

SOFTWARE

Open Access



# PoSE: visualization of patterns of sequence evolution using PAML and MATLAB

Kun Zhao<sup>1\*</sup>, Elizabeth Henderson<sup>1</sup>, Kelley Bullard<sup>2</sup>, M. Steven Oberste<sup>1</sup>, Cara C. Burns<sup>1</sup> and Jaume Jorba<sup>1</sup>

From the 6th Workshop on Computational Advances in Molecular Epidemiology (CAME 2017)  
Boston, MA, USA. 20 August 2017

## Abstract

**Background:** Determining patterns of nucleotide and amino acid substitution is the first step during sequence evolution analysis. However, it is not easy to visualize the different phylogenetic signatures imprinted in aligned nucleotide and amino acid sequences.

**Results:** Here we present PoSE (Pattern of Sequence Evolution), a reliable resource for unveiling the evolutionary history of sequence alignments and for graphically displaying their contents. Substitutions are displayed by category (transitions and transversions), codon position, and phenotypic effect (synonymous and nonsynonymous). Visualization is accomplished using MATLAB scripts wrapped around PAML (Phylogenetic Analysis by Maximum Likelihood), implemented in an easy-to-use graphical user interface. The application displays inferred substitutions estimated by *baseml* or *codeml*, two programs included in the PAML software package. PoSE organizes patterns of substitution in eleven plots, including estimated non-synonymous/synonymous ratios (dN/dS) along the sequence alignment. In addition, PoSE provides visualization and annotation of patterns of amino acid substitutions along groups of related sequences that can be graphically inspected in a phylogenetic tree window.

**Conclusions:** PoSE is a useful tool to help determine major patterns during sequence evolution of protein-coding sequences, hypervariable regions, or changes in dN/dS ratios. PoSE is publicly available at <https://github.com/CDCgov/PoSE>

**Keywords:** Molecular evolution, Bioinformatics, Phylogenetics, MATLAB, PAML

## Background

Most molecular evolution analysis depends on choosing a model of substitution; for example, to estimate genetic distances or infer a phylogenetic tree. This initial step relies on determining patterns of substitutions, which results in the quantitative analysis of the mutations found in an alignment. Although this may be a relatively quick computational step, understanding how substitutions accumulated and visualizing substitution patterns along the alignment provides a wealth of useful information about the dynamics of nucleotide and amino acid change. There are several approaches to track unique changes along a phylogenetic path. Here, we present work using ancestral

reconstruction as implemented in the software package Phylogenetic Analysis using Maximum Likelihood (PAML) [1] for visualizing evolution patterns. The main strengths of PAML lie in the rich repertoire of evolutionary models implemented, which are used to estimate parameters in models of sequence evolution or to test biological hypotheses. Inferred substitutions can be obtained as an optional output in *baseml* and *codeml* programs within PAML, which will generate an additional output file (*rst* file). This file contains the inferred unique substitutions along the phylogenetic tree and is not straightforward to comprehend, particularly for large data sets. In order to visualize the information stored in it, we used MATLAB for capturing, processing, and graphically displaying all changes inferred by PAML. We used this new resource to analyze sequence alignments of rapidly evolving RNA viruses. For example, we examined the pattern of nucleotide and amino acid substitutions to calibrate poliovirus molecular clocks

\* Correspondence: [kzhao@cdc.gov](mailto:kzhao@cdc.gov)

<sup>1</sup>Polio and Picornavirus Laboratory Branch, G-10, Division of Viral Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, 1600 Clifton Rd., N.E, Atlanta, GA 30329, USA  
Full list of author information is available at the end of the article



[2] and to define the evolutionary dynamics of circulating vaccine-derived poliovirus (cVDPV) emergences [3].

### Implementation

PoSE is an open source application package with an easy-to-follow graphical user interface (GUI) built in MATLAB. PoSE benefits from MATLAB's software environment and versatile language syntax for processing large data sets and rendering high-quality graphics. The script also benefits from MATLAB's extensive development for scientific computing, including applications in bioinformatics and genetic data streamlining ([http://www.mathworks.com/company/user\\_stories/centers-for-disease-control-and-prevention-automates-poliovirus-se-quencing-and-tracking.html?by=product](http://www.mathworks.com/company/user_stories/centers-for-disease-control-and-prevention-automates-poliovirus-se-quencing-and-tracking.html?by=product)). Its GUI was coded using procedural programming, which facilitates addition of future features to the script.

