Nematode histone H2A variant evolution reveals diverse histories of retention and loss, and evidence for conserved core-like variants

Swadha Singh[1], Diana Chu[2*], Scott Roy[1,2*]
[1]Quantitative & Systems Biology, University of California, Merced, California, USA.
[2]Department of Biology, San Francisco State University, California, USA.
[*]Corresponding authors: Emails; sroy5@ucmerced.edu; chud@sfsu.edu

ABSTRACT

Though histone variants commonly replace canonical histones, how histone variants arise and evolve is poorly understood. Reconstruction of histone protein evolution is challenging due to high amino acid conservation and large differences in evolutionary rates across gene lineages and sites. Here we combined amino acid sequences and intron position data from 112 nematode genomes to trace the evolutionary histories of the three H2A variants found in *Caenohabditis elegans*: the ancient H2A.Z^{HTZ-1}, the sperm-specific HTAS-1, and HIS-35, which differs by a single glycine-to-alanine C-terminal change. We find disparate evolutionary histories. We find that though H2A.Z^{HTZ-1} is highly conserved, it exhibits recurrent intron loss. This pattern suggests that it is intron presence, rather than position, that may be important to H2A.Z functionality. In contrast, for HTAS-1 and HIS-35, we find that the intron position is conserved across species. HIS-35 arose in the ancestor of *Caenorhabditis* and its sister group, including the genus *Diploscapter*, while the sperm-specific variant HTAS-1 arose more recently in the ancestor of a subset of *Caenorhabditis* species. Both HIS-35 and HTAS-1 exhibit retention in some descendent lineages but also recurrent loss in others, suggesting that histone variant use or functionality is highly flexible. We also find that the single amino acid alanine in HIS-35 that distinguishes it from the glycine found in canonical H2A is ancestral and common across canonical *Caenorhabditis* H2A sequences. Further, we identify one nematode species that bear identical HIS-35 and canonical H2A proteins, suggesting they play redundant roles. We speculate that HIS-35 allows for H2A expression outside

of the S-phase; genes encoding such partially-redundant functions may be functionally important yet relatively replaceable over evolutionary times, consistent with the patchwork pattern of retention and loss of both genes. Our study shows the trajectory for histone variant evolution for distinct functions across species and the utility of intron positions for reconstructing the evolutionary history of gene families, particularly those undergoing idiosyncratic sequence evolution.

INTRODUCTION

All characterized eukaryotic cells compact their DNA into nucleosomes, which consist of an octameric histone complex comprised of two copies each of the core histone proteins- H2A, H2B, H3, and H4, that wrap around ~147 bps of DNA (1-4). Genes encoding canonical core histone (also referred to as 'replication-dependent histones') do not contain introns and are organized into multiple gene copies and in a tandem-repeat structure, which facilitates rapid, coordinated expression of these genes (4-7). The expression of canonical histones is tightly coupled with the S-phase of the cell cycle because of the critical need for the large bulk of histone proteins required to package and compact the newly synthesized DNA Histone variants and epigenetics (8-10).

Histones have evolved variant forms that further regulate chromatin compaction and affect processes like transcription, DNA repair, and development (11-20). Such histone variants are often considered to be a part of the 'histone code' and control distinct sets of genes during specific times and tissues in both animals and plants (11, 12, 21-23). Among all the core histones, H2A is the fastest evolving histone, showing the most diversity in the variants types and expressions (24, 25). For example, the highly conserved H2A.Z is present in every eukaryotic species and ubiquitously expressed. It has been shown in different species to play a variety of roles including transcriptional

regulation, heterochromatin boundaries, DNA repair, DNA replication, and dosage compensation (26-31). In contrast, H2A.X has arisen independently in several lineages, with functions in DNA damage response and transcription (31-33), while macroH2A, which is involved in X inactivation and stress response, is present only in some lineages, making its origins unclear (3, 4, 24, 34-37). Finally, short histone variants, such as H2A.B, H2A.P, H2A.L, and H2A.Q, are expressed in testis (except H2A.B which is also expressed in the brain cells) (25, 38-41). In short, a great wealth of histone H2A variants are observed, ranging from those shared across all eukaryotes to others that are species-specific (3, 4, 18, 19, 36). H2A variants also have distinct functions in a broad range of processes and can be ubiquitously expressed or only expressed only in certain tissues (11, 25, 42, 43). However, with the exception of the well-studied ancient variant H2A.Z, how histone variants have arisen and evolved to play distinct roles in specific species or tissues still remains unknown.

Clues to how histone variants differ may stem from the distinct gene structures and expression patterns compared to canonical histones. Unlike canonical histone genes, variant histone genes are not restricted to S phase expression but can be expressed throughout the cell cycle or in a tissue-specific manner. These histone variants are typically found in a single copy in the genome and contain introns in their pre-mRNAs (5-7, 44, 45). Notably, the roles of these introns in variant function remain poorly studied, including whether the introns themselves impart novel functionality. For instance, sequences found within introns can regulate gene expression (46-50). They can also allow for alternative splicing, which is common in animals and plays important role in development and disease (51-54). Introns have been used in other studies to determine the evolutionary trajectory of other protein families (55-62). Thus, the presence of introns in histone

variants may provide useful clues to exploring the histone evolution and function across closely related species.

Core histone functions are expected to be highly conserved across eukaryotes, given their central roles in ensuring DNA packaging and protection (4, 63). Thus, observed protein changes in core histones are expected to be largely neutral with respect to protein function. On the other hand, amino acid differences between variant histones and core histones are thought to generally lead to functional differences (3, 4, 64). Indeed, protein sequence differences between well-studied variants and their core homologs have been shown to affect chromatin structure and function. HIS-35 provides a particularly interesting example. On the other hand, the functional significance of protein differences between variant histones and their core counterparts in some cases becomes unclear. For instance, the protein sequence of *C. elegans* HIS-35 differs by just one amino acid from the S-phase *C. elegans* H2A, the functional significance of which is unknown. In this study, we focused on the evolutionary histories of the three H2A variants found in *Caenorhabditis elegans:* H2A.Z^{HTZ-1}, HIS-35, and HTAS-1. H2A.Z^{HTZ-1} is an ortholog of the evolutionarily conserved variant, H2A.Z. HIS-35 intriguingly differs from core H2A by a single 'A' instead of a 'G' at the 124th position. HTAS-1 shows a greater divergence from core H2A, particularly at the highly divergent C and N termini, which appears to be expressed only in sperm and has only been reported in *C. elegans* to date (65). The presence of a discrete number of H2A variants with distinct functions and expressions in combination with the availability of sequence data across a large number of nematode species allows a unique opportunity to track the evolutionary trajectory of this histone variant family.

RESULTS

Sequence-based phylogenetic methods fail to reconstruct the H2A gene family evolution

Taking advantage of vast sequencing efforts, we used BLAST searches across 122 reported nematode genomes to identify all annotated copies of H2A and H2A-related gene variants across the diversity of nematodes. After filtering and collapsing identical proteins, we were left with 355 unique sequences. We then used standard phylogenetic methods to reconstruct the evolutionary history of these sequences. However, scrutiny of the recovered phylogenetic tree revealed several bizarre findings (Supplemental Figure 1). For instance, core H2A proteins grouped as separate clades that included very deeply-diverged nematode sequences; on the other hand, species-specific variants were often found grouped far from core proteins from the same or related species. Perhaps the clearest case can be shown by performing phylogenetic reconstruction on core H2A orthologs and on the HIS-35 orthologs identified based on intron sequences (see below). Here, we expect distinct clades of H2A and HIS-35 sequences, yet we recover no such separate clades (Supplemental Figure X).

Some of these anomalies are as expected by errors in phylogenetic reconstruction due to model misspecification (the phenomenon in which differences between the assumed model of sequence evolution and the actual evolutionary process lead to errors in phylogenetic reconstruction (66). The possibility of model misspecification is increased in the case of histone genes by several factors, including extreme differences in rates across sites (some sites are conserved across other eukaryotes, other labile within *Caenorhabditis*); large relative differences in rates across gene lineages (generally extremely slow evolution of core proteins, but substantially more change in some variants); and the generally small number of total sites in short histone proteins. Thus, standard phylogenetic methods based on amino acid sequence alone failed to

provide an accurate representation of the evolutionary history of histone proteins within these 122 nematode species.

Intron position or phase distinguishes the three H2A variants of *C. elegans*.

