Comparative genomics of nidoviruses: towards understanding the biology and evolution of the largest RNA viruses

Gulyaeva, A.

CHAPTER 5

LAMPA, LArge Multidomain Protein Annotator, and its application to RNA virus polyproteins

Anastasia A. Gulyaeva
Andrey I. Sigorskih[#]
Elena S. Ocheredko[#]
Dmitry V. Samborskiy
Alexander E. Gorbalenya

[#]equal contribution

Bioinformatics (2020)
DOI: 10.1093/bioinformatics/btaa065

ABSTRACT

Motivation: To facilitate accurate estimation of statistical significance of sequence similarity in profile-profile searches, queries should ideally correspond to protein domains. For multidomain proteins, using domains as queries depends on delineation of domain borders, which may be unknown. Thus, proteins are commonly used as queries that complicates establishing homology for similarities close to cut-off levels of statistical significance.

Results: In this report we describe an iterative approach, called LAMPA, LArge Multidomain Protein Annotator, that resolves the above conundrum by gradual expansion of hit coverage of multidomain proteins through re-evaluating statistical significance of hit similarity using ever smaller queries defined at each iteration. LAMPA employs TMHMM and HHsearch for recognition of transmembrane regions and homology, respectively. We used Pfam database for annotating 2985 multidomain proteins (polyproteins) composed of more than 1000 amino acid residues, which dominate proteomes of RNA viruses. Under strict cut-offs, LAMPA outperformed HHsearch-mediated runs using intact polyproteins as queries by three measures: number of and coverage by identified homologous regions, and number of hit Pfam profiles. Compared to HHsearch, LAMPA identified 507 extra homologous regions in 14.4% of polyproteins. This Pfam-based annotation of RNA virus polyproteins by LAMPA was also superior to RefSeq expert annotation by two measures, region number and annotated length, for 69.3% of RNA virus polyprotein entries. We rationalized the obtained results based on dependencies of HHsearch hit statistical significance for local alignment similarity score from lengths and diversities of query-target pairs in computational experiments.

Availability: LAMPA 1.0.0 R package is placed on GitHub (https://github.com/Gorbalenya-Lab/LAMPA).

1 INTRODUCTION

Due to high-throughput next-generation sequencing, genomics is outpacing functional and structural characterization of proteins [1]. This gap is especially pronounced and fast growing for viruses, whose discovery and characterization in diverse habitats has been driven by metagenomics over the last ten years [2, 3].

In genomics projects, conceptually translated open reading frames (ORFs) are functionally characterized by bioinformatics tools which use homology recognition for annotation. To improve accuracy of protein annotation, bioinformatics tools use iterative searches of databases of individual sequences (e.g. PSI-BLAST [4] vs GenBank [5]), search profile databases (e.g. HMMER [6] or HHsearch [7, 8] vs Pfam [9], or HHblits [8] vs Uniclust30 [10]), and may involve comparison of query and target secondary structure (e.g. HHsearch vs SCOP [11]). Annotation pipelines favor selectivity over sensitivity by imposing stringent cut-offs on similarity between query and database entries. Scores of similarity are interpreted in statistical frameworks using either expectation values (default cut-off E=0.001, BLAST, HMMER, HHsearch) or homology Probability (default cut-off P=95%, HHsearch).

To recognize distant homologs, popular HHsearch was fine-tuned based on a subset of SCOP 1.63 database with less than 20% pairwise sequence identity of structural domains [7], where mean sequence length is equal 178 aa [11] (Fig. 1), typical of functional and structural domain [12]. Its hit statistical significance increases with score of similarity between query and target, and it depends on sizes and diversities of query and target [13]. Specifically, large size increases likelihood of a hit score emerging by chance, while the opposite is true for small size. Notwithstanding HHsearch training on protein domains, it has been routinely used in analysis of proteins of unknown domain organization. For a single-domain protein, statistical significance of hit similarity must be applicable to its domain, since sizes of both are similar. On the other hand, for multidomain queries, statistical support of a hit associated with individual domain may be underestimated due to inflated search space that encompasses other domains of the query protein [4, 14].

The query size issue could be of little practical consequence for proteins having closely related homologs in sequence databases. However for identification of distant relationships, accurate estimation of statistical significance could be impactful. The above problem may be particularly acute for RNA viruses [15], which typically encode large multidomain proteins (>1000 aa) [16]. (Hereafter and for sake of simplicity, we'll use polyprotein to refer to virus multidomain proteins). They are much larger than most proteins of cellular organisms, whose length distributions resemble lognormal, with a mean below 500 aa [17]. Human immunodeficiency virus, Ebola virus, severe acute

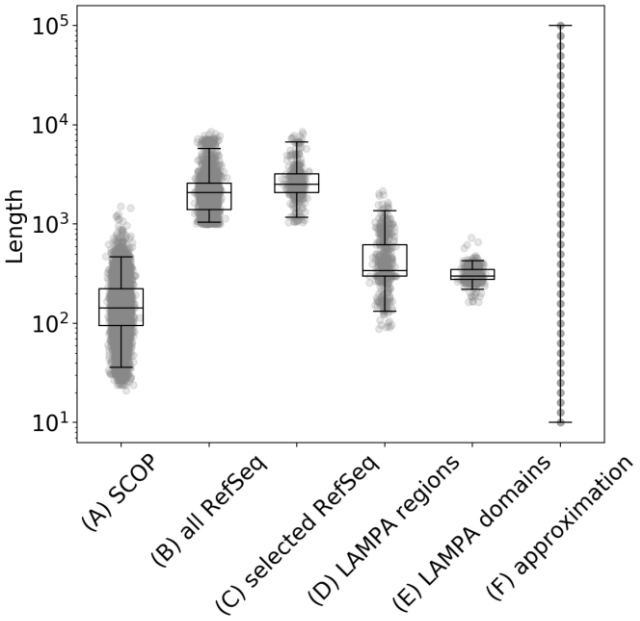

Figure 1 | Length distribution of proteins in datasets relevant to comparison of HHsearch and LAMPA. This plot depicts sizes of six protein datasets labelled from A to F and used or cited in this study. (**A**) 6271 SCOP domains used for HHsearch training (range: 21-1504 aa); (**B**) 2985 RefSeq virus polyproteins (range: 1001-8572 aa); (**C**) 431 RefSeq virus polyproteins which include 507 regions exclusively annotated by LAMPA (range: 1039-8572 aa); (**D**) 507 hit regions generated by LAMPA from 431 RefSeq polyproteins (range: 88-2172 aa); (**E**) 507 domains tentatively demarcated around LAMPA hits (range: 164-732 aa); (**F**) 41 designed sizes of each of three proteins, 123 in total, tested in computational experiments (range: 10 – 100,000 aa).

respiratory syndrome coronavirus, and poliovirus, and very many other eukaryotic viruses encode polyproteins [18, 19]. These polyproteins mediate replication/transcription and promote virus particle formation in either the synthesized form or after being proteolytically processed. Furthermore, the already known proteomes of RNA viruses are exceptionally diverse due to high mutation rate of RNA viruses [20], with many relationships in twilight and midnight zones of homology [21, 22].

In our recent HH-suit-mediated analysis of the largest known polyprotein of RNA virus (PSCNV, 13,556 aa) [23], we initially annotated only three regions by homology (polyprotein 7.1%). To check whether this result could be partially attributed to an underestimation of genuine statistical significance of the similarity between polyprotein domains and target protein profiles, we split the polyprotein using comparative genomics and, indeed, identified three other homologs with high confidence [23].

The above positive experience led us to formalize this approach in R package, called LAMPA, LArge Multidomain Protein Annotator, that we describe in this report. Also we

present proof-of-the principle for LAMPA in study of homology between RNA virus polyproteins and pfamA_31.0 database. It was further supported and expanded by evaluation of dependences of HHsearch statistics for fixed similarity score from lengths and diversities of query and target in computational experiments.

2 METHODS

2.1 Databases and virus protein dataset

We used pfamA_31.0 database [9], accompanying HH-suite [8], as *target* database to identify homology by profile searches and transfer annotation. We were interested in annotating virus proteins and selected a subset of NCBI Viral Genomes Resource database (RefSeq) [1] to serve as *queries* in homology searches and the source of expert annotation (Text S1.1). Only proteins of true RNA viruses that use RNA-dependent RNA polymerase (RdRp), positive and negative single-stranded RNA viruses, (+)ssRNA and (-)ssRNA, respectively, and double-stranded RNA viruses, dsRNA, were included in the query protein dataset (Fig. S1). Protein sequences were obtained from "translation" qualifiers of "CDS" features in RefSeq genome entries. The query database included all 2985 protein sequences of RNA virus genomes listed in "Viral genome browser" table on 2018.07.26 (Table S1), that were 1000 aa or longer (protein length ranged from 1001 to 8572 aa, median=2081 aa; Fig. 1). It was further grouped into 884 clusters using MMseqs2 [24], following the authors recommendations for multidomain proteins and defining sequence identity rate (--cluster-mode 1 --min-seq-id 0.3 --alignment-mode 3) and local alignment coverage (--cov-mode 0 -c 0.8) (see Text S1.2 and Table S1). Most of these proteins are encoded in a single ORF [25]. We parsed RefSeq entries corresponding to the analyzed proteins to extract region annotations from "Region" features [26]. Other annotation features, such as "CDS", "Protein", and "Site", which were not taken into analysis, may overlap with the "Region" or include extra information. For further details about polyprotein query dataset see Text S1.1.

2.2 Comparative sequence analysis

Transmembrane (TM) helices in protein sequences were predicted by TMHMM 2.0c [27]. Secondary structures (SS) of query sequences, regardless of their length, were derived from the predictions made for the respective entire polyproteins by script addss.pl from HH-suite 3.0.0 (2015.03.15) [28], which used PSIPRED 3.5 tool [29]. Query profiles were built and compared to a database by programs HHmake and HHsearch from HH-suite 2.0.16, respectively [7]. In all analyses, parameters of HH-suite programs were left at default values, with the exception of HHmake parameter "-M first", indicating that columns with residue in the first sequence of the FASTA file are considered match states,

and HHsearch three parameters: "-p 0", allowing hits with Probability as low as zero; "-norealign", blocking realignment of reported hits using maximum accuracy (MAC) algorithm; "-alt 10", enabling reporting up to 10 significant alternative alignments between a query and a target profile [14] (Text S1.3). To identify statistically significant hits and homologous regions, HHsearch hits were subjected to post-processing under three cut-offs: Probability >95%, E-value <10, and hit length of >50 aa of the query sequence. Hits satisfying these thresholds and overlapping on query were combined into a cluster, extreme N- and C-terminal residues of which defined boundaries of region in the query that was homologous to target(s). Statistics of the top-scoring hit in the cluster defined the entire cluster, and name of the top-scoring target profile in the cluster annotated the query region. Unless stated otherwise, all reported analyses used the hits post-processing. Also we used HHblits v.3 [8] for analysis of selected polyproteins as detailed in Text S1.4. Analysis and visualization were performed using R 3.3.0 [30].

