Exploring the genome and protein space of viruses

Congyu Lu[1], Yifan Wu[1], Longfei Mao[1], Xingyi Ge[1], Aiping Wu[2,3], Li Deng[4], Fengzhu Sun[5], Yongqiang Jiang[6*], Yousong Peng[1,*]

[1]Bioinformatics Center, College of Biology, Hunan Provincial Key Laboratory of Medical Virology, Hunan University, Changsha 410082, China

[2]Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[3]Suzhou Institute of Systems Medicine, Suzhou, China

[4]Internal Medicine-Neurology, The Third Hospital of Changsha, Changsha 410015, China

[5]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA

[6]State Key Laboratory of Pathogen and Biosecurity，Institute of Microbiology and Epidemiology，Academy of Military Medical Sciences，Academy of Military Sciences，Beijing 100071，China

* To whom correspondence should be addressed. Email: pys2013@hnu.edu.cn (YP) or jiangyq710327@sina.com (YJ)

Abstract

Background: Recent metagenomic studies have identified a huge number of viruses. Yet, the systematic assessment of the true genetic diversity of the whole virus community in our planet remains to be investigated. Here, we explored the genome and protein spaces of viruses by simulating the process of virus discovery in viral metagenomic studies.

Results: This study estimated that there would be at least 3.52e+06 viral Operational Taxonomic Units (vOTUs) and 2.16e+07 viral protein clusters on earth, while only about 30% of the viral genetic diversity has been identified so far. Considering the balance of costs and the detected number of novel viruses, an additional 3.89e+05 samples are required to capture 84% of the total genetic diversity. When analyzing the viral genetic diversity by taxonomy and ecosystem, the estimated viral genetic

diversity were consistent with those mentioned above.

Conclusions: This study for the first time explored the virus genome and protein space and estimated the total number of vOTUs and vPCs on earth when the virus genetic space was saturated. It provides a guide for future sequencing efforts of virus discovery and contributes to a better understanding of viral diversity in nature.

Main text

Viruses are the most abundant and diverse biological entities on earth[1]. According to the Baltimore classification system, viruses can be classified into seven groups, i.e., +ssRNA, -ssRNA, dsRNA, ssDNA, dsDNA, ssRNA-RT, and dsDNA-RT, based on the genetic materials and replication mode[2]. Viruses differ widely in terms of host range, virion morphology, genome type and size[3]. For example, the virus genome size can range from several hundred to millions of bases[4,5]. Viruses not only have a large impact on human health and diseases, but also play an important role in the balance of global ecosystems. Previous studies suggested that viruses kill 20% of bacterial hosts every day in the oceans[1].

An enormous amount of viruses have been identified as the rapid development of Next-Generation-Sequencing (NGS) technology compared to traditional methods of virus identification based on virus isolation[6-8]. For example, the IMG/VR database, which is the largest virus sequence database, has now stored more than 2 million viral genomic sequences[6]. Besides, novel viruses are still being discovered at an unprecedented rate. For example, the Global Ocean Viromes (GOV) 2.0 project has uncovered nearly 200,000 marine viral populations based on sequencing of 145 ocean samples[7]. A natural question arises as to what about the genetic space of viruses. Previous studies by Forest Rohwer estimated that there were 100 million phage species on earth[9]. The Global Virome Project estimated that there were 1.67 million viruses in birds and mammals[10]. Both studies estimated the virus genetic space based on their hosts, assuming that a host species harbors an average of dozens of viruses. They did not consider the actual process of virus discovery through sampling and sequencing.

This study explored the genome and protein spaces of viruses by simulating the process of virus discovery in viral metagenomic studies. Firstly, we investigated the increasing trends of the virus genome and protein space as the number of samples increased. Specifically, each time 50 sequenced samples were randomly selected from the IMG/VR database without replacement, and the accumulated number of viral Operational Taxonomic Units (vOTUs) and viral Protein Clusters (vPCs) (the definition of vOTU and vPC was described in Methods section) identified in accumulated samples selected until that time was counted. This process was repeated until no samples remained (see Methods). As shown in Figure 1A, the number of vOTUs and vPCs increased rapidly when the number of samples selected increased from 1 to 5000; then, the increasing trend slows down. With all the samples in the IMG/VR database, a total of 839,092 vOTUs and 7,617,621 vPCs were identified. Interestingly, the increasing trends of the numbers of vOTUs and vPCs versus the number of samples selected were well fitted by the Sigmoid function, with the R-squared greater than 0.99 for both the vOTUs and vPCs (Figure S1), which were much larger than those obtained using several common functions such as Exponential, Polynomial functions, and so on (Figure S1). Therefore, the increasing trends of vOTUs and vPCs were predicted using the Sigmoid function assuming that more sequenced samples were added in future studies (see the top-left sub-figure in Figure 1A). According to the predictions, the vOTUs and vPCs would continue to expand with a slowing rate.