PoSE has over 8,000 lines of MATLAB source code stored in 5 folders and is optimized for MATLAB version 2015b and later versions. PoSE processes the out-file *rst* file generated by *baseml* or *codeml* and produces eleven graphical results and, in addition, interactively displays inferred nucleotide and amino acid substitutions along the phylogenetic tree. The compiled version of PoSE includes all necessary runtime libraries for execution independently from the MATLAB environment. Users do not need MATLAB in order to run PoSE. The compiled version runs in Windows and Mac (10.10–10.13) environments.

The input file for PoSE is the *rst* file generated after running *baseml* or *codeml* in PAML. This file captures the unique nucleotide and amino acid changes along a phylogenetic tree. PoSE requires *rst* files generated from protein-coding sequence alignments free of gaps and ambiguous bases. The user can refer to the PAML manual for addressing questions related to running *baseml* or *codeml* and for treatment of gaps and ambiguities before running PAML.

Each of the eleven plots can be printed or exported as an image in PDF format. In addition, all substitutions displayed in PoSE can be exported in an Excel spreadsheet that includes a Markov matrix of conditional probabilities of observing each type of nucleotide substitution [4]. After reading the *rst* file from *codeml*, PoSE annotates a phylogenetic tree by mapping all nucleotide and corresponding amino acid substitutions occurring in both external and internal branches of the tree. Displayed trees can be exported as an annotated Newick-format file for further inspection using specialized phylogenetic programs such as FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>).

### Results

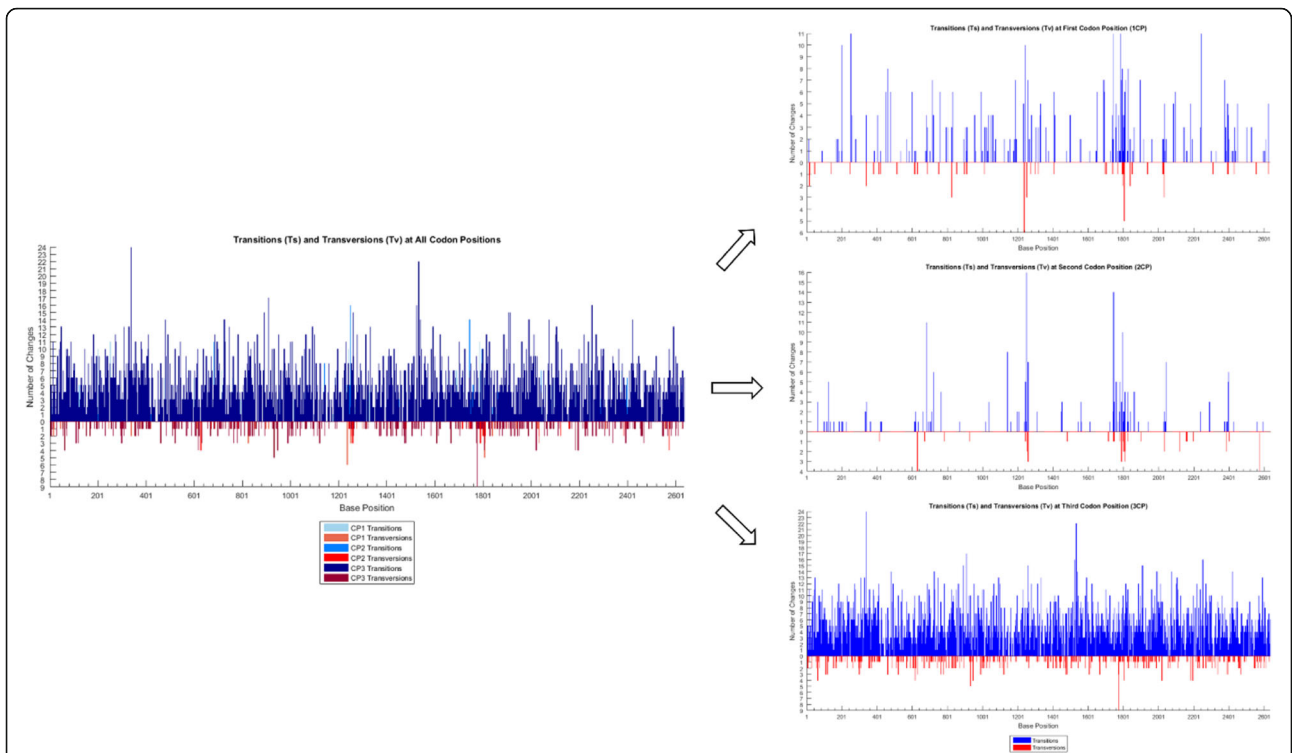
Visualizations include display of nucleotide and amino acid substitutions occurring along a user-defined sequence