As an alternative approach, we used another potential source of phylogenetic information: the position of the spliceosomal introns that disrupt nuclear genes, including variant histones. Intron positions can be conserved over very long times in orthologous genes (55, 57, 58, 60, 61). Whereas early work considered the possibility of intron 'sliding', in which an intron would migrate a few base pairs along a gene, recent work has shown that intron sliding is a very rare occurrence, and thus that intron positions are very often conserved over very long lineages. For example, in *Theileria apicomplexans*, 99.7% of intron positions are conserved between *T. parva* and *T. annulata*, diverging roughly 82 million years ago. In mammals, 99.9% of intron positions are conserved between humans and dogs, diverging around 100 million years ago (55-62).

We began our study of intron-exon structures in H2A variants by obtaining intron-exon structures for all H2A gene family members and performed alignments to determine intron position sharing across genes. We first aligned the three main *C. elegans* H2A variants (Figure 1), HIS-35, HTAS-1, and H2A.Z^{HTZ-1}, each of which has a single intron position (highlighted with a box in Figure 1). Scrutiny of the intron positions showed that the three genes contain introns at different positions, differing in the codon that they interrupt or and/or in the phase at which they disrupt the codon. Interestingly, the intron position in HIS-35 falls very near to that found in H2A.Z^{HTZ-1}. However, these introns are unlikely to represent a shared intron, for two reasons. First, intron positions rarely slide between phases (40). Second, HIS-35's near identity to H2A (see below) strongly suggests that it evolved from intronless H2A by gene duplication and intron gain, and not

from H2A.Z^{HTZ-1}. Thus, the differences in the intron positions between the three *C. elegans* variants suggest that the three variants evolved independently as different duplicates of core H2A, rather than from each other.

[As noted above, the global H2A phylogenetic tree evidences local problems in phylogeny reconstruction, requiring additional information. To leverage phylogenetic information from intron positions,]

Intron position and sequence evidence indicates the origin of HTAS-1 within *Caenorhabditis* and subsequent retention and loss

C. elegans sperm-specific H2A variant, HTAS-1, contains a single intron between codons 26 and 27. It is called a phase 0 intron since it is not interrupting any of the codons. Alignment across all the H2A variants of 122 nematodes revealed 16 genes that share an intron at the exact homologous position as *C. elegans* HTAS-1 (Figure 2). 16/16 of these genes are from species falling within a single subclade of 22 species within the *Caenorhabditis* genus (the clade represented by the red hash mark in Figure 2), suggesting an origin of this intron position within the common ancestor of these species (Figure 2). Scrutiny of the sequence gene tree (Supplementary Fig. 1) revealed that this same set of genes (i.e., those sharing the intron) appears together as a clade, reinforcing the notion that the genes sharing the HTAS-1 intron position represent a set of orthologs. We next sought to identify potential HTAS-1 orthologs in additional *Caenorhabditis* species, both the 6/22 within the HTAS-1-containing subclade that lack intron-containing HTAS-1 candidates as well as species outside this subclade. No genes from any such species were found grouping with the putative HTAS-1 ortholog clade, nor were any 'extra' variants observed in the genomes of these other species (Supplementary Fig. 1); thus, overall, we find no other genes that are candidates for

HTAS-1 orthologs. In total, then, the data is consistent with the origins of HTAS-1 and its gene-specific intron occurring in the same ancestor of a subset of *Caenorhabditis* species.

As noted above, we found several species within the HTAS-1-containing clade of *Caenorhabditis* that did not have a candidate HTAS-1 gene (Figure 2). *C. wallacei*, *C. brenneri* and *C.inopinata* (*sp34*), *C. kamaaina*, *C. panamensis* (*sp28*), *and C. japonica* did not contain annotated genes with either an intron position at the HTAS-1-specific position (a result that was confirmed by searching their genomes for potential genes that were missed in the annotation process) or other genes that grouped with the putative HTAS-1 orthologs (Supplementary Fig. 1). Thus, the data is consistent with a single origin of HTAS-1 and its characteristic intron position within the ancestor of a subset of studied *Caenorhabditis* species, with HTAS-1 having been lost multiple times in at least 6 independent lineages, leading to 16/26 species retaining HTAS-1 (Figure 2).

Intron position conservation suggests the origin of HIS-35 in the *Caenorhabditis-Diploscapter* ancestor and subsequent retention and loss

The variant HIS-35 of *C. elegans* has a phase zero intron which is placed between the 50th and the 51st codon. Alignment across all H2A variants revealed 20 genes that share this intron position (Figure 3, marked with a plus sign). These genes are from species falling in the clades of *Caenorhabditis* and its sister genus *Diploscapter*, suggesting an origin of this intron position within the common ancestor of these *Diploscapter* and *Caenorhabditis* (Figure 3). As with HTAS-1 above, consideration of the 20 putative HIS-35 containing species is consistent with a single origin of *HIS-35* followed by a loss in 5 independent lineages of *Caenorhabditis*.

HIS-35 provides a particularly interesting example for histone protein evolution. In *C. elegans*, the protein sequence of HIS-35 differs by just one amino acid from the S-phase H2A despite substantial evolutionary time, as shown above. Namely, HIS-35 has an "A", while H2A has a "G" at position 124 of the amino acid sequence. If an "A" at this position is an overriding change in the HIS-35 variant, then we expect this change to instigate a different function from the canonical H2A.

However, when we looked at position 124 the canonical H2A sequences of all the *Caenorhabditis* species we actually found out that the "A" is ancestral and highly conserved. This conservation of the "A" at position 124 suggests that HIS-35 likely has not diverged in function from the ancestral H2A. We also found that the predicted protein sequences of HIS-35 and H2A of species *C. kamaaina* are exactly the same.

We next sought evidence for concerted evolution between HIS-35 and core H2A. The multiple copies of core histone genes are known to undergo so-called concerted evolution, with sequences being transferred between paralogs by recombination (67, 68). We, therefore, wondered whether concerted evolution could explain the observation of identical protein sequence changes observed in the H2A and HIS-35 paralogs of some species. Under concerted evolution, the two sequences undergoing concerted evolution are homogenized (one overwrites the other). Consequently, the prediction is that such events should lead the interconverting partners to group on the phylogenetic tree. To search for evidence of concerted evolution, we reconstructed separate phylogenetic trees of exon-1 and exon-2 for all H2A and HIS-35 sequences. Most of the reconstructed tree largely reflected the species tree, suggesting against the possibility of widespread concerted evolution. However, we did observe the grouping of the two gene sequences for *Caenorhabditis sp21* (supplementary fig. 2 and 3), consistent with the concerted evolution of

HIS-35 and H2A in this species. Such occasional concerted evolution of these genes is consistent with a lack of functional differentiation, though the low rate of such events weakens the strength of this argument.

The dynamic history of intron loss and gain in HTZ-1

The ubiquitously expressed *C. elegans* H2A variant HTZ-1 is the ortholog of H2A.Z which is evolutionary conserved across all the eukaryotes (3, 69-71). *C. elegans* H2A.Z^{HTZ-1} has an intron that splits the 57th codon at position 2. Within the alignment across all H2A variants, we searched for genes that share the single intron position of *C. elegans* HTZ-1, revealing 30 genes share this position (Figure 4, marked with a plus sign; Supplemental Figure 1). These are putative HTZ-1 orthologs; consistent with this hypothesis, they grouped together on the tree. Unexpectedly, we found that 22 of these 30 putative H2A.Z genes have two (or more in *Diploscapter*) introns in their genes, one at position 57.2 and the other at position 111.1. Both introns have been repeatedly individually lost in different lineages (including in the lineage leading to the single-intron *C. elegans* H2A.Z^{HTZ-1} gene). Species including *C. elegans*, *C. tropicallis*, *C. sp32*, *C. afra*, *C. guadeloupensis*, *C. virilis* have lost the second H2A.Z intron which is at position 111.1, whereas *C. casteli* and *C. angaria* have lost their first intron. These results are consistent with general results in protein-coding genes, wherein intron loss is common across the *Caenorhabditis* phylogeny (72, 73). Nonetheless, the finding that general trends of intron loss may equally apply to histone variant genes is important in understanding the functional implications of variant introns.

Intron presence is a conspicuous difference between core and variant histones, raising the question of the functional significance of variant histone introns: specifically, are variant histone

introns important for differentiating the specific functions of those variants from their core paralogs? The finding of recurrent loss of introns from variant genes suggests that the specific positions or sequences of introns in histone variant genes may not be of particular importance. Nonetheless, it is of note that all observed H2A.Z orthologs contain at least one intron, suggesting that the presence of at least one intron, at whatever position, could be important for efficient expression of variant histone genes, consistent with their expression through canonical gene expression pathways, in which intron presence often promotes gene expression.