2.3 Statistics

P value of Wilcoxon signed rank test (P_W) was calculated using function "wilcox.test" from R package "stats", with arguments "paired" and "alternative" set to values "TRUE" and "greater", respectively [30].

2.4 Calculation of HHsearch P-value and Probability dependence from lengths and diversities of query-target pair for fixed hit score

HHsearch uses extreme value distribution (EVD) model for estimating hit's P-value, E-value, and Probability from query-target local alignment similarity score. P-value for a given score is defined as:

$$P_{value}(score) = 1 - \exp(-\exp(-\lambda * (score - \mu))) \qquad (1)$$

where λ and μ are the EVD parameters that optimally approximate the score distribution of false positives for a given pair of query and target profiles. E-value is defined as $P_{value}(score)*N_{DB}$, where N_{DB} is the number of searched target profiles in the database. For calculations of λ and μ, HHsearch uses 'profile auto-calibration' that employs two simple artificial neural networks [13]. This default procedure makes use of dependence of λ and μ on four characteristics: profile lengths and sequence diversities of both query and target. The parameters of the neural networks were derived by training on a set of profiles based on 6271 sequences of SCOP20 v1.73 database (minimal, median and maximal protein lengths = 21, 142 and 1504 aa, respectively; 5-to-95% range = 48-to-392 aa) (Fig. 1). Estimation for Probability of detecting homologous relationship (true positives) is also based on the EVD distribution but involves correction by the SS alignment score.

To learn how HHsearch performs on queries of our study with sizes close to or exceeding the largest protein in the training SCOP database, we conducted computational experiments using the HHsearch procedure that generates EVD parameters by adapting corresponding C++ source code into a Python Jupyter notebook (https://github.com/Gorbalenya-Lab/hh-suite-notebooks/tree/LAMPA). We approximated P-value and Probability of hit for fixed local alignment similarity score (including also SS alignment score for Probability) in relation to lengths and/or diversities of the corresponding query and target profiles, one of which may have been set to vary in large range of values (see Text S1.5).

3 RESULTS

3.1 LAMPA, iterative approach for homology recognition and functional annotation of multidomain proteins

LAMPA approach is aimed at improving detection of remote homology in large multidomain proteins (queries). Its multistage iterative procedure includes prediction of TM regions in query by TMHMM at the pre-iteration stage #0 and comparisons of query and its regions with HH-suite profile database(s) (targets) using HHsearch for iterations at stages #1-#3 (Fig. 2). As query, intact protein is used for stages #0 and #1, and various protein regions are used for stages #2 and #3. Iteration is a single execution of a procedure involving protein regions demarcation and submission of regions to HHsearch-mediated homology searches to identify statistically significant hits (values of post-processing cut-offs, specified in 2.2, are default). The approach stages are detailed below:

Stage #0. *Detection of TM regions in original query.* TM region (domain) may include either single or few helices predicted by TMHMM. By default, more than one helix is included in a region if each helix is separated from its neighbor by less than 100 aa. Region boundaries are defined by either helix boundaries (single-helix region) or opposite boundaries of two respective terminal helices (multiple-helix region). TM regions are used to split original query into smaller regions (see stage #2).

Stage #1. *Detection of homology regions in original query.* This is the first iteration of the annotation procedure that uses HHsearch-mediated homology search. Its input and output are the original query and hit annotated regions, respectively.

Stage #2. *Detection of homology regions in split query: query-protein-specific (QP-specific) iterations.* To initiate this stage, the procedure selects regions of the original query that are flanked by either of the following: N- or C-terminus of the original query, TM regions

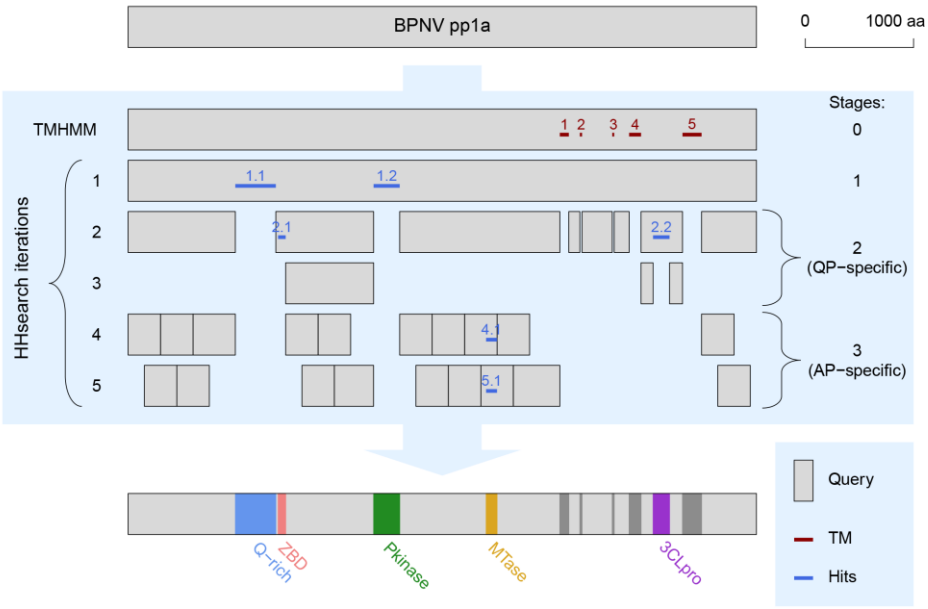

Figure 2 | LAMPA workflow and its application to RNA virus polyprotein. Presented is outline of the LAMPA approach (blue background) applied to polyprotein 1a (pp1a) of ball python nidovirus (BPNV). Grey bars, regions of BPNV pp1a that served as TMHMM or HHsearch queries. Iterations of the procedure and programs used are depicted on the left; stages are indicated on the right. Clusters of TM helices are depicted in dark red, clusters of hits – in dark blue. Hit double digits refer to iteration and hit position on polyprotein from left to right, respectively, except for hits at stage #0 which are labelled with the position only. Hits and annotations obtained on stage #1 represent output of conventional HHsearch. Q-rich, region rich in glutamine residue; ZBD, zinc-binding domain; Pkinase, protein kinase; MTase, methyltransferase; 3CLpro, 3C-like protease. For other details see text.

and hits clusters identified at the stages #0 and #1, respectively. These regions are used as input to HHsearch-mediated homology searches. Obtained hits are used for annotation and to demarcate flanking smaller non-annotated regions. The latter are used to initiate a new iteration in the manner described above. The iterations are repeated until no hits satisfying the cut-offs are identified.

Stage #3. *Detection of homology regions in split query: average-protein-size-specific (AP-specific) iterations.* Non-annotated regions after the stage #2 are split into two overlapping sets of 300 aa queries (default). The most C-terminal queries of both sets are extended to include the remaining part of the respective region, if the remaining part is shorter than 300/2=150 aa (default) and if the extended query does not cover the entire region. The default 300 aa size is close to that of an average protein (AP), hence respective iterations are called AP-specific. Queries are defined starting from either the N-terminus (first AP-specific iteration) or 300/2=150 aa (default) downstream the N-terminus (second

AP-specific iteration) of the non-annotated regions of stage #2. They are run independently. During this stage one and the same region of polyprotein may be found to have homolog and be annotated on both AP-specific iterations, since two sets overlap.

3.2 LAMPA implementation

The above approach was realized as LAMPA 1.0.0 R package (see also Text S1.6) that includes a single command 'LAMPA' with 15 arguments that allow user to specify a single protein query sequence, target database(s), information required to run HH-suit and TMHMM, and parameters of the LAMPA procedure, which are detailed in the package manual. LAMPA package employs two external R packages: seqinr [31] and IRanges [32]. Output of the command is a directory, name of which is identical to the name of the file with query sequence by default. This directory contains a plot (similar to Fig. 2) and two tables summarizing TM predictions and homology annotations made for the query sequence (overlapping with Table S2), as well as files with detailed information about hits constituting each cluster, and a folder with raw data (see package manual for details). Analysis of 2985 virus polyproteins against pfamA_31.0, detailed below, required 2000 min on 16 CPUs for LAMPA to complete (with 0.3 - 2.5 min per query, and approximately extra 1000 min compared to HHsearch). A separate script, not included in the LAMPA package, was used to automate analysis of multiple queries in this study.

3.3 Evaluation of LAMPA performance relative to HHsearch in analysis of RNA virus polyproteins

We evaluated LAMPA performance under default parameter values by querying pfamA_31.0 with 2985 RNA virus polyproteins (see 2.1; Fig. 1). This analysis documents dependence of HHsearch statistics on query size: split protein fragments or regions ('LAMPA') relative to intact proteins ('HHsearch'). Only the most N-terminal cluster of hits was considered in 26 cases of overlapping clusters from the LAMPA AP-specific stage. For annotation-related statistics, we did not consider TM domains (LAMPA stage #0, Fig. 2). The output of the LAMPA stage #1 represented also output of the HHsearch run on intact proteins.

Additionally, HHsearch was also used for further statistical analyses of the difference between outputs of two tools. For these analyses, HHsearch output was not subject to post-processing (see 2.2) that allowed to analyse hits with Probability ≤ 95%, E-value ≥ 10 and size on query ≤ 50 aa (see below). This use of HHsearch was outside the LAMPA framework and required matching of hits obtained by LAMPA and HHsearch for evaluation. We restricted this matching to the top-scoring hits of LAMPA hit clusters and HHsearch that overlapped on query and targeted the same Pfam profile.

