A Polyglot Approach to Bioinformatics Data Integration: A Phylogenetic Analysis of HIV-1

Steven Reisman[1–3], Thomas Hatzopoulos[1,2], Konstantin Läufer[1,2], George K. Thiruvathukal[1,2] and Catherine Putonti[1–3]

[1]Bioinformatics Program, Loyola University Chicago, Chicago, IL, USA. [2]Department of Computer Science, Loyola University Chicago, Chicago, IL, USA. [3]Department of Biology, Loyola University Chicago, Chicago, IL, USA.

ABSTRACT: As sequencing technologies continue to drop in price and increase in throughput, new challenges emerge for the management and accessibility of genomic sequence data. We have developed a pipeline for facilitating the storage, retrieval, and subsequent analysis of molecular data, integrating both sequence and metadata. Taking a polyglot approach involving multiple languages, libraries, and persistence mechanisms, sequence data can be aggregated from publicly available and local repositories. Data are exposed in the form of a RESTful web service, formatted for easy querying, and retrieved for downstream analyses. As a proof of concept, we have developed a resource for annotated HIV-1 sequences. Phylogenetic analyses were conducted for >6,000 HIV-1 sequences revealing spatial and temporal factors influence the evolution of the individual genes uniquely. Nevertheless, signatures of origin can be extrapolated even despite increased globalization. The approach developed here can easily be customized for any species of interest.

KEYWORDS: polyglot programming, RESTful web service, phylogenetics

CITATION: Reisman et al. A Polyglot Approach to Bioinformatics Data Integration: A Phylogenetic Analysis of HIV-1. *Evolutionary Bioinformatics* 2016:12 23–27 doi: 10.4137/EBO.S32757.

TYPE: Original Research

RECEIVED: August 09, 2015. **RESUBMITTED:** October 18, 2015. **ACCEPTED FOR PUBLICATION:** October 25, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

PEER REVIEW: Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,313 words, excluding any confidential comments to the academic editor.

FUNDING: SR is partially supported by the College of Arts and Sciences at Loyola University Chicago. GT and CP are partially supported by Loyola University Chicago's Research Support Grant. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: cputonti@luc.edu

Introduction

The increased throughput, coupled with reduced cost and time, of contemporary sequencing technologies has led to a surge in the number of publicly available, complete, annotated genomic sequences. For smaller viral species, it is now feasible to not only produce a single genome for a species but also capture the diversity present in an ecological niche, the focus of numerous metagenomic studies[1–3] as well as more targeted investigations.[4] Furthermore, next-generation sequencing technologies have tremendous potential for the future of diagnostics and subsequent treatment choices, particularly for viral infections.[5] The sensitivity of deep sequencing can capture even rare variants in mixed infections as well as quasispecies.[6–11] Investigation of the viable variations within a viral species not only provides insight into the evolutionary history of a species but also unveils putative avenues for targeted therapies, such as small interfering RNAs[12–15] and control strategies.

Molecular biology is now plagued with the challenges facing numerous other fields – big data. Cloud-based solutions, eg, CloudBurst,[16] Atlas2,[17] and Rainbow,[18] have provided much needed leverage to meet these demands, facilitating large-scale sequence analyses, while also introducing new difficulties.[19] Furthermore, noSQL databases afford a streamlined solution to both manage large datasets and simplify data retrieval and subsequent analysis. The added benefit of agility and scalability of noSQL databases is ideal for the rapidly advancing trends in DNA sequencing technologies, and it is thus not surprising that noSQL databases have been gaining traction in molecular studies.[20–22]

With the increase in the amount of publicly available genomic sequence data, progress can be stymied by the simple task of collecting sequence data and associated, relevant metadata. In an effort to facilitate the aggregation and management of genomic data for subsequent analyses, we have developed a polyglot approach involving multiple languages (Python and Scala), libraries (Flask [http://flask.pocoo.org] and BioJavaX [http://biojava.org]), and persistence mechanisms (text files and MongoDB NoSQL databases [http://www.mongodb.org]). Individual genes or all genes for a given species can be examined beyond just the sequence itself, including information regarding, for instance, the location and date of isolation.

