ARTICLE

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 and Systems Biology, University of California, Merced, California, United States of America.

[2]Department of Biology, San Francisco State University, California, United States of America.

[*]**Corresponding authors**: Scott Roy (sroy5@ucmerced.edu); Diana Chu (chud@sfsu.edu)

ABSTRACT

Histone variants replace canonical histones in nucleosomes, sometimes changing nucleosome function. Histone variant evolution is poorly characterized, and, as we show here, reconstruction of histone protein evolution can be challenging given large differences in rates across gene lineages and across sites. The positions of introns that interrupt genes can provide complementary phylogenetic information. We combined sequence and intron data to reconstruct the evolution of three histone H2A variants in *Caenorhabditis elegans* to reveal disparate histories. For the variant HIS-35 (which differs from H2A by only a single glycine-to-alanine C-terminal change), we find no evidence for the hypothesis of distinct protein function: the HIS-35 alanine is ancestral and common across canonical *Caenorhabditis* H2A sequences, with one species encoding identical HIS-35 and canonical H2A proteins. We propose instead that HIS-35 allows for H2A expression outside of the S-phase. Genes encoding such "backup" functions could be functionally important yet readily replaceable; consistent with this notion, both HTAS-1 and HIS-35 exhibit phylogenetic patterns that combine long-term evolutionary persistence and recurrent loss. Finally, the H2A.Z homolog, HTZ-1, exhibits recurrent intron loss and gain, suggesting that it is intron presence, rather than a specific intron sequence or position, that may be important in histone variant expression.

Keywords: Gene duplication, Neofunctionalization, Redundancy, Intronless and intron containing histones, Intron position conservation, Histone H2A evolution.

SIGNIFICANCE

Histone variants are proteins that replace the core canonical histones in the nucleosomes and often confer specific structural and functional features. Histone variants represent a unique opportunity to study the functional significance of gene duplicates. The origin of new variants by core histone gene duplication is generally a highly asymmetrical affair marked by variant-specific changes in gene processing signals and accelerated evolution. The functions and evolutionary significance of the diverse variants found in available eukaryotic genomes remain almost entirely obscure, with only a few characterized case studies. By using a novel method that leverages intron position conservation, we reconstructed the evolutionary history of H2A variants within *C. elegans* and relatives. Our findings provide several insights that run counter to current paradigms of histone variant function and evolution.

INTRODUCTION

All characterized eukaryotic cells compact their DNA into nucleosomes, which consist of an octameric histone complex comprised of core canonical histone proteins- H2A, H2B, H3, and H4- in two copies each, around which ~147 bps of DNA is wrapped (Richmond et al. 1997). Genes that encode these 'canonical'

histones are tightly coupled with the cell cycle and DNA replication because of the need for immense numbers of histone proteins to package newly synthesized DNA during the S-phase of the cell cycle (Henikoff and Smith 2015). Genes encoding canonical (or core) histone (sometimes also referred to as the 'replication-dependent histone') do not contain introns and are organized into multiple gene copies and in a tandem-repeat structure, which is thought to facilitate rapid, coordinated expression of these genes (Mei et al. 2017).

In addition, variant histones are also incorporated into nucleosomes. Unlike canonical histone genes, variant histone genes are not restricted to S phase expression but can be expressed constitutively or in a tissue-specific manner (Wolffe 2001). These histone variants are typically found in a single copy in the genome and contain introns in their pre-mRNAs (Marzluff, Wagner, Duronio 2008). Moreover, variant histone proteins typically differ in sequence from their core homologs; upon incorporation in the chromatin, these protein differences can cause changes in the histone-histone and histone-DNA interactions, thereby changing the expression of the genes to which they are bound (Talbert and Henikoff 2010). Histone variant incorporation in chromatin near genes, often referred to as the 'histone code', is one of the most important factors by which specific sets of genes are expressed at specific times and places (Martire and Banaszynski 2020). Despite the striking structural, expression and functional differences between variant and core histone, the

evolutionary history of many variants is still deeply mysterious in the vast majority of organisms.

