ELM: Enhanced lowest common ancestor based method for detecting a pathogenic virus from a large sequence dataset

Keisuke Ueno[1], Akihiro Ishii[2], Kimihito Ito[1]*

[1]Division of Bioinformatics, Research Center for Zoonosis Control, Hokkaido University, Sapporo, Hokkaido, 001-0020, Japan

[2]Hokudai Center for Zoonosis Control in Zambia, Research Center for Zoonosis Control, Hokkaido University, Sapporo, Hokkaido, 001-0020, Japan

*Corresponding to: Kimihito Ito, Division of Bioinformatics, Research Center for Zoonosis Control, Hokkaido University, Sapporo, Hokkaido, 001-0020, Japan. E-mail: itok@czc.hokudai.ac.jp

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.385v1 | CC-BY 4.0 Open Access | received: 11 May 2014, published: 11 May 2014

Abstract

Emerging viral diseases, most of which are caused by the transmission of viruses from animals to humans, pose a threat to public health. Discovering pathogenic viruses through surveillance is the key to preparedness for this potential threat. Next generation sequencing (NGS) helps us to identify viruses without the design of a specific PCR primer. The major task in NGS data analysis is taxonomic identification for vast numbers of sequences. However, taxonomic identification via a BLAST search against all the known sequences is a computational bottleneck. Here we propose an enhanced lowest-common-ancestor based method (ELM) to effectively identify viruses from massive sequence data. To reduce the computational cost, ELM uses a customized database composed only of viral sequences for the BLAST search. At the same time, ELM adopts a novel criterion to suppress the rise in false positive assignments caused by the small database. As a result, identification by ELM is more than 1,000 times faster than the conventional methods without loss of accuracy. We anticipate that ELM will contribute to direct diagnosis of viral infections. The web server and the customized viral database are freely available at http://bioinformatics.czc.hokudai.ac.jp/ELM/.

Keywords: Next generation sequencing, Virus discovery, Diagnostic virology, Virome, Taxonomic identification

Introduction

Most emerging infectious diseases are zoonoses, the pathogens of which are transmitted between humans and animals. The 2009 pandemic H1N1 influenza virus spread worldwide through reassortment that exchanged a gene segment between pigs and humans (Garten et al. 2009). Recently, cases of influenza A virus H7N9 transmitted from birds to humans have been reported (Gao et al. 2013). The 2003 severe acute respiratory syndrome (SARS) outbreak originated from the transmission of a novel bat coronavirus (Li et al. 2005). For the

sporadically endemic Ebola virus, bats are suspected to be the natural reservoir, but this is still controversial (Feldmann et al. 2004). Vector-borne zoonoses caused by transmission of viruses through mosquitoes and ticks have also become a public health concern. The 1999 outbreak of West Nile virus (WNV) that occurred in New York was caused by the transmission of the WNV among birds, horses and humans via mosquitoes (Nash et al. 2001). Similarly, severe fever with thrombocytopenia syndrome (SFTS) was found to be due to a virus transmitted by ticks (Yu et al. 2011).

To prepare for the risk of emerging infectious diseases, we need to identify pathogenic viruses through surveillance of livestock and wild animals. Although universal PCR primers against 16S ribosomal RNA are available for the identification of bacteria, we needed specific PCR primers to identify viruses. In recent years, NGS technologies have become available for identifying novel viruses that cannot be found by Sanger sequencing due to the difficulty of isolation and passage culture (Barzon et al. 2011).

The taxonomic classification of metagenomic sequences is an important task in NGS data analyses (Huson et al. 2007). It has been widely applied to investigate the relationship between human health and the microbiome (Turnbaugh et al. 2006). Recently, a metagenomic analysis of the virome in a monkey infected with simian immunodeficiency virus was conducted, suggesting that the virome was associated with enteropathy caused by HIV (Handley et al. 2012). Through the first screening with NGS, the novel influenza virus H17N10 was identified in bats from metagenomic samples (Tong et al. 2012).

The taxonomic classification of NGS data uses sequence similarity searches such as BLASTX and BLASTN (Altschul et al. 1990) to assign each sequence into a specific taxon based on the hits. However, with the similarity-based approach it is difficult to decide the resolution of assignments because the resolution depends on whether the sequences are conserved or species specific. The metagenome analyzer (MEGAN) employs the lowest

common ancestor (LCA) concept in graph theory to estimate the taxonomical contents of samples (Huson et al. 2007). MEGAN evaluates the resolution of similarity-based assignments as the level of taxonomy based on the LCA.

The LCA is the closest taxon shared among two or more taxa found by a BLAST search for a read. When multiple taxa are found by the BLAST search with sufficiently reliable BLAST scores, the common ancestor is a high-level taxon. The LCA assignments to high-level taxa are associated with conserved sequences. When a single taxon is found by a BLAST search for a read, the common ancestor still remains a low-level taxon. The LCA assignments to low-level taxa are associated with species-specific sequences. Thus, the LCA assignments to low-level taxa are more suitable for resolving closely related organisms than those to high-level taxa.

The SOrt-ITEMS (Monzoorul Haque et al. 2009) and CARMA3 (Gerlach & Stoye 2011) methods extended the LCA using a reciprocal BLAST search to reduce false positives in assignments. CARMA3 introduced the concept of the mutation rate into the LCA algorithm, and reinforced the reciprocal BLAST search to identify a novel taxon, relatives of which are numbered (Gerlach & Stoye 2011).