The code developed is agile; it can be applied for any organism of interest to the user. The approach can be customized for any species of interest. The presented solution is developed with downstream evolutionary analyses in mind, as shown by a proof-of-concept study of the evolution of HIV. Our investigation into the three main HIV genes: gag, pol,

and env, reveals spatial and temporal factors influence the evolution of the individual genes uniquely. The web service for the HIV collection (as well as other datasets investigated by the authors) is publicly available at http://hivdb.cs.luc.edu, and the scripts for generating such a data collection are publicly available at https://github.com/LoyolaChicagoCode/hiv-biojava-scala.

Results and Discussion

Data pipeline for collecting sequence data. Code has been developed to aggregate genomic sequence data and available sequence metadata for subsequent analyses. Figure 1 summarizes this process. All complete and partial genome sequences were parsed and separated into individual folders for each gene via a Scala parser utilizing the BioJavaX (http://biojava.org) library. Each sequence was stored in its gene folder with any relevant metadata available within the GenBank file. The generated folders for each of the parsed genes were then pipelined through several python scripts in order to accomplish several post-processing tasks. First, duplicate gene sequences parsed from the same genome were removed. Second, the gene folders were used to create FASTA-formatted records for each of the gene sequences with any necessary metadata stored in the resulting record's FASTA header. Finally, the PyMongo library (https://pypi.python.org/pypi/pymongo/) was used to insert each of the final FASTA records within our publicly exposed MongoDB database.

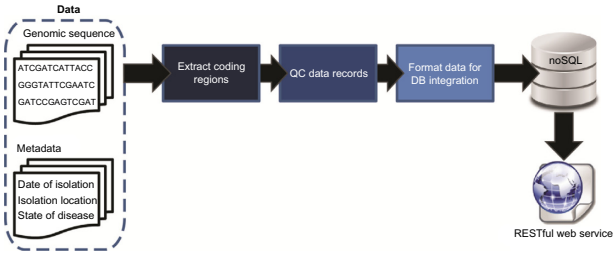

Figure 1. Schematic of data pipeline and access.

Genomic sequence data can then be accessed via a RESTful[23] web service located at http://hivdb.cs.luc.edu. This architecture allows our service to be easily and efficiently accessed by any future data consumers via common web protocols. Data can be queried based upon attributes regarding the source of the data. For example, as shown in Figure 2, the gag gene sequences from strains isolated in the USA can be accessed via the web service. The user can specify search criteria including a year (or range of years) of isolation, the location of isolation, and/or accession number. Sequences meeting the user's search criteria can be returned to the web via the *Query* button or downloaded. All sequence results are in FASTA format for subsequent analysis, such as sequence alignment, primer development, and phylogenetic analysis.

The pipeline has been developed to facilitate users to create repositories for an organism(s) of interest as well as queryable aspects of the sequence annotations. Users need only supply sequences and select attributes and/or genes of interest (otherwise all attributes and genes will be selected). Data are automatically processed. Furthermore, the pipeline is not restricted to publicly available data; any GenBank-formatted file (public or private) can be included. Given the increased