The total numbers of vOTUs and vPCs were estimated to be 3.52e+06 and 2.16e+07 respectively, when the virus genetic space was saturated (Figure 1B) (see Methods). The Proportion of Explored Viral Genetic Space (PEVGS) that was defined as the total number of vOTUs or vPCs identified in the IMG/VR database divided by the estimated number of those when the virus genetic space was saturated, was 24% and 35%, for vOTU and vPC, respectively. As shown in Figure 1A, the number of new viruses identified in new samples decreases as sampling continues. To balance the cost of sampling and the amount of novel viruses found, the sampling endpoint was suggested as the time when less than one novel vOTU was identified in a sample. The

number of samples required in the sampling endpoint was estimated to be 403,844 which could capture 84% of the total vOTUs and vPCs (see the top-left sub-figure in Figure 1A). Unfortunately, only less than 20,000 samples had been sequenced so far according to the IMG/VR database.

We then investigated the taxonomy of the virus genome and protein space.

Analysis of the taxonomy composition of viruses showed that the unclassified viruses (defined as those without any taxonomy information) and the order of *Caudovirales* accounted for 55.2% and 41.9% of all viruses, respectively (Figure S2), and they were identified in 85.1% and 93.0% of samples, respectively (Table S1). In the remaining viral orders, the *Algavirales* accounted for 1.5% of all viruses and were identified in 11.6% of samples in the database (Table S1), while the other viral orders combined accounted for only 1.4% of all viruses. We then investigated the increasing trends of the numbers of vOTUs and vPCs in the unclassified viruses and in multiple viral orders as the number of samples increases. As shown in Figure 2A, the number of vOTUs in unclassified viruses and the order of *Caudovirales* increased most rapidly as the number of samples selected increased, with a much higher rate than that of other viral orders. There are striking differences between the increasing trends of vOTUs from different viral orders. Among the viral orders except the *Caudovirales*, the vOTUs of the *Algavirales* increased more rapidly than other viral orders as the samples increased. At the levels of vPC, the overall pattern of increasing trends was similar to that at the vOTU level (Figure 2B). Interestingly, the order of *Imitervirales* and “Other orders” (all other viral orders except those listed in Figure 2, see Table S1 for details) increased as fast as the *Algavirales* when the the numer of samples selected increased.

We then used the Sigmoid function to fit the increasing trends of vOTU and vPC in the unclassified viruses and in viral orders(Figure 2A & 2B), and estimated the total numbers of vOTUs and vPCs when the virus genetic space was saturated (Figure 2C). According to the estimates, the unclassified viruses and the *Caudovirales* would have comparable genome and protein diversity, with the former having a total of 2 million

119 vOTUs and 12.8 million vPCs, and the latter having a total of 1.1 million vOTUs and
120 7.6 million vPCs. Other viral orders were estimated to have much less genome and
121 protein diversity compared to the *Caudovirales*. For example, the order of
122 *Tubulavirales* was estimated to contain 54 thousand vOTUs and 82 thousand vPCs.
123
124 We then calculated the PEVGS of vOTUs and vPCs for the unclassified viruses and
125 different viral orders (Figure 2C). In most cases, the PEVGS increased from the level
126 of vOTU to vPC, ranging from 20% to 50%. For example, the *Caudovirales* had a
127 PEVGS of 32% and 48% at the level of vOTU and vPC, respectively. Interestingly,
128 the order of *Ligamenvirales* with far smaller genetic diversity than other viral orders,
129 had only a PEVGS of 8% at the vOTU level, while the PEVGSs at the vPC level was
130 80%.
131
132 **We then analyzed the virus genome and protein space by ecosystem.** According to
133 the GOLD (Genomes OnLine database) ecological classification system[11], the
134 samples used in the study were mainly obtained from seven ecosystems including two
135 environmental ecosystems (Aquatic and Terrestrial ecosystems accounting for 38%
136 and 18% of all samples, respectively), three host-associated ecosystems (Mammal,
137 Other-animal and Plant ecosystems accounting for 22%, 4% and 5% of all samples,
138 respectively), and two engineered ecosystems (Wastewater and Built-environment
139 ecosystems accounting for 2% and 6% of all samples, respectively) (Table S2). We
140 investigated the taxonomy composition of viruses in different ecosystems by
141 calculating the proportion of vOTUs of different orders in each ecosystem[12]. As
142 shown in Figure 3A, the taxonomy composition of viruses in different ecosystems
143 was similar, with the order of *Caudovirales* and the unclassified viruses accounting
144 for over 90% of all vOTUs. The order of *Caudovirales* had the largest proportion
145 among all viral orders in all ecosystems, ranging from 41% to 69%. The taxonomy
146 composition of the remaining vOTUs (less than 10%) varied greatly across
147 ecosystems. For example, the order of *Tubulavirales* predominated in the remaining
148 vOTUs of the Built-environment ecosystem, and few other viruses were detected in