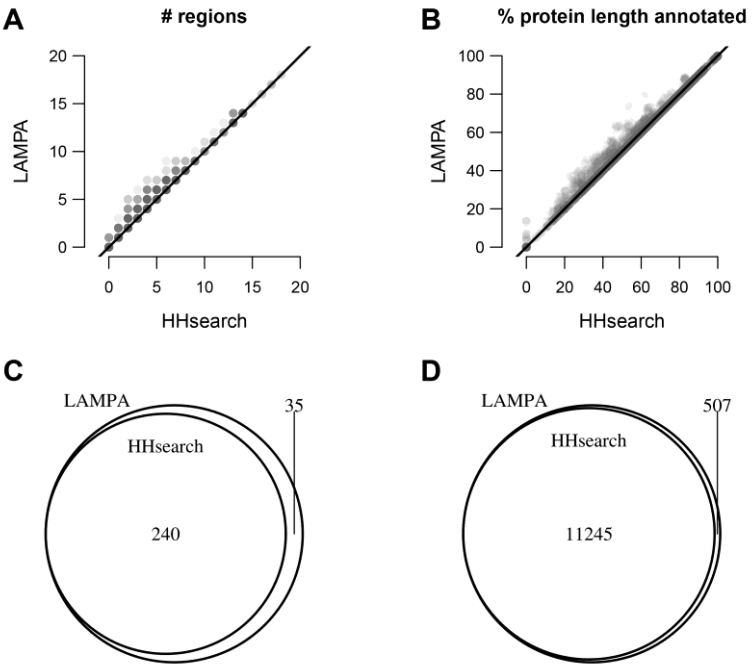

Figure 3 | Gain of homology recognition by LAMPA compared to HHsearch. Presented are four depictions of results of querying pfamA_31.0 with 2985 RNA virus proteins using LAMPA and HHsearch. (**A**) Number of regions (hit clusters) per query protein annotated by the two tools. Each protein is depicted by a transparent grey dot. Since multiple proteins may have the same or similar number of regions annotated by the two tools (X and Y dot coordinates), dots may overlap. Grey density is proportional to the number of overlapping dots. Black line, diagonal. (**B**) Share of protein length (%) annotated by the two tools. For other details see panel A. (**C**) Overlap between Pfam profiles that were linked to RNA virus proteins by the two tools. (**D**) Overlap between RNA virus polypro-tein regions annotated by the two tools.

3.4 LAMPA outperforms HHsearch in recognizing homology and facilitating annotation of RNA virus polyproteins

Neither LAMPA or HHsearch found homology between 163 proteins (5.5% of the dataset) and pfamA_31.0. For 2391 proteins (80.1%), LAMPA and HHsearch hit the same homologous regions, from 1 to 18. For 420 proteins (14.1%), LAMPA annotated from 1 to 3 extra regions on top of 1 to 15 found also by HHsearch (Fig. 3A). For each of the remaining 11 proteins (0.4%), a single region was hit by LAMPA only. Increase in number of annotated regions per protein by LAMPA was statistically significant (P_W=9.5e-86). By design of the procedure, HHsearch outperformed LAMPA for none of the polyproteins. For the three virus genome classes (2273 proteins in total), share of proteins, for which gain in number of annotated regions by LAMPA was observed, varied five-fold: (-)ssRNA viruses (3.1%), dsRNA viruses (10.2%), and (+)ssRNA viruses (15.9%). Among the 712 proteins with unknown virus genome class, LAMPA outperformed HHsearch for 22.2% of polyproteins.

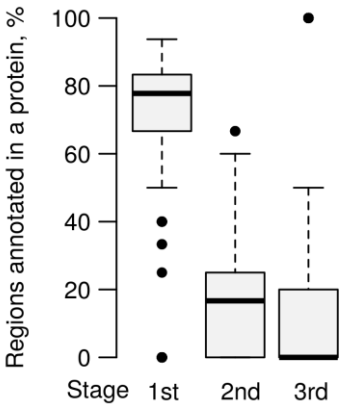

Figure 4 | Contribution of different stages of LAMPA procedure to protein annotation. Contribution of three LAMPA stages to annotation of 431 proteins, including regions exclusively annotated by LAMPA, was measured by percentage of regions annotated in each protein. Total number of regions annotated in each protein was considered 100%, regardless of their actual number and share in the protein. The box-plots, lower and upper limits of the box delimit the first (25%) and third (75%) quartiles, midline limit of the box – median, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box, data beyond that distance are represented by points.

Increase in the number of annotated regions (Fig. 3D) was accompanied by the increase in the polyprotein coverage by annotations, which ranged from 1.0% to 25.5% of polyprotein length (Fig. 3B; P_W=1.18e-72).

Also we compared lists of Pfam profiles hit by LAMPA and HHsearch, and were used for region annotation (Fig. 3C, Table S2). Both tools selected 173 profiles to annotate 5737 virus regions, and extra 67 profiles were used to annotate 5508 and 5947 virus regions by HHsearch and LAMPA, respectively. Also, additional 35 profiles were solely used by LAMPA to annotate 68 virus regions. Key enzymes of RNA viruses (RdRp, helicases, proteases, methyltransferases) dominated the shared part of the LAMPA and HHsearch Pfam profile lists (Fig. S2A). In contrast, the LAMPA-restricted profiles did not include RdRp but included types of enzymes and non-enzymatic proteins not found in the shared list, e.g. seven kinase profiles (Fig. S2B, Table S2). Many protein regions exclusively annotated by LAMPA were from most divergent RNA viruses [33].

3.5 Both QP- and AP-specific stages of LAMPA procedure contributed to gain of annotation

Gain of annotation by LAMPA compared to HHsearch is fully attributed to QP- and AP-specific stages. The gain was observed for 431 polyproteins, with the share of regions exclusively annotated by LAMPA varying from 6.2% to 100.0% (mean = 27.2%) of all recognised regions. Mean percentage of regions annotated in these proteins during the stages #1-#3 were 72.8%, 17.1% and 10.2%, respectively (Fig. 4). During QP- and AP-

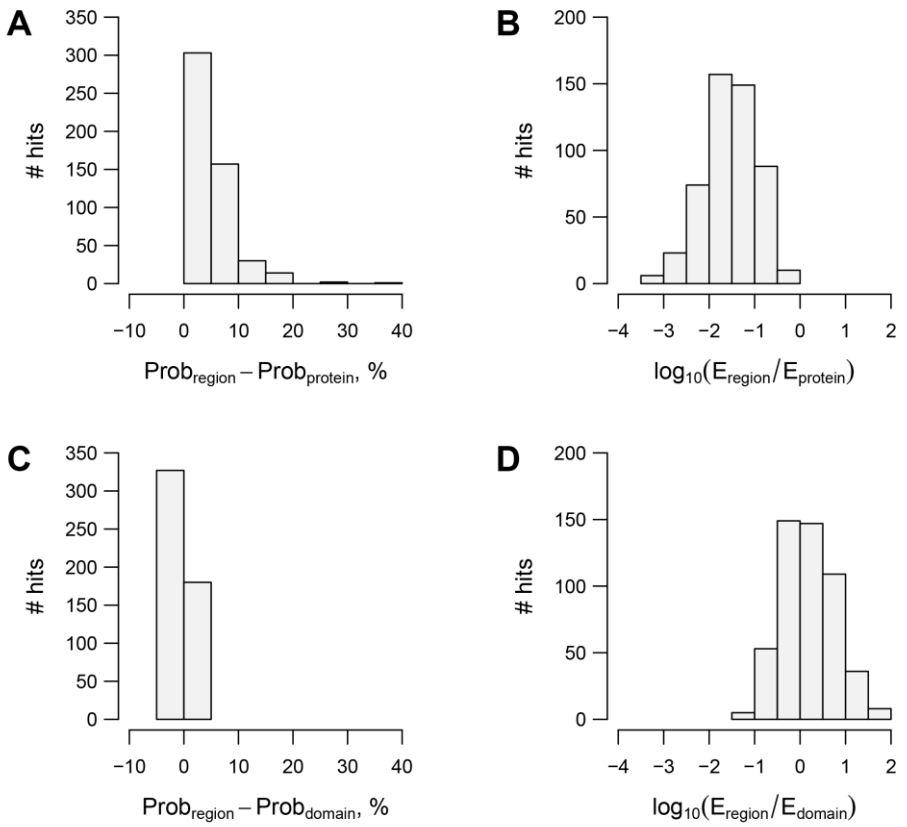

Figure 5 | Gain of hit statistical significance by LAMPA compared to HHsearch. LAMPA hits to region queries, obtained during the QP-specific and AP-specific stages of LAMPA procedure, are compared with matching HHsearch hits to polyprotein queries, in respect to hit Probability (**A**) and E-value (**B**); and with matching HHsearch hits to putative domain queries (operational definition, see text for details), in respect to hit Probability (**C**) and E-value (**D**). Analysed HHsearch hits were not subject to post-processing.

specific stages, regions were identified in 322 proteins (10.8% of the whole dataset) and 126 proteins (4.2%), respectively.

3.6 Increase of hit statistical significance by LAMPA com-pared to HHsearch is modest but common

LAMPA identified 507 clusters of hits on 431 proteins, HHsearch counterparts of which were removed by post-processing under the used thresholds (see 2.2; Fig. 3D). We used the top-scoring hits in these clusters to estimate the gain of statistical significance (Probability and E-value) by LAMPA compared to HHsearch and represent clusters in all analyses described below. We identified matching HHsearch hits for all 507 LAMPA hits (Table S2), with 437 hits (86.2%) having identical coordinates on query. In each pair of hits,

LAMPA hit was characterised by higher Probability and lower E-value (Fig. 5A and 5B). Probability increase by LAMPA compared to HHsearch was in the range from 0.5% to 37.6%, with mean 5.3% (Fig. 5A). Decimal logarithm of LAMPA to HHsearch E-values ratio ranged from -3.4 to -0.2 with mean -1.5 (Fig. 5B). Positive correlation between Probability and −logE-value was accompanied by E-value variation around two orders of magnitude for most Probabilities before and after they were elevated above the cut-off by LAMPA (Fig. S3). Likewise, for E-values around 10^{-1}, Probability varied approximately ±5%, illustrating that choice of statistic in addition to significance cut-off may affect output.

3.7 LAMPA-demarcated regions may approximate authentic domains for purpose of homology detection

The LAMPA region queries may still be (much) larger than the actual domains, natural borders of which remain unknown. Because of this uncertainty, we reasoned that the gain of statistical significance by LAMPA compared to HHsearch might provide only a lower estimate for the actual difference between Probabilities and E-values of the respective hits obtained for the polyprotein and expected for its domains. To improve understanding about how close the obtained LAMPA Probabilities and E-values for protein regions may be to those of the actual domains, we adopted an operational definition of polyprotein domain in relation to homology hit and used it to approximate borders of the actual domains; in total 507 hits on 431 polyproteins (see above) were considered for this purpose. Operational domain was demarcated as LAMPA hit that was extended by 100 aa to the N- and C-terminus; if distance to the polyprotein terminus was less than 100 aa, extension was adjusted accordingly (which was used in 48 of 507 cases). The demarcated domain sizes ranged from 164 to 732 aa (mean=315 aa) that was close to dominant domain size in public databases and narrower compared to the range of 88 to 2172 aa (mean=479 aa) of region queries that produced the original LAMPA hits (Fig. 1). For each of 507 hits, we then compared Probability and E-value values, assigned by LAMPA, to those obtained by HHsearch for a matching hit in a separate analysis that used demarcated domains as queries and involved no hits post-processing (see 2.2; Table S2).