The histone code, which is crucial to the complex and subtle regulation of genes, has repeatedly been diversified by the evolution of new histone variants, each of which has a unique evolutionary trajectory from its canonical counterpart (Yun et al. 2011), (Bönisch and Hake 2012). Among all the core histones, H2A is the fastest evolving histone, and it shows the most diversity in the variants. It has some widely distributed variants, in particular H2A.Z, which is present in every eukaryotic species, but may act distinctly in different species and at different genomic locations (Deal and Henikoff 2011; Fan et al. 2004; Meneghini, Wu, Madhani 2003). H2A also has some variants, such as H2A.X, which have arisen independently in several lineages, whereas variants such as macroH2A are present only in some lineages (Bönisch and Hake 2012). In short, there is a great wealth of histone variants observed, ranging from variants shared across all eukaryotes to variants that are species-specific; and from variants that are ubiquitously expressed to variants that are expressed only in certain tissues (Buschbeck and Hake 2017; Dryhurst, Thambirajah, Ausió 2004; Inui, Martello, Piccolo 2010; Talbert and Henikoff 2010).

Here, the evolutionary trajectories of H2A variants of *Caenorhabditis elegans* have been reconstructed in nematodes. There are three H2A variants present in *C.*

elegans namely: 1) the ubiquitously expressed variant, HTZ-1, which is an ortholog of the evolutionarily conserved variant, H2A.Z.; 2) variant HIS-35, not much is known about them; and 3) Sperm-specific variant HTAS-1, which has only been reported in *C. elegans* to date (Whittle et al. 2008). By combining protein sequence and intron position conservation, we reconstructed the origins and evolution of H2A variants. We find surprising patterns suggesting unprecedented functions for histone variants.

RESULTS AND DISCUSSION

Phylogenetic methods fail to reconstruct H2A gene family evolution

We used BLAST searches to identify all annotated copies of H2A and H2A-related gene variants across 168 available nematode genomes. After filtering and collapsing identical proteins, we were left with 355 unique sequences. We used standard phylogenetic methods to reconstruct the evolutionary history of these sequences. However, scrutiny of the recovered phylogenetic tree revealed several bizarre findings. For instance, core H2A proteins formed clades that included very deeply-diverged nematode sequences; on the other hand, species-specific variants often grouped far from proteins from the same or related species. Some of these anomalies are as expected by errors in phylogenetic reconstruction due to model

misspecification, which is very likely here given the extreme between-lineage and between-site variations in evolutionary rates.

Introns as an additional source of phylogenetic information

Another potential source of phylogenetic information is 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 (Irimia and Roy 2008; Roy, Fedorov, Gilbert 2003). This approach is validated by individual instances of intron conservation, for instance showing conservation of introns in the ancient H2A variant HTZ-1 between distantly related nematodes (Supplementary Figure 1). To leverage phylogenetic information from intron positions, we obtained intron-exon structures for all H2A gene family members and performed alignments to determine intron position sharing across genes.

The three H2A variants of *C. elegans* differ either in intron positions or phases in which the introns interrupt the codon.

We first aligned the three main *C. elegans* H2A variants (Figure 1), HIS-35, HTAS-1, and HTZ-1, each of which has a single intron position (highlighted with a box in Figure 1). Scrutiny of intron positions showed that the three genes either contain introns at different positions, differing both in the codon that they interrupt or in the phase at which they disrupt the codon. Interestingly, the intron position in HIS-35

falls very near to that found in HTZ-1. However, these introns are unlikely to represent a shared intron, given that intron positions rarely slide between phases (Sêton Bocco and Csűrös 2016) and given that HIS-35's near identity to H2A (see below) strongly suggests that it is derived from H2A and not from HTZ-1 (and thus is not expected to share its intron position).