interval (summary plots via *baseml*) and along the phylogenetic tree (via *codeml*). Transition (Ts: A↔G, C↔T) and transversion (Tv: A↔C, A↔T, G↔C, G↔T) substitutions are analyzed by codon position (Fig. 1) and then frequency plots summarize the overall accumulation of Ts and Tv and the accumulation of each substitution within transitions and transversions (Fig. 2).

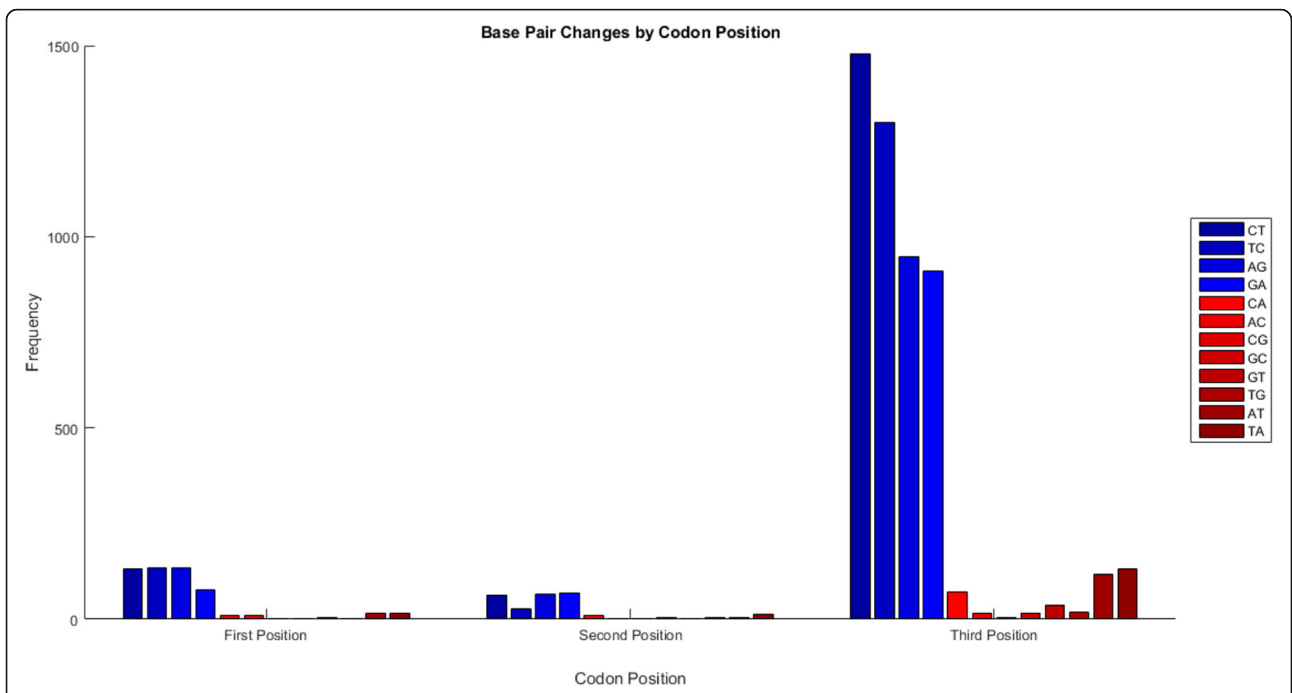
Phylogenetic signatures along the sequence interval are visualized by inspecting each type of substitution according to the phenotypic effect: 1) substitutions not leading to an amino acid change, synonymous transitions (As) and synonymous transversions (Bs), and 2) substitutions leading to an amino acid change, nonsynonymous transitions (Aa) and nonsynonymous transversions (Ba). Occurrence of these four signals can be graphically visualized at each site along the sequence (Fig. 3). PAML estimates quantities for As, Bs, Aa, and Ba according to the nucleotide evolution model set in the control file. PoSE extracts estimated substitutions and calculates the average of these estimations on user-determined sliding sequence window intervals at shifting step sizes. Total synonymous substitutions ( $dS = As + Bs$ ) and total nonsynonymous substitutions ( $dN = Aa + Ba$ ) are calculated for  $dN/dS$  ratio. The distribution pattern of As, Bs, Aa, Ba, and  $dN/dS$  ratio is summarized in subsequent plots (Fig. 4). Sequence windows and step sizes can be dynamically changed in order to refine the plots or explore different parameters. Likewise, the  $dN/dS$  ratio is plotted for identifying sequence regions under putative selection (Fig. 5). The last two plots displayed in PoSE show the cumulative number of As, Bs, Aa, Ba, and total number of substitutions (Kt) along the sequence region in user-specified step sizes.

