**The “HistoneDB 2.0 - with Variants” is a phylogeny-based resource to analyze histones and their variants**

Eli J. Draizen1#, Alexey K. Shaytan1#, Leonardo Marino-Ramirez2, Paul B. Talbert3, David Landsman2\*, Anna R. Panchenko1\*

1Computational Biophysics Group, Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, MSC 6075, Bethesda, MD 20894-6075, USA, 2Bioinformatics of Chromatin Structure Group, Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, MSC 6075, Bethesda, MD 20894-6075, USA, 3Howard Hughes Medical Institute, Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA.

# - authors contributed equally to this work

Corresponding authors:

David Landsman ([landsman@ncbi.nlm.nih.gov](mailto:landsman@ncbi.nlm.nih.gov)), Anna Panchenko ([panch@ncbi.nlm.nih.gov](mailto:panch@ncbi.nlm.nih.gov))

**Abstract**

Database URL: <http://www.ncbi.nlm.nih.gov/projects/HistoneDB>

**Introduction**

Nucleosomes constitute the elementary building blocks of chromatin and play important functional roles in epigenetic regulation of transcription, replication and reprogramming. Each nucleosome is composed of a nucleosome ‘core’ and linker DNA; if a linker histone is present, it is referred to as ‘chromatosome’. The structure of the nucleosome core is very conserved from throughout eukaryotes irrespective of histone sequence variants, mutations and post-translational modifications. It includes a 147 bp segment of DNA wrapped around an octamer composed of two copies each of four core histone proteins (H3, H4, H2A, H2B). All four types of histones share the same histone fold while the sequence identity between them might not exceed 25%. A linker histone H1 has a different fold and makes a unique set of interactions with the linker DNA. As a result, the chromatin fiber is compacted through the interaction of a linker histone with the linker DNA to form higher order chromatin structures.

In most eukaryotes, there are paralogs, or ‘histone variants’, for some or all of the core histone proteins (H2A, H2B, H3, H4, H1) Nucleosomes may employ different sets of histone variants, and post-translational modifications, which may be essential to perform their diverse functions and respond to environmental stimuli. Recent data have been accumulating about the functionally diverse histone variants. Some of the most striking examples include H2A.Z histone acetylation and deposition in memory formation, and modulation of olfactory neurons life span by histone variant H2B.E.

Classification of histone paralogs is a daunting task. They are usually subdivided into ‘canonical’ replication-dependent histones that are expressed during the S-phase of cell cycle and replication-independent histone ‘variants’, constitutively expressed during the cell cycle. This division is based on the history of their discovery, however, and has its limitations. ‘Canonical’ histones in plants and animals usually encompass a variety of structurally distinct paralogs of H2A and H2B, while in many unicellular organisms, there is no special set of ‘canonical’ replication-coupled paralogs, and ‘variants’ fulfill their role, rendering the distinction between these classes meaningless in these organisms.In animals, genes encoding canonical histones are typically clustered along the chromosome, lack introns, and employ a specific type of regulation at the RNA level with a stem loop structure at the 3’ end instead of polyA tail. On the other hand, genes encoding histone variants are usually not clustered, have introns, and their mRNAs are regulated with polyA tails similar to the mRNAs of most genes. In plants, canonical histone genes lack introns, but they are not clustered and the mRNAs are polyadenylated. Remarkably, more complex multicellular organisms typically have a higher number of histone variants providing a variety of different functions.

Each histone variant has characteristic sequence and structural features which account for its specific function. The similarity between canonical histones and other histone variants can be very substantial with only very few amino acid differences and with overall conservation of most structural features. However, in some cases histone variants differ from canonical counterparts in sequence as much as different families of canonical histones differ from each other (~25% identity). Variants may have shortened or extended secondary structures of the histone fold (for four types of variants), and may have characteristic regions with physico-chemical properties drastically different from the canonical histones. Many of these features and their functional implications are largely unknown or poorly annotated. The phylogenetic origin of histone paralogs has been addressed in several studies which pointed to a monophyletic origin for some variants while others, including canonical histones, were found to originate repeatedly in evolution. Although some histone variants can have unique post-translational modification patterns, the majority of them remain to be characterized for these modifications.