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 H2A variants revealed 16 genes that share an intron at the exact homologous position as HTAS-1 (Figure 2). 16/16 of these genes are from species falling within a single clade of *Caenorhabditis* species, 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) is reconstructed as a clade, suggesting that phylogenetic reconstruction was successful for this region of the tree (since such correspondence is not expected from random errors). Because this clade of species shares the *C. elegans* HTAS-1 intron position, it suggests that these genes represent orthologs of HTAS-1. However, the lack of this intron in the genes of other

Caenorhabditis species is not conclusive, since the intron could have arisen after the origins of the HTAS-1 gene in an earlier ancestor. However, no candidates for these alternative HTAS-1 orthologs in other *Caenorhabditis* species were found: no additional *Caenorhabditis* genes were found to group with the candidate HTAS-1 clade in the sequence tree (and, in particular, no genes from the species lacking the HTAS-1-like intron position), and in general no 'extra' variants were observed in these more distantly-related species that were candidates for being orthologs to HTAS-1 (Supplementary Fig. 1). Thus, all available data are consistent with a single origin of intron-containing HTAS-1 within the ancestor of a subclade of *Caenorhabditis* nematodes.

Interestingly, 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 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) and did not contain gene copies that grouped with the candidate HTAS-1 clade in the sequence tree (Supplementary Fig. 1). Moreover, as with *Caenorhabditis* species outside of this clade, they lacked additional H2A variants which might represent HTAS-1 orthologs that have lost the characteristic HTAS-1 intron. 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, HTAS-1 has 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*.

Core histone functions are expected to be highly conserved across eukaryotes, given their central roles in ensuring DNA packaging and protection (Alberts et al. 2002). 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. 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. The protein sequence of HIS-35 differs by just one amino acid from the S-phase H2A. 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 HIS2A. The multiple copies of core histone genes are known to undergo so-called concerted evolution, with sequences being transferred between paralogs by recombination (Nei and Rooney 2005; Scienski, Fay, Conant 2015). 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. We reconstructed separate phylogenetic trees of exon-1 and exon-2 for all H2A and HIS-35 sequences. While most of the reconstructed tree largely reflected the species tree, we observed the grouping of the two gene sequences for *Caenorhabditis sp21* (supplementary fig. 2 and 3). Occasional concerted evolution of these genes is consistent with a lack of functional differentiation.

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

The ubiquitously expressed *C. elegnas* H2A variant HTZ-1 is the ortholog of H2A.Z which is evolutionary conserved across all the eukaryotes. *C. elegans* HTZ-1 has an intron that splits the 57th codon at position 2. Alignment across all H2A variants revealed 30 genes that share an intron at the exact homologous position (Figure 4, marked with a plus sign). Interestingly, most HTZ-1 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* HTZ-1 gene). Species including *C. elegans*, *C. tropicallis*, *C. sp32*, *C. afra*, *C. guadeloupensis*, *C. virilis* have lost the second HTZ-1 intron which is at position 111.1, whereas *C. casteli* and *C. angaria* have lost their first intron.

Discussion

This study represents the first to date to examine the evolution of variant histones by looking at their intron position conservation. Previous studies have shown the intron position conservation among widely diverged eukaryotic species (Irimia and Roy 2008; Roy, Fedorov, Gilbert 2003). For instance, intron position conservation in humans, mice, and fish (Irimia and Roy 2008). Thus, intron positions contain a record of evolutionary history that can facilitate insights into gene history.

In addition to tracing the origins and subsequent history of gene loss and retention, our results provide unexpected insights into the probable functions of histone H2A variants. The most striking case involves HIS-35. HIS-35 differs by a single amino acid ('A' at the 124th position) from the S-phase H2A; the notion that histone variants represent functionally distinct proteins thus predicts that HIS-35 is functionally different as compared to its canonical counterpart because of that single amino acid change. 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, points us to the fact that the variant HIS-35 might have 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. This pattern gave us another reason to believe that the changes present in HIS-35 are not functionally differentiated from H2A at the protein level, but instead may be used as a backup

for canonical H2A whenever it is needed outside the 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.

Our data also have somewhat ambivalent implications for HTAS-1 function. As with HIS-35 above, on the one hand, HTAS-1 has been maintained by selection for long periods of time in many lineages, but has been lost in multiple independent lineages. On the other hand, the much larger degree of protein sequence differences between HTAS-1 and H2A would seem to decrease the probability that HTAS-1 protein is functionally identical to H2A protein, particularly given the remarkable conservation of core H2A proteins. Greater divergence of sperm-specific HTAS-1 from core H2A is consistent with the rapid evolution of male reproductive proteins (Kasimatis and Phillips 2018; Swanson and Vacquier 2002; Turner and Hoekstra 2004; Wilburn and Swanson 2016).