149 this ecosystem. While in the Other-animal ecosystem, the order of *Petitvirales* and
150 *Ortervirales* predominated in the remaining vOTUs. Interestingly, the order of
151 *Algavirales* was enriched in the Aquatic ecosystem where their hosts (algae)
152 predominate[12].
153 We then analyzed the extent of shared viral genetic diversity among different
154 ecosystems. As shown in Figure 3B, the proportion of shared vPCs varied widely
155 among different ecosystems. For two environmental ecosystems, i.e., the Aquatic and
156 Terrestrial ecosystems, they shared 14% and 32% of the vPCs with other ecosystems,
157 respectively. The Aquatic ecosystem has the largest viral genetic diversity of any
158 ecosystems. For example, 3.88e+06 vPCs were identified in the Aquatic ecosystem,
159 accounting for 26%, 6%, 28%, 28%, 39%, and 35% of vPCs identified in the
160 ecosystems of Terrestrial, Mammal, Other-animal, Plant, Wastewater, and
161 Built-environment, respectively. For the host-associated ecosystems, the Plant and
162 Other-animal ecosystems shared 47% and 43% of the vPCs with other ecosystems,
163 respectively, while the Mammal ecosystem had only 9% of the vPCs shared with
164 others. For the Built-environment and Wastewater ecosystems, a large percentage
165 (60% and 46%, respectively) of vPCs were shared with other ecosystems. The
166 proportion of shared vOTUs in different ecosystems was also analyzed (Figure S3).
167 The shared proportion of viral genetic diversity increased from the level of vOTU
168 (range: 0.4%-6.8%) to vPC (range: 9%-60%). The median shared proportion in all
169 ecosystems were 2% and 43% at the level of vOTU and vPC, respectively.
170
171 In order to analyze the increasing trend of virus genetic diversity by ecosystem as the
172 number of samples increases, the random sampling was conducted as mentioned
173 above for each ecosystem using only samples of the corresponding ecosystem (see
174 Table S2 for the number of samples in different ecosystems). The increasing trends of
175 vOTU and vPC in each ecosystem were analyzed and fitted by the Sigmoid functions.
176 As shown in Figure 3C, the increasing trends of vOTU in all ecosystems were similar
177 which showed slowing rate as the number of samples selected increased. The number
178 of vOTUs of the Aquatic ecosystem increased most rapidly as the number of samples

179 selected increased, while that of other ecosystems increased much slowly. At the
180 levels of vPC, the overall increasing trends were similar to those observed at the
181 vOTU level (Figure 3D). However, the differences between the Aquatic and other
182 ecosystems, especially between the Aquatic and Terrestrial ecosystem, in increasing
183 trends of vPC were much smaller than those observed at the level of vOTU.
184
185 We then estimatedthe total numbers of vOTUs and vPCs in different ecosystems
186 when the viral genetic space was saturated. The Aquatic ecosystem was predicted to
187 have the largest number of vOTUs (2.2 million) and vPCs (10.9 million), followed by
188 Mammal (0.4 million vOTUs and 2.7 million vPCs) and Terrestrial ecosystems (0.3
189 million vOTUs and 5.8 million vPCs). The four remaining ecosystems were predicted
190 to contain much less viral diversity. For example, the Plant and Wastewater
191 ecosystems were predicted to contain only 76.7 thousand and 88.8 thousand vOTUs,
192 respectively.
193 The PEVGSs were calculated for each ecosystem at levels of vOTU and vPC (Figure
194 3E). For all ecosystems, the PEVGSs increased from the level of vOTU (12%-27%)
195 to vPC (21%-38%). The host-associated ecosystems had higher PEVGSs than
196 environmental and engineered ecosystems. For example, the Mammal ecosystem had
197 PEVGSs of 25% and 37% at levels of vOTU and vPC, respectively, while those were
198 12% and 21% for the Built-environment ecosystem.
199
200 **Discussion**
201 In the last few years, a number of large projects aimed at the whole virome in a
202 variety of specific ecological environments have been carried out and have discovered
203 huge numbers of novel viruses. However, there was still a lack of a comprehensive
204 understanding about the ultimate amount of virus genetic diversity on earth. This
205 study provided the first estimate of potential virus genome and protein space based on
206 the current virus genetic diversity, and found that there would be a total of 3.52e+06
207 vOTUs and 2.16e+07 vPCs on earth. Given that a vOTU has an average of 6 vPCs,
208 this estimate of virus genetic space was consistent at both the genome and protein

