproducing this article:  
<http://www.sciencemag.org/about/permissions.dtl>

# Collection and Motif-Based Prediction of Phosphorylation Sites in Human Viruses

Daniel Schwartz\*† and George M. Church

(Published 31 August 2010; Volume 3 Issue 137 rs2)

Although various databases have been established that are designed to compile an ever-growing list of protein phosphorylation sites in plants and animals, no such repository exists for viruses. Here, we developed the viral posttranslational modification (virPTM) database, which contains a comprehensive list of 329 accurately localized phosphorylation sites in proteins from 52 human viruses published between 1986 and the present. Additionally, to aid in the detection of new viral phosphorylation sites, we used the *scan-x* tool to make thousands of high-specificity serine, threonine, and tyrosine phosphorylation predictions in 229 viruses that replicate in human cells. By cross-validating our prediction results with the literature-based entries in the virPTM database, we highlight the effectiveness of the *scan-x* tool with viral data and extrapolate the existence of at least 4000 as yet unidentified phosphorylation sites on hundreds of viral proteins. Together, these results imply a substantial role for human kinases in mediating viral protein functions and suggest, more generally, that viral primary structure may provide important clues to aid in the rational design of therapeutic agents.

## INTRODUCTION

Over the past 35 years, the world has seen the emergence or reemergence of more than 30 human viruses, including HIV-1, HIV-2, West Nile virus, severe acute respiratory syndrome (SARS) coronavirus, and most recently, H1N1 influenza virus (1). Our understanding of viral evolution and pathogenesis, as well as our ability to combat viruses with rationally designed therapeutic agents, relies on a molecular-level knowledge of the mechanisms by which viruses redirect normal cellular resources for the purpose of viral replication. Most viruses have evolved to accomplish the task of hijacking cellular machinery through targeted interactions between their own proteins and those of their host. Although the interactions between viral and host-cell proteins are often difficult to elucidate experimentally, the primary structures of viral proteins hold direct clues in the form of short, linear protein motifs. Such linear amino acid residue signatures can serve as molecular beacons for various protein-protein interactions, which often result in posttranslational modification (PTM) of proteins, proteolytic cleavage, and altered protein localization (2).

Among the most widely studied motifs are those that mediate protein phosphorylation by kinases. To date, about one-fifth of the 518 human kinases have known associated motifs (3). Typically, these motifs are represented as linear sequence patterns by the single-letter abbreviations for the 20 amino acids. For example, the phosphorylation motif of the kinase Akt can be described by the pattern RxRxx[ST], where “x” denotes any amino acid residue, the underlined residues denote the phosphorylation site, and the brackets indicate that the kinase can phosphorylate either serine (S) or threonine (T) residues.

The chemical and structural diversity afforded by protein phosphorylation is used by nearly all cellular processes, with deficiencies in phosphorylation contributing to various human diseases (4). Due to technological advances in both enrichment strategies and high-throughput peptide se-

quencing by tandem mass spectrometry, the number of known phosphorylation sites across all studied species has grown exponentially over the past decade. To date, the literature cites more than 100,000 phosphorylation sites from organisms representing all domains of life. In response to the increase in the number of known phosphorylation sites, various phosphorylation site databases have emerged (5–13). These databases serve as organism-specific repositories for phosphorylation data and enable the research community to better assess the current state of knowledge.

Many viruses have evolved to use phosphorylation as a means of enhancing viral replication and inhibiting normal cellular functions, which has led to explorations of human kinases as potential antiviral drug targets (14). Phosphorylation in viruses is most often achieved through the action of host-cell kinases, although some viral families (for example, herpesviruses) encode their own protein kinases. Although specific phosphorylation sites on the most widely studied human viral proteins were uncovered starting in the mid 1980s, databases of viral phosphorylation sites have yet to be created. Thus, the extent of viral phosphorylation across all human virus families has been difficult to assess. Additionally, the widespread use of low-throughput techniques in the discovery and confirmation of phosphorylation sites in viral proteins suggests that most sites likely remain undiscovered.

Here, we addressed two of the aforementioned deficiencies. First, we compiled the first database of viral phosphorylation sites. Second, from a large collection of known human phosphorylation sites ( $n = 24,293$ ) gathered from various databases, in conjunction with the *scan-x* prediction tool (15), we have performed the first high-specificity prediction of viral phosphorylation sites in viruses that infect humans. By cross-validating these prediction results against our literature-based viral phosphorylation database, we were also able to extrapolate a minimum expected number of phosphorylation sites across all sequenced human viruses.

## RESULTS

### Composition of a viral phosphorylation database

Our database of viral phosphorylation sites, named “virPTM” for viral posttranslational modification, was compiled through a comprehensive literature

Department of Genetics, Harvard Medical School, Boston, MA 02115, USA.

\*Present address: Department of Physiology and Neurobiology, University of Connecticut, Storrs, CT 06269, USA.

†To whom correspondence should be addressed. E-mail: daniel.schwartz@uconn.edu

review of 229 viruses and virus families known to replicate in human hosts. The database currently contains 329 unique viral phosphorylation sites on 111 proteins from 52 viruses and is composed of precisely localized sites detected both *in vivo* and *in vitro*. In total, 150 journal references were used to curate the database. Phosphorylation sites that are known to occur, but that have not been precisely localized, were not included in the present version of virPTM. Each entry in the database contains (i) the virus name, (ii) the viral protein name, (iii) the viral protein sequence, (iv) the site of the modification, (v) the local primary structure (that is, seven amino acid residues upstream and downstream of the modification site), and (vi) the reference in which experimental detection of the site is described. In addition to existing in its entirety online (<http://virptm.hms.harvard.edu>), the entries of the database (with the exception of whole-protein sequences) are provided here (table S1).

Inspection of the virPTM database indicates the steady increase in the number of literature-based phosphorylation sites in human viruses between 1986 and the present (Fig. 1). Forty percent of all precisely localized viral phosphorylation sites come from the following five well-studied viruses: human herpesvirus 4 (HHV-4; 32 sites on seven proteins), human adenovirus 5 (HAdV-5; 27 sites on six proteins), HHV-1 (26 sites on nine proteins), HIV-1 (25 sites on nine proteins), and HHV-5 (24 sites on five proteins) (Table 1). At present, the virPTM database contains 250 serine phosphorylation sites (76% of the total), 65 threonine phosphorylation sites (20%), and 14 tyrosine phosphorylation sites (4%). Because most viruses do not encode their own protein kinases, most sites contained within virPTM are expected to be the result of phosphorylation by human kinases; however, the database also contains 14 sites that are phosphorylated by viral kinases and 11 sites that are autophosphorylated by viral kinases.

### Prediction of viral phosphorylation sites with *scan-x*

Because the number of experimentally verified phosphorylation sites in human viruses that are contained within virPTM likely represents only a small fraction of all viral phosphorylation sites, much experimental work is needed to uncover the complete phosphoproteome of human viruses.

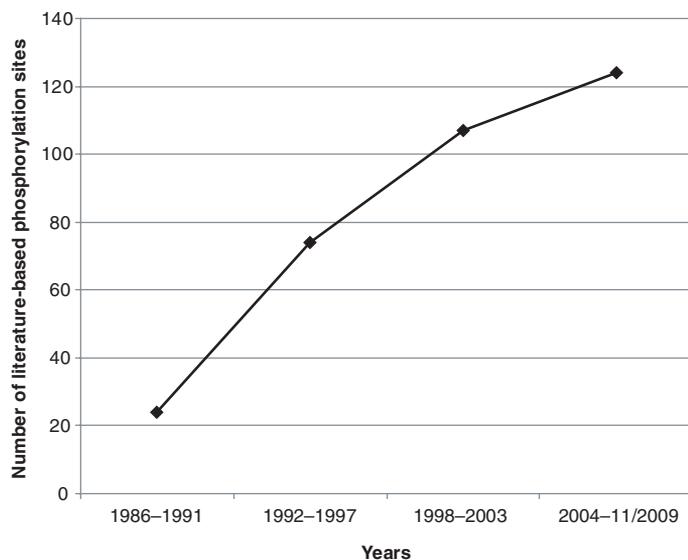


Fig. 1. New viral phosphorylation sites reported in the literature over time. Number of new precisely localized human viral phosphorylation sites cited in the literature between 1986 and November 2009.