In this paper we present a database HistoneDB2.0 that collects canonical histones and histone variants, their sequences, and structural and functional features. This database is the successor to a previous Histone Database which represented a curated collection of sequences and structures of histones and non-histone proteins containing histone folds (1). The HistoneDB2.0 consists of two parts. First, we compile a manually curated set of histone variants and their multiple alignments with the annotated characteristic features and descriptions of their function. Second, we construct profile Hidden Markov Models (HMM) based on these alignments and use them to search any sequence of interest or all sequences from the non-redundant (NR) database, given that they pass rigorous criteria established in this study. As a result, automatic annotations are produced. Moreover, HistoneDB2.0 allows comparison of variants to each other or to match the canonical histones with histone variants for any given organism. The phylogenetic tree of histone variants presents another important evolutionary aspect of the database so that it is feasible to browse through the lineage-specific or universally conserved variants and decipher their characteristic features. The database promotes the new nomenclature for histone variants proposed recently.

**Database construction**

For each histone type, H2A, H2B, H3, H4, and H1, we collected sequences from the previous manual classification described in …(Ref) (2) (so called “curated sequences” set). These sequences were aligned using MUSCLE program (3) and alignments were further checked manually to make sure they had a wide taxonomic span and did not contain insertions or deletions in the core histone fold regions (“curated alignments” set). Alignments were used to train Hidden Markov Models, using HMMER 3.1b2 (4), creating one HMM for each variant. Next, all of the variant models were combined into one file and pressed using HMMER 3.1b2 hmmpress. Curated sequences and alignments were manually annotated with respect to the location of the structural, sequence and functional features characteristics for a given histone variant. These features were extracted from the literature, were inferred from the analysis of variant nucleosome structures and were obtained by using the automatic software.

**Database content: curated sequence and alignments**

HistoneDB contains annotations for 22 histone variants for the five histone families H2A, H2B, H3 and H1. Note that almost no characterized variants are available for H4 histone. In addition, HistoneDB has models for all four types of canonical histones. Below we briefly describe different histone families and the main features of their variants.

***H2A family***

Histone H2A has the highest number of known variants (six models in the database), some of which are relatively well characterized:

* H2A.X is the most common H2A variant, with the defining sequence motif ‘SQ(E/D)Y’, becomes phosphorylated during the DNA damage response, chromatin remodeling, and X chromosome inactivation in somatic cells. H2A.X and canonical H2A havediverged several times in phylogenetic history, but each H2A.X version is characterized by similar structure and function, suggesting it may represent the ancestral state.
* H2A.Z regulates transcription, DNA repair, suppression of antisense RNA, and Polymerase II recruitment. Notable features of H2A.Z include a large hydrophobic patch, a sequence motif ‘DEELD,’ a one amino acid insertion in loop1,and a one amino acid deletion in the docking domain. Isoform H2A.Z.2 was shown to be driving the progression of malignant melanoma (26051178). (5)
* macroH2A contains a histone fold domain and an extra C-terminal macro domain which can bind ADP. This histone variant is used in X-inactivation and transcriptional regulation. Structures of both domains are available, but the inter-domain linker is too flexible to be crystalized.
* H2A.B is a rapidly evolving Barr body deficient variant, known for its involvement in spermiogenesis. H2A.B has a shortened docking domain, which wraps a shorter DNA region. It is closely related to H2A.L and to H2A.M, a recently discovered mammalian-specific variantthat binds to huntingtin protein M.

Other less extensively studied H2A variants include H2A.J, which is very similar to canonical H2A, testis-specific TS H2A.1 and H2A.Q. There are also species-specific canonical variants, such as human H2A.1 through H2A.10.