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 in related species, and evolutionary turnover of those introns, suggests that specific observed introns may not have crucial roles in the expression of histone variants, and that intron importance in variant histones may simply reflect general expression-promoting roles of introns rather than histone-specific roles. 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.

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 (Altschul et al. 1997). 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 (Thompson, Higgins, Gibson 1994). 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 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 (Edgar 2004; Nguyen et al. 2015). VT+R9, variable time method, was selected by IQtree for our data. The tree is also submitted in a "newick format" as supplementary Data -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.

Acknowledgements

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

References

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

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research 25(17):3389-402.

Bönisch C and Hake SB. 2012. Histone H2A variants in nucleosomes and chromatin: More or less stable? Nucleic Acids Research 40(21):10719-41.

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

Chu DS, Liu H, Wu TF, Ralston EJ, Nix P, Meyer BJ, Yates III JR. 2006. Sperm chromatin proteomics identifies evolutionarily conserved fertility factors. Nature (London) 443(7107):101-5.

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

Dryhurst D, Thambirajah AA, Ausió J. 2004. New twists on H2A. Z: A histone variant with a controversial structural and functional past. Biochemistry and Cell Biology 82(4):490-7.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32(5):1792-7.

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

Henikoff S and Smith MM. 2015. Histone variants and epigenetics. Cold Spring Harbor Perspectives in Biology 7(1):a019364.

Inui M, Martello G, Piccolo S. 2010. MicroRNA control of signal transduction. Nature Reviews Molecular Cell Biology 11(4):252-63.

Irimia M and Roy SW. 2008. Spliceosomal introns as tools for genomic and evolutionary analysis. Nucleic Acids Research 36(5):1703-12.

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

Mariño-Ramírez L, Kann MG, Shoemaker BA, Landsman D. 2005. Histone structure and nucleosome stability. Expert Review of Proteomics 2(5):719-29.

Martire S and Banaszynski LA. 2020. The roles of histone variants in fine-tuning chromatin organization and function. Nature Reviews. Molecular Cell Biology 21(9):522-41.

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

Mei Q, et al. 2017. Regulation of DNA replication-coupled histone gene expression. Oncotarget 8(55):95005.

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

Nei M and Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annual Review of Genetics 39(1):121-52.

Nguyen L, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32(1):268-74.

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

Roy SW, Fedorov A, Gilbert W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proceedings of the National Academy of Sciences - PNAS 100(12):7158-62.

Scienski K, Fay JC, Conant GC. 2015. Patterns of gene conversion in duplicated yeast histones suggest strong selection on a coadapted macromolecular complex. Genome Biology and Evolution 7(12):3249-58.

Sêton Bocco S and Csűrös M. 2016. Splice sites seldom slide: Intron evolution in oomycetes. Genome Biology and Evolution 8(8):2340-50.

Swanson WJ and Vacquier VD. 2002. The rapid evolution of reproductive proteins. Nature Reviews Genetics 3(2):137-44.

Talbert PB and Henikoff S. 2010. Histone variants - ancient wrap artists of the epigenome. Nature Reviews. Molecular Cell Biology 11(4):264-75.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22(22):4673-80.

Turner LM and Hoekstra HE. 2004. Causes and consequences of the evolution of reproductive proteins. Int J Dev Biol 52(5-6):769-80.

Whittle CM, McClinic KN, Ercan S, Zhang X, Green RD, Kelly WG, Lieb JD. 2008. The genomic distribution and function of histone variant HTZ-1 during C. elegans embryogenesis. PLoS Genetics 4(9):e1000187.

Wilburn DB and Swanson WJ. 2016. From molecules to mating: Rapid evolution and biochemical studies of reproductive proteins. Journal of Proteomics 135:12-25.

Wolffe AP. 2001. Histone genes. In: Brenner's online encyclopedia of genetics, four-volume set. Elsevier Inc. 948 p.