Testing each serine, threonine, and tyrosine residue for phosphorylation on candidate viral proteins is often prohibitive; therefore, a means to prioritize sites with a higher likelihood of phosphorylation is required. To this end, we have also included the results of our high-confidence human phosphorylation predictor, *scan-x* (15), on the virPTM Web site (the complete set of predictions is also provided in table S2). By running *scan-x* on the complete set of 229 human viral proteomes (which consist of 3744 proteins) available from the National Center for Biotechnology Information (NCBI) Web site (<http://www.ncbi.nlm.nih.gov/sites/genome/>), we predicted 6723 high-confidence serine and threonine phosphorylation sites and 3202 high-confidence tyrosine phosphorylation sites on 2881 viral proteins. These results provide the residues with the highest likelihood of modification (as deemed by *scan-x*) in the total set of 211,018 serine, threonine, and tyrosine residues contained within all human viral proteins. Because each potential modification site is given a *scan-x* score, these sites may be further prioritized by investigators to yield predictions even beyond the high-specificity thresholds used in our analyses.

### Sensitivity and specificity of *scan-x* on viral data

Although extensive cross-validation was performed during the development of the *scan-x* tool, and its improved performance relative to other human phosphoprediction software has been described previously (15), we sought to assess the expected predictive performance of *scan-x* in the context of viral phosphorylation data. The assessment of a predictive tool such as *scan-x* typically involves calculations of sensitivity and specificity. In the present application, sensitivity can be described as the proportion of phosphorylation sites called “positive” by the predictor relative to all known phosphorylation sites, whereas specificity can be described as the proportion of phosphorylation sites called “negative” by the predictor relative to all known unphosphorylated sites. Thus, accurate calculations of sensitivity and specificity require lists of known phosphorylated and unphosphorylated sites, which effectively serve as positive and negative controls for the method. Such lists were obtained with the sequence information already contained in the virPTM database.

For the positive control, we started with the complete set of 329 known literature-based phosphorylation sites. Because *scan-x* requires a minimum of seven residues on each side of the phosphorylation site to make a prediction, phosphorylation sites in the database that were within seven residues of the N or C terminus were removed. Furthermore, given that *scan-x* predictions are based on the human kinome, viral phosphorylation sites that are the result of virally encoded kinases were also removed. After this filtering procedure, 284 viral phosphorylation sites remained, including 272 serine and threonine and 12 tyrosine phosphorylation sites (table S3). This positive control data set enabled us to compute the sensitivity of *scan-x* on viral phosphorylation data. *scan-x* correctly predicted 116 of the 272 serine and threonine phosphorylation sites, thus yielding a serine and threonine phosphorylation sensitivity of 42.6%. With regard to the tyrosine data, *scan-x* correctly predicted 9 of the 12 sites, thus yielding a tyrosine phosphorylation sensitivity of 75.0% (Table 2).

Computing the specificity of *scan-x* with respect to viral phosphorylation data required a list of unphosphorylated serine, threonine, and tyrosine residues (that is, negative phosphorylation sites) to serve as a negative control; however, obtaining such a list of unphosphorylated residues is difficult for several reasons. First, one cannot be certain that a particular residue does not get phosphorylated under all possible cellular conditions, and second, viral phosphorylation has not been comprehensively studied to date. To overcome these challenges, we used a proxy for negative phosphorylation sites on the basis of the assumption that most serine, threonine, and tyrosine residues in viral proteins remain in an unphosphorylated state. Thus, all serine, threonine, and tyrosine residues from previously studied

**Table 1.** Viral phosphorylation sites and the corresponding references contained in the virPTM database for the five viruses that contain the greatest number of phosphorylation sites. EBV, Epstein-Barr virus; HCMV,

human cytomegalovirus. Abbreviations for the amino acids are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