Inferred nucleotide and amino acid substitutions along the path of the phylogenetic tree can be inspected by processing PAML's *codeml* results (Fig. 6). In order to visualize the evolutionary pattern calculated from PAML, the tree is annotated with inferred substitutions as synonymous and nonsynonymous substitutions are displayed separately. The program offers the option of highlighting the substitution types inferred at the tips of the tree. The interactive menu allows further tracking of the substitution pattern along the tree; for example, tracking synonymous transitions in particular branches or exploring nonsynonymous substitutions for particular amino acid changes within a branch.

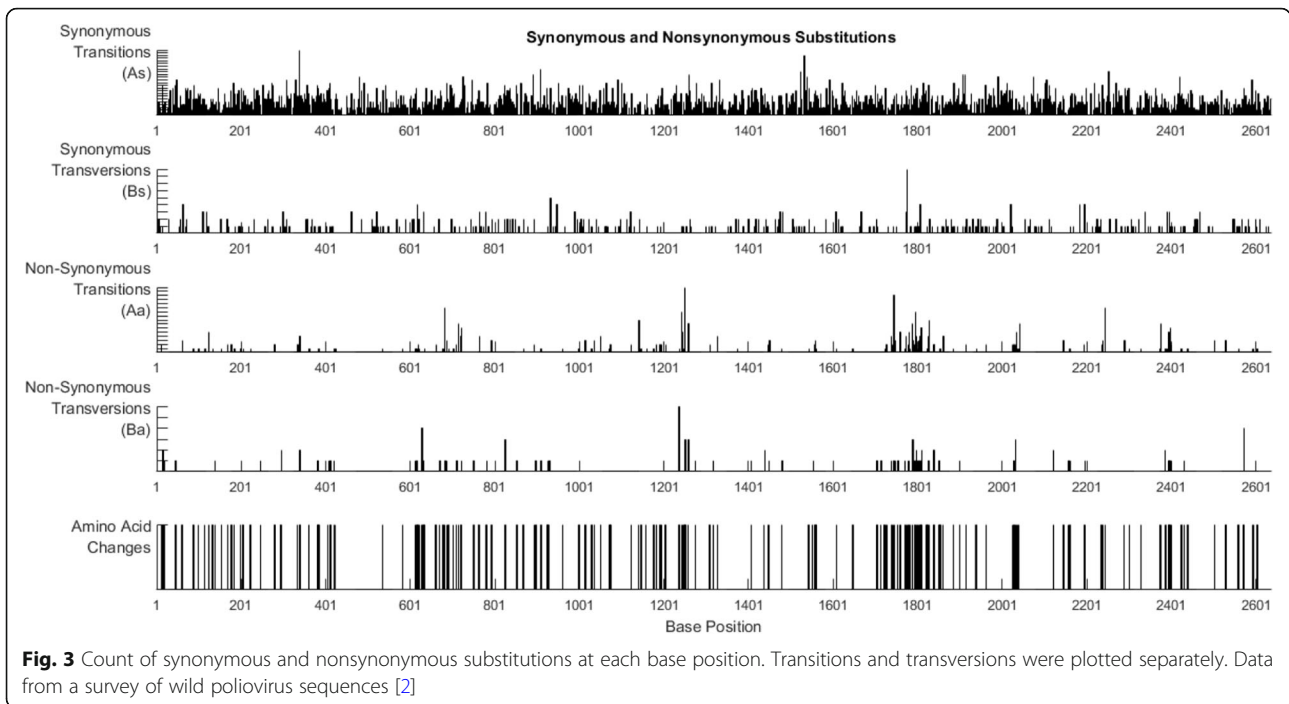
PoSE exports two annotated phylogenetic trees in Newick format; one with inferred amino acid changes and another with corresponding nonsynonymous nucleotide changes. In addition, PoSE generates two reports in Excel format documenting all the data displayed in the plots and in the phylogenetic tree, including a Markov matrix of relative frequencies of specific base changes.