Yun M, Wu J, Workman JL, Li B. 2011. Readers of histone modifications. Cell Research 21(4):564-78.

Figure Legends

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 red box: HTAS-1 has a phase 0 intron between 26th and 27th amino acid; HIS-35 has a phase 0 intron between 50th and 51st amino acid. HTZ-1 has a phase 2 intron splitting the 57th amino acid.

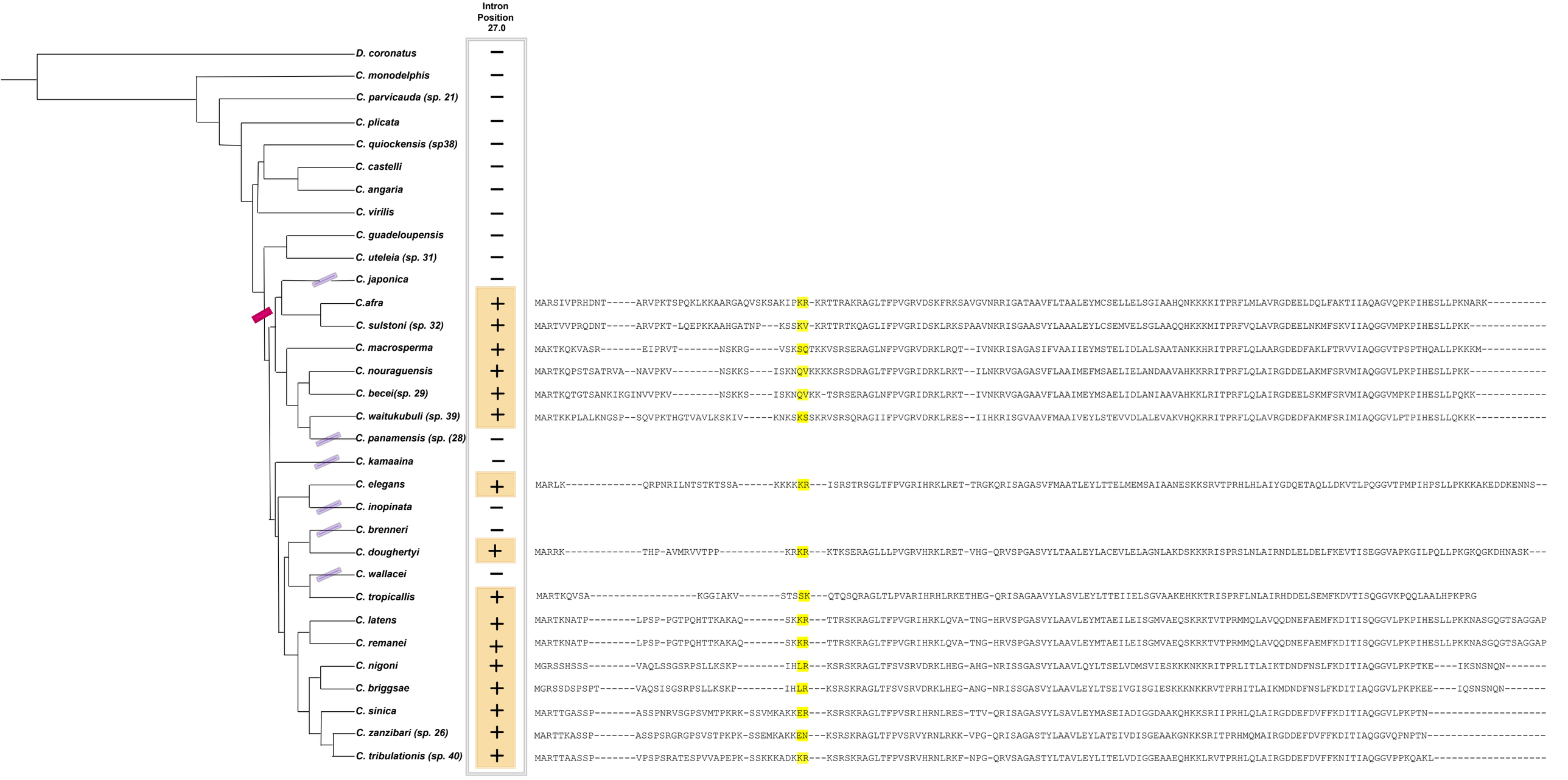
Fig. 2: Intron position and sequence evidence indicates the origin of HTAS-1 within *Caenorhabditis* evolution and subsequent loss. On the left is the previously reconstructed species tree of *Caenorhabditis*. The likely origin of HTAS-1 is indicated with a dark pink bar. HTAS-1 presence in a species is indicated by the plus sign whereas the losses have been shown with the minus sign. 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).