Viral protein	Residue	Sequence*	Reference	Viral protein	Residue	Sequence*	Reference
<i>HHV-4 or EBV</i>				<i>HHV-1</i>			
BMRF1	S <sup>337</sup>	HTVSPSPSPPPPPRT	(19)	Glycoprotein B	T <sup>887</sup>	VMRKRRTNYTQVFN	(46)
	T <sup>344</sup>	SPPPPRTPTWESPA	(19)	ICP0	S <sup>224</sup>	GHTVRALSPTHPEPT	(47)
	S <sup>349</sup>	PRTPTWESPARPETP	(19)		T <sup>226</sup>	TVRALSPTHPEPTTD	(47)
	T <sup>355</sup>	ESPARPETPSPAIPS	(19)		T <sup>231</sup>	SPTHPEPTTDEDDDD	(47)
BZLF1	S <sup>186</sup>	RYKNRVASRKCKRAF	(20)		T <sup>232</sup>	PTHPEPTTDEDDDD	(47)
	S <sup>173</sup>	ESLEECDSELEIKRY	(21)		S <sup>365</sup>	NRDPVIVISDPPASP	(47)
	S <sup>167</sup>	RKPQQPESELECDSE	(21)		S <sup>367</sup>	DPVIVISDPPASPHR	(47)
	T <sup>14</sup>	TSQDVKFTDPYQVP	(22)		S <sup>371</sup>	ISDPPASPHRPPAA	(47)
	S <sup>209</sup>	EVASAKSSENDRLRL	(23)		S <sup>508</sup>	RPRKRRGSGQENPSP	(47)
EBNA-1	S <sup>334</sup>	GGRGRGGSGGRGRGG	(24)		S <sup>514</sup>	GSGQENPSPQSTRPP	(47)
	T <sup>8</sup>	MSDEGPGTGPNGGLG	(25)		S <sup>517</sup>	QENPSPQSTRPPLAP	(47)
	T <sup>20</sup>	GLGKGDTSQPEGSG	(25)		T <sup>518</sup>	ENPSPQSTRPPLAPA	(47)
	S <sup>21</sup>	LGEKGDTSQPEGSGG	(25)	ICP4	T <sup>180</sup>	WRPSASSSSSDSGSS	(48)
	S <sup>60</sup>	RPGAPGSGSGRHR	(25)	ICP22	Y <sup>116</sup>	DDTSDWSYDDIPPRP	(49)
	S <sup>62</sup>	GAPGGSGSGRHRDG	(25)	ICP27(UL54)	S <sup>114</sup>	RLGARRPSCSPEQHG	(50)
	S <sup>78</sup>	RRPQKRPSCIGCKGT	(25)		S <sup>16</sup>	IDLGLDLSDSLDLDD	(50)
	S <sup>365</sup>	RERARGGSRERARGR	(25)		S <sup>18</sup>	LGLDLSDDLDEDDPP	(50)
	S <sup>383</sup>	RGEKRRPSPSSQSSS	(25)	VP16	S <sup>375</sup>	RTKNNYGSTIEGLLD	(51)
	S <sup>393</sup>	SQSSSSGSPRRPPP	(25)		S <sup>411</sup>	AGHTRRLSTAPPTDV	(51)
EBNA-2	S <sup>469</sup>	FETTESPSSDEYVE	(26)		S <sup>18</sup>	DMNADGASPPPPRPA	(52)
	S <sup>243</sup>	TELQPTFSPPRMHLF	(27)		S <sup>353</sup>	SYSSFTTSPSEAVMR	(52)
EBNA-LP	S <sup>35</sup>	LRRHRSPSPTRGGQE	(28)		S <sup>452</sup>	DMLGDGDSGPGGFTF	(52)
LMP-1	S <sup>313</sup>	LPHSPSDSAGNDGGP	(29)	UL34	T <sup>195</sup>	AITRRRRTRRSREAY	(53)
	T <sup>324</sup>	DGGPPQLTEEVENKG	(29)		S <sup>198</sup>	RRRTRRSREAYGAE	(53)
	S <sup>211</sup>	PQQATDDSGHESDSN	(30)	UL11	S <sup>40</sup>	FDVVDIESEEEGNFY	(54)
	S <sup>215</sup>	TDDSGHESDSNSNEG	(30)	US3	S <sup>147</sup>	PPGIRRRSRDEIGAT	(55)
LMP-2A	S <sup>15</sup>	PMGAGPPSPGGDPDG	(31)				
	S <sup>102</sup>	GLPPPPYSPRDDSSQ	(31)				
	Y <sup>112</sup>	DDSSQHIYEEAGRGS	(31)	CAp24	S <sup>109</sup>	SDIAGTTSTLQEQIG	(56)
	Y <sup>74</sup>	NGDRHSDYQPLGTQD	(32)		S <sup>149</sup>	VRMYSPTSLDIRQG	(56)
	Y <sup>85</sup>	GTQDQSLYLGLQHDG	(32)		S <sup>178</sup>	TLRAEQASQEVKNWM	(56)
	Y <sup>64</sup>	PPPYEDPYWGNQDRH	(33)	Gag	S <sup>111</sup>	IEEEQNKSKKKAQQA	(57)
					Y <sup>132</sup>	SNQVSQNY-	(58)
				Nef	T <sup>15</sup>	SSVIGWPTVREMRMR	(59)
100k	Y <sup>365</sup>	KLEETHLYTFRQGYV	(34)		S <sup>6</sup>	-MGGKWSKSSVIGW	(60)
	Y <sup>682</sup>	LRKGRGVYLDPQSGE	(34)		S <sup>9</sup>	GGKWSKSSVIGWPTV	(60)
DBP	T <sup>12</sup>	EEEQRETTPERGRGA	(35)	Rev	S <sup>99</sup>	SGTQGVGSPQILVES	(61)
	S <sup>31</sup>	PTMEDVSSPSPSPPP	(35)		S <sup>5</sup>	-MAGRSQSDSEEL	(62)
	S <sup>33</sup>	MEDVSSPSPSPPPPR	(35)		S <sup>8</sup>	MAGRSQSDSEELIRT	(62)
	S <sup>35</sup>	DVSSPSPSPPPPRAP	(35)	Tat	S <sup>62</sup>	RRRAHQNSQTHQASL	(63)
	S <sup>70</sup>	ALVPRTPSPRPSTSA	(35)		T <sup>64</sup>	RAHQNSQTHQASLSK	(63)
	T <sup>75</sup>	TPSPRPSTSAADLAI	(35)		S <sup>68</sup>	NSQTHQASLSKQPTS	(63)
	S <sup>76</sup>	PSPRPSTSAADLAI	(35)	Vif	S <sup>144</sup>	AGHNKVGSLQYLALA	(64)
	S <sup>92</sup>	KKKKKRPSKPERPP	(35)		T <sup>155</sup>	LALAALITPKKIKPP	(64)
	S <sup>100</sup>	PKPERPSPPEVIVDS	(35)		T <sup>188</sup>	KGHRGSHTMNGH-	(64)
	S <sup>107</sup>	SPEVIVDSEEREDV	(35)		T <sup>96</sup>	WRKKRYSTQVDPELA	(65)
E1A	S <sup>219</sup>	AILRRPTSPVSRASN	(36)		S <sup>165</sup>	KIKPPLPSVTKLTD	(65)
	S <sup>231</sup>	ECNSSTDSCDSGPSN	(36)	Vpr	S <sup>79</sup>	FRIGCRHSRIGVTRQ	(66)
	S <sup>89</sup>	TFPPAPGSPPEPHLS	(37)		S <sup>94</sup>	RRRNGASRS-	(66)
	S <sup>96</sup>	SPEPPHLSRQPEQPE	(38)		S <sup>96</sup>	ARRNGASRS-	(66)
	S <sup>132</sup>	HEAGFPSPDDEDEEG	(39)	Vpu	S <sup>53</sup>	LIERAEDSGNESEGE	(67)
	S <sup>185</sup>	TCGMFVYSPVSEPEP	(40)		S <sup>57</sup>	AEDSGNESEGEISAL	(67)
	S <sup>188</sup>	MFVYSPVSEPEPEPE	(40)		S <sup>62</sup>	NESEGEISALVEMGV	(68)
	S <sup>227</sup>	PVSRECNSTSDSCDS	(41)				
	S <sup>228</sup>	VSRECNSTSDSCDSG	(41)				
	S <sup>234</sup>	SSTSDSCDSGSPNTTP	(41)	Glycoprotein B	S <sup>900</sup>	GYRHLKDSDEEENV-	(69)
	S <sup>237</sup>	DSCDSGFSNTTPEIH	(41)	IE1	S <sup>399</sup>	LATAGVSSSDSLVSP	(70)
E1B	S <sup>164</sup>	RRRRQEQSPWNPRAG	(42)		S <sup>402</sup>	SSSDSLVSPPEPVP	(70)
	T <sup>495</sup>	FGSSDEDTD-	(43)		S <sup>406</sup>	SLVSPPEPVPATIP	(70)
E3 RID-β	S <sup>116</sup>	ANQPRPPSPPTPEIS	(44)		S <sup>423</sup>	SVIVAENSQDEESEQ	(70)
Pol	S <sup>67</sup>	TSGGSPASPLLDASS	(45)		S <sup>428</sup>	ENSDQEESEQSDEEE	(70)
<i>HAAdV-5</i>				<i>HHV-5 or HCMV</i>			
100k	Y <sup>365</sup>	KLEETHLYTFRQGYV	(34)				
	Y <sup>682</sup>	LRKGRGVYLDPQSGE	(34)				
DBP	T <sup>12</sup>	EEEQRETTPERGRGA	(35)				
	S <sup>31</sup>	PTMEDVSSPSPSPPP	(35)				
	S <sup>33</sup>	MEDVSSPSPSPPPPR	(35)				
	S <sup>35</sup>	DVSSPSPSPPPPRAP	(35)				
	S <sup>70</sup>	ALVPRTPSPRPSTSA	(35)				
	T <sup>75</sup>	TPSPRPSTSAADLAI	(35)				
	S <sup>76</sup>	PSPRPSTSAADLAI	(35)				
	S <sup>92</sup>	KKKKKRPSKPERPP	(35)				
	S <sup>100</sup>	PKPERPSPPEVIVDS	(35)				
	S <sup>107</sup>	SPEVIVDSEEREDV	(35)				
E1A	S <sup>219</sup>	AILRRPTSPVSRASN	(36)				
	S <sup>231</sup>	ECNSSTDSCDSGPSN	(36)				
	S <sup>89</sup>	TFPPAPGSPPEPHLS	(37)				
	S <sup>96</sup>	SPEPPHLSRQPEQPE	(38)				
	S <sup>132</sup>	HEAGFPSPDDEDEEG	(39)				
	S <sup>185</sup>	TCGMFVYSPVSEPEP	(40)				
	S <sup>188</sup>	MFVYSPVSEPEPEPE	(40)				
	S <sup>227</sup>	PVSRECNSTSDSCDS	(41)				
	S <sup>228</sup>	VSRECNSTSDSCDSG	(41)				
	S <sup>234</sup>	SSTSDSCDSGSPNTTP	(41)				
	S <sup>237</sup>	DSCDSGFSNTTPEIH	(41)				
E1B	S <sup>164</sup>	RRRRQEQSPWNPRAG	(42)				
	T <sup>495</sup>	FGSSDEDTD-	(43)				
E3 RID-β	S <sup>116</sup>	ANQPRPPSPPTPEIS	(44)				
Pol	S <sup>67</sup>	TSGGSPASPLLDASS	(45)				

continued on next page

Downloaded from [stke.sciencemag.org](http://stke.sciencemag.org) on September 30, 2010