levels. Besides, the estimated virus genetic space was also consistent with the sum of viral genetic diversity estimated individually at each ecosystem, and that estimated individually by taxonomy. These results suggested the robustness of the estimates on viral genome and protein space.

Only about one-third of the viral genetic space has been revealed until now. Many more virome projects are needed to capture the viral genetic diversity. Unfortunately, the number of novel virus identified in new samples becomes lower as more viruses are identified. We estimated that an additional 3.89e+05 samples would be needed to capture 84% of the whole viral genetic diversity. According to our analysis, the samples from the Aquatic ecosystem captured much larger viral genetic diversity than that in other ecosystems, and should be put in priority in efforts of virus discovery. In addition, some viral orders had preferences for some ecosystems. For example, the *Algavirales* was mainly enriched in the Aquatic ecosystem. Sequencing studies based on such preference can help precise identification of some specific kinds of viruses.

Although the ICTV system and Baltimore system were used in viral taxonomy, more than half of viruses analyzed in the study were unclassified and lack taxonomy information. It is in great need to develop an expandable classification framework for all viruses. The ideal classification framework should have the ability to classify all viruses, easy to use, and scalable. Previous studies by Jang and colleagues have developed the vConTACT for classification of prokaryotic viruses based on genomic sequences[13]. In the study, a gene sharing network of viruses based on the shared protein clusters among viral genomes was built, and a distance-based clustering and metric were integrated to provide the measures of clustering confidence. vConTACT is a scalable and automated classification tool as it can build classifications for more than 10,000 sequences which were previously unclassified in oceanic viromics studies. Moreover, Jang's study provided a novel and important framework for organizing viral genetic diversity, and demonstrated that the gene content-based methods were suitable for a unified classification of all viruses. Unfortunately, it was not widely used in virome studies. There was still a lack of a feasible classification framework for the whole virome on earth.

239 There are several limitations to this study. Firstly, the estimated virus protein space
240 (number of vPCs) may be influenced by the parameter settings in protein clustering.
241 Fortunately, the estimated total number of vPCs was relatively stable when the
242 parameters varied, suggesting the robustness of our estimate on virus protein space
243 (Table S3). Secondly, there are diverse ecosystems on earth, while samples used in the
244 study were mainly derived from seven ecosystems. Besides, the samples had bias in
245 the ecosystem. The ecosystems that are closely related to humans were sampled more
246 frequently than other ecosystems, such as the Mammal ecosystem (Table S2), while
247 some ecosystems may be under-represented such as the Other-animal ecosystem and
248 the extreme environments. This may lead to the under-estimation of the viral genetic
249 diversity. Thirdly, the samples may have bias in region or country as developed
250 countries generally sequenced much more samples than the developing countries,
251 especially those in Africa and South America, although the detailed location of
252 samples were unknown in the study. Fourthly, the RNA viruses were severely
253 under-represented as most viruses collected in the IMG/VR database were identified
254 from the metagenomic studies that favor DNA viruses. Combined together, our
255 estimate of the virus genetic diversity should be considered as the lower bound of that
256 in nature.
257

258 **Conclusion**

259 In summary, this study for the first time explored the virus genome and protein space
260 and estimated the total number of vOTUs and vPCs on earth when the virus genetic
261 space was saturated. It provides a guide for future sequencing efforts of virus
262 discovery and contributes to a better understanding of viral diversity in nature.
263

264 **Methods**

265 **Data retrieval**

266 The viral genome and protein sequences were downloaded from the IMG/VR
267 database (v3) in November, 2021[6]. The meta information of metagenomic samples
268 from which viruses were identified was downloaded from the IMG/M database[14]. The

samples derived from experimental cultures were removed because they can not reflect the natural virus diversity. A total of 14,998 samples were kept for further analysis which included a total of 2,021,802 viral genomic sequences, 839,092 vOTUs (a vOTU was defined as a cluster of genomic sequences that had an average of 95% or higher sequence identity on more than 85% of genomic regions according to the IMG/VR database[6]), and 48,121,348 protein sequences.