While taxonomic classification of metagenomic sequences has been developed with respect to accuracy, NGS technologies continue to improve sequencing throughput, and require considerable computational time and resources to perform taxonomic classification. The throughput of Roche 454 sequencing is 700 Mb with an average length of 400-800 bases. The present throughputs of NGS have become over 1 Gb with Illumina sequences of 600 Gb and an average length of ~100 bases, SOLiD sequences of 20 Gb with an average length of ~50 bases, and Ion Torrent PGM sequences of 1 Gb with an average length of ~200 bases (Barzon et al. 2011). These massive sequencing data prevent the fast detection of infecting viruses from metagenomic samples.

To reduce the computational time, we constructed a customized database composed only of viruses for the BLAST search. However, customized databases also increase accidental hits, i.e. the match of host sequences to viral genomic sequences. Here, we introduce ELM with a customized viral database for taxonomic identification. The method is based on the assumption that valid hits, the match of viral sequences to viral genomic sequences, raise the probability of finding other similar genomic sequences in the BLAST search. In other words, true assignments with the LCA should be sensitive to the threshold of the bit score in the BLAST search. Consequently, ELM can suppress the rise of false positive assignments while saving computational time and resources.

Materials and Methods

The ELM server performs taxonomic identification of viral sequences from NGS datasets via three steps (Figure 1). In step one, the server carries out a BLASTN search for a customized database of viral genomic sequences. In step two, the server performs the LCA-based taxonomic assignments using MEGAN software (Huson et al. 2007) with default parameters. In step three, the server iterates the LCA assignments with different parameters for the threshold of the bit scores for the BLAST hits and investigates the taxa in which the number of assigned reads is significantly changed. In this step, the server provides a novel criterion for evaluating the LCA assignments.

BLAST search for customized database

To reduce the computational time and save disk space, we constructed a customized database composed only of viral genomic sequences for a BLASTN search. First, the RefSeq genomic sequences were downloaded from the NCBI. Then a total of 3,336 viral genomic sequences were selected using a custom-made script program and converted into BLAST databases by the formatdb command in the NCBI BLAST package. We used the BLASTN program in the

NCBI BLAST+ version 2.2.26 package with the default parameters to search for similar sequences. The hits with an *E*-value under 10^{-3} were used for subsequent analyses.

LCA analysis for taxonomic classification

The LCA method assigns sequence reads to taxa with a criterion for the resolution of assignments (Huson et al. 2007). $h(q, s)$ is the set of taxa found by a BLAST search for a sequence read q under the threshold of the bit score s. For a set of taxa $h(q, s)$, the common ancestor located farthest from the root of the taxonomic tree defines the LCA as the representative taxon. Thus, the LCA allows the assignment of a read to a single taxon. At the same time, the taxonomic levels indicate the resolution of assignments because the LCA allows broad hits to be assigned as high-level taxa but specific hits to be assigned as low-level taxa. It also means that the number of the LCA assigned reads depends on the thresholds of the bit scores for BLAST hits.

We use MEGAN software version 4.62.5 for the LCA analysis (Huson et al. 2007). MEGAN assigns sequence reads into taxa at ten hierarchical levels: Kingdom, phylum, class, order, family, varietas, genus, species group, subspecies, and species in the taxonomic ordering relation.

ELM for evaluating the LCA assignments

To introduce an additional criterion for the taxonomic assignment, ELM repeats the LCA analysis further under different top percent score filters for the BLAST hits and compares these LCA assignments with the reference assignment under the top 10% score filter (Figure 2). Here, the top x percent score filter retains the BLAST hits whose bit scores lie within x% of the best score (Huson et al. 2007). $n(x)$ is the total number of the LCA assigned reads for a taxon and its descendants under top x percent score filter. Then the difference Δn from that under the reference top 10% score filter is given by:

$$\Delta n = n(10) - n(x) \qquad (1)$$

Here, Δn indicates to what extent the assigned reads are shifted into upper taxa as increasing x greater than 10%. We analyzed the increase of Δn, which is associated with sequence similarity to relatives, to discriminate between true and false assignments. In the statistical analysis of Δn, we introduce the inflation index *IF*, which is the *Z* score for outlier detection, to compare the effect of top percent score filters on the taxonomic assignments. The *IF* for a taxon is given by:

$$IF = \frac{\Delta n - \mu}{\sigma} \qquad (2)$$

where μ is the average of Δn for all assigned taxa, and σ is the standard deviation. Since multiple comparisons in *IF*s under top percent score filters ranging from 20% to 100% are performed nine times at 10% intervals, a *P* value of less than 0.05/9 is accepted for statistical significance after Bonferroni correction. Accordingly, *IF* >2.54 (one-tailed) is accepted with statistical significance.

Benchmark tests for NGS datasets

To evaluate the ability of ELM to detect pathogenic viruses from large sequence datasets, five real datasets were used. Dataset 1 consisted of 4,449,766 unassembled reads from a rodent sample in Zambia (Ishii et al. 2011). Reads with an average length of 236 bases were obtained by Ion Torrent Personal Genome Machine (PGM) sequencing. Dataset 2 consisted of 4,146,547 unassembled reads from a reptile sample (SRR: 527074) deposited in the NCBI Sequence Read Archive (SRA). Reads with an average length of 200 bases were obtained by Illumina sequencing (Stenglein et al. 2012). Dataset 3 consisted of 12,393,506 unassembled reads from a simian sample (SRR: 167721) deposited in the SRA. Reads with an average length of 73 bases were obtained by Illumina sequencing (Chen et al. 2011). We selected these three datasets to evaluate the effects of the read length, host and NGS platform. Furthermore, we applied ELM to fecal samples including multiple virus and phage taxa in dataset 4 (SRR: 1055974 for 12-day-old piglets) and dataset 5 (SRR: 1055972 for 54-day-old

piglets). Reads with an average length of 291 bases in dataset 4 and 400 bases in dataset 5 were obtained by 454 GS FLX Titanium sequencing (Sachsenroder et al. 2014). In these benchmark tests, the BLAST searches were performed on a workstation with an Intel Sandy Bridge CPU 2.6 GHz processor. We compared the result of the BLASTN search for the customized database with that for the NCBI NT database.