***H2B family***

The H2B variants in mammals include testis-specific H2B.1, H2B.W, subH2B, and newly characterized variant H2B.E. (6)

* H2B.W is involved in spermiogenesis, telomere associated functions in sperm and is found in spermatogenic cells. It is characterized by the extension of the N-terminal tail.
* subH2B participates in regulation of spermiogenesis and is found in non-nucleosomal particle in the subacrosome of spermatozoa. This variant has a bipartite nuclear localization signal.
* Recently discovered variant H2B.E is involved in regulation of olfactory neuron function in mice. It is very similar to the canonical H2B.
* As for H2A, canonical H2Bs include a variety of distinct protein sequences

**H3 family**

* cenH3, or centromeric H3, replaces canonical H3 in centromeric nucleosomes and is essential for kinetochore formation in most eukaryotes. . cenH3s typically have only about 50-60% amino acid identity to canonical H3 in the histone fold domain and no conservation of the N-terminal. cenH3s have an extended loop1 and usually replace Phe84 in canonical H3 with Trp, and Thr 107 with Ala, Cys, or Ser. (7) The structure of cenH3 nucleosomes has been controversial and differs between species. In budding yeast cenH3 nucleosomes have a single H4 molecule and appear to be hemisomes with one molecule each of cenH3, H4, H2A and H2B, whereas fission yeast cenH3 nucleosomes have two H4 molecules and may be similar to octameric canonical nucleosomes. H3.3 refers to replication-independent H3s, which typically differ from canonical H3s by only a few amino acids that are necessary for replication-independent assembly. H3.3 and canonical H3s diverged independently in plants, animals and ciliates. Fungi typically have H3s that undergo both replication-coupled and replication-independent assembly, similar to H3.3s, which may represent the ancestral state. s
* Other H3 variants are lineage-specific, and often include germ cell or pollen variants

**H1 family**

H1 variants lack histone folds and typically have a short basic amino-terminal domain, a globular winged-helix domain and a lysine-rich carboxy-terminal domain. They evolve faster than the core histone families and are abundant and diverse. Many seem to have redundant functions, while some show germline specificity. (8)