Viral protein	Residue	Sequence*	Reference
IE2	S <sup>431</sup>	DQEESQSDDEEEEG	(70)
	S <sup>448</sup>	EEREDTVSVKSEPV	(70)
	S <sup>451</sup>	EDTVSVKSEPVSEIE	(70)
	S <sup>455</sup>	SVKSEPVSEIEEVAP	(70)
	T <sup>27</sup>	SKVPRPETPVTKATT	(71)
	S <sup>144</sup>	ASATPELSPRKKPRK	(71)
	T <sup>233</sup>	GSGTPRVTSPTHPLS	(71)
	S <sup>234</sup>	SGTFRVTSPTHPLSQ	(71)
	S <sup>472</sup>	EEDTDESDNEIHNP	(72)
	pp65 UL97	S <sup>2</sup>	-MSALRSRA
S <sup>3</sup>		-MSSALRSRAR	(73)
S <sup>11</sup>		ALRSRARSASLGT	(73)
S <sup>13</sup>		RSRARSASLGT	(73)
T <sup>16</sup>		ARSASLGT	(73)
T <sup>18</sup>		SASLGT	(73)
S <sup>133</sup>		LRRPVVPTSSRGS	(73)
T <sup>134</sup>		RRPVVPTSSRGSAA	(73)
T <sup>190</sup>		SVRRALFTGGSDPSD	(73)

\*Bold residues denote phosphorylation sites and hyphens denote protein termini.

**Table 2.** Sensitivity and specificity of the *scan-x* predictor on viral phosphorylation data.

	Sensitivity (%)	Specificity (%)*
Serine and threonine phosphorylation	42.6	92.9
Tyrosine phosphorylation	75.0	90.9

\*Specificity values represent a lower bound on specificity.

viral proteins bearing at least one phosphorylation site, but not previously shown to be phosphorylated, were used as negative phosphorylation sites. It should be noted a priori, however, that this approach is conservative given that a percentage of the sites deemed negative are in fact true phosphorylation sites that have yet to be identified. With the proxy, calculated values of specificity represent the lower bound on the true specificity of the *scan-x* methodology. This approach resulted in 5360 negative serine and threonine phosphorylation sites and 1040 negative tyrosine phosphorylation sites among the viral proteins in the virPTM database (table S3). Among these, *scan-x* incorrectly predicted 383 serine and threonine phosphorylation sites and 96 tyrosine phosphorylation sites, thus yielding specificities of 92.9 and 90.9%, respectively. However, as mentioned earlier, we can reasonably assume that some percentage of these incorrectly predicted phosphorylation sites are in fact true phosphorylation sites, which suggests that the actual specificities of *scan-x* are greater than the reported values.

### Description of the virPTM version 1.0 Web site

The virPTM Web site is intended to serve as a dynamic repository for the research community of PTM sites in viral proteins. At present, the Web site contains information on viral phosphorylation sites; however, additional viral PTMs are expected to be added in the future. The Web site can be subdivided into two main sections: (i) literature-based viral phosphorylation sites and (ii) *scan-x* viral phosphorylation predictions. Links to both of these sections are visible on the main Web site page. The main page also contains a link to the positive and negative control data sets that were used to evaluate the sensitivity and specificity of the *scan-x* predictor.

Clicking on the “literature-based viral phosphorylation sites” link on the main page enables users to view the experimentally validated phosphorylation sites contained in the database that have been collated from the literature (Fig. 2A). The viruses are sorted alphabetically by their acronyms.

The format of the entries is the same as shown here (Table 1 and table S1), but also includes links to the complete protein sequence and the corresponding reference for the phosphorylation site in PubMed. Phosphorylation sites are highlighted in red with dashes indicating the N or C terminus of a protein. Asterisks in the “Residue & Position” column of the database denote phosphorylation sites that are due to virally encoded kinases.

To access the predictive section of the virPTM Web site, users should click on the “*scan-x* viral phosphorylation predictions” link, which provides an alphabetically sorted list of the 229 viruses known to replicate in human cells about which phosphorylation predictions have been made. Clicking on the name of a virus opens a new Web page that contains the *scan-x* prediction results for that virus (see screenshot in Fig. 2B). The *scan-x* prediction results page can be divided into two regions. A list of viral proteins about which phosphorylation predictions have been made is enclosed in two horizontal lines below the heading of the page. Each protein has a total score (equal to the sum of all site-specific scores on the protein) and a maximum score (equal to the highest-scoring predicted site on the protein) located beside it. Proteins on the *scan-x* results page are sorted according to their maximum score. Numbered links beside each protein, which represent the overall rank of the protein in the entire viral prediction data set, enable users to jump down to the prediction results for that protein. The virPTM logo in the upper right-hand corner of the screen provides a link back to the main virPTM Web page.

On a sample screenshot of the *scan-x* prediction results for the NP-1 protein of the human bocavirus, the numbers “164/2881” indicate that a predicted phosphorylation site had the 164th highest maximum *scan-x* score among the total set of 2881 viral proteins that were scanned (Fig. 2C). The numbers “186.167” and “29.560” indicate the total and maximum *scan-x* scores on the protein, respectively. Predicted phosphorylation sites correspond to the central residues of the 15–amino acid peptides located beneath the viral protein sequence. The number after the 15–amino acid peptide sequence indicates the *scan-x* score of the peptide. The number in parentheses after the *scan-x* score indicates the position of the predicted phosphorylation site in the context of the overall protein sequence. Residues in the 15–amino acid peptides may be colored green, red, or blue. Green residues indicate a positive contribution to the overall prediction, whereas red residues indicate a negative contribution to the overall prediction. Blue residues highlight the *motif-x*–derived phosphorylation motif under which the prediction was made. Phosphorylation sites may be predicted with multiple motifs. Thus, Ser<sup>18</sup> in the NP-1 protein of human bocavirus is predicted to be phosphorylated under the RxxS, SP, RxxSP, SPxR, and SPxxxxK motifs (Fig. 2C). Viewing the overall prediction results for the NP-1 protein indicates seven total phosphorylation sites predicted by *scan-x* including Ser<sup>18</sup>, Ser<sup>35</sup>, Ser<sup>37</sup>, Ser<sup>39</sup>, Tyr<sup>52</sup>, Ser<sup>194</sup>, and Thr<sup>202</sup>. Although no specific phosphorylation sites have been previously mapped to the NP-1 protein of human bocavirus, phosphorylation of the homologous NP-1 protein in bovine parvovirus (from which human bocavirus partially derives its name: bovine-canine virus) has been known for more than two decades (16), thus lending support to the *scan-x* phosphorylation predictions for this protein.

### DISCUSSION

Given the recent increase in the amount of phosphorylation data because of protein sequencing enabled by tandem mass spectrometry, various general and organism-specific phosphorylation databases have emerged (5–13). Yet, in the field of human virology, no such resource has previously been compiled. To address this need, we have collated a database of in vitro and in vivo viral phosphorylation sites based on a thorough literature review. The database currently contains information on precisely localized phosphorylation sites in viruses known to infect humans; however, it should be

considered a work in progress. Future improvements to the database are likely to consist of (i) the inclusion of other PTMs that act on viral proteins (for example, acetylation, glycosylation, and SUMOylation, among others); (ii) the inclusion of proteins that are modified but whose sites of modification have not been determined; (iii) the inclusion of viruses that infect hosts other than humans; and (iv) notes on the conditions under which the modifications were detected (whether in vivo or in vitro, identifying the modifying kinase involved, etc.).

Because viral phosphorylation sites can often provide unique insights into interacting host proteins due to their motif signatures, high-confidence

predictions of viral phosphorylation sites may be used to better understand the biological pathways exploited by these disease-causing agents. To this end, we have also provided the research community with the first database of predicted viral phosphorylation sites with the *scan-x* algorithm. Because no viral phosphorylation sites were used in the training of the *scan-x* predictor, the literature-based viral phosphorylation sites contained within the virPTM database served as an unbiased test set to characterize the sensitivity and specificity of the *scan-x* tool on viral phosphorylation data. Furthermore, this data set may also be used to benchmark existing and future human phosphorylation prediction software. The virPTM Web site provides a

set of positive and presumed negative phosphorylation sites, which will be regularly updated.