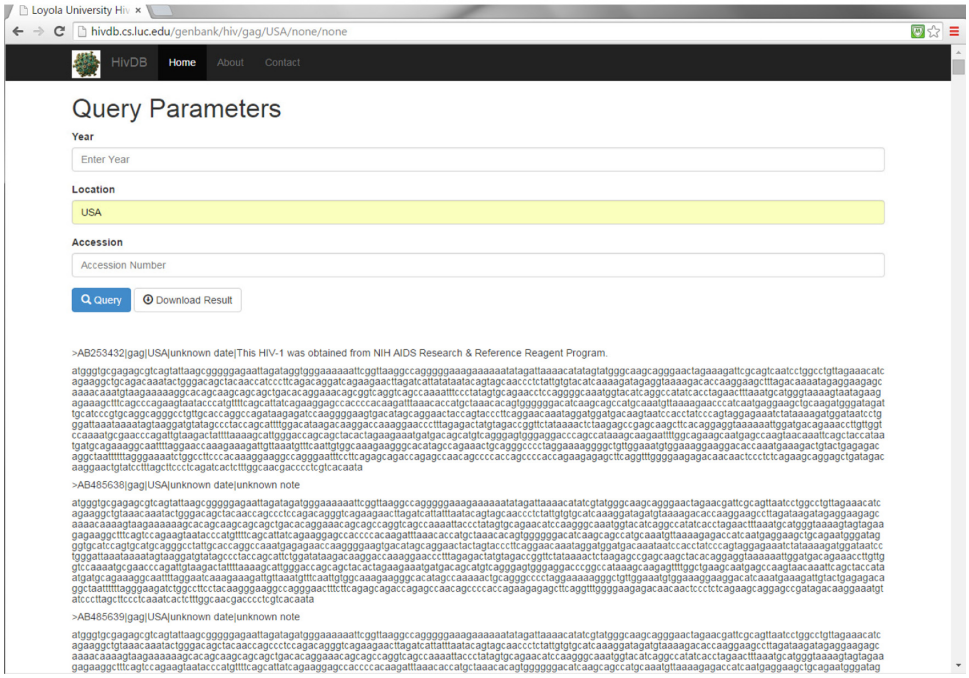

Figure 2. Gene sequence data presented through RESTful web service. Users can query for specific information, eg, as shown here are HIV isolates from the USA, via the Query button or download sequence files in FASTA format meeting their search criteria.

throughput of contemporary sequencing technologies and the decreased cost in sequencing runs, whole genome sequencing is being conducted at unprecedented rates. As such, researchers sequencing novel strains or isolates can incorporate their strains into the data repository once GenBank files are generated. Although this pipeline has been employed by the authors for the analysis of several different taxa, the RESTful web service presented here includes publicly available data for HIV-1 sequences.

Case study: investigation of the evolution of HIV-1. All publicly available complete and near-complete HIV-1 genomic sequences, totaling more than 6,000 sequences, were retrieved from the National Center for Biotechnology Information (NCBI) and processed by our pipeline (see the "Methods and materials" section). Individual gene sequences are publicly available at http://hivdb.cs.luc.edu. Data are accessible in FASTA format to facilitate downstream analyses. To incorporate the metadata collected, including country and date of isolation, this information has been integrated into the FASTA record header. HIV-1 sequences were selected as a proof of concept for this tool as HIV sequences are among the most well curated, thanks in large part to efforts such as those at the Los Alamos National Laboratory's HIV sequence database (http://www.hiv.lanl.gov/).

Previously, phylogenetics has shed light on the origin of HIV and played a key role in identifying recombination events.[24,25] As previous molecular studies have shown, the evolutionary history of the HIV-1 lineage includes three groups (M, N, and O) representative of separate transfers from chimpanzees.[26] Focusing on the three HIV genes gag, pol, and env, the hivdb data repository was queried for coding regions isolated from the same country as well as globally over a particular time period. Host, immunological and antiretroviral drug selection pressures have shaped much of the diversity observed within these three genes.[27] For instance, the investigation of the HIV gag gene sequences from the USA (2,048 sequences: 1990–2011) and Thailand (872 sequences: 2000–2011) is shown in Figure 3A and B, respectively. The phylogenic trees derived for different geographic regions revealed different tree topologies as expected. Sequences isolated during 2005 in the USA exhibit significant sequence variation, including a number of sequences which are distinctly different from sequences isolated during any other year (Fig. 3A). These two gag trees reveal a general trend observed for other countries and other genes: sequences isolated during the same year do not necessarily group together or exhibit a ladder-like topology frequently observed for intra-host HIV phylogenies[28]; this is in concordance with previous HIV survey findings that multiple lineages coexist at any given time.[29]