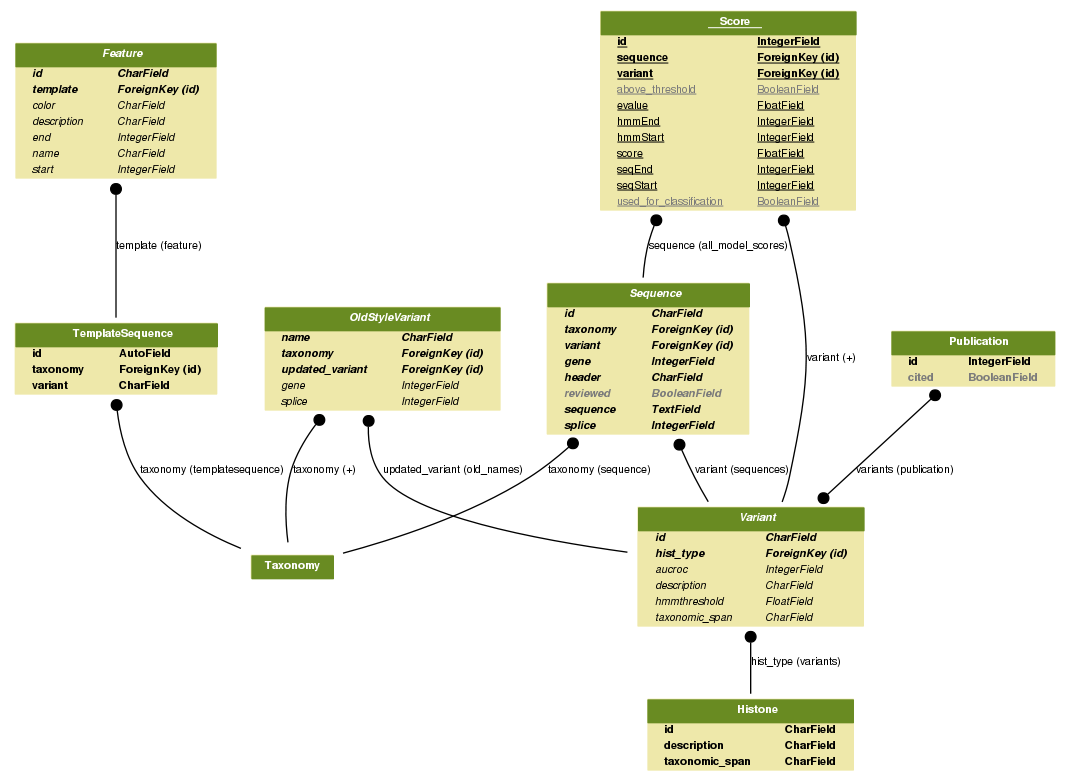
**Automatic annotation of sequences**

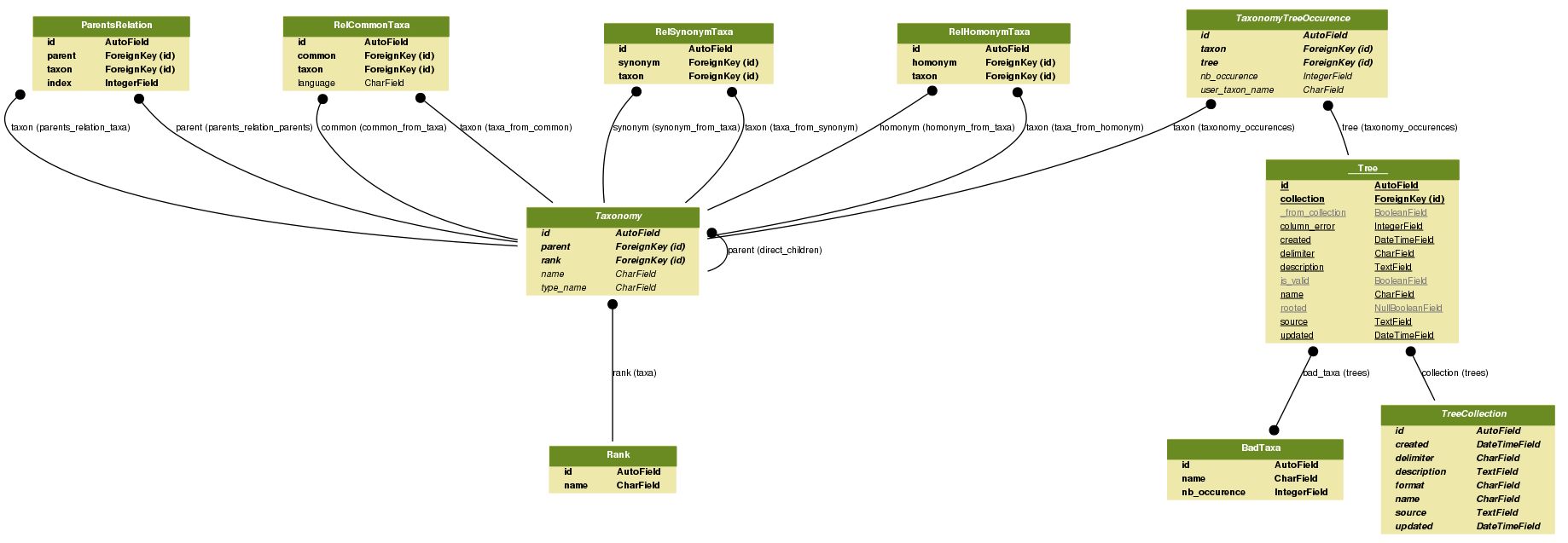
All sequences from the non-redundant (nr) database have been classified by our models (variants and canonical) and were added into the HistoneDB as “automatically extracted sequences” set. We assigned a model with a maximum HMMer score for a given sequence, in addition this score was required to cross a certain threshold identified as follows. For any given model, curated sequences that were used to construct its curated alignment were treated as positives and curated sequences of other models as negatives. Then we performed HMMsearch of the database of all models against a set of curated sequences and calculated the specificity of retrieval of each sequence by its native model. Specificity of retrieval was estimated based on the number of true positives (TPs; sequences correctly predicted by their native model) and false positives (FPs; incorrectly predicted sequences) found above each score HMMER score cutoff. The specificity was calculated as TN/(FP+TN). Then we interpolated an inverse curve of score cutoffs versus specificity, and obtained the score cutoff value corresponding to the specificity of 0.95. Note that different models had different cutoff scores (see Supplematary Information for details about evaluating each variant model). As can be seen on Figure X, 95% specifity threshold also results in a very high sensitivity where X% of all positive case are classified correctly.