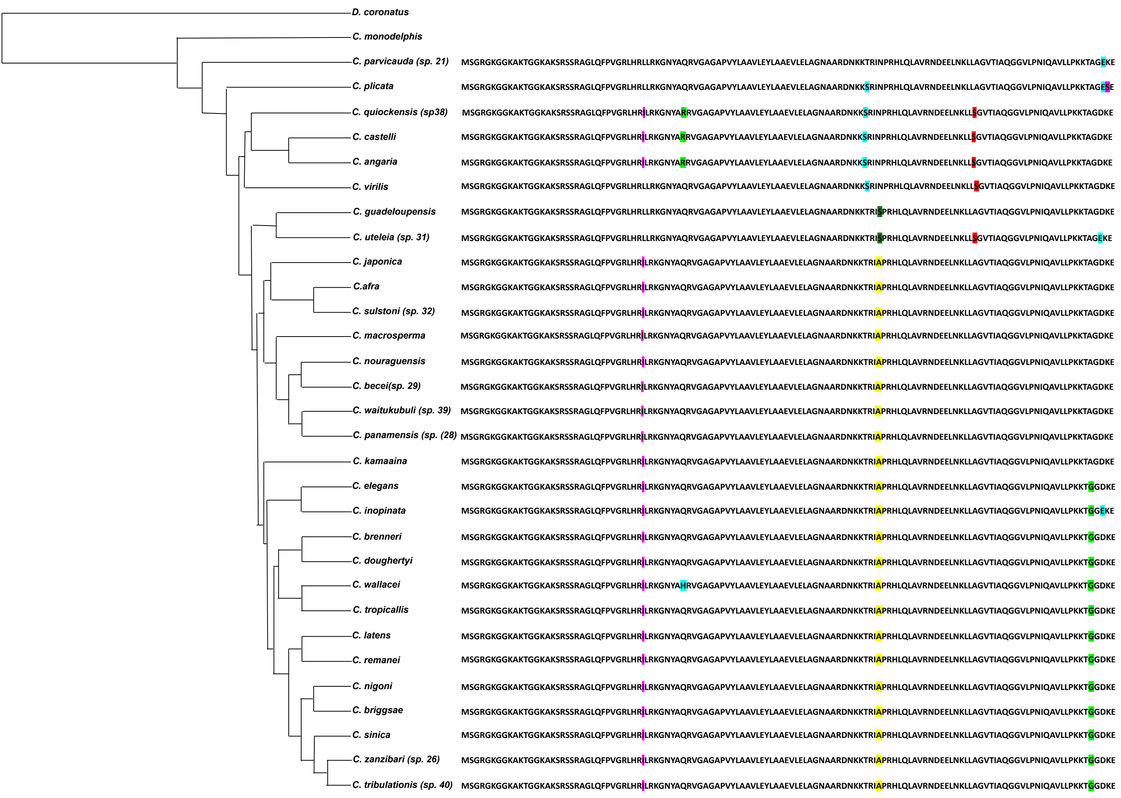
Fig. 4: The dynamic history of intron loss and gain in HTZ-1. On the left is the previously reconstructed species tree of *Caenorhabditis*. HTZ-1 characteristic intron presence and absence in a species is indicated by the plus and minus sign.