Results

Identification of infecting viruses using the LCA with BLASTN-NT

To identify infecting viruses, we performed conventional LCA-based assignment using the results of a BLASTN search of the NCBI NT database (Figure 3). The taxa assigned at the varietas level in dataset 1 showed that this rodent host was infected with *Old world arenavirus* (Figure 3A). A previous study showed that the rodent host was infected with *Luna virus*, which belongs to the *Old world arenaviruses* (Ishii et al. 2011). Totally, 99.9% of the sequences were derived from eukaryotes, including sequences from the rodent host. The reptile host in dataset 2 was infected with *Lymphocytic choriomeningitis virus* (Figure 3B). This result was consistent with the closest virus described in the literature (Stenglein et al. 2012). In dataset 2, 99.5% of the sequences were probably derived from the reptile host. According to the literature concerning dataset 3, the simian host was infected with a novel simian adenovirus, which is close to *Simian adenovirus 3*, *Simian adenovirus 18* and *Simian adenovirus 21* with about 55% pairwise nucleotide identity (Chen et al. 2011). We found *Simian adenovirus 49*, *Simian adenovirus 18* and *Simian adenovirus 1* in dataset 3 (Figure 3C), suggesting results similar to those in the literature. Similarly, in dataset 3, most of the sequences (95.1%) were likely derived from the simian host.

To assess the required computational resources, we measured the elapsed time for the BLAST search (Table 1). As seen in Table 1, we found that the elapsed time for the BLAST search depended on the number of reads and hits. Although multiple threads and parallel jobs

reduced the computational time, we needed at least one day with 8 threads and 32 parallel jobs. The sizes of the resulting tabulated format files ranged from 60-648 gigabytes, possibly affecting the elapsed time for the LCA analysis.

Taxonomic classification using the LCA with BLASTN viruses

According to the literature on taxonomic classification, the sequence similarity search of BLAST is a computational bottleneck (Gerlach & Stoye 2011). Therefore, we used the customized viral database for a BLAST search to investigate how much the computational time was reduced and whether the conventional LCA could identify the infecting viruses. In Figure 4, the top 3 assigned reads show the capturing of infecting viruses. However, most of the assigned taxa were false positives (Figure 4). At the varietas level of assignments in dataset 1, 98.7% of the reads were assigned into *Choristoneura occidentalis granulovirus* and *Spodoptera litura granulovirus*, and only 0.4% (1,518/383,939) of them were assigned into *Luna virus* (Figure 4A). In the case of BLAST-NT, we failed to identify *Luna virus* but detected *Old world arenaviruses*, with 1,245 reads at the family level, including the following relatives: *Mobala virus*, 125 reads; *Morogoro virus*, 73 reads; and *Mopeia virus*, 56 reads. In dataset 2, 8,387 reads were assigned into 141 viral taxa at the genus level, and 573 were assigned into *Lymphocytic choriomeningitis virus* (Figure 4B). In the case of BLAST-NT, 454 reads were assigned into *Lymphocytic choriomeningitis virus*. These results showed consistency between BLAST viruses and BLAST-NT. Of the 5,952 reads assigned into viral taxa at the genus level in dataset 3, 468 were assigned into *Simian adenovirus 49* (Figure 4C). The most assigned taxon in BLAST viruses was *Simian adenovirus 49*, but 99 reads in BLAST-NT were assigned into the closest relative, *Simian adenovirus 18*. Although the assignments with BLAST-NT were more favorable than those with BLAST viruses, the coverage of the identified *Simian adenovirus 49* was sufficient to perform the subsequent analysis. These results showed that the sensitivity of the LCA with BLASTN viruses

outperformed the LCA with BLASTN-NT, suggesting that the BLAST search of the viral database was sufficient for subsequent analysis.

Next, we investigated whether the elapsed time for the BLASTN search was effectively reduced (Table 2). The elapsed time in the BLAST search was reduced to 0.03%-0.05% (Tables 1 and 2). This showed the synergy effect of the reduction of the custom database to 0.1% (from 38 Gb to 49 Mb) for the size of FASTA files and to 1.4%-8.8% for the number of BLASTN hits (Table 1 and 2). Furthermore, the elapsed time for the LCA analysis was also reduced despite the additional nine assignments for ELM analysis.