We obtained data for all 507 hits, with 457 hits (90.1 %) having identical coordinates on query in LAMPA and HHsearch analyses. The difference between the two Probability values ranged from -1.8% to 4.6% with mean and median close to zero (both were equal -0.2%); absolute value of the difference didn't exceed 2% in 99.8% of cases (Fig. 5C). Decimal logarithm of the E-values ratio ranged from -1.3 to 1.8, mean 0.2 (Fig. 5D). These differences were evenly distributed and much smaller than those observed in comparison of LAMPA hits to region queries and HHsearch hits to polyprotein queries (Fig. 5A and 5B). Based on these results we concluded that sizes of queries used by LAMPA during iterative

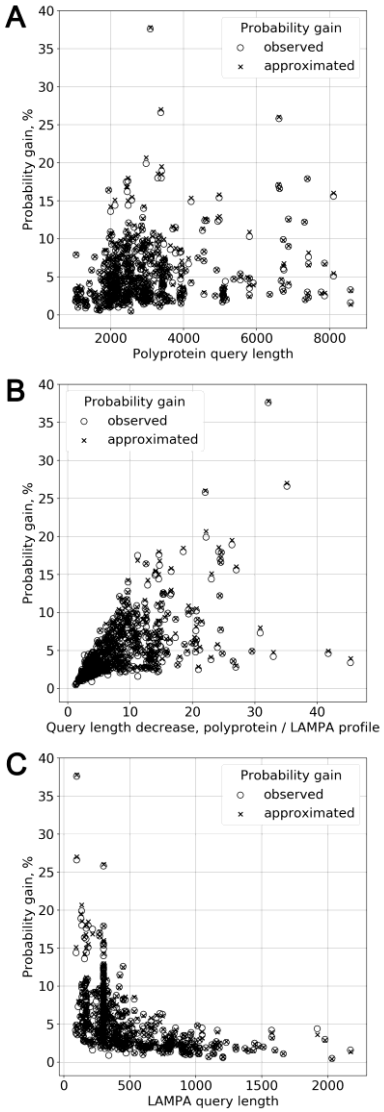

Figure 6 | Relationship between Probability gain by LAMPA and query lengths. Difference between Probabilities of hit to region query (LAMPA stages #2 or #3) vs polyprotein query (HHsearch without hits post-processing) (empty circle), is compared with difference between the respective approximated Probabilities for the matching hit in computational experiments (cross) at the Y axis, for 507 hits in total. These values are plotted against values of three characteristics of respective queries at the X axis: (**A**) polyprotein length (stage #1), (**B**) ratio of polyprotein to query region length (stage #1 vs stage #2/3), and (**C**) query region length (stage #2/3).

stages may be close to those of the respective authentic domains for the purpose of statistical evaluation of homology and annotation transfer under the employed cut-off.

3.8 Increase of statistical significance of hits by LAMPA compared to HHsearch is proportional to respective decrease of query length

We then asked how LAMPA-based increase of statistical significance in 507 hits of 431 proteins in 504 pairs of polyprotein and Pfam profile depended on lengths of polyprotein (original query, varied between 1039 and 8572 aa) and its fragments (queries varied between 88 to 2172 aa at LAMPA stages #2 and #3) (Fig. 1). We observed steady but highly uneven increase of Probability gain for polyproteins in the size range between 1001 and approximately 3000 aa which then levelled (Fig. 6A). That positive dependence was stronger and more common when Probability gain was plotted against relative length decrease in queries of LAMPA compared to HHsearch, which varied in the range from 1x to 45.3x, with 68.2% of the decreases of query length being in the 1-10x range (Fig. 6B). Accordingly, Probability gain fall steeply with increase of the LAMPA query length up to 2172 aa; it was below 10% and 5% for LAMPA queries including more than 448 aa and 747 aa, respectively (Fig. 6C).

3.9 Estimation of hits Probability by LAMPA may be approximated in computational experiment

Non-uniform dependence of Probability gain from query length (Fig. 6A, C) implied other characteristics be involved. Indeed, besides query length, target length and diversities of query and target are used by HHsearch for the calculation of λ and μ that affect hit score P-value (see 2.4). Accordingly, we analysed the relationship between estimates of hit statistical significance and possible lengths of the corresponding query and target profiles systematically using computational experiments. They used local alignment similarity score of HHsearch hit of *full-length* query-target pair for approximating hit Probability on queries of *other observed and computationally generated sizes*, assuming that hit score may not change with query size. This assumption proved to be accurate within a margin of error (see below).

We used the HHsearch neural networks to generate EVD parameters, followed by calculation of Probability, as well as P-value, of hit to polyprotein region from local alignment similarity score of this hit in every full-length query-target pair for which hit Probability gain was observed (in total 507 hits; Figs. 3D and 6; for details see https://github.com/Gorbalenya-Lab/hh-suite-notebooks/tree/LAMPA). First we noted good agreement between gains of Probabilities obtained in computational experiments and LAMPA runs (Fig. 6). They are within of +0.7%/-0.4% deviation of Probability gain estimation by LAMPA for the 95 percentile of hit scores in the dataset (Fig. S4A). The modest difference between the two values is explained by respective deviation of the underlying similarity score of the pairwise HHsearch hit alignment for polyprotein, which was fixed in computational experiments, from region-specific score that is calculated for

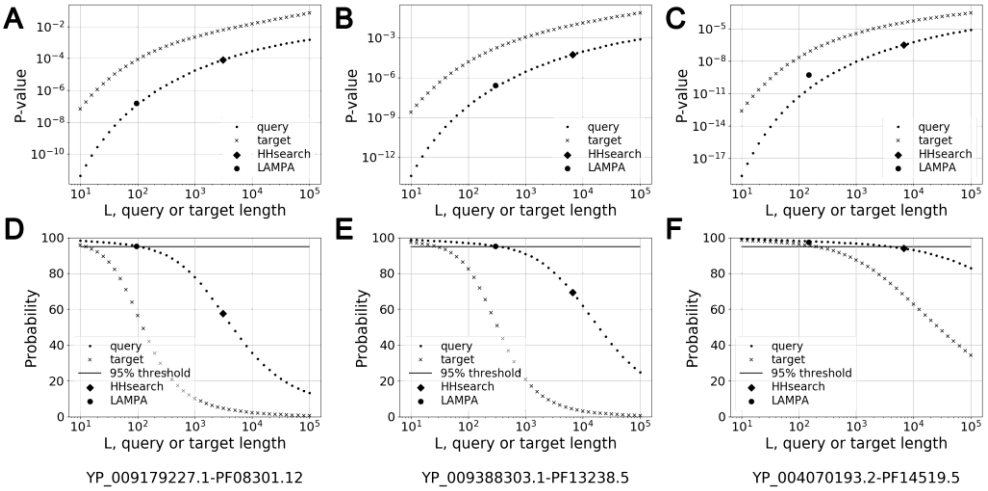

Figure 7 | Relationship between hit statistical significance and profile lengths in computational experiments. HHsearch hit P-value (**A-C**) and Probability (**D-F**) were estimated for 41 designed lengths of *query* or *target*, each of which was equidistant from its immediate neighbour on base 10 logarithmic scale (see Text S1). The 41 pairs of values were plotted to reveal relationship between two characteristics. These plots used hit score values of three query-target pairs, which are specified at the bottom of the figure and whose respective hit statistics values at the #1 stage (HHsearch), and #2 or #3 stages (LAMPA) are also depicted.

actual query and target profiles by LAMPA. Thus, by default, the same hit alignment involving polyprotein and its part as queries might have slightly different scores and also coordinates, further contributing to difference between the respective Probabilities (and P-values, Fig. S4B) in computational experiments.

3.10 P-value and Probability of HHsearch hits depend non-linearly on the lengths and diversities of query and tar-get profiles in computational experiments

The increase of the hit Probability during QP- and AP-specific iterations (Fig. 6) is likely explained by the use of query length in the auto-calibration procedure of HHsearch (see 2.4). We then conducted four computational experiments for three selected query-target pairs (Text S1.5) that were characterized by the largest Probability gain of LAMPA hit at stages #2 (37.6%) and #3 (25.8%), respectively, and associated with the largest decrease of query size (47 fold) (Fig. 7, Fig. S5 and Table S3). They also represent considerable ranges of hit scores (40.2, 41.1, and 67.2 for three pairs) and target diversities (6.7, 11.5, and 7.7). Forty one computationally designed lengths of each of three queries were tested (Fig. 1; Text S1.5).

In the three query-target pairs, both P-value and Probability showed strong non-linear dependence on designed sizes of query and target (Fig. 7) (hereafter we use “designed” to distinguish computational experiment from LAMPA). Specifically, P-value changed steeply,

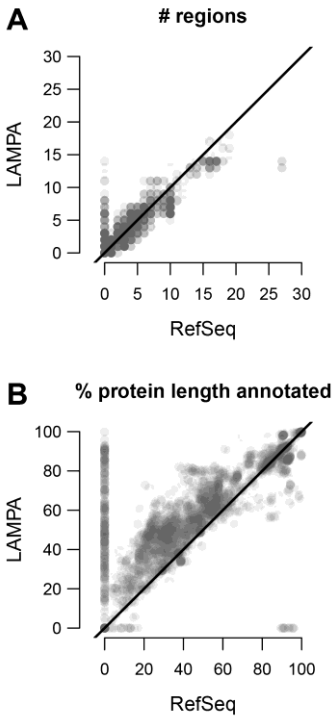

Figure 8 | Summary statistic of annotation coverage by LAMPA and RefSeq experts. Comparison of the number of regions per protein (**A**) or percentage of protein length (protein coverage) (**B**) annotated by LAMPA (stages #1-3) and RefSeq experts, based on analysis 2985 RNA virus proteins. Each protein is represented by a transparent grey dot; dot density is proportional to the number of proteins with identical characteristics. Black line, diagonal.

with curves of designed queries and targets running in parallel relative to each other (Fig. 7A-C). In the designed length range from 100 to 10000 aa, which encompasses most queries and targets of this study, P-value increased by approximately four orders of magnitude for queries of three pairs. This increase was limited to two orders of magnitude for the three selected queries illustrating LAMPA gain versus HHsearch. In contrast, dependence of Probability on length of designed queries and targets followed inverted logistic curve and differed between target and query as well as between the three pairs (Fig. 7D-F). Dependence of Probability on designed query size was most no-ticeable only below the 95% threshold, where it followed growth phase of logistic. The selected LAMPA and HHsearch queries were at different places of this growth phase in two query-target pairs (Fig. 7D,E) and outside the growth phase in third pair (Fig. 7F) which explained different Probability gains of LAMPA hit in these pairs. Hit score and target diversity contributed to variable Probability gain in three pairs (Text S1.5).