Generation of protein clusters

To reduce the computational cost in generation of protein clusters, firstly, all viral protein sequences were clustered by the linclust of MMseqs2 with both the coverage and identity set to be 0.6[15]. A total of 11,862,236 elementary protein clusters were obtained. Then, the representative sequences of these elementary protein clusters were compared with each other by 'all vs all' blast using DIAMOND BLASTP[16]. A protein similarity network was built by connecting the representative protein sequences with reciprocal blast E-value smaller than 1e-05. The edge weights were calculated by the negative logarithms of the average of E-values in the reciprocal blast between two nodes. The protein similarity network was then clustered with HipMCL with the inflation value set to be 3[17]. A total of 7,617,621 clusters were obtained, with each cluster defining a viral Protein Cluster (vPC) containing one or more elementary protein clusters. The influence of the parameter selection on the generation of vPCs were shown in Table S4.

Simulation of the sampling in viral metagenomic studies

To simulate the process of virus discovery in viral metagenomic studies, each time 50 sequenced samples were randomly selected from all samples without replacement, and the number of vOTUs and vPCs identified in accumulated samples selected until that time was counted. This sampling process was repeated until no samples left. The increasing trend of the number of accumulated vOTUs and vPCs versus the number of accumulated samples selected were analyzed. The simulation process was repeated 100 times to reduce the randomness in sampling.

Predicting the size of vOTUs and vPCs when the genetic space was saturated

Several mathematical functions including Sigmoid, Power, Exponential, Second-order

Polynomial, Triple-order polynomial, Logarithmic, and Inverse proportional functions were used to fit the increasing trend of the number of accumulated vOTUs versus the common logarithm of the number of accumulated samples selected in the simulation process with the scipy.optimize.curve_fit package in Python (The formulas for these functions were listed in Supplementary materials)[18]. Three criteria were used to select the optimal function. The first is the R-squared of the function in fitting the increasing trend of vOTUs[19]. The higher, the better. Both the Sigmoid and Power functions had R-squared of greater than 0.99 which were much larger than the R-squared for other functions (Figure S1A). The second is that the derivative of the function which reflects the increasing rate of vOTUs, should decrease as the number of selected samples increased since the number of newly discovered vOTUs per sample decreased with the continued sampling. The third is that the function should converge since the total number of viruses on earth is not infinite. Only the Sigmoid function met these criteria (Figure S1). Therefore, the Sigmoid function was selected to fit the increasing trend of vOTUs and vPCs (Figure S1).

Since the Sigmoid function is a bounded function, the upper bound of the function was considered to be the ultimate size of viral genetic space when using the Sigmoid function in fitting the increasing trend of vOTUs and vPCs.

Key Points

- This study provided the first estimate of potential virus genome and protein space based on the current virus genetic diversity, and found that there would be a total of 3.52e+06 vOTUs and 2.16e+07 vPCs on earth.
- This study investigated the virus genome and protein space by ecosystem and taxonomy.
- This study provides a guide for future sequencing efforts of virus discovery and contributes to a better understanding of viral diversity in nature.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors gave their consent for publication.

Availability of data and materials

All data used in the study are derived from the IMG/VR and IMG/M databases.

Competing interests

The authors declare that they have no conflict of interest.

Funding

This work was supported by the National Natural Science Foundation of China (32170651), Hunan Provincial Natural Science Foundation of China (2020JJ3006), and Double-First Class Construction Funds of Hunan University (521119400156)

Authors' contributions

Y.P. conceived and designed the analysis. Y.P., C.L. and Y.W. performed the analysis and prepared all figures. Y.P., C.L., Y.J., F.S., L.D., A.W., X.G. and L.M. wrote the main manuscript text. All authors reviewed the manuscript.

Acknowledgements

We would like to thank members in Peng's lab for helpful suggestions.

Biographical note

Congyu Lu and Yifan Wu are doctoral students at Hunan University.

Longfei Mao, Xingyi Ge and Yousong Peng are Professors at Hunan University.

Aiping Wu is a Professor at Suzhou Institute of Systems Medicine.

Li Deng is a doctor at Third Hospital of Changsha.

Fengzhu Sun is a Professor at University of Southern California.

Yongqiang Jiang is a Professor at Academy of Military Medical Sciences.

360 **References**

361 1 Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nature reviews*
362 *microbiology* **5**, 801-812 (2007).

363 2 Lefkowitz, E. J. *et al.* Virus taxonomy: the database of the International Committee on
364 Taxonomy of Viruses (ICTV). *Nucleic acids research* **46**, D708-D717 (2018).