Discussion

Introns as sources of phylogenetic information

This study is among the first to leverage the information present in intron positions to decipher the evolutionary history of histone variants (62). Previous studies have shown intron position conservation among widely diverged eukaryotic species (55, 57, 59, 60, 62). For instance, intron positions are highly conserved between humans, mice, and fish (59). Thus, intron positions contain a record of evolutionary history that can facilitate insights into gene history. The utility of introns here is different for the different variants, and different questions addressed. The clearest case comes in HIS-35, in which nearly complete evolutionary conservation of H2A and HIS-35 sequences leads to a lack of phylogenetic signal within proteins. At the other end of the spectrum lies HTAS-1, which shows much more rapid sequence evolution. However, here, the discrepancy between rates between HTAS-1 and H2A, and likely between different sites in the two proteins (i.e., the N and C termini are highly constrained in H2A but fast-evolving for HTAS-1) makes it impossible to define a single evolutionary model across the gene family. As expected by general long branch attraction consideration, this leads to fast-evolving HTAS-1 incorrectly grouping far

away from *Caenorhabditis* core H2A sequences. This generally undermines our confidence in sequence-based phylogenetic reconstruction of HTAS-1 genes. Ironically, intron-defined HTAS-1 orthologs do group as a clade, indicating that sequence-based phylogenetic reconstruction was likely successful for this group; however, without intron position information, our general lack of confidence in the methods' success for HTAS-1 would lead us to disbelieve this finding. Thus, for this case, even though the sequence-based phylogenetic methods apparently correctly identified the HTAS-1 clade, the orthologous information from intron positions was necessary for us to be confident in the sequence-based phylogeny. Finally, intron positions were indispensable in distinguishing the origins of HTAS-1. Because HTAS-1 arose in an ancestor containing two genes encoding nearly identical proteins (H2A and HIS-35), it is very difficult to determine whether H2A evolved from the preexisting variant HIS-35 or *de novo* from H2A. The fact that all candidate HTAS-1 genes lack the HIS-35 intron strongly suggests the latter.

Fast evolution of sperm-specific variant HTAS-1

Sperm-specific proteins show generally elevated rates of evolution, consistent with strong selection on sperm functions because of sperm competition (74, 75). The current data show that this is decisively the case for the sperm-specific variant HTAS-1. We show that *C. elegans* HTAS-1's greater divergence from core H2A is not because of differences in an evolutionary age, but very much despite it: HTAS-1 is most divergent variant protein despite being the most recent to diverge from core H2A, having since evolved at a rate many times higher than any other H2A paralog.

What is the functional significance of recently-evolved histone variants?

In addition to tracing the origins and subsequent history of gene loss and retention, our results provide insights into the possible functions of histone H2A variants. For example, HIS-35 differs by a single amino acid ('A' at the 124th position) from the S-phase H2A. Given that characterized histone variants are thought to largely represent functionally distinct proteins, one hypothesis is that this single difference functionally differentiates HIS-35 protein from H2A. One way that single amino acid changes could have outsize effects is through altering the landscape of posttranslational modifications, which are key to histone function. For instance, this is the case with H3 variants, H3.1 and H3.2 which differ from one another by single amino acid (76).

Given these possibilities for small amino acid differences to change function, we approached the single H2A/HIS-35 difference with the hypothesis that the single difference led to functionally different proteins. However, when we looked at the H2A sequences of all the *Caenorhabditis* species, we found an 'A' at position 124 to be ancestral. Considering the presence of an A at position 124 in other canonical H2As, suggests the variant HIS-35 might has the same function as the canonical H2A. This hypothesis is also supported by the case of *C. kamaanina*, in which the encoded HIS-35 and H2A protein sequences are exactly the same. While these results do not disprove the hypothesis that H2A and HIS-35 encode proteins with important functional differences (except in the case of *C. kamaanina*), we propose instead that HIS-35 protein's functional importance lies in allowing for the expression of proteins (nearly) functionally identical to canonical H2A outside of the S-phase, given that canonical H2A is restricted to S-phase. Such a potential semi-redundancy could help to find the ambivalent phylogenetic pattern, in which retention of HIS-35 in most species suggests functional importance whereas loss in 5 independent lineages suggests conditional expendability. Interestingly, a similar pattern of lineage-specific

loss has been observed for H2B variants, which has also been interpreted as encoding functionally important but partially redundant functions (74).

Our data also have somewhat ambiguous implications for HTAS-1 function. On the one hand, HTAS-1 has been maintained for long periods of time in many lineages but has been lost in multiple independent lineages, as found for HIS-35. On the other hand, the much larger degree of protein sequence difference between HTAS-1 and H2A would seem to decrease the probability that HTAS-1 protein is functionally identical to H2A protein particularly given the extended C & N terminus of HTAS-1 which has previously been reported to play a vital role in DNA compaction, chromosome segregation, and fertility (65). Moreover, the particular chromatin constraints of sperm production raise the possibility that HTAS-1 proteins could encode distinct functions relative to H2A proteins, for instance by leading to greater sperm compaction; however, it is also possible that a distinct H2A paralog is simply necessary to ensure timely expression of H2A proteins outside of S-phase during spermatogenesis. More study of the functional significance of HTAS-1 is clearly needed.

Concluding remarks

To summarize, these results show exceptions to previously reported patterns, challenging sometimes implicit assumptions about non-core histones. First, whereas protein sequence differences between core and variant histone paralogs are often assumed to reflect differences in protein function, here we show that the variant protein HIS-35 is likely to have a redundant function with core H2A despite the sequence difference. Second, while all *C. elegans* H2A variants have a single intron, our observation of multi-intron variants and of recurrent intron loss, suggests that specifical introns may not have crucial roles in the expression of histone variants.

Instead, the role of introns in variant histones may simply lie in introns' general roles in promoting expression. Third, the combination of conservation and loss of variant histones points to potentially lineage-specific, partially redundant, or easily replaced roles of some histone variants. Future studies should explore the generality of these patterns across other lineages of eukaryotes. In addition to our specific findings for histone variant biology, these results highlight that introns can be useful in the reconstruction of the histories of complex gene families.

MATERIAL AND METHODS

Data source

Genomic sequences and gene feature format files of 168 Nematode species were obtained from WormBase (https://wormbase.org/) and *Caenorhabditis* database (http://caenorhabditis.org/).

Data mining and processing

All the known genes of 168 Nematode species with characterized exon-intron structures were fetched from their genome using their respective 'gene feature format' file. We then noted the positions of the introns in the header of their respective genes and translated the gene sequences.

To identify the homologs of H2A and their variants, BLASTP, version 2.9.0+, was performed using standard parameters while treating the translated gene sequences (of 168 Nematode species) as the database and H2A and variant (HTZ-1, HTAS-1, HIS-35) protein sequences as the query (77). Using a maximum e-value of 1e-10, 8003 hits were retrieved which were the homologs of H2A and H2A-variant genes. We then removed dubious genes encoding proteins more than 200 amino acids long, because histone proteins are generally shorter. We

collapsed the genes whose introns align at the position and have an identical protein sequence. After filtering off the genes we were left with 355 distinct entries.

Previous studies have shown the intron position conservation among widely diverged eukaryotic species. Therefore, to assess the intron position conservation among the putative H2A variant genes, we performed a Multiple Sequence Alignment (MSA) using the default parameters of CLUSTALW (78). We mapped the intron positions of each gene onto the corresponding protein CLUSTALW alignment, allowing us to identify as potential HTAS-1, HTZ-1, and HIS-3 orthologs (79, 80) those genes with intron positions matching *C. elegans* intron positions in those genes.

Phylogenetic Analysis

Multiple sequence alignment of 355 H2A variants homologs was performed using default parameters of MUSCLE and generated a phylogenetic tree (supplementary fig. 1) using IQtree, which does an automatic selection of the model by doing a model fit test and likelihood scoring (79, 80). VT+R9, the variable time method, was selected by IQtree for our data. The tree is also submitted in a "Newick format" as supplementary material -1. The tree didn't yield a clear phylogenetic signal for HIS-35 or HTZ-1, with homologs exhibiting the *C. elegans* HIS-35 or HTZ-1 intron positions scattered over the tree (supplementary fig. 1; IP 234 and IP 233 respectively). However, when we took a closer look at the HIS-35 characteristic intron-containing genes, we see that genes containing an intron at the *C. elegans* HIS-35 intron position (supplementary fig. 1, IP 234) are restricted to most species of *Caenorhabditis* and its sister genus *Diploscapter*. A clear clade of species was seen which had HTAS-1 characteristic intron position (supplementary fig. 1, IP 152).