Identification of infecting viruses using ELM with BLASTN viruses

To reduce the false assignments of the BLAST search for the customized viral database, we compared true and false assignments to confirm whether the true assignments altered into high-level taxa in ELM analysis (Figure 5). As shown for the varietas level assignments of dataset 1 in Figure 5A and D, the assignment of *Luna virus* was significantly changed (*IF* > 2.54, ranging from 20% to 100%), suggesting that ELM correctly identified the infecting virus. In the genus level assignments of dataset 2, the assignment of *Lymphocytic choriomeningitis virus* was most changed (*IF* > 9, ranging from 20% to 100%) but, at the varietas level, those of *unclassified Tospovirus*, *Tomato spotted wilt virus* and *Impatiens necrotic spot virus* were only slightly changed (Figure 5B and E). Figure 5C and F show that, in the genus level assignments of dataset 3, the assignment of *Simian adenovirus 49* was significantly changed (*IF* > 10, ranging from 20% to 100%). However, at the varietas level, the assignments of *Ictalurid herpesvirus 1*, *Simian adenovirus 3* and *Human adenovirus 54* were also changed, suggesting that ELM failed to exclude the false assignment of *Ictalurid herpesvirus 1*. On the other hand, of the taxa shown in Figure 4, the false assignments were little changed, suggesting that ELM excluded the false assignments dependent on the customized database (not shown in Figure 5). These results suggested that the combination

with the taxonomic level was better than only the inflation indices. The results for viruses identified using ELM with BLASTN viruses were assembled using SSAKE v3.8.1 (Warren et al. 2007) and are summarized in Table 3.

Next, we evaluated the effect of the BLAST hit score on the inflation indices. The results showed that the inflation indices had little association with the *E*-value in the BLAST search (Additional file 1: Figure S1). We also investigated the coverage of the BLAST hits. Valid hits in dataset 1 were distributed across target genomic sequences but not a specific genomic sequence, something not seen in datasets 2 and 3 (Additional file 1: Figure S2).

Virome analyses using ELM with BLASTN viruses

To investigate whether ELM could detect multiple viral taxa, we analyzed the fecal virome of piglets using ELM with BLASTN viruses. We identified the shift of *Kobuvirus* in dataset 4 to *Bocavirus* and *Dependovirus* in dataset 5, which depended on the age of the piglets (Table 4). These results were consistent with abundant virus genera described in the literature (Sachsenroder et al. 2014). However, we failed to identify *pig stool-associated small circular DNA virus* in dataset 5 (Table 4). This virus belongs to the *single-stranded circular DNA viruses*. The members of this family show extensive genetic diversity (Cheung et al. 2013). The results suggested that, in this case, the inflation index was not preferable for evaluating the LCA assignments.

Discussion

ELM with a specific database drastically reduced the computational time and saved disk space. Furthermore, ELM was effective even for short reads. Though short reads can reduce the accuracy of BLAST searches, in this study we verified ELM for average lengths of between 73 and 400 bases. The results showed no difference between the capabilities for taxonomic assignment.

One approach to reduce the computational time needed for the BLAST search is the subtraction of reads by mapping host-derived reads onto reference sequences (Chang et al. 1994; Simons et al. 1995). This approach might be considered effective for reducing analyzed sequence data but is limited to known hosts. It is not suitable for surveillance of wild animals or metagenomic analysis because the host sequences have yet to be deposited in databases. Therefore, we need to decide a moderate threshold for NGS data before the mapping.

For ELM we adopted another approach using specific databases composed only of target sequences to reduce the computational cost. The difficulty in applying this approach directly to virus identification was the increase of false positive assignments (Figure 4). We tested several ways to solve this problem. As shown in Additional file 1: Figure S1, changing the threshold of the *E*-value dependent on the size of the database is probably not effective for discriminating between true and false assignments. The criteria for evaluating breadth coverage, i.e. the proportion of reads mapped across the hit genomes and the depth coverage (the number of reads mapped at a position), also failed to identify the target viruses (Additional file 1: Figure S2). On the other hand, ELM analyzed how sequence similarity to the relatives changes. This extension of the LCA method suppressed the rise in false positive assignments. A limitation of ELM would be the false-negative errors because ELM cannot detect viruses distantly related to other relatives (Table 4). Therefore, viruses without relatives should be carefully handled without the inflation index.

Conclusions

ELM is especially useful for the first screening of infectious diseases caused by viruses. In surveillance for pathogenic viruses, taxonomic assignment of the host sequence is not necessary for the initial screening. For this, sensitivity for detecting viruses is particularly required. Our results suggest that ELM recovers most reads assigned to target viruses.

Therefore, we can apply these results to further sophisticated analyses. ELM will contribute to analyses of NGS data for limited targets such as the direct diagnosis of viral infections.

Abbreviations

NGS, Next generation sequencing; LCA, Lowest common ancestor; ELM, Enhanced LCA-based method.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KU designed, implemented, tested and evaluated the method, and wrote the manuscript. AI provided the experimental NGS data analyzed in the manuscript. KI participated in the design and evaluation of the method and collaborated in writing the manuscript. All authors read and approved the manuscript.

Acknowledgements

We thank Stephan Schuster and Daniel Huson for allowing us to use MEGAN software for this work.

References

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403-410.

Barzon L, Lavezzo E, Militello V, Toppo S, and Palu G. 2011. Applications of next-generation sequencing technologies to diagnostic virology. *Int J Mol Sci* 12:7861-7884.

Chang Y, Cesarman E, Pessin MS, Lee F, Culpepper J, Knowles DM, and Moore PS. 1994. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* 266:1865-1869.

Chen EC, Yagi S, Kelly KR, Mendoza SP, Tarara RP, Canfield DR, Maninger N, Rosenthal A, Spinner A, Bales KL, Schnurr DP, Lerche NW, and Chiu CY. 2011. Cross-species transmission of a novel adenovirus associated with a fulminant pneumonia outbreak in a new world monkey colony. *PLoS Pathog* 7:e1002155.

Cheung AK, Ng TF, Lager KM, Bayles DO, Alt DP, Delwart EL, Pogranichniy RM, and Kehrli ME, Jr. 2013. A divergent clade of circular single-stranded DNA viruses from pig feces. *Arch Virol* 158:2157-2162.