This classification algorithm was applied to classify any sequence of interest. A user can enter a sequence in FASTA format to classify through the “Analyze Your Sequence” tab. A sequence is classified using HMMER against all variant models and by running BLAST against all sequences in the HistoneDB . The results are shown first with the most likely variant or variants, a comparison between the analyzed sequence and the canonical of the similar species determined by BLAST along with the new feature annotations. HMMER scores against all models are listed in modal window.

**Methods and Software**

The HistoneDB 2.0 is written in Django, a high-level Python Web Framework, with a MySQL backend. The project has two applications, ‘browse,’ and ‘djangophylcore.’ Browse contains the HistoneDB models (equivalent to database tables), views (python functions to render each page), and templates (HTML files). Djangophylocore is a previously developed django application to store the NCBI taxonomy database in a Django relational database using an algorithm similar to Modified Preorder Tree Traversal (9). The website layout is based on Twitter Bootstrap, with four important pages: Main browse of all histone types, Individual histone type browse, Variant browse, Analyze, and Search.





**Browsing the variants**

The front page allows choosing one of the five histone types, by selecting a color coded 3D model for each variant. After a histone core type is chosen, a summary is displayed that contains a list of variants with their alternate names, taxonomic spans, and sequence counts in curated and automatically extracted sequence sets. One can also see a phylogenetic tree constructed using jsPhyloSVG software (10), which shows if variants have mono- or polyphyletic origins and analyzes the relationship between different variants of the same histone type. The tree is created by aligning all curated sequences for a given histone type using MUSCLE v3.8.31 (3) and by further applying the Neighbor-Joining procedure implemented in CLUSTALW 2.1 (11). The trees were converted to PhyloXML using BioPython and were edited to add colors and variant and taxonomy labels in jsPhyloSVG.

The next tab displays a set of curated sequences with their corresponding identifiers, sequence descriptions and taxonomic classification. By clicking on the row of this table, a user can select a variant sequence to be aligned to the canonical sequence of the same histone type. This allows highlight the variant specific features. If one or more sequences are selected, there are options to view a multiple sequence alignment, export it in FASTA format or add to a basket to combine sequences from different variants.

“Curated alignments” tab depicts a multiple sequence alignment of all curated sequences for the selected histone type. Alignments are annotated with the key sequence, structural and functional features, such as secondary structures, key arginines, acidic patch and features specific for a given variant. Alignments are displayed using BioJS MSA Viewer (12).

The fourth tab contains a list of all sequences automatically extracted using HHMER (see previous section) assigned to a given variant type. For each sequence, GenBank identifier, variant’s name, taxonomy, a current GenBank description, HMMer Bitscore and E-value are provided. If one or more sequences are selected, there are options to view a multiple sequence alignment or add them to a basket. For more advanced users, there is an option view the scores for selected sequences against all HMM models (“Advanced”-“Score against all HMMs”). This option might be useful for examining if a sequence is similar to several models or if there was a classification error. This can be illustrated using sequence GI 121989 from Drosophila melanogaster as an example. It is classified as H2A.Z, but contains the motif, which is common in H2A.X, and therefore exhibits the features of both variants. Further data would be needed to correctly classify this sequence. It is also possible to filter the sequences by taxonomy, GenBank headers, sequence motifs, GI number, and sequences found in RefSeq.

The last tab “Inferred Taxonomy Distribution” allows browsing the taxonomy of automatically extracted sequences and view the corresponding classification of a sequence within the selected taxa with respect to all database models. The score represents the average HMM score of sequences from a given taxa. Scores are calculated for rank 'order,' the leaves of this tree, and it's parents scores/colors are extrapolated by averaging the colors of its children.

**Searching the database**

There are two types of search, basic search and advanced search. The basic filter allows to search based on the key words from the taxonomic classification while an advanced filter provides options to search for specific histone types with a particular sequence motifs and to show only unique sequences from a given organism.