In addition to looking at the viral diversity from isolates collected within the same country, we also investigated the variants present globally at a given time. Again sequences were retrieved for gag (725 sequences), pol (818 sequences), and env (427 sequences) coding regions. In this example, sequences were retrieved if annotated as being isolated between 2000 and 2005. As shown in Figure 4, there are three main groups within the tree, regardless of the gene being considered. Sequences isolated from Asia are typically found within the same clade, as are sequences from Africa. The third group includes sequences from Europe, North America, and South America. There are, of course, deviations from this trend; these deviations can be the result of multiple introductions, group, or the presence of more than one subtype or recombinant form in circulation, a factor that has been observed by other studies.[26,29–32] Sequences isolated from the same geographic region within the same clade, however, suggest that signatures of origin can be extrapolated even despite increased globalization.

Conclusions

Genomic studies often must consider not only sequences but also the metadata surrounding those sequences. One barrier to such studies is a simple method to collect and organize sequences such that their metadata is also easily accessible. We have taken a polyglot approach to develop a tool which pipelines the process of collecting genomic data and organizing it as automated. As a proof of concept, our pipeline

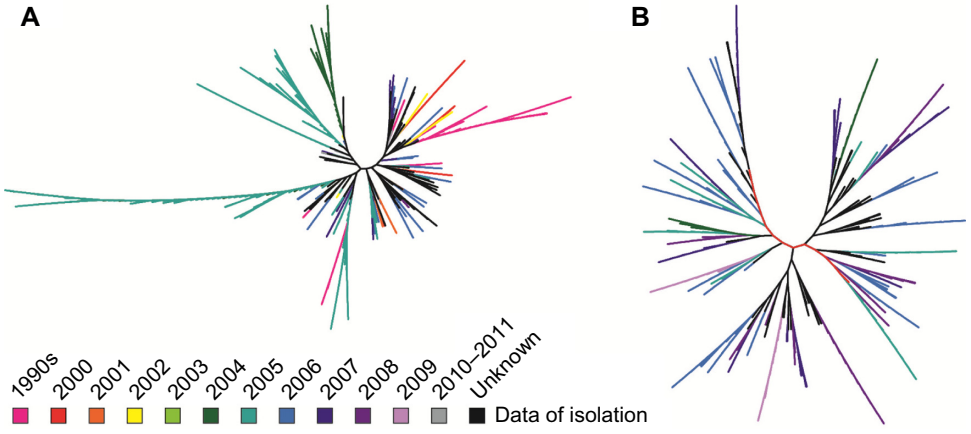

Figure 3. Phylogenetic analysis of the HIV gag coding region from strains isolated before 2012 in (**A**) the USA and (**B**) Thailand.