3.11 LAMPA can significantly expand RefSeq expert annotation of RNA virus polyproteins

Finally, we compared annotations of the RNA virus polyproteins by LAMPA and HHsearch versus RefSeq experts (Fig. 8, Fig. S6). Concerning the *number* of annotated regions per polyprotein, LAMPA and HHsearch were as good as RefSeq for 38.8 and 41.4% of polyproteins, respectively, while RefSeq expert or LAMPA/HHsearch outperformed the other for 23.3/27.0% and 37.9/31.6% of polyproteins, respectively (Fig. 8A, Fig. S6A). Notably, LAMPA and HHsearch annotated regions in 298 and 291 out of 426 polyproteins with no RefSeq annotation and increased the number of annotated region(s) for further 833 and 652 polyproteins. Increase in the number of annotated regions per protein by LAMPA but not HHsearch was statistically significant (P_W=3.11e-08 and 0.752, respectively). LAMPA and HHsearch annotations covered larger share of polyprotein (mean region length was 312, 321 and 265 aa for LAMPA, HHsearch and RefSeq annotation, respectively). This coverage increase was observed for 78.7 and 77.5% proteins, respectively, (Fig. 8B, Fig. S6B) and was statistically significant (P_W=1.07e-291 and 3.81e-273). We note that the above numbers apply to annotation in the “Region” fields of RefSeq entries. Other fields may record non-redundant annotation which is particularly likely for RefSeq entries with zero regions annotated in the “Region” field. These entries are in minority in the dataset. In summary, LAMPA expands further HHsearch annotation that may already improve RefSeq annotation of RNA virus polyproteins.

4 DISCUSSION

In this report we present an iterative LAMPA pipeline for advanced homology detection in large multidomain proteins and proof-of-the-principle for LAMPA in its application to RNA virus polyproteins. Statistical apparatus of HHsearch, used in LAMPA, was trained on a dataset of structurally defined domains with the median size of 142 aa to ascertain high sensitivity and selectivity, although HHsearch is used for annotation of proteins, regardless of their domain composition and size. This expanded application of HHsearch is due to two factors: 1) in contrast to sequence diversity of query (profile) (see HHblits), domain composition of query received relatively little attention in relation to HHsearch sensitivity; 2) considerable complexity and uncertainty of domain delineation in protein sequences. We have addressed both aspects in this study and offer a practical solution to the detection of distant homology in multidomain proteins using conventional profile-based tools in the LAMPA pipeline, which could be particularly useful in the on-going exploration of the Virosphere [2, 3, 23].

Length along with diversity are the two characteristics of query and target that determine hits Probability and P-value in HHsearch profiles' auto-calibration procedure [13]. We employed this procedure in computational experiments of high accuracy to plot the dependence of hits Probability and P-value from designed query/target lengths of several query-target pairs over a large size range that was beyond those used for tuning the auto-calibration procedure (12 to 1504 aa) and this study (1001 to 8572 aa) (Fig. 1). The produced plots revealed constrained statistic-specific shape of considerable variation for the two statistics characterizing a hit score in relation to query size (Fig. 7). Due to training of the auto-calibration procedure on the *domain* dataset, this variation informs about hit score statistics in application to *single-domain* proteins. When applied to *multidomain* proteins, like those used in this study, it illustrates how statistical significance of hit scores may be underappreciated depending on difference of sizes of the intact protein and its domains. This underappreciation is realized regardless of multidomain protein size, although it may be consistently considerable only for large proteins.

In line with the formula 1 (see 2.4), the computational experiments revealed also complex dependencies of statistical significance of HHsearch hits on designed target length and profile diversities of query and target (Fig. 7, Fig. S5). These dependencies explained variable gains of hit statistical significance by LAMPA compared to HHsearch in different query-target pairs. They also provide theoretical foundation for further efforts of improving the homology recognition by LAMPA through enriching queries using HHblits and targeting several databases, as is discussed below.

For queries including single domain or larger, false positive rate of LAMPA may not be different from that of HHsearch [7, 8], which is used for calculation of hit statistical significance. Our results were obtained with Probability cut-off of 95%, which was chosen to ascertain homology detection and suppress false positives [14]. The user may use E-value instead of Probability or lower the cut-off that will trade confidence in homology detection for increasing polyprotein coverage. We expect LAMPA to outperform HHsearch at these lower cut-offs as well. Due to logistic dependence between Probability and query length (Fig. 7D-F), Probability gains with under 95% cut-offs could be bigger than reported here.

We used TMHMM and HHsearch to functionally annotate polyproteins on structural grounds and by homology, respectively; they were used by LAMPA to delimit uncharacterized polyprotein regions that queried Pfam 31.0 further. (As discussed in Text S1.3, the use of HHsearch in the LAMPA framework was adjusted for analysis of RNA virus polyproteins). Once this iterative query-specific characterization at the QP-stage was exhausted, we used average protein domain size to delimit the remaining non-annotated regions during further database searches. This AP-stage has elements of arbitrariness

which were partially addressed *ad hoc* by using two alternative starting points for query delimitation.

This aspect and the entire pipeline may be advanced further. At the stage #0, other programs in addition to TMHMM may assist with functional annotation, e.g. mapping disordered regions, or regions anomalously enriched with certain amino acid residues, or cleavage sites for particular proteases like it was demonstrated in our recent study [23]. In that study, HHsearch was used to scan several databases, and this provision is also available in the LAMPA 1.0.0 package. Also, iterative profile programs, e.g. PSI-BLAST or HHblits, could be incorporated in the LAMPA to enrich query and improve homology recognition by targeting proteins that are not part of curated profile databases. These improvements could increase relative share of the QP-stage in homology detection and region annotation. In theory, the LAMPA may identify all domains at the #1 and QP-stage, with the AP-stage generating no hits, either due to the lack of queries or homology. Notwithstanding future advances, the current LAMPA version may already complement HHblits, the current top homology search tool. Indeed, under the 95% Probability cut-off HHblits failed to annotate 195 of 507 regions that LAMPA but not HHsearch annotated in 431 polyproteins of this study (Table S2, Text S1.4).

The reported gain of hit statistical significance by LAMPA compared to HHsearch was modest but sufficient to elevate many hits above the Probability 95% cut-off. It improved homology detection and hit coverage in 14.4% of polyproteins which were enriched with sequences that share not more than 30% identity with others in the dataset. Thus, gain of hit statistical significance by LAMPA compared to HHsearch could be larger for viruses that prototype genera or higher rank taxa rather than species dominating our dataset (see Text S1.2).

LAMPA annotation was most frequent for (+)ssRNA viruses, which correlates with their abundance and expanded diversity relative to dsRNA and (-)ssRNA viruses. Most newly detected homologs may already be known in other related viruses, which is evident from names and descriptions of hit Pfam profiles that often refer to viruses and their proteins (Table S2). However, they also include those not reported in literature, e.g. ZBD and MTase domains in pp1a (YP_009052476.1) of BPNV, python tobanivirus (Fig. 2; Table S2). The detection of the MTase domain, which is apparently conserved in the distantly related fish WBV (YP_803214.1) in this genome location, is particularly intriguing. These viruses and other nidoviruses with genomes > 20 kb are known to encode one or two MTases far downstream in the pp1b part of the pp1ab polyprotein [23, 34, 35] that were implicated in the 5'-end mRNA cap formation [36]. These and other functional assignments (Table S2) could be used to direct experimental research and in reconstruction of evolution of RNA viruses.

LAMPA facilitates homology detection and may be used to improve annotation coverage by other tools and experts in genomic projects, as well as in curated databases, including RefSeq. However, other factors besides detection of homology may affect quality of annotation [37, 38] and they were outside the scope of this study.

ACKNOWLEDGEMENTS

We thank Andrey M. Leontovich and Igor A. Sidorov for discussions and assistance.

FUNDING

This work has been supported by the EU Horizon2020 EVAg 653316 project and the LUMC MoBiLe program; AEG was a Leiden University Fund (LUF) Professor.

Conflict of Interest: none declared.

SUPPLEMENTARY INFORMATION

Text S1.1 Virus protein dataset

The RefSeq database was chosen to compile the query virus database for three reasons. First, it is one of the best representations of the known RNA virus genome diversity that is publicly available. Second, RefSeq maintains proper taxonomic representation of viruses that alleviates considerable biases of genome sequencing toward selected viruses of societal significance. Third, RefSeq curates annotation of genome records, which could be used as a standard to compare to [1].

Most viruses are represented by a single polyprotein in our query dataset, but large RNA viruses may encode several, either overlapping or not. Non-overlapping polyproteins are encoded in separate ORFs on single or multiple genome segments (see Table S1). In contrast, polyproteins of some viruses, notably those of nidoviruses and alphaviruses, are expressed from two ORFs using either ribosomal frame-shifting signal or read-through terminal codon [25]. Often, a RefSeq genome entry contains a “CDS” feature attributed to the combination of the two such ORFs, alongside a “CDS” feature attributed to the first ORF. A “CDS” feature attributed to the second ORF may also be included, even though it may not be expressed independently of the first ORF. These extra “CDS” features constitute a source of redundancy, as our query dataset was created by extracting protein

sequences ≥1000 aa from “translation” qualifiers of all “CDS” features of the selected RefSeq genome entries.

Proteins of (+)ssRNA viruses accounted for 47.1% of the query dataset, length of the proteins ranged from 1001 to 8572 aa (polyprotein of a flavi-like Gamboa mosquito virus [39]), median length was 2168 aa. Proteins of (-)ssRNA viruses accounted for 18.2% of the dataset, length of the proteins ranged from 1003 to 4403 aa (L protein of Shayang Spider Virus 1 from the order *Bunyavirales* [40]), median length was 2122 aa. Proteins of dsRNA viruses accounted for 10.9% of the dataset, length of the proteins ranged from 1002 to 7391 aa. Two dsRNA viruses with largest protein sizes, 6359 and 7391 aa, and possibly others with similar large sizes may in fact be (+)ssRNA viruses (polyproteins of Gentian Kobu-sho-associated virus [41, 42] and Ceratobasidium endornavirus D [43, 44]). Median length of the dsRNA virus proteins, included in the dataset, was 1274 aa. For the remaining 23.9% proteins of the dataset, genome type was not specified in the corresponding genome entries, while their lengths ranged from 1001 to 7421 aa, median=1963 aa.