Confirmation of H2A variant losses

We found a loss of HTAS-1, HTZ-1, and HIS-35 characteristic introns in a few lineages (marked by a minus sign in figures 2, 3, and 4). To know whether this is a real loss or reflected errors in gene annotation, tblastn searches were performed across the genome of these species. This manual curation led to the variant's characteristic intron splice sites being identified by eye in a few species due to alignment gaps at the exact intron position, indicating that these species truly contain the variant and that failure to initially identify the variant is due to a failure of the annotation to include these genes.

Acknowledgments

We thank our lab members for their helpful discussions. This work was supported by NSF:1616878; NSF STC DBI-1548297; NSF MCB RUI 1817611; NIH NICHD1R03HD093990A1

FIGURES

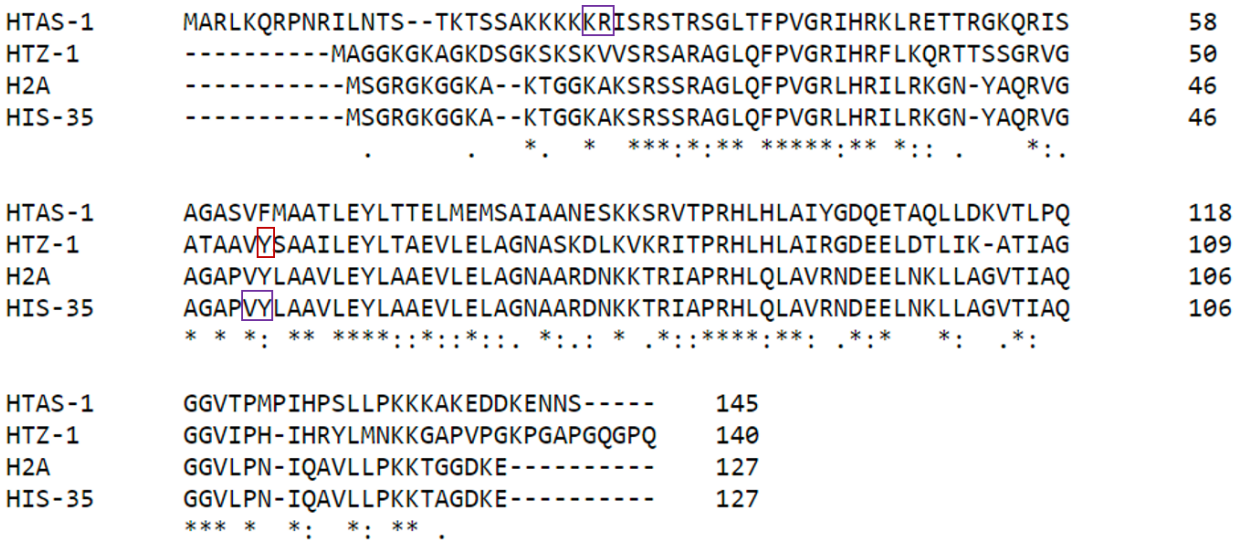


Fig. 1. Sequence and intron position comparison of three H2A variants of *C. elegans. C. elegans* core histone H2A and its three variant paralogs, contain different sequences and intron positions. The variants differ either in intron positions or the phases in which they interrupt the codon. As marked with a purple box: HTAS-1 has a phase 0 intron between the 26th and 27th amino acid; HIS-35 has a phase 0 intron between 50th and 51st amino acid. HTZ-1 (marked with red box) has a phase 2 intron splitting the 57th amino acid.

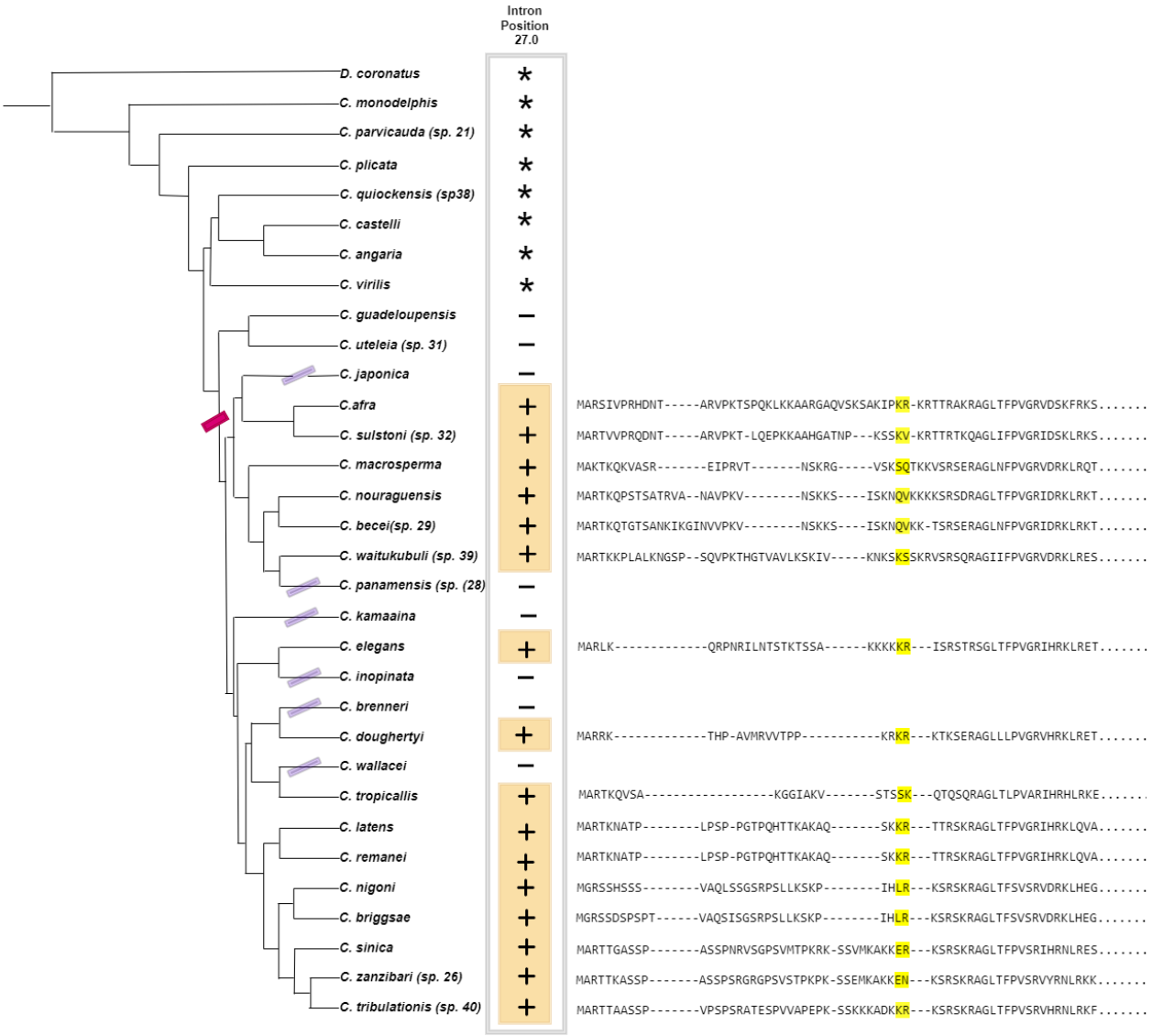


Fig. 2: Intron position and sequence evidence indicate the origin of HTAS-1 within *Caenorhabditis* evolution and subsequent loss. On the left is the previously reconstructed species tree of *Caenorhabditis*. The likely origin of HTAS-1 is indicated with a dark pink bar. HTAS-1 presence in a species is indicated by the plus sign whereas the HTAS-1 losses have been shown with the minus sign. The species where HTAS-1 was not incorporated are indicated as an asterisk. On the right is the multiple sequence alignment of HTAS-1 proteins showing the aligned intron positions (highlighted with yellow).