Feldmann H, Wahl-Jensen V, Jones SM, and Stroher U. 2004. Ebola virus ecology: a continuing mystery. *Trends Microbiol* 12:433-437.

Gao R, Cao B, Hu Y, Feng Z, Wang D, Hu W, Chen J, Jie Z, Qiu H, Xu K, Xu X, Lu H, Zhu W, Gao Z, Xiang N, Shen Y, He Z, Gu Y, Zhang Z, Yang Y, Zhao X, Zhou L, Li X, Zou S, Zhang Y, Yang L, Guo J, Dong J, Li Q, Dong L, Zhu Y, Bai T, Wang S, Hao P,

Yang W, Han J, Yu H, Li D, Gao GF, Wu G, Wang Y, Yuan Z, and Shu Y. 2013. Human Infection with a Novel Avian-Origin Influenza A (H7N9) Virus. *N Engl J Med*.

Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, Sessions WM, Xu X, Skepner E, Deyde V, Okomo-Adhiambo M, Gubareva L, Barnes J, Smith CB, Emery SL, Hillman MJ, Rivailler P, Smagala J, de Graaf M, Burke DF, Fouchier RA, Pappas C, Alpuche-Aranda CM, Lopez-Gatell H, Olivera H, Lopez I, Myers CA, Faix D, Blair PJ, Yu C, Keene KM, Dotson PD, Jr., Boxrud D, Sambol AR, Abid SH, St George K, Bannerman T, Moore AL, Stringer DJ, Blevins P, Demmler-Harrison GJ, Ginsberg M, Kriner P, Waterman S, Smole S, Guevara HF, Belongia EA, Clark PA, Beatrice ST, Donis R, Katz J, Finelli L, Bridges CB, Shaw M, Jernigan DB, Uyeki TM, Smith DJ, Klimov AI, and Cox NJ. 2009. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* 325:197-201.

Gerlach W, and Stoye J. 2011. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res* 39:e91.

Handley SA, Thackray LB, Zhao G, Presti R, Miller AD, Droit L, Abbink P, Maxfield LF, Kambal A, Duan E, Stanley K, Kramer J, Macri SC, Permar SR, Schmitz JE, Mansfield K, Brenchley JM, Veazey RS, Stappenbeck TS, Wang D, Barouch DH, and Virgin HW. 2012. Pathogenic simian immunodeficiency virus infection is associated with expansion of the enteric virome. *Cell* 151:253-266.

Huson DH, Auch AF, Qi J, and Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Res* 17:377-386.

Ishii A, Thomas Y, Moonga L, Nakamura I, Ohnuma A, Hang'ombe B, Takada A, Mweene A, and Sawa H. 2011. Novel arenavirus, Zambia. *Emerg Infect Dis* 17:1921-1924.

Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, Wang H, Crameri G, Hu Z, Zhang H, Zhang J, McEachern J, Field H, Daszak P, Eaton BT, Zhang S, and Wang LF. 2005. Bats are natural reservoirs of SARS-like coronaviruses. *Science* 310:676-679.

Monzoorul Haque M, Ghosh TS, Komanduri D, and Mande SS. 2009. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* 25:1722-1730.

Nash D, Mostashari F, Fine A, Miller J, O'Leary D, Murray K, Huang A, Rosenberg A, Greenberg A, Sherman M, Wong S, and Layton M. 2001. The outbreak of West Nile virus infection in the New York City area in 1999. *N Engl J Med* 344:1807-1814.

Sachsenroder J, Twardziok SO, Scheuch M, and Johne R. 2014. The general composition of the faecal virome of pigs depends on age, but not on feeding with a probiotic bacterium. *PLoS One* 9:e88888.

Simons JN, Pilot-Matias TJ, Leary TP, Dawson GJ, Desai SM, Schlauder GG, Muerhoff AS, Erker JC, Buijk SL, Chalmers ML, and et al. 1995. Identification of two flavivirus-like genomes in the GB hepatitis agent. *Proc Natl Acad Sci U S A* 92:3401-3405.

Stenglein MD, Sanders C, Kistler AL, Ruby JG, Franco JY, Reavill DR, Dunker F, and Derisi JL. 2012. Identification, characterization, and in vitro culture of highly divergent arenaviruses from boa constrictors and annulated tree boas: candidate etiological agents for snake inclusion body disease. *MBio* 3:e00180-00112.

Tong S, Li Y, Rivailler P, Conrardy C, Castillo DA, Chen LM, Recuenco S, Ellison JA, Davis CT, York IA, Turmelle AS, Moran D, Rogers S, Shi M, Tao Y, Weil MR, Tang K, Rowe LA, Sammons S, Xu X, Frace M, Lindblade KA, Cox NJ, Anderson LJ, Rupprecht CE, and Donis RO. 2012. A distinct lineage of influenza A virus from bats. *Proc Natl Acad Sci U S A* 109:4269-4274.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, and Gordon JI. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444:1027-1031.

Warren RL, Sutton GG, Jones SJ, and Holt RA. 2007. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* 23:500-501.

Yu XJ, Liang MF, Zhang SY, Liu Y, Li JD, Sun YL, Zhang L, Zhang QF, Popov VL, Li C, Qu J, Li Q, Zhang YP, Hai R, Wu W, Wang Q, Zhan FX, Wang XJ, Kan B, Wang SW, Wan KL, Jing HQ, Lu JX, Yin WW, Zhou H, Guan XH, Liu JF, Bi ZQ, Liu GH, Ren J, Wang H, Zhao Z, Song JD, He JR, Wan T, Zhang JS, Fu XP, Sun LN, Dong XP, Feng ZJ, Yang WZ, Hong T, Zhang Y, Walker DH, Wang Y, and Li DX. 2011. Fever with thrombocytopenia associated with a novel bunyavirus in China. *N Engl J Med* 364:1523-1532.