**Fig. 1** Transition and transversion substitutions (y-axis) at each base position (x-axis) stratified to three codon positions. Data from a survey of wild poliovirus sequences [2]



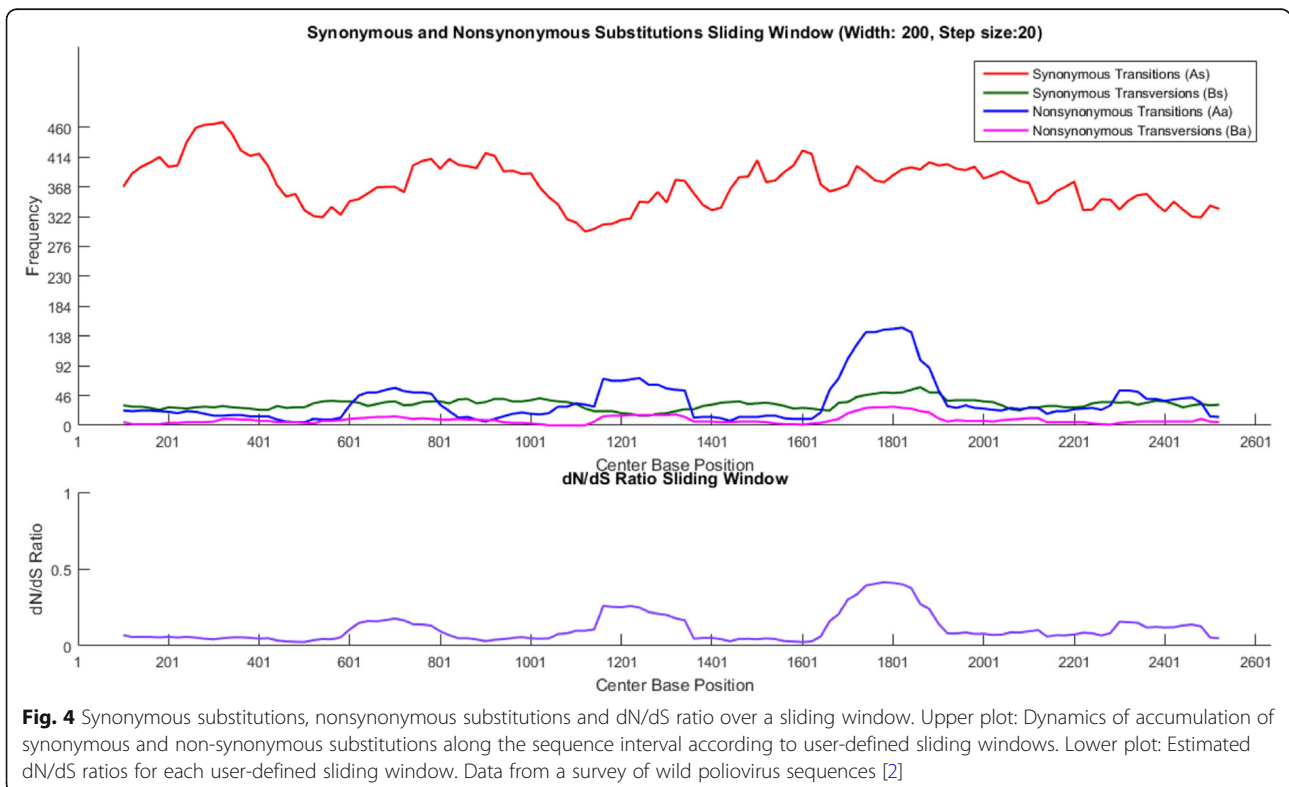
**Fig. 2** Distribution of all types of transitions and transversion substitutions at each codon position. Data from a survey of wild poliovirus sequences [2]