The evaluation of the *scan-x* predictor with literature-based viral phosphorylation sites contained in the virPTM database has also provided the ability to extrapolate the breadth of viral protein phosphorylation more generally. In the total set of 229 human viral proteomes, *scan-x* predicted 6723 serine and threonine phosphorylation sites and 3202 tyrosine phosphorylation sites. Assuming the same sensitivity and specificity levels achieved in the literature-based data, we can expect that of these sites, at a minimum, 1563 serine and threonine residues and 274 tyrosine residues are correctly assigned as phosphorylated by *scan-x*. We can further extrapolate that a minimum of 3677 serine and threonine and 365 tyrosine phosphorylation sites should currently exist on all proteins from human viruses (see Materials and Methods), most of which (>90%) remain undiscovered. Whereas these statistics suggest that, at worst, only 23% (that is, 1563 of 6723) of the *scan-x* serine and threonine predictions and 9% (274 of 3202) of the *scan-x* tyrosine predictions may be correct, when compared to the expected odds of identifying phosphorylated serine and threonine, as well as tyrosine residues at random in all viral proteins (2 and 1%, respectively), the utility of *scan-x* is apparent.

Finally, because *scan-x* operates at the level of phosphorylation sequence motifs (often unique to kinases or kinase families), the inference of the modifying kinase for known or predicted phosphorylation sites can also potentially be made. A proof of principle is provided by simian virus 40 (SV40), which contains

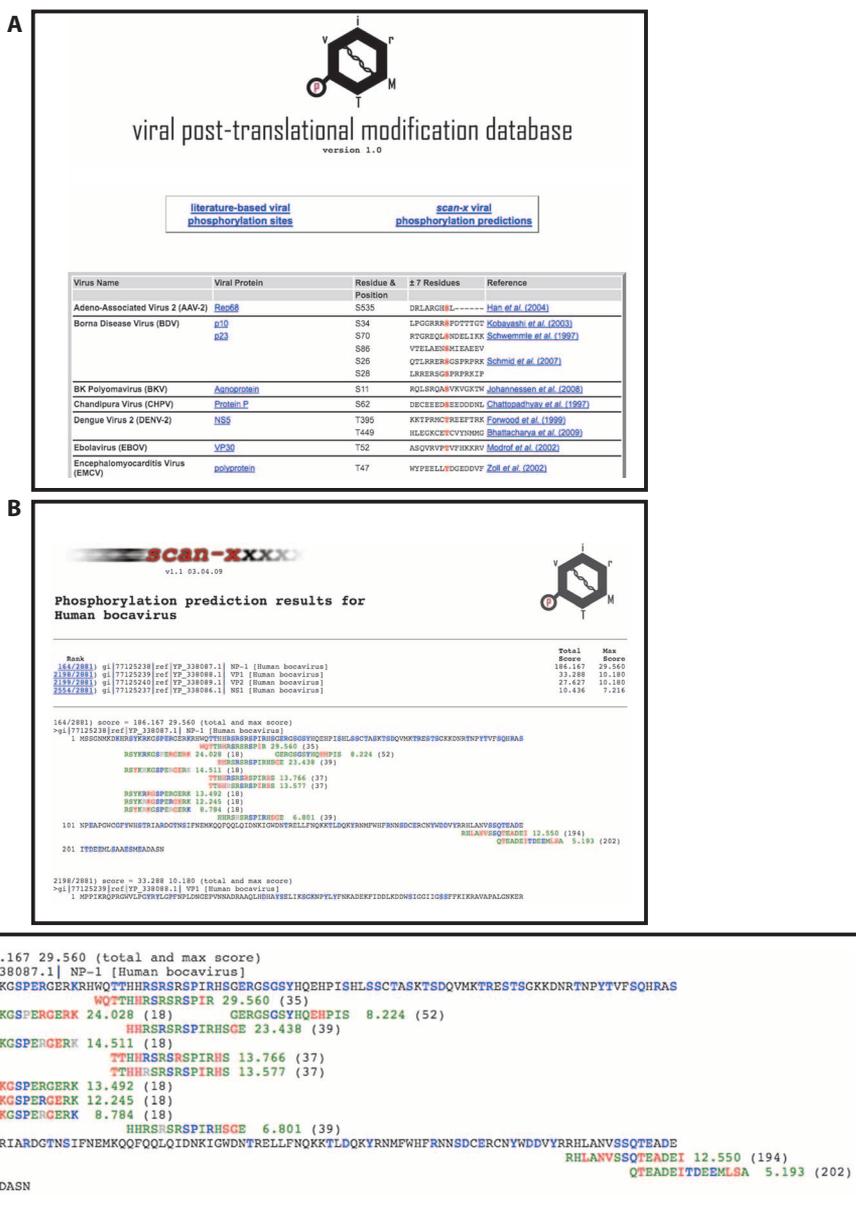


Fig. 2. Screenshots from the virPTM Web site. (A) Web page with table showing literature-based viral phosphorylation sites. (B) Sample *scan-x* prediction results page for human bocavirus proteins. (C) Detailed enlargement of *scan-x* phosphorylation predictions on the NP-1 protein of the human bocavirus highlighting seven predicted phosphorylation sites at positions Ser<sup>18</sup>, Ser<sup>35</sup>, Ser<sup>37</sup>, Ser<sup>39</sup>, Tyr<sup>52</sup>, Ser<sup>194</sup>, and Thr<sup>202</sup> (see text for greater detail).

13 annotated phosphorylation sites in the virPTM database (8 of which were correctly predicted by *scan-x*), as well as an additional 10 *scan-x* serine and threonine phosphorylation predictions that have yet to be validated. Among the complete list of 23 known and predicted serine and threonine phosphorylation sites are 6 sites (23%) that conform to an “SQ” motif, a known phosphorylation signature for the kinase ataxia-telangiectasia mutated (ATM). The large proportion of SQ motifs on SV40 proteins suggests that ATM might be an important interacting protein for SV40. That previous research has demonstrated that ATM controls the replication of SV40 *in vivo* (17) lends strong support to the notion that known and predicted phosphorylation motifs on viral proteins may be used as starting points to explore and understand interactions between viral proteins and host kinases. Although the present study has dealt exclusively with phosphorylation, the concept of exploring viral primary structure as a means of synthesizing hypotheses regarding viral-host interactions is broadly generalizable and may ultimately be used in the rational design and testing of antiviral drugs.

## MATERIALS AND METHODS

### Curation of the virPTM database

To find the subset of viruses with sequenced genomes and proteomes that are able to replicate in human hosts, we started with a complete list of sequenced viruses ( $n = 2947$ ) from the NCBI Web site (<http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=10239&type=5&name=Viruses>) and performed literature searches to determine the host range of each virus. This analysis resulted in the retrieval of 229 viruses (and viral proteomes) that are able to replicate in human cells, which represented a starting point for further data mining. The virPTM database was then collated through directed Internet searches on Google, PubMed, and various scientific journal Web sites with keywords specific to each of the 229 human viruses (for example, virus name) and the subject of protein phosphorylation (for example, “kinase,” “phosphorylation,” “phosphorylation site,” etc.). Retrieved journal articles were also further mined for relevant citations relating to viral phosphorylation. Phosphorylation sites were only added to the database if the experimental evidence unambiguously localized the site of modification. For example, if the mutation of two adjacent potential phosphorylation sites showed a decrease in the overall extent of protein phosphorylation, but these sites were not disambiguated individually, then neither site was included in the database. Amino acid sequences of proteins bearing phosphorylation sites that are contained in the virPTM database were preferentially obtained directly from the reference (if published) or from the reference sequence for the viral protein obtained on the NCBI or the ViralZone Web site (<http://www.expasy.ch/viralzone/>). So as not to introduce bias, all decisions regarding the inclusion of specific phosphorylation sites in the virPTM database were made blind with regard to their prediction status by *scan-x*. In cases in which the same modification site was detected by multiple references, the reference with the earliest publication date was used in the database. The process of collation was performed until repeated Internet or literature searches yielded no additional phosphorylation sites beyond those already included in the database. The online version of the database will be updated periodically.