Figures

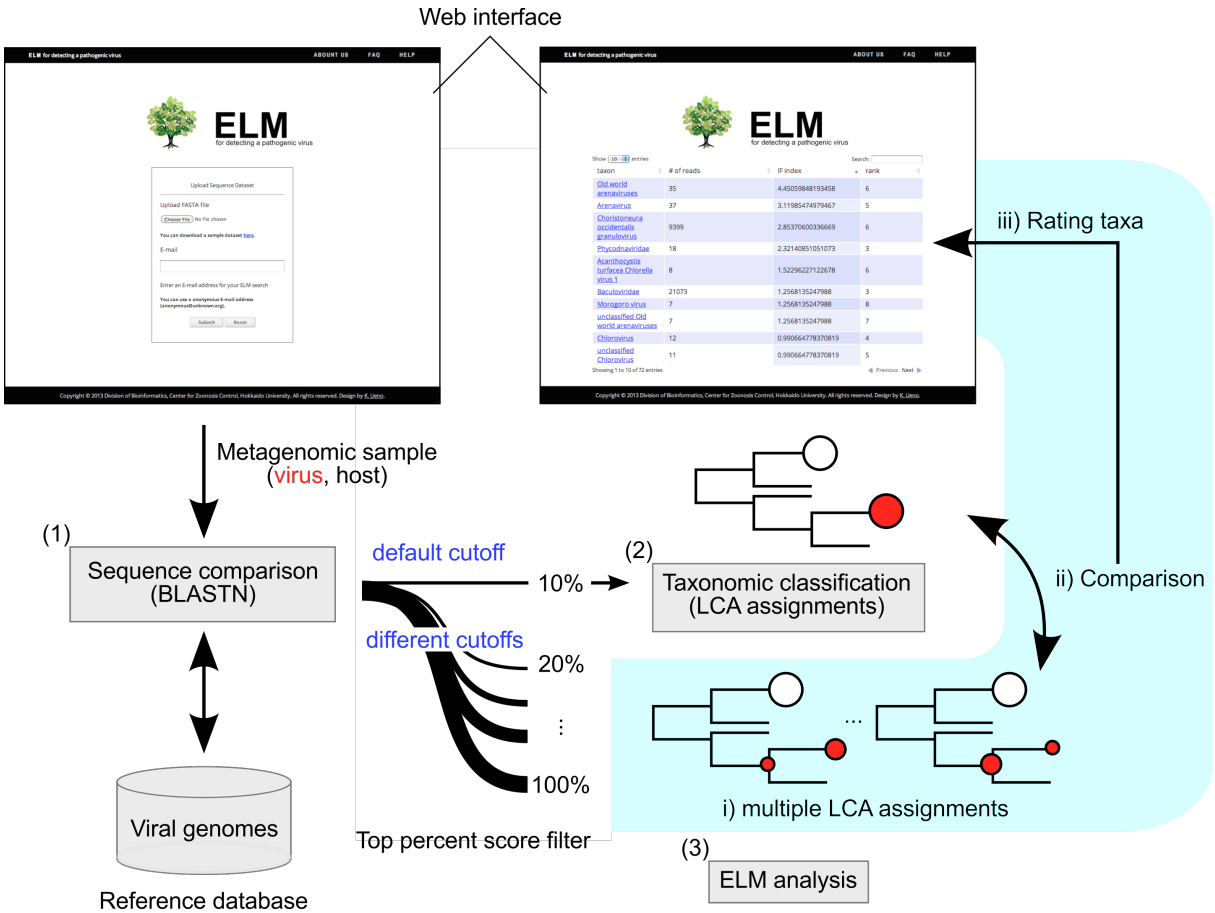
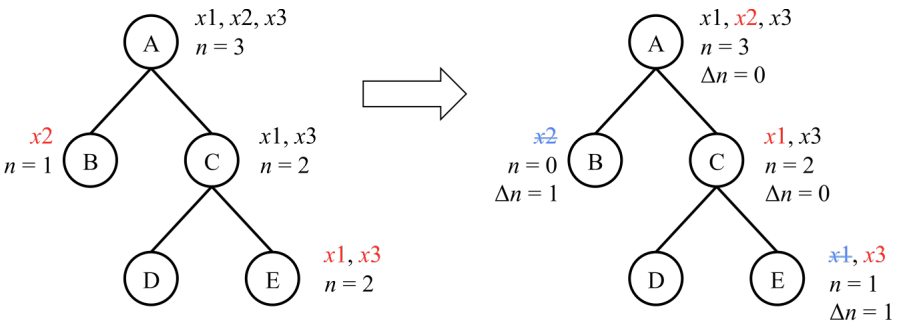


Figure 1 - Overview of ELM server and web interface.

Users input NGS data as a zip file (~ 1Gb). (1) The web server matches NGS reads against known viral genomes using BLAST (Altschul et al. 1990). (2) Taxonomic classification based on the LCA is performed using MEGAN under the top 10% score filter (Huson et al. 2007). (3) In ELM analyisis, multiple comparisons of the LCA assignments are performed under different top percent score filters. The server displays the results with the ratings of taxa.