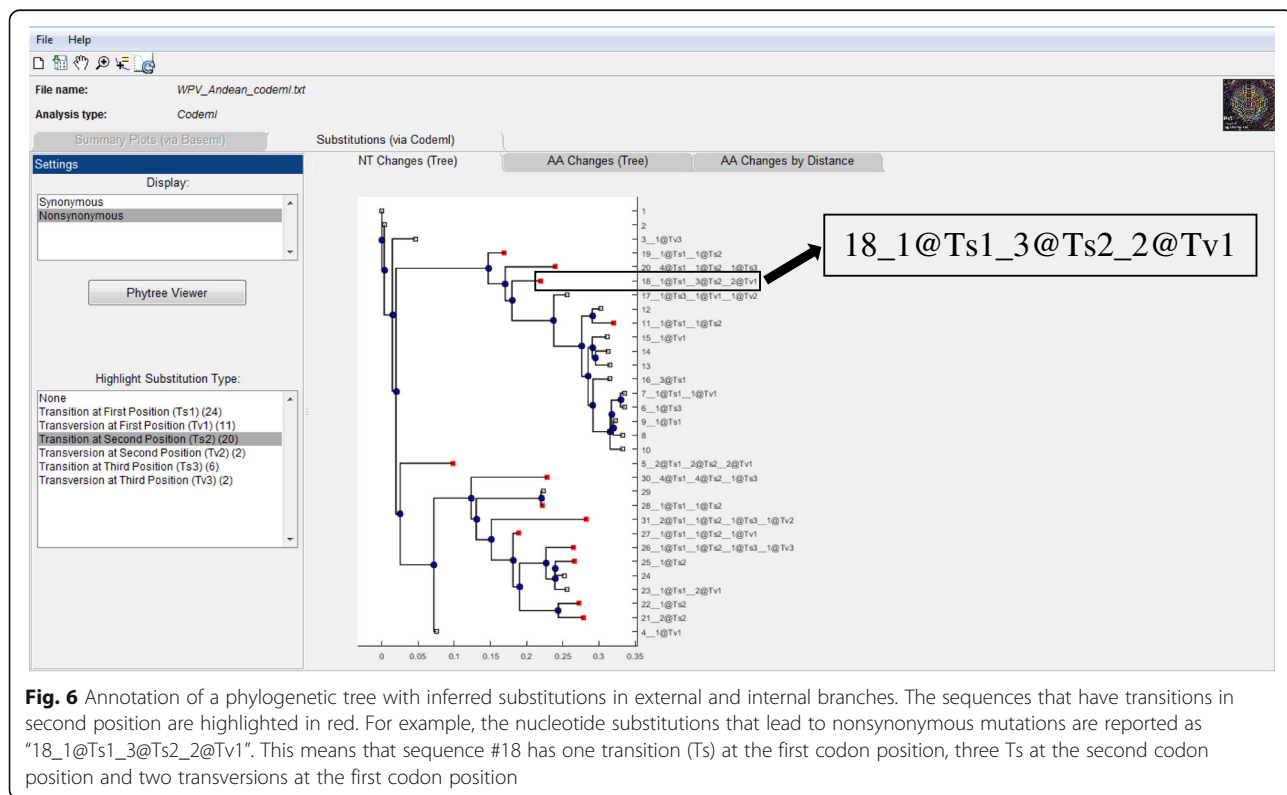
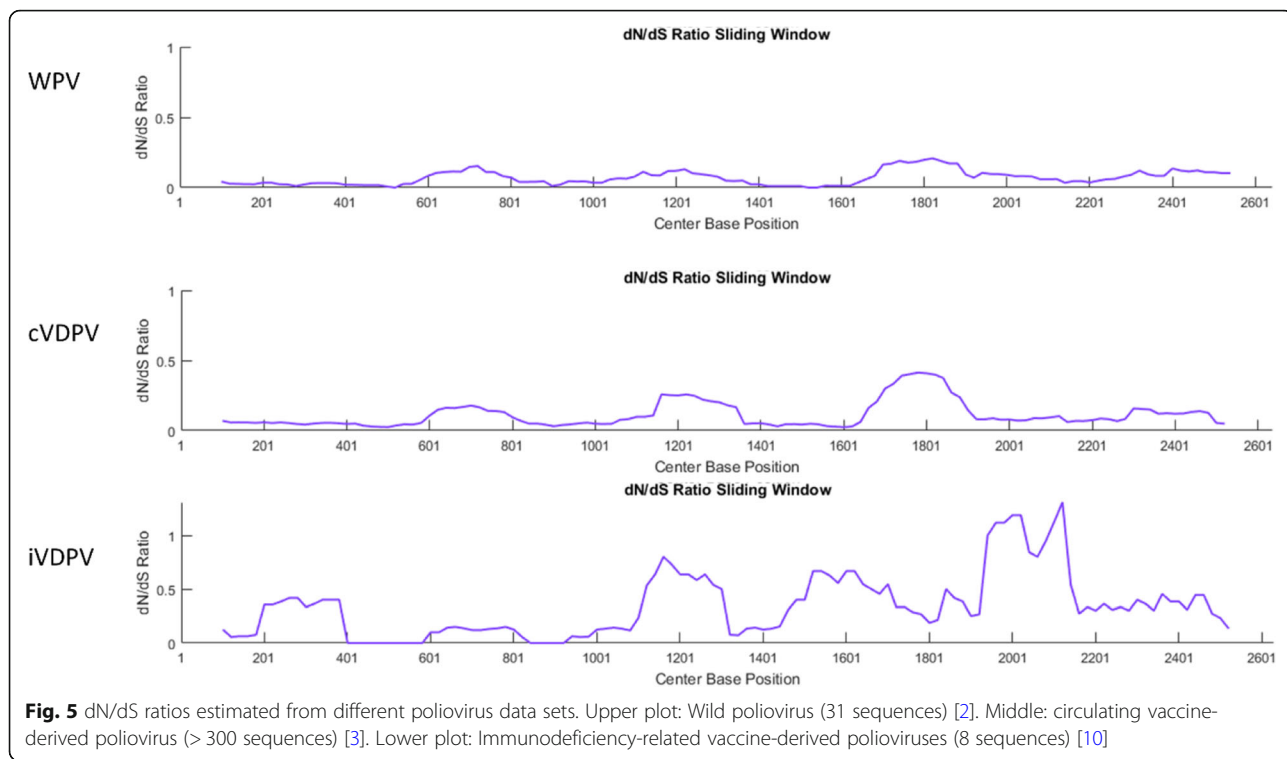