### Prediction of viral phosphorylation sites

All viral phosphorylation sites predicted in the present study (that is, from the virPTM database, from whole proteomes of human viruses, or from the proteins on which at least one literature-based phosphorylation site was detected) were predicted with the previously described human *scan-x* predictor (15). Briefly, the *scan-x* predictor works by scanning the motifs extracted by *motif-x* (18) from a large database ( $n = 24,293$ ) of human ser-

ine, threonine, and tyrosine phosphorylation sites against a sequence of interest, thus deriving a *scan-x* score. Details of the *scan-x* scoring procedure and the list of human motifs used in the creation of the *scan-x* predictor were provided previously (15). Serine and threonine sites matching a motif with at least three fixed positions, two fixed positions, or one fixed position (including the central residue) that received minimum *scan-x* scores of 2.2, 12.2, or 42.2, respectively, were deemed predicted serine and threonine phosphorylation sites. Similarly, tyrosine sites matching a motif with at least three fixed positions, two fixed positions, or one fixed position (including the central residue) that received minimum *scan-x* scores of  $-3.3$ , 6.7, or 36.7, respectively, were deemed predicted tyrosine phosphorylation sites. The aforementioned *scan-x* score cutoffs were used because they corresponded to the 95% specificity level by a 10-fold cross-validation procedure performed on the human phosphorylation data set (15).

### Sensitivity, specificity, and projection calculations

Sensitivity was calculated as (number of true positives)/(number of true positives + number of false negatives), and specificity was calculated as (number of true negatives)/(number of true negatives + number of false positives). The minimum number of correct *scan-x* identifications was calculated by multiplying the total number of viral *scan-x* predictions by the positive predictive value [PPV = (number of true positives)/(number of true positives + number of false positives)], which was calculated with the virPTM data set. To project the minimum total number of expected phosphorylation events in all human viruses, we divided the minimum number of correct *scan-x* identifications by the sensitivity (calculated with the virPTM database). The expected odds of identifying a phosphorylated residue at random was calculated as the projected minimum number of total viral phosphorylation sites divided by the total number of residues that may be phosphorylated in all viral proteomes.

## SUPPLEMENTARY MATERIALS

[www.sciencesignaling.org/cgi/content/full/3/137/rs2/DC1](http://www.sciencesignaling.org/cgi/content/full/3/137/rs2/DC1)

Table S1. Literature-based viral phosphorylation sites contained in the virPTM database version 1.0.

Table S2. *scan-x* phosphorylation prediction results on 229 human viral proteomes.

Table S3. Positive and negative control data sets used in the determination of sensitivity and specificity of the *scan-x* predictor on viral phosphorylation data.

## REFERENCES AND NOTES

1. R. Zell, A. Krumbholz, P. Wutzler, Impact of global warming on viral diseases: What is the evidence? *Curr. Opin. Biotechnol.* **19**, 652–660 (2008).
2. V. Neduva, R. B. Russell, Peptides mediating interaction networks: New leads at last. *Curr. Opin. Biotechnol.* **17**, 465–471 (2006).
3. M. L. Miller, L. J. Jensen, F. Diella, C. Jørgensen, M. Tinti, L. Li, M. Hsiung, S. A. Parker, J. Bordeaux, T. Sicheritz-Ponten, M. Olhovskiy, A. Pasculescu, J. Alexander, S. Knapp, N. Blom, P. Bork, S. Li, G. Cesareni, T. Pawson, B. E. Turk, M. B. Yaffe, S. Brunak, R. Linding, Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.* **1**, ra2 (2008).
4. P. Cohen, The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *Eur. J. Biochem.* **268**, 5001–5010 (2001).
5. T. S. Prasad, K. Kandasamy, A. Pandey, Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.* **577**, 67–79 (2009).
6. P. V. Hornbeck, I. Chabra, J. M. Kornhauser, E. Skrzypek, B. Zhang, PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* **4**, 1551–1561 (2004).
7. N. Farriol-Mathis, J. S. Garavelli, B. Boeckmann, S. Duvaud, E. Gasteiger, A. Gateau, A. L. Veuthey, A. Bairoch, Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics* **4**, 1537–1550 (2004).
8. F. Gnäd, S. Ren, J. Cox, J. V. Olsen, B. Macek, M. Oroshi, M. Mann, PHOSIDA (phosphorylation site database): Management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* **8**, R250 (2007).
9. F. Diella, C. M. Gould, C. Chica, A. Via, T. J. Gibson, Phospho.ELM: A database of phosphorylation sites—update 2008. *Nucleic Acids Res.* **36**, D240–D244 (2008).