**Conclusion**

Despite considerable progress in sequencing and understanding the functions of histone variants, many of them remain poorly annotated. Moreover, the specific molecular mechanisms of variants’ action, deposition and eviction are still unknown. One of the reasons is that the histone variants are highly specific and at the same time multi-functional and context dependent. The importance of studying the histone variants is difficult to overestimate; they are involved in regulation of many cellular processes and represent emerging key players in cancer (26051178). To analyze all known histone variants, compare them and interpret their functions would be challenging and timely. To fulfill this goal we designed the HistoneDB2.0: to organize histones by variant; to provide reference alignments for each variant; to offer curated annotations of variant specific features; to understand how variants evolved; to find orthologs of variants in other species; and finally, to promote the new histone variant nomenclature. This database allows finding and comparing variant sequences from different organisms or between histone variants and corresponding canonical histones. This in turn will help to understand the origin of functional specificity of variants and to guide the 3D modelling of variant nucleosomes. The database can be easily extended to include new variants and annotate them based on the similarity to the existing annotated variants.

**Acknowledgements**

This work was supported by the Intramural Research Program of the National Library of Medicine, NIH. AS was supported in part by the US-Russia Collaboration in the Biomedical Sciences NIH visiting fellows program. We would like to thank Franco Simonetti for discussions about Django and the IT support team at NCBI for the troubleshooting the server related issues.

Fig 1. The Nucleosome core particle. H2A is yellow, H2B is red, H3 is blue, and H4 is green. H1 is not shown. Figure created by \ref{Shaytan2015}

Figures: slide#4, slide#6, #7, phylogenetic tree, ROC curve, database schema

**References**

1. Mariño-Ramírez,L., Levine,K.M., Morales,M., Zhang,S., Moreland,R.T., Baxevanis,A.D. and Landsman,D. (2011) The Histone Database: an integrated resource for histones and histone fold-containing proteins. *Database (Oxford).*, **2011**, bar048.

2. Talbert,P.B., Ahmad,K., Almouzni,G., Ausió,J., Berger,F., Bhalla,P.L., Bonner,W.M., Cande,W.Z., Chadwick,B.P., Chan,S.W.L., *et al.* (2012) A unified phylogeny-based nomenclature for histone variants. *Epigenetics Chromatin*, **5**, 7.

3. Edgar,R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

4. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.

5. Eirín-López,J.M., González-Romero,R., Dryhurst,D., Ishibashi,T. and Ausió,J. (2009) The evolutionary differentiation of two histone H2A.Z variants in chordates (H2A.Z-1 and H2A.Z-2) is mediated by a stepwise mutation process that affects three amino acid residues. *BMC Evol. Biol.*, **9**, 31.

6. Shaytan,A.K., Landsman,D. and Panchenko,A.R. (2015) Nucleosome adaptability conferred by sequence and structural variations in histone H2A-H2B dimers. *Curr. Opin. Struct. Biol.*, **32C**, 48–57.

7. Postberg,J., Forcob,S., Chang,W.-J. and Lipps,H.J. (2010) The evolutionary history of histone H3 suggests a deep eukaryotic root of chromatin modifying mechanisms. *BMC Evol. Biol.*, **10**, 259.

8. Thatcher,T.H. and Gorovsky,M.A. (1994) Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. *Nucleic Acids Res.*, **22**, 174–179.

9. Ranwez,V., Clairon,N., Delsuc,F., Pourali,S., Auberval,N., Diser,S. and Berry,V. (2009) PhyloExplorer: a web server to validate, explore and query phylogenetic trees. *BMC Evol. Biol.*, **9**, 108.

10. Smits,S. a. and Ouverney,C.C. (2010) jsPhyloSVG: A javascript library for visualizing interactive and vector-based phylogenetic trees on the web. *PLoS One*, **5**, 6–9.

11. Larkin,M. a., Blackshields,G., Brown,N.P., Chenna,R., Mcgettigan,P. a., McWilliam,H., Valentin,F., Wallace,I.M., Wilm, a., Lopez,R., *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

12. Gómez,J., García,L.J., Salazar,G. a., Villaveces,J., Gore,S., García,A., Martín,M.J., Launay,G., Alcántara,R., Del-Toro,N., *et al.* (2013) BioJS: An open source JavaScript framework for biological data visualization. *Bioinformatics*, **29**, 1103–1104.