We used RefSeq annotation of the virus sequences as a standard in our study. Although it is useful, the RefSeq remains a project in progress, and its annotation is subject to frequent update and revision. Much of its annotation is based on profile analysis involving Pfam, CDD or other databases. In this respect, our findings using strict significance cut-offs are equally reliable and can be considered true to the extent we could transfer Pfam profiles descriptions to the identified homologous regions of query proteins.

Text S1.2 Redundancy of the virus protein dataset in relation to comparison of LAMPA and HHsearch

Majority of the 2985 polyproteins of the query dataset are encoded by viruses that prototype virus species, which is a main criterion for their selection by RefSeq team to address redundancy problem and ensure their relevance for research and applications. However, known species are distributed highly unevenly among virus families that creates a bias. To evaluate how similar polyproteins of these species are in the protein distance space, we have clustered 2985 sequences using MMseqs2 software (0.8 coverage and 30% identity; single-linkage clustering mode) in analysis that delineated 884 clusters (with number of sequences per cluster varying from 1 to 124; average and median number of sequences per cluster 3.4 and 11, respectively) (Table S1). Inspection of virus taxonomy of these clusters indicates that they correspond loosely to taxa or a subset of taxa of classified viruses at genus/subfamily rank, depending on virus family. We found that 431 polyproteins, for which LAMPA outperformed HHsearch (Fig. 1, C dataset), represent a disproportionally large share of the total number of clusters (14.4% sequences found in 26.1% clusters) and were enriched with polyproteins representing less populated clusters

(231 clusters, average/median: 1.9/4.0 sequences per cluster). Thus, LAMPA outperformed HHsearch for annotation of a larger share of sequences in the clustered dataset than in the original dataset. This observation implies that the main observations and conclusions of our study were not undermined by selection of the RefSeq virus polyproteins as queries, without prior clustering.

Text S1.3 The use of HHsearch in the LAMPA framework for analysis of RNA virus polyproteins

The application of HHsearch to analysis of virus polyproteins in the LAMPA framework required non-default values for two parameters. The first parameter, "-norelaign", was used to switch off maximum accuracy (MAC) realignment algorithm, the postprocessing step at which the hit alignment is improved and the hit's span can be also adjusted, while hit scores (E-value/Probability) remain intact [14]. Although this postprocessing may improve alignment, we observed hit degradation and even its complete loss due to MAC use. A solution to this problem was suggested (https://github.com/soedinglab/hh-suite/issues/153). Second parameter, "-alt 10", increased the maximal number of reported alternative alignments between query and the same target profile to ten. The default maximum of two alternative alignments was found to be problematic, as RNA virus polyproteins may include more than two paralogs.

Also HHsearch may be prone to the overestimation of statistical significance of hits (false positives), if query size is at the low extreme of the size range of the training dataset. In the LAMPA framework, short queries smaller than domain may indeed be used at stage #2, if the query is flanked, from one or both sides, by hits that cover only a portion of the respective domain. These considerations prompted a limit on hit length (>50 aa by default) that also defined minimal length of query at stage #2.

Text S1.4 The use of HHblits to evaluate LAMPA gain of RNA virus polyproteins annotation

We used 431 polyproteins, which include 507 regions annotated by LAMPA but not HHsearch, as queries for HHblits to see whether this tool could annotate these regions. The polyprotein queries were initially enriched with homologs by running HHblits v.3 [8] against Uniclust30_2018_08 database [10] with 1, 2, or 3 search iterations and default other options (i.e. 0.001 for E-value cutoff for alignment extension and max. diversity threshold Neff=20, which stops further iterations). The enriched queries were then used for HHblits search in PfamA database with default options and only one search iteration (subsequent search iterations showed no significant improvement of hit score and coverage). Finally the obtained HHblits hits on PfamA profiles were mapped on corresponding LAMPA hits to 507 regions (Table S2). HHblits hit was considered as

matching, if it had Probability value above 95% and covered more than 70% of query region of respective LAMPA hit alignment. We observed that 195 of 507 regions were either not reported by HHblits at all (37) or were attributed with Probability value under the 95% cut-off (155) or had low query coverage (3).

Text S1.5 Dependence of P-value and Probability of fixed HHsearch hit score from size and diversity in query-target pairs of LAMPA analysis

We conducted several computational experiments using HHsearch neural networks. First, we assessed dependence of the Probability gain on query length, using different measures, in 507 query-target pairs from the hit list of LAMPA analysis of RNA virus polyproteins (Table S2). The obtained results were compared with those obtained in the LAMPA analysis and presented on Fig.6. Then, we selected three query-target pairs from the above list (Table S3) and conducted four in-depth computational experiments (for details see https://github.com/Gorbalenya-Lab/hh-suite-notebooks/tree/LAMPA). In first three experiments, diversities of query and target profiles were fixed at their respective real values (hereafter, the 'real' refers to characteristics of the full-length query or target profile). In the first experiment, we estimated P-value and Probability for computationally generated 41 different lengths of *query*, each of which was equidistant from its immediate neighbour on base 10 logarithmic scale in the query length space that ranged from 10^1 to 10^5 aa, with the target length fixed at its real value. In complementary second experiment, we estimated values of two statistics for the 41 length variants of the *target*, as specified above, and with the query length fixed at its real value. Results of these two experiments for three selected query-target pairs (Table S3) were combined separately for P-value and Probability, respectively (Fig. 7). In the third experiment, we estimated Probability for all combinations of the 41 length variants of the *query* and *target*. Results of this experiment were visualised using contour plots that depict change of Probability in the query length vs target length space (Fig. S5A-C). In the fourth experiment, lengths of query and target profiles were fixed at their respective real values. Then, we estimated Probability for all combinations of computationally generated 43 diversities of *query* and *target*, each of which was equidistant from its immediate neighbour on linear scale in the diversity space that ranged from 1 to 15. Results of this experiment were visualised using contour plots that depict change of Probability in the query diversity vs target diversity space (Fig. S5D-F).

Several factors contributed to variable Probability gain by LAMPA in the three query-target pairs (Table S3). In the YP_004070193.2-PF14519.5 pair, it was limited to 3.4% because of high HHsearch hit score = 67.2 that defined Probability = 94% which was close to the LAMPA 95% cut-off (Figs. 7F and S5C). Likewise relatively low scores, 40.2 and 41.1, defined high Probability gains in pairs YP_009179227.1-PF08301.12 and YP_009388303.1-

PF13238.5 (Figs. 7D,E and S5A,B). This gain was smaller in the second pair because of higher diversity of its target profile (PF13238.5 vs PF08301.12 – 11.5 and 6.1, respectively) (Fig. S5D,E).

The dependence of Probability on lengths and diversities of the query and target profiles is complex and remarkably symmetrical (Fig. S5). The actual Probability values strongly depend on the external parameters (hit score, query and target lengths for Fig. S5D-F plots). Notably, it can show non-monotonous changes for a fixed query or target diversity over most of the range values. In the present study, query profiles were based on a single sequence (diversity = 1), with Probability estimation only increasing with further increase of the observed diversity in three target profiles (Fig. S5D-F).

Text S1.6 Instructions regarding the usage of LAMPA R package

The package is provided on GitHub: https://github.com/Gorbalenya-Lab/LAMPA. It can be installed using R commands *library(devtools); install_github('Gorbalenya-Lab/LAMPA')* and loaded using R command *library(LAMPA)*. The package contains a single user-level function, that is called also LAMPA. To display detailed information about the usage of this function, use R command *help(LAMPA)*.

While we run the analysis of RNA virus polyproteins using HHmake and HHsearch programs from HH-suite 2.0.16 and script addss.pl from HH-suite 3.0.0 against pfamA_31.0, the package is expected to work with other versions of these HH-suite programs and scripts as well, provided that they have the same input and output data formats. Other databases compatible with the HH-suite programs can also be used. Running LAMPA based solely on HH-suite v.3.x is technically possible but may be affected by HHsearch v.3.x issue which leads to overuse of random access memory (RAM) during searches of large databases and could cause job crushed (https://github.com/soedinglab/hh-suite/issues/124).

Single run of the LAMPA function conducts the annotation procedure for a single query sequence. To apply the function to multiple query sequences, user can employ R loop *for* iterating over query sequences and running the LAMPA function for each query sequence in succession [30]; the number of central processing units (CPUs) utilized in HHsearch searches can be regulated via the LAMPA argument *cpu*. Alternatively, user can employ R package doParallel to run the LAMPA function for multiple query sequences in parallel [45]; it is recommended to set value of the LAMPA argument *cpu* to 1 in this case.

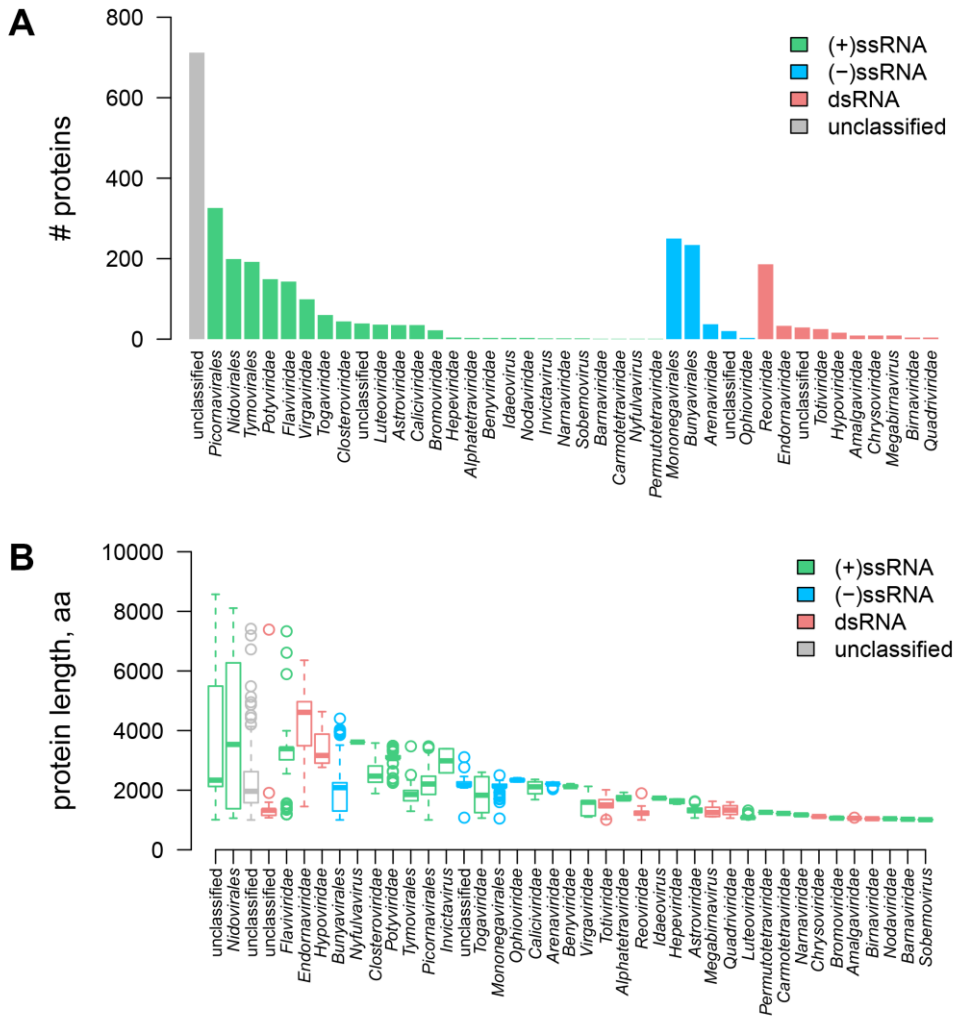

Figure S1 | Composition of the analysed RNA virus polyprotein dataset. (**A**) Number of proteins belonging to different taxonomic groups. (**B**) Length of proteins from different taxonomic groups. Virus taxonomy for each protein were derived from the corresponding genome RefSeq entry; only the most senior taxonomic rank specified in the entry is shown for each protein.