Read ID	BLAST hit	Hit score
x1	E	100
x1	D	80
…	…	…

Read ID	BLAST hit	Hit score
x2	B	100
x2	C	80
…	…	…

Read ID	BLAST hit	Hit score
x3	E	100
x3	D	20
…	…	…

Figure 2 - Schematic representation of the ELM algorithm.

An example of the LCA assigned NGS reads into target viral taxa. The LCA assignment is affected by top percent score filters—that is, the BLAST hits for the similar sequences in the relatives. ELM evaluates this effect on the assignments. Circled A to E represent viral taxa on a taxonomic tree. The reads assigned as the LCA are shown in red. The reads corresponding to the reads assigned to descendant taxa as the LCA are shown in black. The total number of the LCA assigned reads for each taxon and its descendants is denoted as n. Δn indicates the differences in n as varying thresholds of top percent score filters. The reads with strikeouts (blue) are the LCA assignments shifted into the upper taxon.

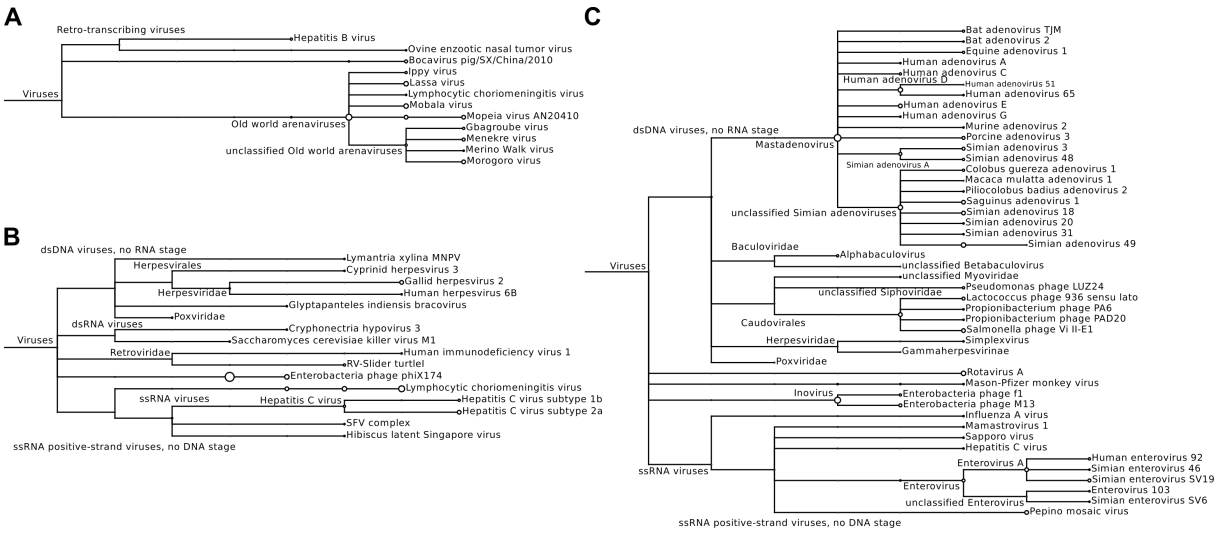


Figure 3 - Taxonomic identification using the LCA with BLASTN-NT.

The taxonomic trees for (A) rodent, (B) reptile and (C) simian samples. The circle sizes indicate the relative numbers of assigned reads. These trees were created using MEGAN (Huson et al. 2007). Here, only the viral taxa are illustrated.

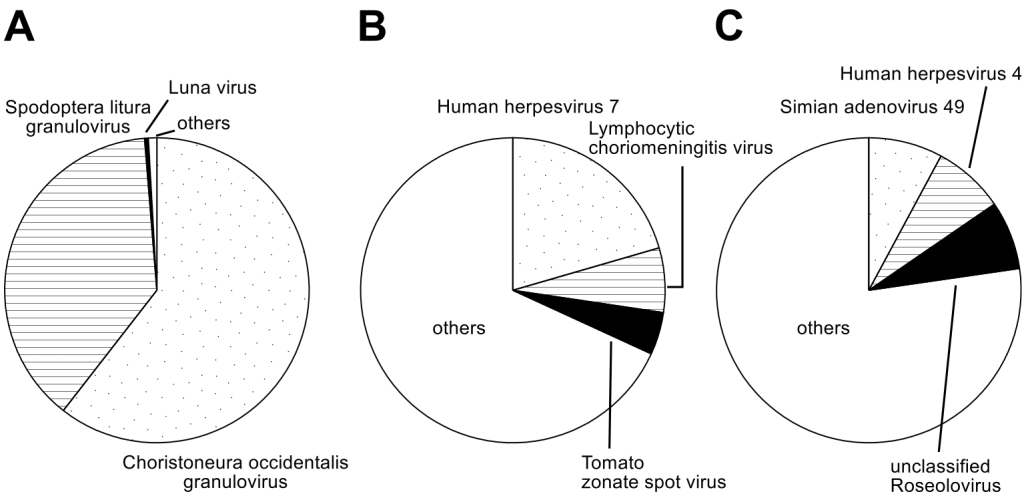


Figure 4 - Taxonomic classification using the LCA with BLASTN viruses.
The pie charts illustrate the number of reads assigned to taxa for (A) the rodent sample at the varietas level, (B) the reptile sample at the genus level and (C) the simian sample at the genus level. Here, only the top three taxa are denoted.