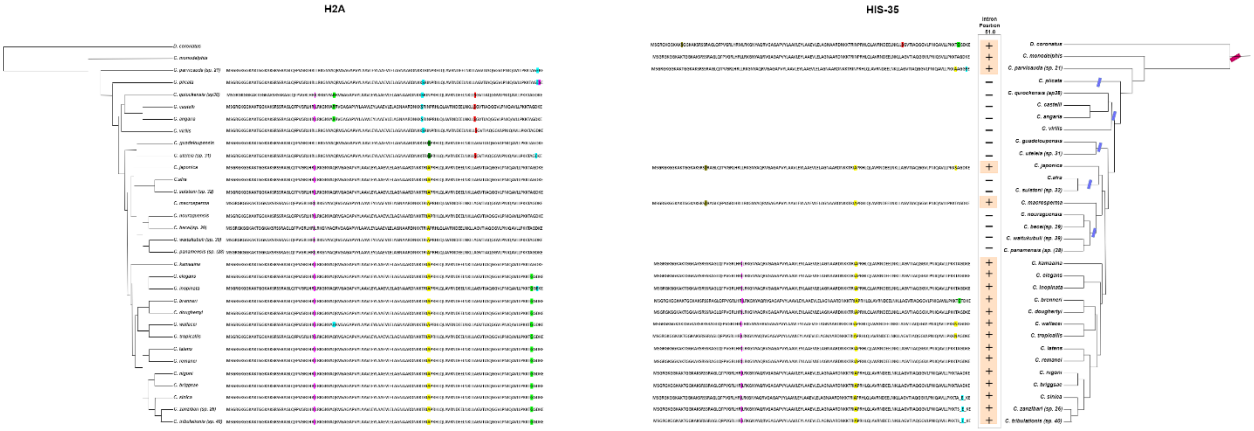


Fig. 3: Comparison of canonical H2A (left) and HIS-35 (right) sequences across *Caenorhabditis* species. Inferred sequence changes relative to the reconstructed ancestral sequence are highlighted with different colors (pink, red, yellow, blue, green), with identical changes colored identical colors across the figure (i.e., the L-to-R change at position 35th is colored purple on both protein alignments)

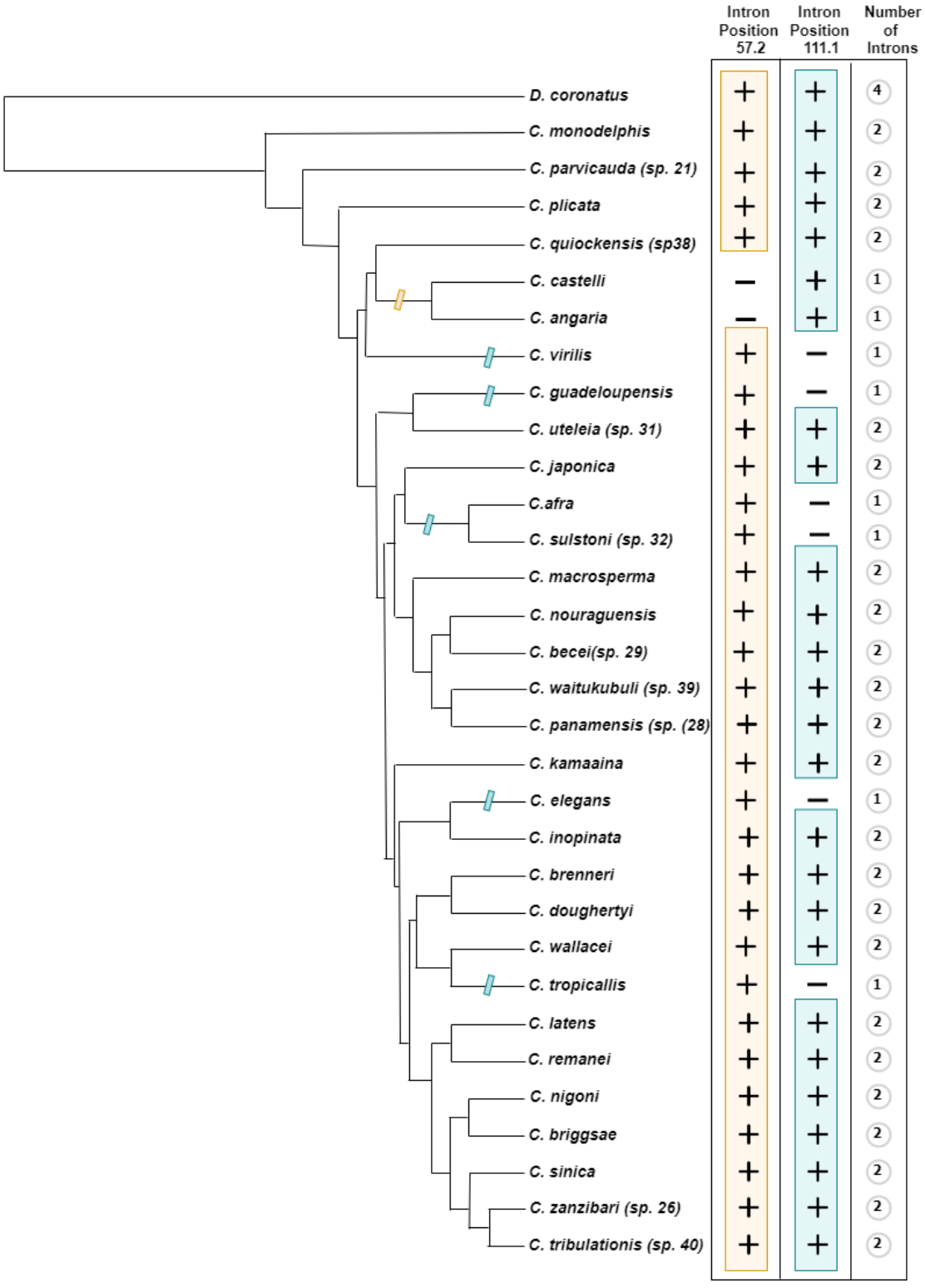
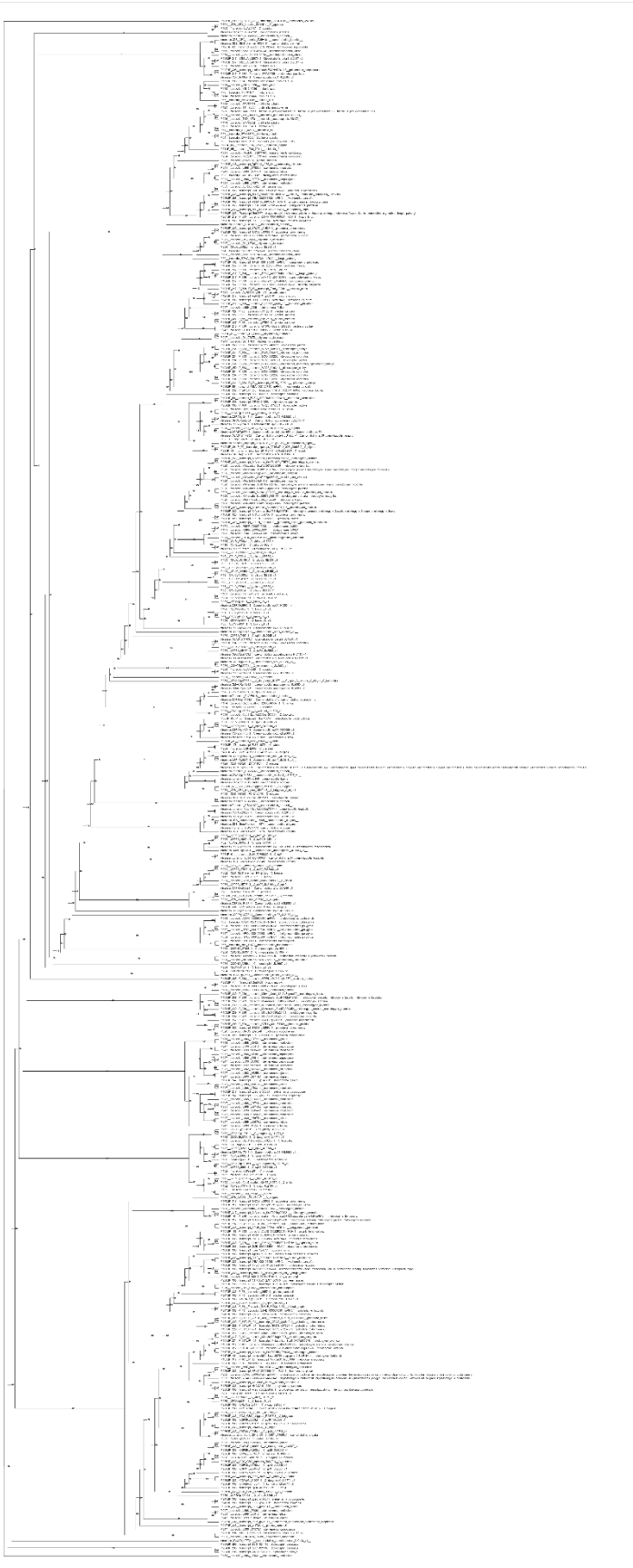
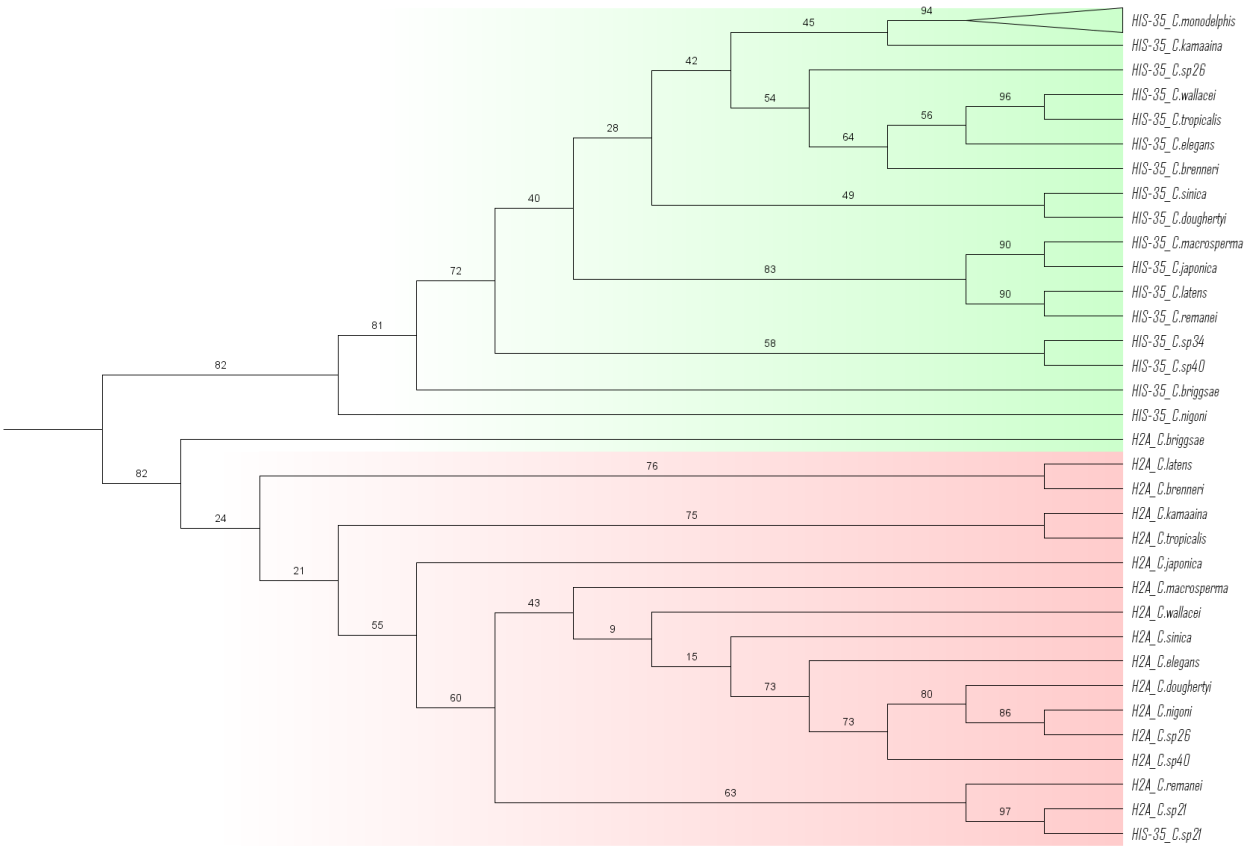