**Discussion**

PoSE is a new user-friendly MATLAB script which organizes and graphically displays data from results obtained in *baseml* and *codeml* included in the software package PAML. PAML can be run natively using the command-line

or using a GUI interface [5]. PoSE can quickly process large data sets (> 1,000 recorded substitutions). However, prior knowledge of the methods used in these two programs is highly recommended. For example, gaps or ambiguities in protein-coding sequence alignments may







produce out-of-frame results. Future versions of PoSE will incorporate a single pipeline from PAML to PoSE, including scanning for potential gaps, ambiguities, or misalignments.

There are numerous software packages and bioinformatics resources for estimating genetic distances and inferring phylogenetic trees from sequence data. However, little is known about the nature and dynamics of accumulation of mutations over a sequence region. PoSE provides visualization of the actual changes occurring in phylogenetically related homologous sequences. For example, inspection of transition and transversion changes per site provides information about patterns in the mode of evolution [6, 7]. Also, graphical visualization of current changes provides a higher level of granularity than that observed in sequence alignments; such as detection of hypervariable regions due to increased number of non-synonymous substitutions (Fig. 3).

The molecular clock model of evolution is of particular interest in studies related to rapidly evolving viruses. Inference from sequence data of the tempo and mode of virus transmission can be readily determined using well-known bioinformatics tools such as BEAST [8]. However, mutation saturation due to multiple substitutions per site can underestimate divergence dates [9]. By analyzing all unique nucleotide substitutions, PoSE scans for putative regions of sequence saturation and provides clues about potential substitutions involved in mutation saturation (Fig. 3).