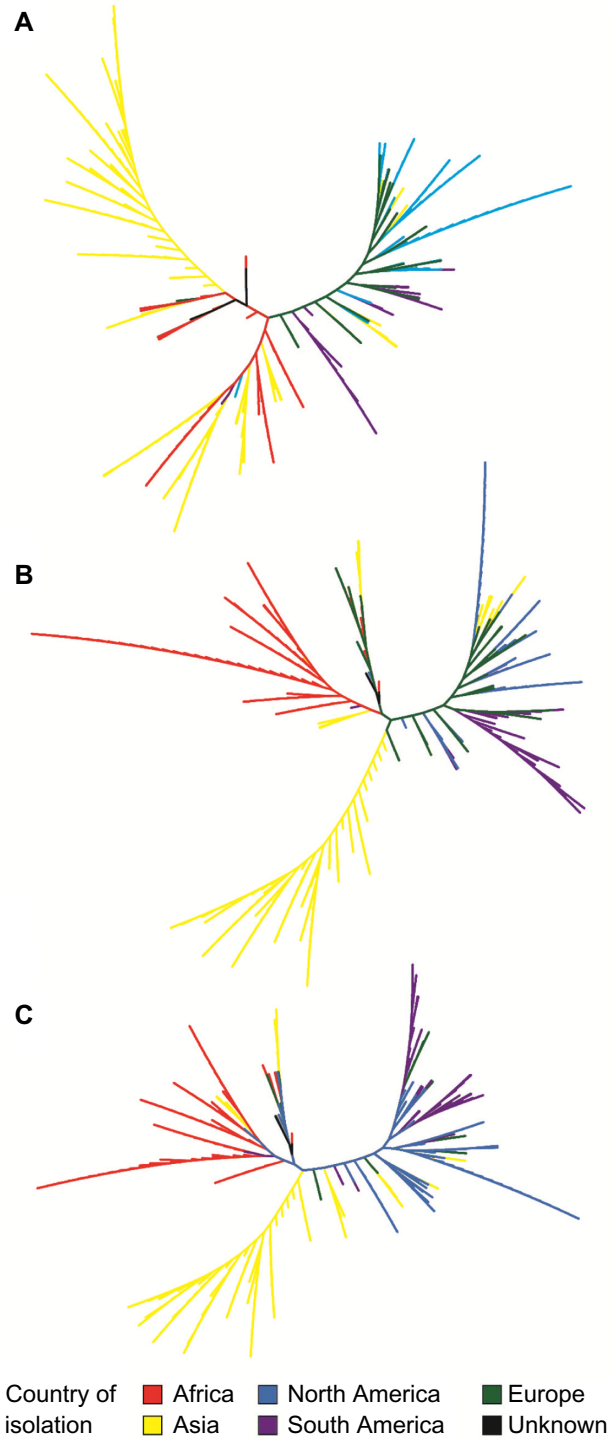

Figure 4. Phylogenetic analysis of the HIV (**A**) gag, (**B**) pol, and (**C**) env coding regions from all genomic sequences isolated between 2000 and 2005. Branches are colored according to the continent from which they were isolated.

process has been applied to an evolutionary study of HIV-1. Phylogenetic analysis of the HIV genes gag, pol, and env finds both spatial and temporal factors uniquely influence the evolution of the individual genes; a finding that is in congruence with prior studies of the evolutionary history of the virus. More importantly, the case study highlights the abilities of the tool. Although utilized for the investigation of a virus here, the approach can be applied to any species of interest.

Methods and Materials

Database development. All generated FASTA-formatted files are stored within the document-based noSQL MongoDB (http://www.mongodb.org). Since metadata from publicly accessible genome data are often not uniformly written, such a system allows each file to contain its own attributes with MongoDB's key value documents. As a result, updated information can easily be added to any given FASTA file, without needing to change the structure of our database. By default, each document contains keys titled "sequence," "country," "accession," "date," "gene," and "note" which map to their corresponding values. A RESTful web service[23] has been created using the Flask Python microframework (http://flask.pocoo.org); this permits users to query stored documents via several parameters, including country of isolation, date, and accession number. Although the queries can be completed through standard HTTP GET and POST requests, a user interface has also been developed for accessing the data.

Extracting information from GenBank files. Scala parsers were developed to extract metadata from NCBI GenBank files. The parser utilizes each GenBank file's CDS tags in order to retrieve information about each gene sequence. Then with the start and end nucleotide found in the tags, the gene sequence is taken from the genome within the GenBank file. Post-processing of the records was performed using Python scripts developed in-house; these scripts remove any duplicate records (an artifact of duplicate gene annotations within the GenBank file) as well as create FASTA-formatted sequence files. The PyMongo library (https://pypi.python.org/pypi/pymongo/) was used to insert the data into the MongoDB. All scripts can be found online at https://github.com/LoyolaChicagoCode/hiv-biojava-scala.