365 3 Masson, P. *et al.* ViralZone: recent updates to the virus knowledge resource. *Nucleic acids*
366 *research* **41**, D579-D583 (2012).

367 4 Philippe, N. *et al.* Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that
368 of parasitic eukaryotes. *Science* **341**, 281-286 (2013).

369 5 Collins, R., Gellatly, D., Sehgal, O. & Abouhaidar, M. Self-cleaving circular RNA associated with
370 rice yellow mottle virus is the smallest viroid-like RNA. *Virology* **241**, 269-275 (1998).

371 6 Roux, S. *et al.* IMG/VR v3: an integrated ecological and evolutionary framework for
372 interrogating genomes of uncultivated viruses. *Nucleic acids research* **49**, D764-D775 (2021).

373 7 Gregory, A. C. *et al.* Marine DNA viral macro-and microdiversity from pole to pole. *Cell* **177**,
374 1109-1123. e1114 (2019).

375 8 Lu, C. & Peng, Y. Computational Viromics: Applications of the Computational Biology in
376 Viromics Studies. *Virologica Sinica* **36**, 1256-1260, doi:10.1007/s12250-021-00395-7 (2021).

377 9 Rohwer, F. Global phage diversity. *Cell* **113**, 141 (2003).

378 10 Carroll, D. *et al.* The global virome project. *Science* **359**, 872-874 (2018).

379 11 Mukherjee, S. *et al.* Genomes OnLine Database (GOLD) v. 8: overview and updates. *Nucleic*
380 *acids research* **49**, D723-D733 (2021).

381 12 Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools.
382 *Database* **2020,** baaa062 (2020).

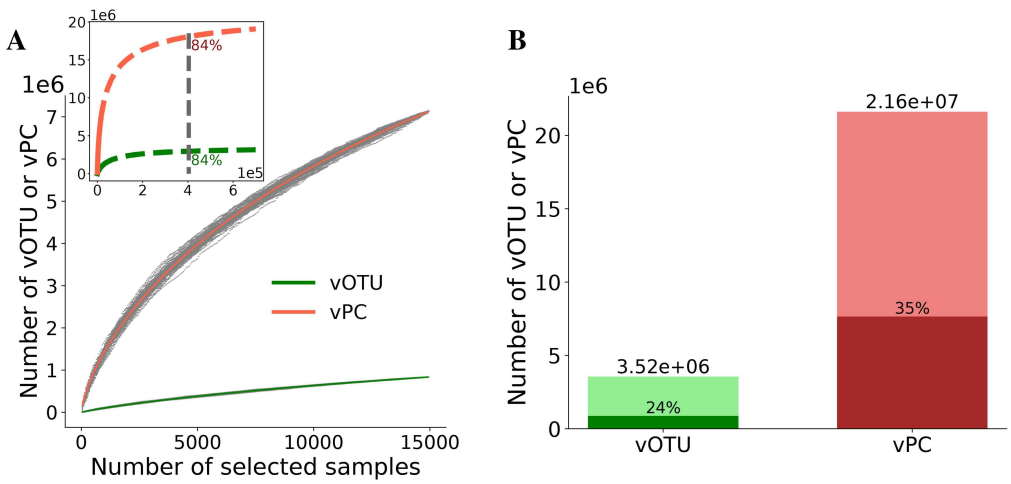
383 13 Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is
384 enabled by gene-sharing networks. *Nature biotechnology* **37**, 632-639 (2019).

385 14 Chen, I.-M. A. *et al.* The IMG/M data management and analysis system v.6.0: new tools and
386 advanced capabilities. *Nucleic Acids Research* **49**, D751-D763, doi:10.1093/nar/gkaa939
387 (2020).

388 15 Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the
389 analysis of massive data sets. *Nature biotechnology* **35**, 1026-1028 (2017).

390 16 Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using
391 DIAMOND. *Nature methods* **18**, 366-368 (2021).
392 17 Azad, A., Pavlopoulos, G. A., Ouzounis, C. A., Kyrpides, N. C. & Buluç, A. HipMCL: a
393 high-performance parallel implementation of the Markov clustering algorithm for large-scale
394 networks. *Nucleic acids research* **46**, e33-e33 (2018).
395 18 Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python.
396 *Nature methods* **17**, 261-272 (2020).
397 19 Lewis-Beck, C. & Lewis-Beck, M. *Applied regression: An introduction*. Vol. 22 (Sage
398 publications, 2015).
399
400
401 **Figure legend**

402 **Figure 1 Estimating the virus genome and protein space.** (A) Fitting the increasing
403 trend of vOTU and vPCs versus the number of selected samples. The gray points
404 referred to the number of vOTUs and vPCs identified in the selected samples in 100
405 times of simulating sequencing studies (see Methods). The solid line referred to the
406 fitted curve of the Sigmoid function for the vOTU (green) and vPC (red). The top-left
407 sub-figure referred to the prediction of the Sigmoid function. The dashed vertical line
408 referred to the suggested sampling endpoint. (B) The estimated total number of
409 vOTUs and vPCs when the virus genetic space was saturated, and the proportion of
410 explored viral genetic space at levels of vOTU and vPC.