10. H. Li, X. Xing, G. Ding, Q. Li, C. Wang, L. Xie, R. Zeng, Y. Li, SysPTM: A systematic resource for proteomic research on post-translational modifications. *Mol. Cell. Proteomics* **8**, 1839–1849 (2009).
11. B. Bodenmiller, D. Campbell, B. Gerrits, H. Lam, M. Jovanovic, P. Picotti, R. Schlappbach, R. Aebersold, PhosphoPeP—a database of protein phosphorylation sites in model organisms. *Nat. Biotechnol.* **26**, 1339–1340 (2008).
12. P. Durek, R. Schmidt, J. L. Heazlewood, A. Jones, D. MacLean, A. Nagel, B. Kersten, W. X. Schulze, PhosPhAT: The *Arabidopsis thaliana* phosphorylation site database. An update. *Nucleic Acids Res.* **38**, D828–D834 (2010).
13. P. A. Grimsrud, D. den Os, C. D. Wenger, D. L. Swaney, D. Schwartz, M. R. Sussman, J. M. Ane, J. J. Coon, Large-scale phosphoproteomic analysis in *Medicago truncatula* roots provides insight into in vivo kinase activity in legumes. *Plant Physiol.* **152**, 19–28 (2010).
14. L. M. Schang, A. Bantly, M. Knockaert, F. Shaheen, L. Meijer, M. H. Malim, N. S. Gray, P. A. Schaffer, Pharmacological cyclin-dependent kinase inhibitors inhibit replication of wild-type and drug-resistant strains of herpes simplex virus and human immunodeficiency virus type 1 by targeting cellular, not viral, proteins. *J. Virol.* **76**, 7874–7882 (2002).
15. D. Schwartz, M. F. Chou, G. M. Church, Predicting protein post-translational modifications using meta-analysis of proteome scale data sets. *Mol. Cell. Proteomics* **8**, 365–379 (2009).
16. M. Lederman, J. T. Patton, E. R. Stout, R. C. Bates, Virally coded noncapsid protein associated with bovine parvovirus infection. *J. Virol.* **49**, 315–318 (1984).
17. Y. Shi, G. E. Dodson, S. Shaikh, K. Rundell, R. S. Tibbetts, Ataxia-telangiectasia-mutated (ATM) is a T-antigen kinase that controls SV40 viral replication in vivo. *J. Biol. Chem.* **280**, 40195–40200 (2005).
18. D. Schwartz, S. P. Gygi, An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol.* **23**, 1391–1398 (2005).
19. P. W. Yang, S. S. Chang, C. H. Tsai, Y. H. Chao, M. R. Chen, Effect of phosphorylation on the transactivation activity of Epstein-Barr virus BMRF1, a major target of the viral BGLF4 kinase. *J. Gen. Virol.* **89**, 884–895 (2008).
20. M. Baumann, H. Mischak, S. Dammeier, W. Kolch, O. Gires, D. Pich, R. Zeidler, H. J. Delecluse, W. Hammerschmidt, Activation of the Epstein-Barr virus transcription factor BZLF1 by 12-*O*-tetradecanoylphorbol-13-acetate-induced phosphorylation. *J. Virol.* **72**, 8105–8114 (1998).
21. A. S. El-Guindy, G. Miller, Phosphorylation of Epstein-Barr virus ZEBRA protein at its casein kinase 2 sites mediates its ability to repress activation of a viral lytic cycle late gene by Rta. *J. Virol.* **78**, 7634–7644 (2004).
22. A. S. El-Guindy, S. Y. Paek, J. Countryman, G. Miller, Identification of constitutive phosphorylation sites on the Epstein-Barr virus ZEBRA protein. *J. Biol. Chem.* **281**, 3085–3095 (2006).
23. R. Asai, A. Kato, Y. Kawaguchi, Epstein-Barr virus protein kinase BGLF4 interacts with viral transactivator BZLF1 and regulates its transactivation activity. *J. Gen. Virol.* **90**, 1575–1581 (2009).
24. K. Shire, P. Kapoor, K. Jiang, M. N. Hing, N. Sivachandran, T. Nguyen, L. Frappier, Regulation of the EBNA1 Epstein-Barr virus protein by serine phosphorylation and arginine methylation. *J. Virol.* **80**, 5261–5272 (2006).
25. S. J. Duellman, K. L. Thompson, J. J. Coon, R. R. Burgess, Phosphorylation sites of Epstein-Barr virus EBNA1 regulate its function. *J. Gen. Virol.* **90**, 2251–2259 (2009).
26. B. Kwiatkowski, S. Y. Chen, W. H. Schubach, CKII site in Epstein-Barr virus nuclear protein 2 controls binding to hSNF5/Ini1 and is important for growth transformation. *J. Virol.* **78**, 6067–6072 (2004).
27. W. Yue, J. Shackelford, J. S. Pagano, cdc2/cyclin B1-dependent phosphorylation of EBNA2 at Ser243 regulates its function in mitosis. *J. Virol.* **80**, 2045–2050 (2006).
28. A. Yokoyama, M. Tanaka, G. Matsuda, K. Kato, M. Kanamori, H. Kawasaki, H. Hirano, I. Kitabayashi, M. Ohki, K. Hirai, Y. Kawaguchi, Identification of major phosphorylation sites of Epstein-Barr virus nuclear antigen leader protein (EBNA-LP): Ability of EBNA-LP to induce latent membrane protein 1 cooperatively with EBNA-2 is regulated by phosphorylation. *J. Virol.* **75**, 5119–5128 (2001).
29. R. K. Moorthy, D. A. Thorley-Lawson, Biochemical, genetic, and functional analyses of the phosphorylation sites on the Epstein-Barr virus-encoded oncogenic latent membrane protein LMP-1. *J. Virol.* **67**, 2637–2645 (1993).
30. K. Y. Chien, Y. S. Chang, J. S. Yu, L. W. Fan, C. W. Lee, L. M. Chi, Identification of a new in vivo phosphorylation site in the cytoplasmic carboxyl terminus of EBV-LMP1 by tandem mass spectrometry. *Biochem. Biophys. Res. Commun.* **348**, 47–55 (2006).
31. C. G. Panousis, D. T. Rowe, Epstein-Barr virus latent membrane protein 2 associates with and is a substrate for mitogen-activated protein kinase. *J. Virol.* **71**, 4752–4760 (1997).
32. S. Fruehling, R. Longnecker, The immunoreceptor tyrosine-based activation motif of Epstein-Barr virus LMP2A is essential for blocking BCR-mediated signal transduction. *Virology* **235**, 241–251 (1997).
33. L. V. Matkova, C. Helmstetter, R. J. Ingham, G. Gish, C. K. Lindholm, I. Ernberg, T. Pawson, G. Winberg, The Shb signalling scaffold binds to and regulates constitutive signals from the Epstein-Barr virus LMP2A membrane protein. *Oncogene* **26**, 4908–4917 (2007).
34. Q. Xi, R. Cuesta, R. J. Schneider, Regulation of translation by ribosome shunting through phosphotyrosine-dependent coupling of adenovirus protein 100k to viral mRNAs. *J. Virol.* **79**, 5676–5683 (2005).
35. V. Cleghon, A. Piderit, D. E. Brough, D. F. Klessig, Phosphorylation of the adenovirus DNA-binding protein and epitope mapping of monoclonal antibodies against it. *Virology* **197**, 564–575 (1993).
36. M. L. Tremblay, C. J. McGlade, G. E. Gerber, P. E. Branton, Identification of the phosphorylation sites in early region 1A proteins of adenovirus type 5 by amino acid sequencing of peptide fragments. *J. Biol. Chem.* **263**, 6375–6383 (1988).
37. C. L. Smith, C. Deboucq, M. Rosenberg, J. S. Culp, Phosphorylation of serine residue 89 of human adenovirus E1A proteins is responsible for their characteristic electrophoretic mobility shifts, and its mutation affects biological function. *J. Virol.* **63**, 1569–1577 (1989).
38. D. J. Dumont, R. C. Marcellus, S. T. Bayley, P. E. Branton, Role of phosphorylation near the amino terminus of adenovirus type 5 early region 1A proteins. *J. Gen. Virol.* **74**, 583–595 (1993).
39. S. G. Whalen, R. C. Marcellus, D. Barbeau, P. E. Branton, Importance of the Ser-132 phosphorylation site in cell transformation and apoptosis induced by the adenovirus type 5 E1A protein. *J. Virol.* **70**, 5373–5383 (1996).
40. S. G. Whalen, R. C. Marcellus, A. Whalen, N. G. Ahn, R. P. Ricciardi, P. E. Branton, Phosphorylation within the transactivation domain of adenovirus E1A protein by mitogen-activated protein kinase regulates expression of early region 4. *J. Virol.* **71**, 3545–3553 (1997).
41. A. S. Tumell, R. J. Grand, C. Gorbea, X. Zhang, W. Wang, J. S. Mymryk, P. H. Gallimore, Regulation of the 26S proteasome by adenovirus E1A. *EMBO J.* **19**, 4759–4773 (2000).
42. C. J. McGlade, M. L. Tremblay, P. E. Branton, Mapping of a phosphorylation site in the 176R (19 kDa) early region 1B protein of human adenovirus type 5. *Virology* **168**, 119–127 (1989).
43. D. Takayasu, J. G. Teodoro, S. G. Whalen, P. E. Branton, Characterization of the 55K adenovirus type 5 E1B product and related proteins. *J. Gen. Virol.* **75**, 789–798 (1994).
44. D. L. Lichtenstein, P. Krajcsi, D. J. Esteban, A. E. Tollefson, W. S. Wold, Adenovirus RID $\beta$  subunit contains a tyrosine residue that is critical for RID-mediated receptor internalization and inhibition of Fas- and TRAIL-induced apoptosis. *J. Virol.* **76**, 11329–11342 (2002).
45. M. Ramachandra, R. Nakano, P. M. Mohan, A. B. Rawitch, R. Padmanabhan, Adenovirus DNA polymerase is a phosphoprotein. *J. Biol. Chem.* **268**, 442–448 (1993).
46. A. Kato, J. Arai, I. Shiratori, H. Akashi, H. Arase, Y. Kawaguchi, Herpes simplex virus 1 protein kinase Us3 phosphorylates viral envelope glycoprotein B and regulates its expression on the cell surface. *J. Virol.* **83**, 250–261 (2009).
47. D. J. Davido, W. F. von Zagorski, W. S. Lane, P. A. Schaffer, Phosphorylation site mutations affect herpes simplex virus type 1 ICP0 function. *J. Virol.* **79**, 1232–1243 (2005).
48. K. Xia, N. A. DeLuca, D. M. Knipe, Analysis of phosphorylation sites of herpes simplex virus type 1 ICP4. *J. Virol.* **70**, 1061–1071 (1996).
49. C. R. Brandt, A. W. Kolb, Tyrosine 116 of the herpes simplex virus type 1 IeA22 protein is an ocular virulence determinant and potential phosphorylation site. *Invest. Ophthalmol. Vis. Sci.* **44**, 4601–4607 (2003).
50. Y. Zhi, R. M. Sandri-Goldin, Analysis of the phosphorylation sites of herpes simplex virus type 1 regulatory protein ICP27. *J. Virol.* **73**, 3246–3257 (1999).
51. D. O'Reilly, O. Hanscombe, P. O'Hare, A single serine residue at position 375 of VP16 is critical for complex assembly with Oct-1 and HCF and is a target of phosphorylation by casein kinase II. *EMBO J.* **16**, 2420–2430 (1997).
52. S. Ottosen, F. J. Herrera, J. R. Doroghazi, A. Hull, S. Mittal, W. S. Lane, S. J. Triezenberg, Phosphorylation of the VP16 transcriptional activator protein during herpes simplex virus infection and mutational analysis of putative phosphorylation sites. *Virology* **345**, 468–481 (2006).
53. A. Kato, M. Yamamoto, T. Ohno, H. Kodaira, Y. Nishiyama, Y. Kawaguchi, Identification of proteins phosphorylated directly by the Us3 protein kinase encoded by herpes simplex virus 1. *J. Virol.* **79**, 9325–9331 (2005).
54. J. S. Loomis, J. B. Bowzard, R. J. Courtney, J. W. Wills, Intracellular trafficking of the UL11 tegument protein of herpes simplex virus type 1. *J. Virol.* **75**, 12209–12219 (2001).
55. A. Kato, M. Tanaka, M. Yamamoto, R. Asai, T. Sata, Y. Nishiyama, Y. Kawaguchi, Identification of a physiological phosphorylation site of the herpes simplex virus 1-encoded protein kinase Us3 which regulates its optimal catalytic activity in vitro and influences its function in infected cells. *J. Virol.* **82**, 6172–6189 (2008).
56. C. Cartier, P. Sivad, C. Tranchat, D. Decimo, C. Desgranges, V. Boyer, Identification of three major phosphorylation sites within HIV-1 capsid. Role of phosphorylation during the early steps of infection. *J. Biol. Chem.* **274**, 19434–19440 (1999).
57. G. Yu, F. S. Shen, S. Sturch, A. Aquino, R. I. Glazer, R. L. Felsted, Regulation of HIV-1 gag protein subcellular targeting by protein kinase C. *J. Biol. Chem.* **270**, 4792–4796 (1995).
58. P. Gallay, S. Swingler, C. Aiken, D. Trono, HIV-1 infection of nondividing cells: C-terminal tyrosine phosphorylation of the viral matrix protein is a key regulator. *Cell* **80**, 379–388 (1995).