HIV data collection. HIV-1 genomes were downloaded as GenBank files from the NCBI nucleotide database specifying the following: "Human immunodeficiency virus 1" (porgn:_txid11676) AND (8000:11000[Sequence Length]). This query collects all full-length and near full-length genomic sequences. Data were collected from the NCBI on February 26, 2013, obtaining 4,724 individual sequence records. Records missing country of isolation and/or collection date information, totaling 1,622 records, were manually curated via one of two sources. Records retrieved which are also available via Los Alamos National Laboratory (LANL)'s HIV database (http://www.hiv.lanl.gov/) were referenced to ascertain whether the isolation/date information was available. In the event that these data were also missing from LANL, publications referenced in the GenBank file were evaluated. The database was updated at a later date (September 25, 2015), further exemplifying the ease of use for the proposed method of data aggregation and exposure in the form of a RESTful web service. This update

expanded the sequence database to include an additional 1,342 sequences (6,066 total). Data are stored through figshare (figshare.com) and can be retrieved via wget http://files.figshare.com/2304758/hiv.tar.gz.

Phylogenetic analysis. Sequences retrieved from the HIV database were examined following one of two strategies. First, sequences were aligned via ClustalW, and neighbor-joining trees with partial deletion (site coverage cutoff, 75%) were computed with the maximum composite likelihood model using the MEGA 5 software tool[33]; trees were visualized using the tool Phylowidget[34] and produced using the Adobe Illustrator. The trees shown in Figures 3 and 4 were created using this strategy. In parallel, a second strategy was employed: sequences were aligned using MUSCLE and maximum likelihood trees using the Jukes–Cantor and generalized time-reversible models that were generated via FastTree[35] within the Geneious tool (Biomatters Ltd.). In deriving these trees, support values were computed. Trees were visualized using Geneious; Supplementary Files 1 and 2 contain the phylogenies (derived using the Jukes–Cantor model and with support values shown) for the same set of sequences as shown in Figure 3A and B, respectively. Newick format files can be found for all five trees (Figs. 3 and 4) derived using this second strategy with the Jukes–Cantor model in Supplementary File 3.

Acknowledgment

The authors would like to thank Mr. Yousef Aleneze for his preliminary work on the project.

Author Contributions

Conceived and designed the experiments: GT, CP. Contributed to the development of the code: SR, TH, KL, GT. Analyzed the data: SR, TH, CP. Wrote the first draft of the manuscript: SR, CP. Contributed to the writing of the manuscript: TH, KL, GT. Agreed with the manuscript results and conclusions: SR, TH, KL, GT, CP. All authors reviewed and approved the final manuscript.

Supplementary Materials

Supplementary File 1.
Supplementary File 2.
Supplementary File 3.