Fig. 4: The dynamic history of intron loss and gain in HTZ-1. On the left is the previously reconstructed species tree of *Caenorhabditis*. HTZ-1 characteristic intron presence and absence in a species is indicated by the plus and minus sign.



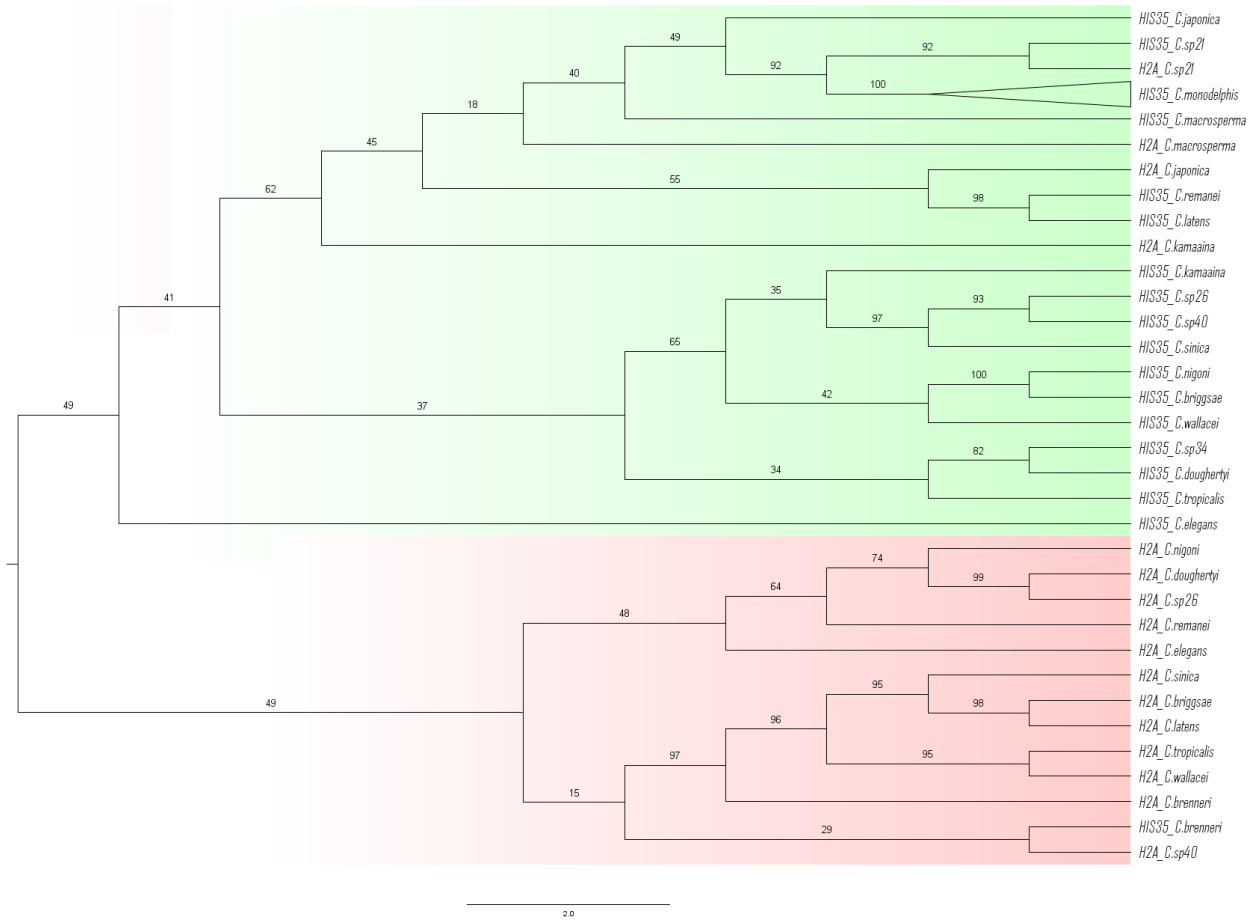
Supplementary Fig. 1

Phylogenetic tree reconstructed from H2A and H2A-related gene variants across 168 available nematode genomes. The tree is made up of 355 unique protein sequences. On the tree, “IP” followed by a number stands for the position at which the intron was aligned during the multiple sequence alignment. The intron position of HTAS-1 is 152 (IP 152). The intron position of HIS-35 is 234 (IP 234) and the intron position of the variant HTZ-1 is 236 (IP 236). The canonical H2A’s are indicated as “intronless”.