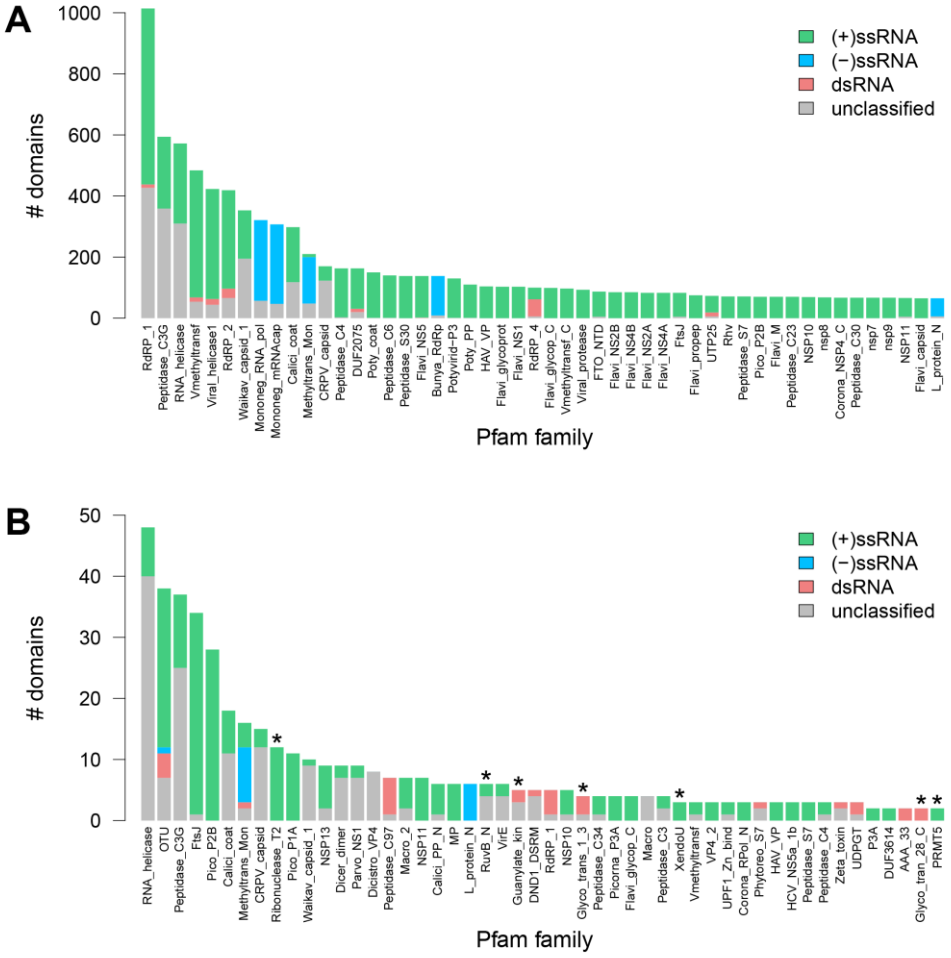

Figure S2 | Target profiles that dominated LAMPA hit lists of RNA virus polyproteins. Fifty Pfam profiles that were most frequently hit by RNA virus polyproteins during (**A**) stage #1 and (**B**) stages #2-#3 of the LAMPA procedure. Pfam profiles, not hit at stage #1 (unique to LAMPA compared to conventional HHsearch), are highlighted with asterisks.

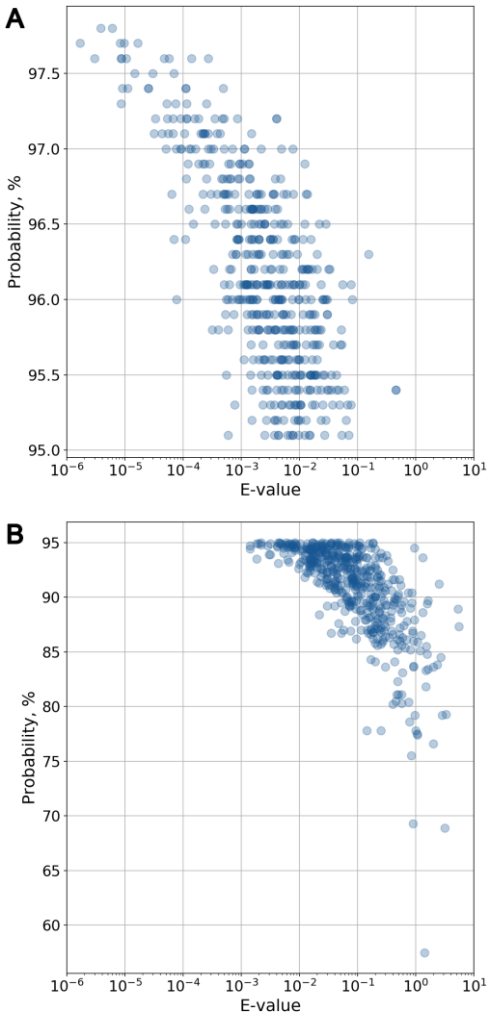

Figure S3 | Relationship between Probability and E-value for HHsearch hits. The plots show relationship between Probability and E-value for 507 hits that were elevated above 95% Probability cut-off by LAMPA at stages #2 and #3 (**A**) compared to stage #1 that is equivalent to HHsearch output (**B**). Probabilities and E-values of hits are inversely related, and this relationship is modulated by hits' secondary structure scores that are distributed in a wide range (from -3.6 to 18.8) and affect Probability but not E-value. Variation of Probability values decreases and E-values in logarithmic scale increases after hits were elevated above 95% Probability cut-off. Both these trends are determined by the properties of hit score auto-calibration procedure; in particular by the observed dependence of Probability and P-,E-value on query profile length, see Figure 7.

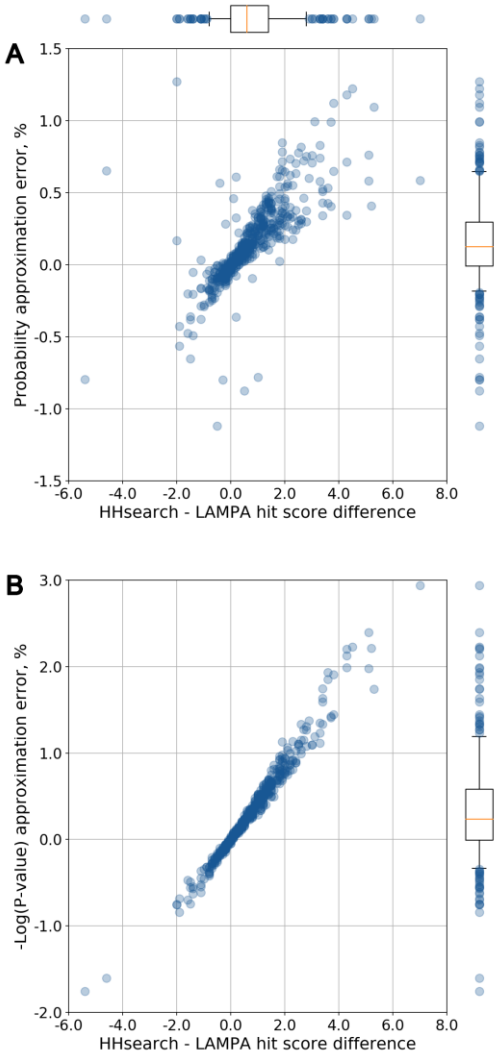

Figure S4 | Statistic approximation error and its dependence on hit score accuracy of query in computational experiments. In computational experiments, hit statistics were calculated for each query, regardless of its length, using fixed hit score(s) obtained for respective intact polyprotein. The depicted plots show relationship between deltas of hit statistic (Y axis) and its score (X axis) calculated for polyprotein and its region, which were used as queries at stages #1 vs #2 and #3 of LAMPA. The delta of hit statistic, Probability (panel **A**) and P-value (panel **B**), is equal to error of statistic approximated in computational experiments. Hit score used to calculate Probability but not P-value is composite and includes secondary structure score. Box-and-whisker summary statistic for two variables: box, 25%-75% range, whiskers 2.5%-97.5% range.