REFERENCES

1. Fierer N, Breitbart M, Nulton J, et al. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl Environ Microbiol*. 2007;73:7059–66.
2. Holmfeldt K, Solonenko N, Shah M, et al. Twelve previously unknown phage genera are ubiquitous in global oceans. *Proc Natl Acad Sci U S A*. 2013;110:12798–803.
3. Hurwitz BL, Sullivan MB. The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One*. 2013;8:e57355.
4. Deng L, Ignacio-Espinoza JC, Gregory AC, et al. Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature*. 2014;513:242–5.
5. Barzon L, Lavezzo E, Militello V, et al. Applications of next-generation sequencing technologies to diagnostic virology. *Int J Mol Sci*. 2011;12:7861–84.
6. Wang C, Mitsuya Y, Gharizadeh B, et al. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res*. 2007;17:1195–201.
7. Ramakrishnan MA, Tu ZJ, Singh S, et al. The feasibility of using high resolution genome sequencing of influenza A viruses to detect mixed infections and quasispecies. *PLoS One*. 2009;4:e7105.
8. Solmone M, Vincenti D, Prosperi MCF, et al. Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen. *J Virol*. 2009;83:1718–26.
9. Abdelrahman T, Hughes J, Main J, et al. Next generation sequencing sheds light on the natural history of hepatitis C infection in patients that fail treatment. *Hepatology*. 2015;61:88–97.
10. Verheyen J, Litau E, Sing T, et al. Compensatory mutations at the HIV cleavage sites p7/p1 and p1/p6-gag in therapy-naive and therapy-experienced patients. *Antivir Ther*. 2006;11:879–87.
11. Quiñones-Mateu ME, Avila S, Reyes-Teran G, et al. Deep sequencing: becoming a critical tool in clinical virology. *J Clin Virol*. 2014;61:9–19.
12. Lares MR, Rossi JJ, Ouellet DL. RNAi and small interfering RNAs in human disease therapeutic applications. *Trends Biotechnol*. 2010;28:570–9.
13. Truong NP, Gu W, Prasadam I, et al. An influenza virus-inspired polymer system for the timed release of siRNA. *Nat Commun*. 2013;4:1902.
14. Paul AM, Shi Y, Acharya D, et al. Delivery of anti-viral siRNA with gold-nanoparticles inhibits dengue virus infection in vitro. *J Gen Virol*. 2014;95:1712–22.
15. Jin F, Li S, Zheng K, et al. Silencing herpes simplex virus type 1 capsid protein encoding genes by siRNA: a promising antiviral therapeutic approach. *PLoS One*. 2014;9:e96623.
16. Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics*. 2009;25:1363–9.
17. Evani US, Challis D, Yu J, et al. Atlas2 Cloud: a framework for personal genome analysis in the cloud. *BMC Genomics*. 2012;13(suppl 6):S19.
18. Zhao S, Prenger K, Smith L, et al. Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing. *BMC Genomics*. 2013;14:425.
19. Schatz MC, Langmead B, Salzberg SL. Cloud computing and the DNA data race. *Nat Biotechnol*. 2010;28:691–3.
20. Borozan I, Wilson S, Blanchette P, et al. CaPSID: a bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. *BMC Bioinformatics*. 2012;13:206.
21. Hird SM. lociNGS: a lightweight alternative for assessing suitability of next-generation loci for evolutionary analysis. *PLoS One*. 2012;7:e46847.
22. Ningthoujam SS, Choudhury MD, Potsangbam KS, et al. NoSQL data model for semi-automatic integration of ethnomedicinal plant data from multiple sources. *Phytochem Anal*. 2014;25:495–507.
23. Fielding RT. *Architectural styles and the design of network-based software architectures* [PhD thesis]. Irvine: University of California, Information and Computer Science Department; 2000.
24. Gao F, Bailes E, Robertson DL, et al. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature*. 1999;397:436–41.
25. Gao F, Yue L, Robertson DL, et al. Genetic diversity of human immunodeficiency virus type 2: evidence for distinct sequence subtypes with differences in virus biology. *J Virol*. 1994;68:7433–47.
26. Rambaut A, Posada D, Crandall KA, et al. The causes and consequences of HIV evolution. *Nat Rev Genet*. 2004;5:52–61.
27. Brenner B, Wainberg MA, Roger M. Phylogenetic inferences on HIV-1 transmission: implications for the design of prevention and treatment interventions. *AIDS*. 2013;27:1045–57.
28. Shankarappa R, Margolick JB, Gange SJ, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol*. 1999;73:10489–502.
29. Castro-Nallar E, Pérez-Losada M, Burton GF, et al. The evolution of HIV: inferences using phylogenetics. *Mol Phylogenet Evol*. 2012;62:777–92.
30. Ahumada-Ruiz S, Flores-Figueroa D, Toala-González I, et al. Analysis of HIV-1 pol sequences from Panama: identification of phylogenetic clusters within subtype B and detection of antiretroviral drug resistance mutations. *Infect Genet Evol*. 2009;9:933–40.
31. Jung M, Leye N, Vidal N, et al. The origin and evolutionary history of HIV-1 subtype C in Senegal. *PLoS One*. 2012;7:e33579.
32. Abubakar YF, Meng Z, Zhang X, et al. Multiple independent introductions of HIV-1 CRF01_AE identified in China: what are the implications for prevention? *PLoS One*. 2013;8:e80487.
33. Tamura K, Peterson D, Peterson N, et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 2011;28:2731–9.
34. Jordan GE, Piel WH. PhyloWidget: web-based visualizations for the tree of life. *Bioinformatics*. 2008;24:1641–2.
35. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010;5:e9490.