411
412 **Figure 2** Analysis of the viral genome and protein space by taxonomy. (A) and (B)

referred to the increasing trends of vOTUs and vPCs, respectively, as the number of selected samples in different viral taxonomy. (C) The estimated total number (histogram) of vOTUs and vPCs in each taxonomy unit when the virus genetic space was saturated, and the PEVGS of each taxonomy at levels of vOTU and vPC (blue dashed lines).

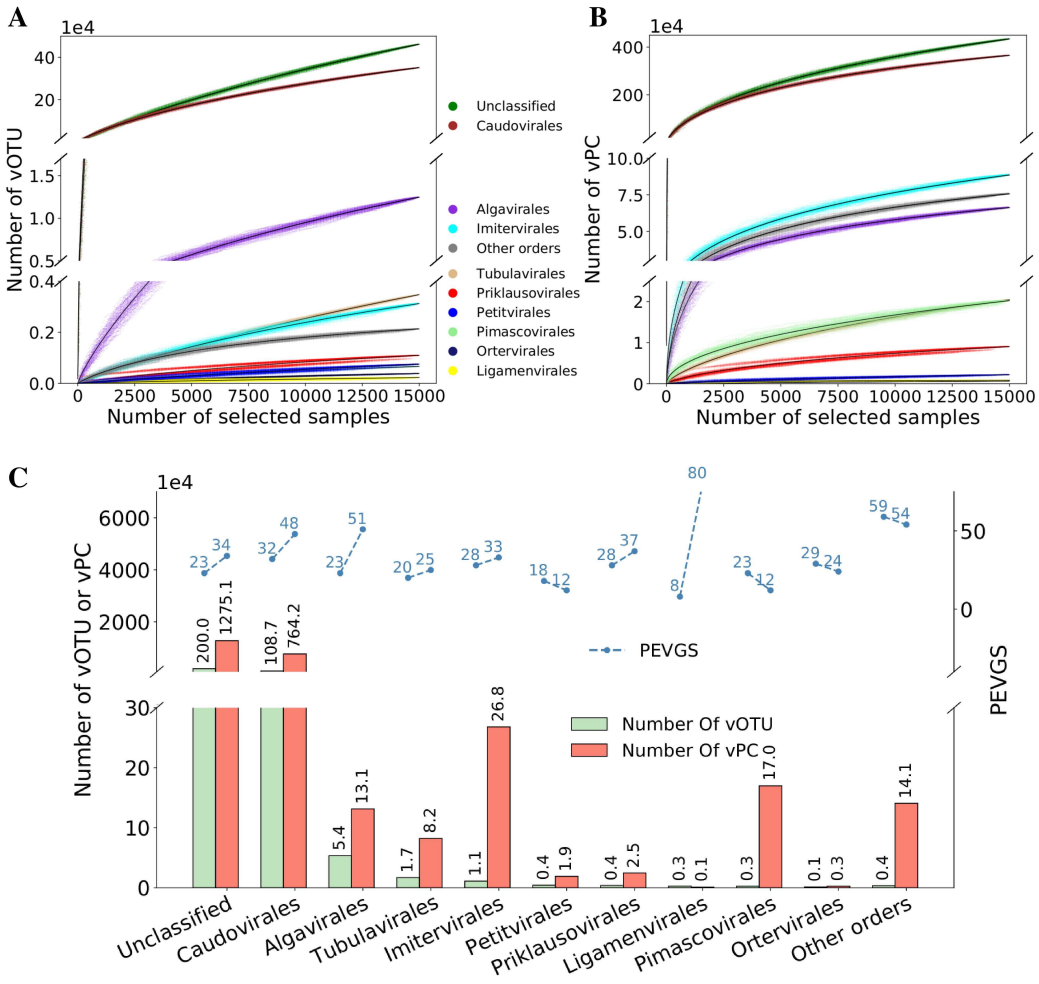
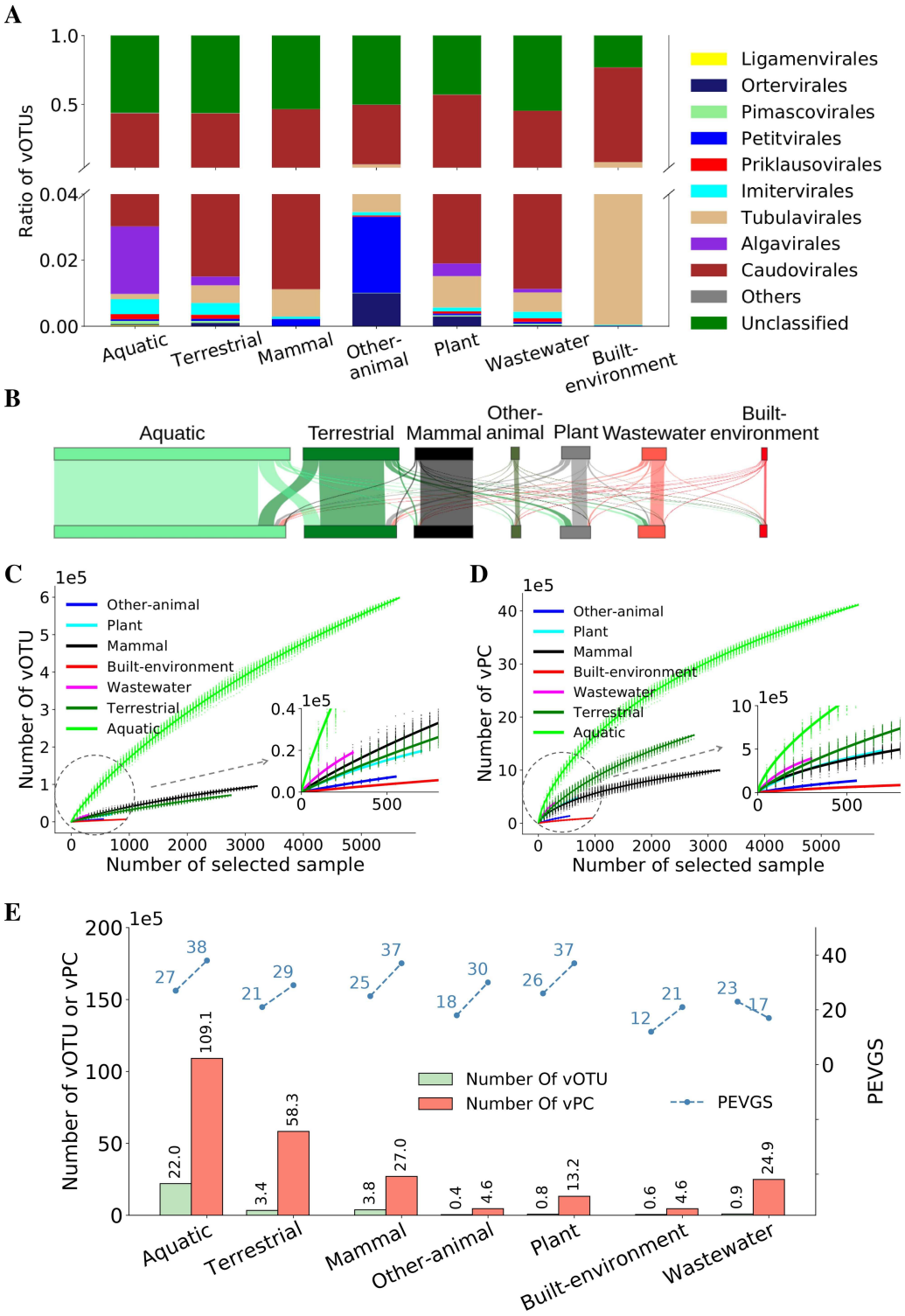


Figure 3 Analysis of the viral genome and protein space by ecosystem. (A) The taxonomy composition of viruses identified in different ecosystems; (B) The number of vPCs shared between different ecosystems. The thickness of lines between ecosystems is proportional to the number of vPCs shared between ecosystems. (C) and (D) referred to the increasing trends of the numbers of vOTUs and vPCs, respectively, as the number of selected samples in different ecosystems. The simulation process was conducted based on the ecosystem. (E) The estimated total

426 numbers (histogram) of vOTUs and vPCs in each ecosystem when the virus genetic
427 space was saturated, and the PEVGS of viruses in each ecosystem at levels of vOTU
428 and vPC (blue dashed lines).



429