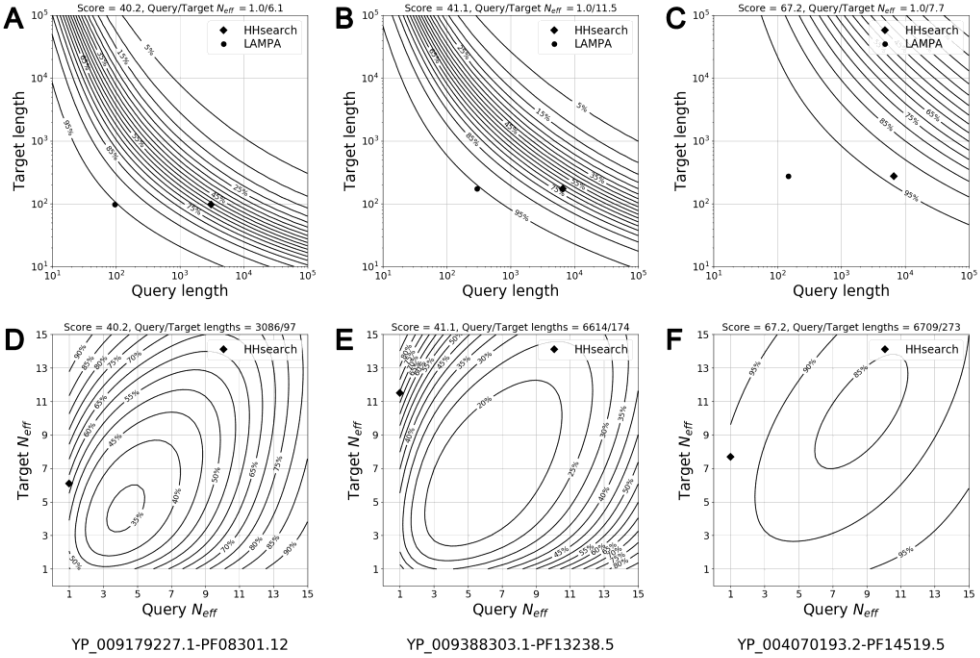

Figure S5 | Relationship of hit Probability to query and target lengths and diversities in computational experiments. Presented are results of estimation of HHsearch hit Probability for different combinations of either query and target lengths (**A-C**) or query and target profile diversities (**D-F**), which were computationally generated. Diamond and circle labels in A-C panels indicate lengths of profiles used to detect the hit by HHsearch (without hits post-processing) and LAMPA (stage #2 or #3), respectively. Diamond label in D-F panels indicates real values of target and query diversities. Three query-target pairs used for panels A and D, B and E, and C and F are indicated at the bottom.

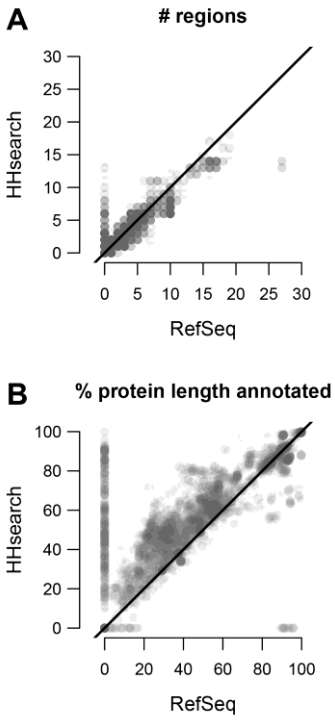

Figure S6 | Summary statistic of annotation coverage by HHsearch and RefSeq experts. Comparison of the number of regions per protein (**A**) or percentage of protein length (protein coverage) (**B**) annotated by HHsearch (LAMPA stage #1) and RefSeq experts, based on analysis 2985 RNA virus proteins. Each protein is represented by a transparent grey dot; dot density is proportional to the number of proteins with identical characteristics. Black line, diagonal.

Table S1 | RNA virus polyproteins used for testing LAMPA.

Table is available from https://doi.org/10.1093/bioinformatics/btaa065

Table S2 | Hits between RNA virus polyproteins and PfamA profiles identified during QP-specific and AP-specific stages of LAMPA.

Table is available from https://doi.org/10.1093/bioinformatics/btaa065

Table S3 | Characteristics affecting estimation of statistical significance of similarity in three query-target pairs.

Table is available from https://doi.org/10.1093/bioinformatics/btaa065

REFERENCES

1. Brister JR, Ako-Adjei D, Bao Y, Blinkova O: **NCBI viral genomes resource**. *Nucleic Acids Res* 2015, **43**(Database issue):D571-577.
2. Suttle CA: **Marine viruses--major players in the global ecosystem**. *Nat Rev Microbiol* 2007, **5**(10):801-812.
3. Zhang YZ, Chen YM, Wang W, Qin XC, Holmes EC: **Expanding the RNA Virosphere by Unbiased Metagenomics**. *Annu Rev Virol* 2019.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
5. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I: **GenBank**. *Nucleic Acids Res* 2019, **47**(D1):D94-D99.
6. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W29-37.
7. Söding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**(7):951-960.
8. Remmert M, Biegert A, Hauser A, Söding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment**. *Nat Methods* 2012, **9**(2):173-175.
9. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A *et al*: **The Pfam protein families database in 2019**. *Nucleic Acids Res* 2019, **47**(D1):D427-D432.
10. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Soding J, Steinegger M: **Uniclust databases of clustered and deeply annotated protein sequences and alignments**. *Nucleic Acids Res* 2017, **45**(D1):D170-D176.
11. Fox NK, Brenner SE, Chandonia JM: **SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures**. *Nucleic Acids Res* 2014, **42**(Database issue):D304-309.
12. Wheelan SJ, Marchler-Bauer A, Bryant SH: **Domain size distributions can predict domain boundaries**. *Bioinformatics* 2000, **16**(7):613-618.
13. Remmert M: **Fast, sensitive protein sequence searches using iterative pairwise comparison of hidden Markov models**. *Doctoral dissertation.* Munich: Ludwig Maximilian University; 2011.
14. Söding J, Remmert M, Hauser A: **User Guide 2.0.15: HH-suite for sensitive protein sequence searching based on HMM-HMM alignment**. In: *HH-suite.* 2012.
15. Baltimore D: **Expression of animal virus genomes**. *Bacteriol Rev* 1971, **35**(3):235-241.

16. Das K, Arnold E: **Negative-Strand RNA Virus L Proteins: One Machine, Many Activities**. *Cell* 2015, **162**(2):239-241.

17. Zhang J: **Protein-length distributions for the three domains of life**. *Trends Genet* 2000, **16**(3):107-109.

18. Dougherty WG, Semler BL: **Expression of virus-encoded proteinases: functional and structural similarities with cellular enzymes**. *Microbiol Rev* 1993, **57**(4):781-822.

19. Gorbalenya AE, Snijder EJ: **Viral cysteine proteinases**. *Perspectives in Drug Discovery and Design* 1996, **6**(1):64-86.

20. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R: **Viral mutation rates**. *J Virol* 2010, **84**(19):9733-9748.

21. Kuchibhatla DB, Sherman WA, Chung BY, Cook S, Schneider G, Eisenhaber B, Karlin DG: **Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently "orphan" viral proteins**. *J Virol* 2014, **88**(1):10-20.

22. Habermann BH: **Oh Brother, Where Art Thou? Finding Orthologs in the Twilight and Midnight Zones of Sequence Similarity**. In: *Evolutionary Biology: Convergent Evolution, Evolution of Complex Traits, Concepts and Methods.* Edited by Pontarotti P. Cham: Springer International Publishing; 2016: 393-419.

23. Saberi A, Gulyaeva AA, Brubacher JL, Newmark PA, Gorbalenya AE: **A planarian nidovirus expands the limits of RNA genome size**. *PLoS Pathog* 2018, **14**(11):e1007314.

24. Steinegger M, Soding J: **MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets**. *Nat Biotechnol* 2017, **35**(11):1026-1028.

25. Firth AE, Brierley I: **Non-canonical translation in RNA viruses**. *J Gen Virol* 2012, **93**(Pt 7):1385-1409.

26. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D *et al*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. *Nucleic Acids Res* 2016, **44**(D1):D733-745.

27. Sonnhammer EL, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences**. *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:175-182.

28. Steinegger M, Meier M, Mirdita M, Vohringer H, Haunsberger SJ, Soding J: **HH-suite3 for fast remote homology detection and deep protein annotation**. *BMC Bioinformatics* 2019, **20**(1):473.

29. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices**. *J Mol Biol* 1999, **292**(2):195-202.

30. R Core Team: **R: A Language and Environment for Statistical Computing**. In. Vienna, Austria: R Foundation for Statistical Computing; 2018.

31. Charif D, Lobry JR: **SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis**. In: *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations.* Edited by Bastolla U, Porto M, Roman HE, Vendruscolo M. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007: 207-232.

32. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for computing and annotating genomic ranges**. *PLoS Comput Biol* 2013, **9**(8):e1003118.

33. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS *et al*: **Redefining the invertebrate RNA virosphere**. *Nature* 2016, **540**:539-543.

34. Schutze H, Ulferts R, Schelle B, Bayer S, Granzow H, Hoffmann B, Mettenleiter TC, Ziebuhr J: **Characterization of White bream virus reveals a novel genetic cluster of nidoviruses**. *J Virol* 2006, **80**(23):11598-11609.

35. Stenglein MD, Jacobson ER, Wozniak EJ, Wellehan JF, Kincaid A, Gordon M, Porter BF, Baumgartner W, Stahl S, Kelley K *et al*: **Ball python nidovirus: a candidate etiologic agent for severe respiratory disease in Python regius**. *MBio* 2014, **5**(5):e01484-01414.

36. Decroly E, Ferron F, Lescar J, Canard B: **Conventional and unconventional mechanisms for capping viral mRNA**. *Nat Rev Microbiol* 2011, **10**(1):51-65.

37. Punta M, Ofran Y: **The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function**. *PLoS Comput Biol* 2008, **4**(10):e1000160.

38. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A *et al*: **A large-scale evaluation of computational protein function prediction**. *Nat Methods* 2013, **10**(3):221-227.

39. Shi M, Lin XD, Vasilakis N, Tian JH, Li CX, Chen LJ, Eastwood G, Diao XN, Chen MH, Chen X *et al*: **Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses**. *J Virol* 2016, **90**(2):659-669.

40. Li CX, Shi M, Tian JH, Lin XD, Kang YJ, Chen LJ, Qin XC, Xu J, Holmes EC, Zhang YZ: **Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses**. *Elife* 2015, **4**.

41. Kobayashi K, Atsumi G, Iwadate Y, Tomita R, Chiba K-i, Akasaka S, Nishihara M, Takahashi H, Yamaoka N, Nishiguchi M *et al*: **Gentian Kobu-sho-associated virus: a tentative, novel double-stranded RNA virus that is relevant to gentian Kobu-sho syndrome**. *Journal of General Plant Pathology* 2013, **79**(1):56-63.

42. Bekal S, Domier LL, Gonfa B, McCoppin NK, Lambert KN, Bhalerao K: **A novel flavivirus in the soybean cyst nematode**. *J Gen Virol* 2014, **95**(Pt 6):1272-1280.

43. Ong JWL, Li H, Sivasithamparam K, Dixon KW, Jones MGK, Wylie SJ: **Novel Endorna-like viruses, including three with two open reading frames, challenge the membership criteria and taxonomy of the Endornaviridae**. *Virology* 2016, **499**:203-211.

44. Valverde RA, Khalifa ME, Okada R, Fukuhara T, Sabanadzovic S, Ictv Report C: **ICTV Virus Taxonomy Profile: Endornaviridae**. *J Gen Virol* 2019.

45. Microsoft Corporation, Weston S: **doParallel: Foreach Parallel Adaptor for the 'parallel' Package**. In.; 2018.