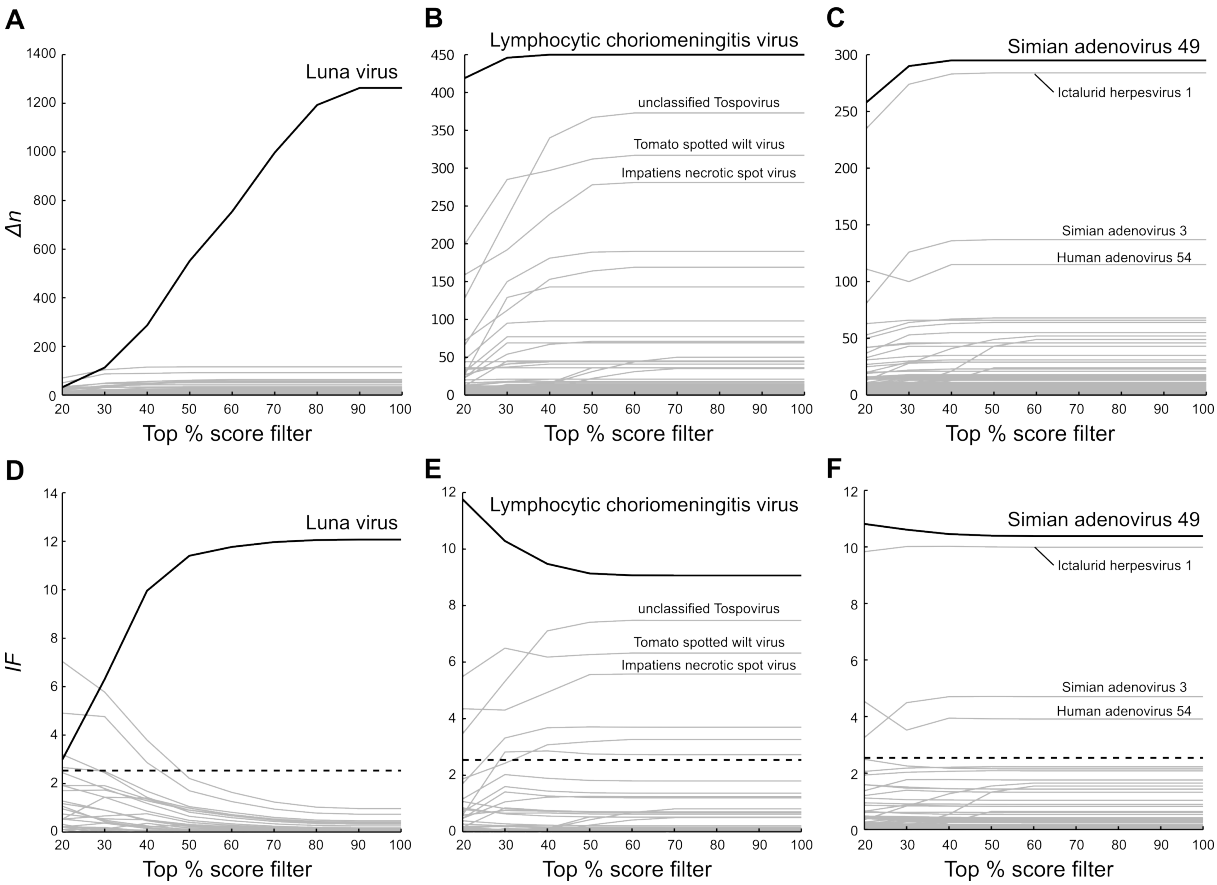


Figure 5 - ELM analyses of BLASTN viruses below the varietas level.

The solid lines depict the differences between the number of the LCA assigned reads for (A) rodent, (B) reptile and (C) simian samples and the inflation indices for (D) rodent, (E) reptile and (F) simian samples. The dashed lines indicate the inflation indices under the null hypothesis.

Tables

Table 1 - Elapsed time for the LCA with BLASTN-NT.

Dataset No.	# of reads	# of BLASTN hits	CPU time	
			BLAST	LCA
1	4,449,766	4,424,602	12,179h	96m
2	4,146,547	2,754,210	8,704h	22m
3	12,393,506	10,674,129	23,313h	82m

Table 2 - Elapsed time for ELM with BLASTN viruses.

Dataset No.	# of BLASTN hits	CPU time	
		BLAST	LCA
1	387,271	4h	5m
2	38,948	4h	1m
3	379,780	6h	5m

Table 3 - Detection of the viral genomes using ELM with BLASTN viruses.

Dataset No.	Virus[a]	# of reads	# of contigs	Average contig length
1	*Luna virus*	1,518	405	454 nt
2	*LCMV*	573	33	117 nt
3	*SAdV-49*	468	11	89 nt

Contigs were assembled using SSAKE v3.8.1 (Warren et al. 2007). [a]*LCMV*, *Lymphocytic choriomeningitis virus*; *SAdV-49*, *Simian adenovirus* 49.

Table 4 - Detection of abundant virus genera in fecal viromes of piglets using ELM with BLASTN viruses.

Dataset No.	ELM with BLASTN viruses (# of reads)	LCA with BLAST-NT (# of reads)
4	*Kobuvirus*[a] (6,449)	*Kobuvirus* (6,446)
5	*Dependovirus*[a] (759), *Bocavirus* (133)	*Dependovirus* (754), *Bocavirus* (528), *Chimpanzee stool associated circular ssDNA virus*[b] (106)

[a]The genus includes descendant taxa *IF* > 2.54. [b]According to the literature (Sachsenroder et al. 2014), this virus is the novel *pig stool-associated single-stranded DNA virus*, which is not assigned to a specific genus.

Additional files

Additional file 1 – Supplementary data

File “supplementary.pdf” contains Figure S1 and Figure S2.