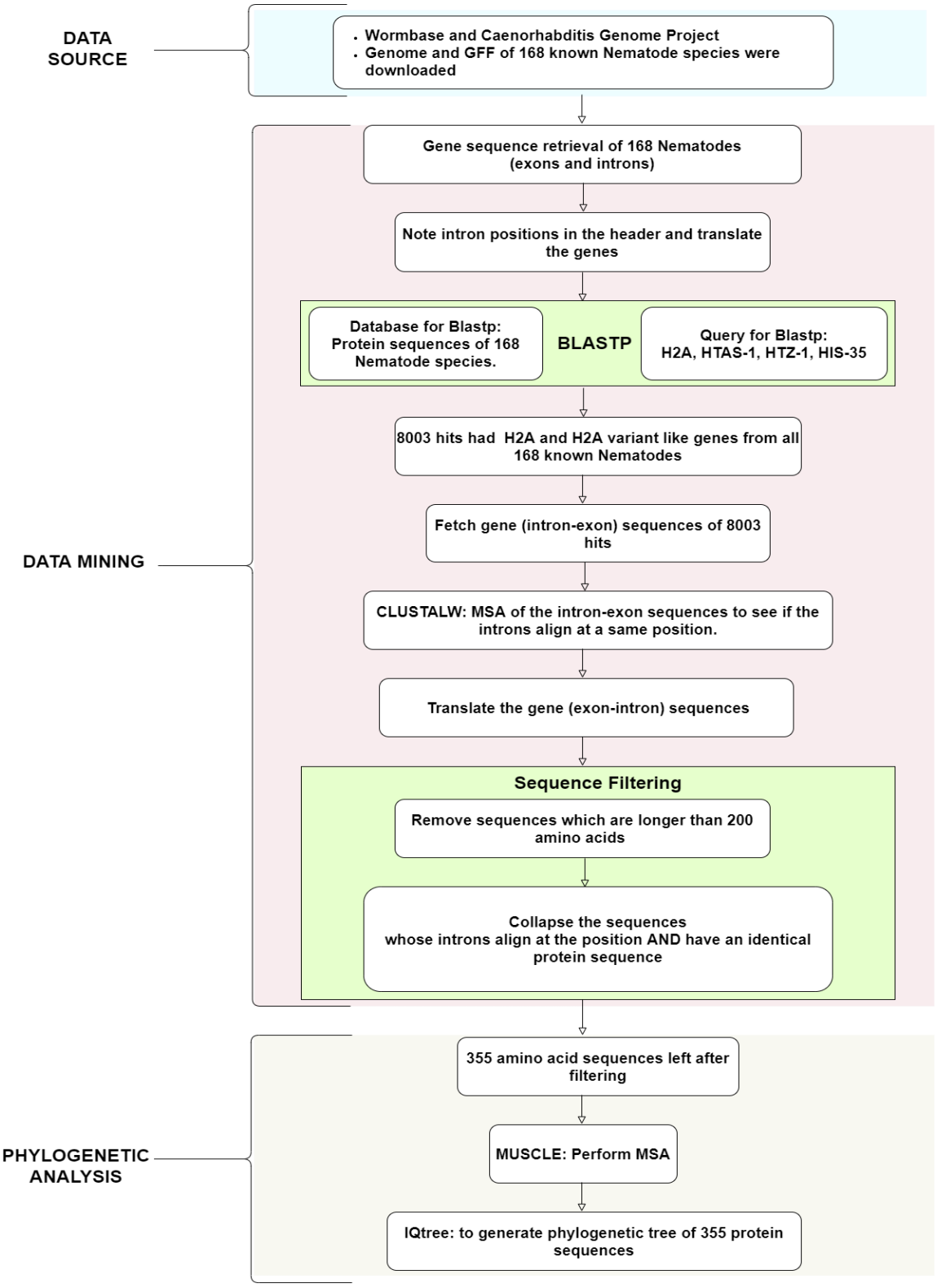
Supplementary Fig. 2

Phylogenetic tree reconstructed from nucleotide sequences of canonical *Caenorhabditis* H2A gene sequences and first exons of candidate HIS-35 gene sequences. We see a clear cluster of H2A nucleotide sequences and HIS-35 with a few exceptions, in particular *Caenorhabditis sp 21.*



Supplementary Fig. 3

Phylogenetic tree reconstructed from nucleotide sequences of canonical *Caenorhabditis* H2A gene sequences and second exons of candidate HIS-35 gene sequences.



Supplementary Fig. 4

Material and method Flowchart

References

1. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature. 1997;389(6648):251-60.

2. Luger K, Dechassa ML, Tremethick DJ. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? Nature reviews Molecular cell biology. 2012;13(7):436-47.

3. Talbert PB, Henikoff S. Histone variants - ancient wrap artists of the epigenome. Nature reviews. Molecular cell biology. 2010 Apr;11(4):264-75.

4. Malik HS, Henikoff S. Phylogenomics of the nucleosome. Nature structural & molecular biology. 2003;10(11):882-91.

5. Mei Q, Huang J, Chen W, Tang J, Xu C, Yu Q, et al. Regulation of DNA replication-coupled histone gene expression. Oncotarget. 2017 -10-16;8(55):95005.

6. Pandey NB, Chodchoy N, Liu T, Marzluff WF. Introns in histone genes alter the distribution of 3′ ends. Nucleic Acids Res. 1990;18(11):3161-70.

7. Romeo V, Schümperli D. Cycling in the nucleus: regulation of RNA 3′ processing and nuclear organization of replication-dependent histone genes. Curr Opin Cell Biol. 2016;40:23-31.

8. Biterge B, Schneider R. Histone variants: key players of chromatin. Cell Tissue Res. 2014;356(3):457-66.

9. Henikoff S, Smith MM. Histone variants and epigenetics. Cold Spring Harbor perspectives in biology. 2015 Jan 05,;7(1):a019364.

10. Talbert PB, Ahmad K, Almouzni G, Ausió J, Berger F, Bhalla PL, et al. A unified phylogeny-based nomenclature for histone variants. Epigenetics & chromatin. 2012;5(1):1-19.

11. Martire S, Banaszynski LA. The roles of histone variants in fine-tuning chromatin organization and function. Nature reviews. Molecular cell biology. 2020 Sep;21(9):522-41.

12. Weber CM, Henikoff S. Histone variants: dynamic punctuation in transcription. Genes Dev. 2014;28(7):672-82.

13. Talbert PB, Henikoff S. Environmental responses mediated by histone variants. Trends Cell Biol. 2014;24(11):642-50.

14. Filipescu D, Szenker E, Almouzni G. Developmental roles of histone H3 variants and their chaperones. Trends in Genetics. 2013;29(11):630-40.

15. Maze I, Noh K, Soshnev AA, Allis CD. Every amino acid matters: essential contributions of histone variants to mammalian development and disease. Nature Reviews Genetics. 2014;15(4):259-71.

16. Otero S, Desvoyes B, Gutierrez C. Histone H3 dynamics in plant cell cycle and development. Cytogenetic and genome research. 2014;143(1-3):114-24.

17. Volle C, Dalal Y. Histone variants: the tricksters of the chromatin world. Curr Opin Genet Dev. 2014;25:8-14.

18. Herchenröther A, Wunderlich TM, Lan J, Hake SB. Spotlight on histone H2A variants: From B to X to Z. Seminars in Cell & Developmental Biology; Elsevier; 2022.

19. Talbert PB, Henikoff S. Histone variants at a glance. J Cell Sci. 2021;134(6):jcs244749.

20. Phillips EO, Gunjan A. Histone variants: The unsung guardians of the genome. DNA repair. 2022;112:103301.

21. Tessarz P, Kouzarides T. Histone core modifications regulating nucleosome structure and dynamics. Nature reviews Molecular cell biology. 2014;15(11):703-8.

22. Deal RB, Henikoff S. Histone variants and modifications in plant gene regulation. Curr Opin Plant Biol. 2011;14(2):116-22.

23. Lei B, Berger F. H2A variants in Arabidopsis: versatile regulators of genome activity. Plant Communications. 2020;1(1):100015.

24. Bönisch C, Hake SB. Histone H2A variants in nucleosomes and chromatin: more or less stable? Nucleic acids research. 2012 Nov;40(21):10719-41.

25. Molaro A, Young JM, Malik HS. Evolutionary origins and diversification of testis-specific short histone H2A variants in mammals. Genome Res. 2018;28(4):460-73.

26. Faast R, Thonglairoam V, Schulz TC, Beall J, Wells JR, Taylor H, et al. Histone variant H2A. Z is required for early mammalian development. Current Biology. 2001;11(15):1183-7.

27. Meneghini MD, Wu M, Madhani HD. Conserved histone variant H2A. Z protects euchromatin from the ectopic spread of silent heterochromatin. Cell. 2003;112(5):725-36.

28. Fan JY, Rangasamy D, Luger K, Tremethick DJ. H2A. Z alters the nucleosome surface to promote HP1α-mediated chromatin fiber folding. Mol Cell. 2004;16(4):655-61.

29. Colino-Sanguino Y, Clark SJ, Valdes-Mora F. The H2A. Z-nuclesome code in mammals: emerging functions. Trends in Genetics. 2021.

30. Petty EL, Collette KS, Cohen AJ, Snyder MJ, Csankovszki G. Restricting dosage compensation complex binding to the X chromosomes by H2A. Z/HTZ-1. PLoS genetics. 2009;5(10):e1000699.

31. Long J, Carter B, Johnson ET, Ogas J. Contribution of the histone variant H2A. Z to expression of responsive genes in plants. Seminars in Cell & Developmental Biology; Elsevier; 2022.