## Conclusions

PoSE is an ongoing bioinformatics project aiming at analyzing patterns of sequence evolution in protein-coding sequence alignments. It was developed from studies of large data sets of poliovirus genomes. Compatible with the rapid nature of RNA virus evolution, PoSE facilitates processing and visualization of very large data sets containing thousands of inferred nucleotide and amino acid substitutions. PoSE complements inference of genetic distances and phylogenetic trees by contributing detailed information about the nature, distribution, and dynamics of mutations in easy-to-grasp graphical representations.

## Availability and requirements

**Project name:** PoSE.

**Project home page:** <https://github.com/CDCgov/PoSE>

**Operating system(s):** Windows and Mac (10.10–10.13).

**Programming language:** MATLAB.

**Other requirements:** MATLAB Runtime for free execution of PoSE.

**License:** GNU GPL.

**Any restrictions to use by non-academics:** none.

## Abbreviations

Aa: nonsynonymous transitions; As: synonymous transitions; Ba: nonsynonymous transversions; BEAST: Bayesian Evolutionary Analysis by

Sampling Trees; Bs: synonymous transversions; cVDPV: circulating vaccine-derived poliovirus; PAML: Phylogenetic Analysis using Maximum Likelihood; PoSE: Visualization of Patterns of Sequence Evolution

## Acknowledgments

The authors would like to thank Ms. Anita Gajjala (Mathworks) for her technical assistance on MATLAB coding. The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the Centers for Disease Control and Prevention. The use of trade names is for identification only and does not imply endorsement by the CDC or the U.S. government.

## Funding

All work and publication costs were funded by the Centers for Disease Control and Prevention.

## Availability of data and materials

PoSE installation files, example files and a user's manual about PoSE can be found at <https://github.com/CDCgov/PoSE>

## About this supplement

This article has been published as part of *BMC Bioinformatics Volume 19 Supplement 11, 2018: Proceedings from the 6th Workshop on Computational Advances in Molecular Epidemiology (CAME 2017)*. The full contents of the supplement are available online at <https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-11>.

## Authors' contributions

JJ conceived the software. KZ, EH, and KB designed algorithmic solutions and wrote the code. KZ and JJ wrote the manuscript. MSO, CCB, and JJ contributed to the experimental design and edited the manuscript. All authors have read and approved the manuscript.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Author details

<sup>1</sup>Polio and Picornavirus Laboratory Branch, G-10, Division of Viral Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, 1600 Clifton Rd., N.E, Atlanta, GA 30329, USA. <sup>2</sup>IHRC Inc., Atlanta, GA, USA.

Published: 22 October 2018

## References

1. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007;24(8):1586–91.
2. Jorba J, Campagnoli R, De L, Kew O. Calibration of multiple poliovirus molecular clocks covering an extended evolutionary range. *J Virol.* 2008; 82(9):4429–40.
3. Burns CC, Shaw J, Jorba J, Bukbuk D, Adu F, Gumedé N, Pate MA, Abanida EA, Gasasira A, Iber J, et al. Multiple independent emergences of type 2 vaccine-derived polioviruses during a large outbreak in northern Nigeria. *J Virol.* 2013;87(9):4907–22.
4. Allman ES, Rhodes JA. *Mathematical models in biology: an introduction*. New York: Cambridge University Press; 2004.
5. Xu B, Yang Z. PAMLX: a graphical user interface for PAML. *Mol Biol Evol.* 2013;30(12):2723–4.
6. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol.* 1993;10:512–26.

7. Zhao K, Jorba J, Shaw J, Iber J, Chen Q, Bullard K, Kew OM, Burns CC. Are circulating type 2 vaccine-derived polioviruses (VDPVs) genetically distinguishable from immunodeficiency-associated VDPVs? *Comput Struct Biotechnol J*. 2017;15:456–62.
8. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol*. 2012;29(8):1969–73.
9. Duchene S, Di Giallonardo F, Holmes EC. Substitution model adequacy and assessing the reliability of estimates of virus evolutionary rates and time scales. *Mol Biol Evol*. 2016;33(1):255–67.
10. Yang CF, Chen HY, Jorba J, Sun HC, Yang SJ, Lee HC, Huang YC, Lin TY, Chen PJ, Shimizu H, et al. Intratypic recombination among lineages of type 1 vaccine-derived poliovirus emerging during chronic infection of an immunodeficient patient. *J Virol*. 2005;79(20):12623–34.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