59. B. Guy, M. P. Kiery, Y. Riviere, C. Le Peuch, K. Dott, M. Girard, L. Montagnier, J. P. Lecocq, HIV F/3' orf encodes a phosphorylated GTP-binding protein resembling an oncogene product. *Nature* **330**, 266–269 (1987).
60. P. L. Li, T. Wang, K. A. Buckley, A. L. Chenine, S. Popov, R. M. Ruprecht, Phosphorylation of HIV Nef by cAMP-dependent protein kinase. *Virology* **331**, 367–374 (2005).
61. M. H. Malim, S. Bohnlein, J. Hauber, B. R. Cullen, Functional dissection of the HIV-1 Rev *trans*-activator—derivation of a *trans*-dominant repressor of Rev function. *Cell* **58**, 205–214 (1989).
62. F. Meggio, D. M. D'Agostino, V. Ciminale, L. Chieco-Bianchi, L. A. Pinna, Phosphorylation of HIV-1 Rev protein: Implication of protein kinase CK2 and pro-directed kinases. *Biochem. Biophys. Res. Commun.* **226**, 547–554 (1996).
63. S. R. Brand, R. Kobayashi, M. B. Mathews, The Tat protein of human immunodeficiency virus type 1 is a substrate and inhibitor of the interferon-induced, virally activated protein kinase, PKR. *J. Biol. Chem.* **272**, 8388–8395 (1997).
64. X. Yang, J. Goncalves, D. Gabuzda, Phosphorylation of Vif and its role in HIV-1 replication. *J. Biol. Chem.* **271**, 10121–10129 (1996).
65. X. Yang, D. Gabuzda, Mitogen-activated protein kinase phosphorylates and regulates the HIV-1 Vif protein. *J. Biol. Chem.* **273**, 29879–29887 (1998).
66. Y. Zhou, L. Ratner, Phosphorylation of human immunodeficiency virus type 1 Vpr regulates cell cycle arrest. *J. Virol.* **74**, 6520–6527 (2000).
67. U. Schubert, T. Schneider, P. Henklein, K. Hoffmann, E. Berthold, H. Hauser, G. Pauli, T. Porstmann, Human-immunodeficiency-virus-type-1-encoded Vpu protein is phosphorylated by casein kinase II. *Eur. J. Biochem.* **204**, 875–883 (1992).
68. E. Estrabaud, E. Le Rouzic, S. Lopez-Vergès, M. Morel, N. Belaidouni, R. Benarous, C. Transy, C. Berlioz-Torrent, F. Margottin-Goguet, Regulated degradation of the HIV-1 Vpu protein through a  $\beta$ TrCP-independent pathway limits the release of viral particles. *PLoS Pathog.* **3**, e104 (2007).
69. M. A. Jarvis, T. R. Jones, D. D. Drummond, P. P. Smith, W. J. Britt, J. A. Nelson, C. J. Baldick, Phosphorylation of human cytomegalovirus glycoprotein B (gB) at the acidic cluster casein kinase 2 site (Ser<sub>900</sub>) is required for localization of gB to the *trans*-Golgi network and efficient virus replication. *J. Virol.* **78**, 285–293 (2004).
70. J. Reinhardt, G. B. Smith, C. T. Himmelheber, J. Azizkhan-Clifford, E. S. Mocarski, The carboxyl-terminal region of human cytomegalovirus IE1491aa contains an acidic domain that plays a regulatory role and a chromatin-tethering domain that is dispensable during viral replication. *J. Virol.* **79**, 225–233 (2005).
71. N. Y. Harel, J. C. Alwine, Phosphorylation of the human cytomegalovirus 86-kilodalton immediate-early protein IE2. *J. Virol.* **72**, 5481–5492 (1998).
72. H. Pande, T. D. Lee, M. A. Churchill, J. A. Zaia, Structural analysis of a 64-kDa major structural protein of human cytomegalovirus (Towne): Identification of a phosphorylation site and comparison to pp65 of HCMV (AD169). *Virology* **178**, 6–14 (1990).
73. M. C. Baek, P. M. Krosky, D. M. Coen, Relationship between autophosphorylation and phosphorylation of exogenous substrates by the human cytomegalovirus UL97 protein kinase. *J. Virol.* **76**, 11943–11952 (2002).
74. **Acknowledgments:** We thank M. Chou for his continued work in developing the *scan-x* algorithm and software and the Harvard Medical School Research Information Technology Group for hosting the virPTM Web site. **Funding:** This work was supported in part by NIH grant GM068763 and the U.S. Department of Energy Genome Sciences Program. **Author contributions:** D.S. curated the database, designed and performed analyses, and wrote the manuscript; G.M.C. provided conceptual advice and edited the manuscript. **Competing interests:** Technology transfer and commercial scientific advisory roles for G.M.C. can be found at: <http://arep.med.harvard.edu/gmc/tech.html>

Submitted 19 April 2010

Accepted 13 August 2010

Final Publication 31 August 2010

10.1126/scisignal.2001099

**Citation:** D. Schwartz, G. M. Church, Collection and motif-based prediction of phosphorylation sites in human viruses. *Sci. Signal.* **3**, rs2 (2010).