32. Herchenröther A, Wunderlich TM, Lan J, Hake SB. Spotlight on histone H2A variants: From B to X to Z. Seminars in Cell & Developmental Biology; Elsevier; 2022.

33. Rogakou EP, Pilch DR, Orr AH, Ivanova VS, Bonner WM. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. J Biol Chem. 1998;273(10):5858-68.

34. Rivera-Casas C, Gonzalez-Romero R, Cheema MS, Ausió J, Eirin-Lopez JM. The characterization of macroH2A beyond vertebrates supports an ancestral origin and conserved role for histone variants in chromatin. Epigenetics. 2016;11(6):415-25.

35. Sun Z, Bernstein E. Histone variant macroH2A: from chromatin deposition to molecular function. Essays Biochem. 2019;63(1):59-74.

36. Osakabe A, Molaro A. Histone renegades: Unusual H2A histone variants in plants and animals. Seminars in Cell & Developmental Biology; Elsevier; 2022.

37. Guberovic I, Farkas M, Corujo D, Buschbeck M. Evolution, structure and function of divergent macroH2A1 splice isoforms. Seminars in Cell & Developmental Biology; Elsevier; 2022.

38. Contrepois K, Coudereau C, Benayoun BA, Schuler N, Roux P, Bischof O, et al. Histone variant H2A. J accumulates in senescent cells and promotes inflammatory gene expression. Nature communications. 2017;8(1):1-18.

39. Shaytan AK, Landsman D, Panchenko AR. Nucleosome adaptability conferred by sequence and structural variations in histone H2A–H2B dimers. Curr Opin Struct Biol. 2015;32:48-57.

40. Jiang X, Soboleva TA, Tremethick DJ. Short histone H2A variants: small in stature but not in function. Cells. 2020;9(4):867.

41. Hoghoughi N, Barral S, Vargas A, Rousseaux S, Khochbin S. Histone variants: essential actors in male genome programming. The Journal of Biochemistry. 2018;163(2):97-103.

42. Millar CB. Organizing the genome with H2A histone variants. Biochem J. 2013;449(3):567-79.

43. Buschbeck M, Hake SB. Variants of core histones and their roles in cell fate decisions, development and cancer. Nature reviews Molecular cell biology. 2017;18(5):299-314.

44. Marzluff WF, Wagner EJ, Duronio RJ. Metabolism and regulation of canonical histone mRNAs: life without a poly (A) tail. Nature Reviews Genetics. 2008;9(11):843-54.

45. Wolffe AP. Histone Genes. In: Brenner's Online Encyclopedia of Genetics, Four-Volume Set. Elsevier Inc; 2001. p. 948-52.

46. Le Hir H, Nott A, Moore MJ. How introns influence and enhance eukaryotic gene expression. Trends Biochem Sci. 2003;28(4):215-20.

47. Dixon RJ, Eperon IC, Samani NJ. Complementary intron sequence motifs associated with human exon repetition: a role for intragenic, inter-transcript interactions in gene expression. Bioinformatics. 2007;23(2):150-5.

48. Rose AB. Intron-mediated regulation of gene expression. Nuclear pre-mRNA processing in plants. 2008:277-90.

49. Casas-Mollano JA, Lao NT, Kavanagh TA. Intron-regulated expression of SUVH3, an Arabidopsis Su (var) 3-9 homologue. J Exp Bot. 2006;57(12):3301-11.

50. Callis J, Fromm M, Walbot V. Introns increase gene expression in cultured maize cells. Genes Dev. 1987;1(10):1183-200.

51. Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T. Epigenetics in alternative pre-mRNA splicing. Cell. 2011;144(1):16-26.

52. Armstrong JF, Pritchard-Jones K, Bickmore WA, Hastie ND, Bard JB. The expression of the Wilms' tumour gene, WT1, in the developing mammalian embryo. Mech Dev. 1993;40(1-2):85-97.

53. Cieply B, Carstens RP. Functional roles of alternative splicing factors in human disease. Wiley Interdisciplinary Reviews: RNA. 2015;6(3):311-26.

54. Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes Dev. 2003;17(4):419-37.

55. Roy SW, Fedorov A, Gilbert W. Large-Scale Comparison of Intron Positions in Mammalian Genes Shows Intron Loss but No Gain. Proceedings of the National Academy of Sciences - PNAS. 2003 Jun 10,;100(12):7158-62.

56. Sêton Bocco S, Csűrös M. Splice Sites Seldom Slide: Intron Evolution in Oomycetes. Genome biology and evolution. 2016 Aug 25,;8(8):2340-50.

57. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Current Biology. 2003;13(17):1512-7.

58. Coulombe-Huntington J, Majewski J. Characterization of intron loss events in mammals. Genome Res. 2007;17(1):23-32.

59. Irimia M, Roy SW. Spliceosomal introns as tools for genomic and evolutionary analysis. Nucleic acids research. 2008 Mar;36(5):1703-12.

60. Sakharkar MK, Tan TW, de Souza SJ. Generation of a database containing discordant intron positions in eukaryotic genes (MIDB). Bioinformatics. 2001;17(8):671-5.

61. Rogozin IB, Lyons-Weiler J, Koonin EV. Intron sliding in conserved gene families. Trends in Genetics. 2000;16(10):430-2.

62. Van Daal A, White EM, Elgin SC, Gorovsky MA. Conservation of intron position indicates separation of major and variant H2As is an early event in the evolution of eukaryotes. J Mol Evol. 1990;30(5):449-55.

63. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Chromosomal DNA and its packaging in the chromatin fiber. In: Molecular Biology of the Cell. 4th edition. Garland Science; 2002.

64. Draizen EJ, Shaytan AK, Mariño-Ramírez L, Talbert PB, Landsman D, Panchenko AR. HistoneDB 2.0: a histone database with variants—an integrated resource to explore histones and their variants. Database. 2016;2016.

65. Chu DS, Liu H, Wu TF, Ralston EJ, Nix P, Meyer BJ, et al. Sperm chromatin proteomics identifies evolutionarily conserved fertility factors. Nature (London). 2006 Sep 07,;443(7107):101-5.

66. Kelchner SA, Thomas MA. Model use in phylogenetics: nine key questions. Trends in Ecology & Evolution. 2007;22(2):87-94.

67. Nei M, Rooney AP. CONCERTED AND BIRTH-AND-DEATH EVOLUTION OF MULTIGENE FAMILIES. Annual review of genetics. 2005 Dec 15,;39(1):121-52.

68. Scienski K, Fay JC, Conant GC. Patterns of Gene Conversion in Duplicated Yeast Histones Suggest Strong Selection on a Coadapted Macromolecular Complex. Genome biology and evolution. 2015 Nov 11,;7(12):3249-58.

69. Scacchetti A, Becker PB. Variation on a theme: Evolutionary strategies for H2A. Z exchange by SWR1-type remodelers. Curr Opin Cell Biol. 2021;70:1-9.

70. Ibarra-Morales D, Rauer M, Quarato P, Rabbani L, Zenk F, Schulte-Sasse M, et al. Histone variant H2A. Z regulates zygotic genome activation. Nature communications. 2021;12(1):1-14.

71. Jackson JD, Falciano VT, Gorovsky MA. A likely histone H2A. F/Z variant in Saccharomyces cerevisiae. Trends Biochem Sci. 1996;21(12):466-7.

72. Kiontke K, Gavin NP, Raynes Y, Roehrig C, Piano F, Fitch DH. Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proceedings of the National Academy of Sciences. 2004;101(24):9003-8.

73. Irimia M, Rukov JL, Penny D, Garcia-Fernandez J, Vinther J, Roy SW. Widespread evolutionary conservation of alternatively spliced exons in Caenorhabditis. Mol Biol Evol. 2008;25(2):375-82.

74. Raman P, Rominger MC, Young JM, Molaro A, Tsukiyama T, Malik HS. Novel classes and evolutionary turnover of histone H2B variants in the mammalian germline. Mol Biol Evol. 2022;39(2):msac019.

75. Kasimatis KR, Phillips PC. Rapid gene family evolution of a nematode sperm protein despite sequence hyper-conservation. G3: Genes, Genomes, Genetics. 2018;8(1):353-62.

76. Hake SB, Allis CD. Histone H3 variants and their potential role in indexing mammalian genomes: the “H3 barcode hypothesis”. Proceedings of the National Academy of Sciences. 2006;103(17):6428-35.

77. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997 Sep;25(17):3389-402.

78. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research. 1994 Nov 11,;22(22):4673-80.

79. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004 Mar 01,;32(5):1792-7.

80. Nguyen L, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular biology and evolution. 2015 Jan;32(1):268-74.