High resolution figures are pasted below. Figure 1-4 respectively

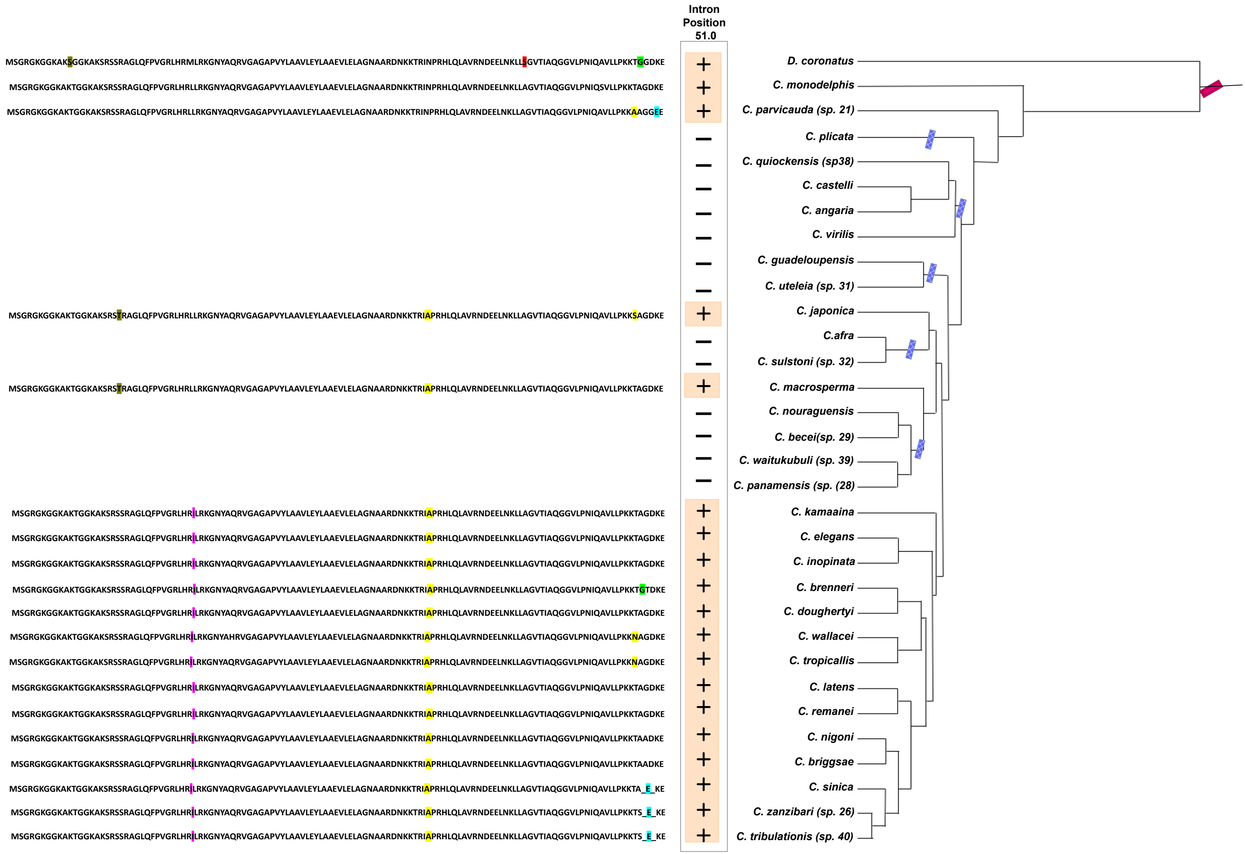
```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


HTAS-1          MARLKQRPNRILNTSTKTSSAKKKKRISRSTRSGLTFPVGRIHRKLRETTRGKQRISAG
H2A             -MSGRGKGGK----------AKTGGKAKSRSSRAGLQFPVGRLHRILRKGNYA-QRVGAG
HIS-35          -MSGRGKGGK----------AKTGGKAKSRSSRAGLQFPVGRLHRILRKGNYA-QRVGAG
HTZ-1           MAGGKGKAGK--------DSGKSKSKVVSRSARAGLQFPVGRIHRFLKQRTTSSGRVGAT
                    . . ..          .*.  *  ***:*:** *****:** *.: . .  *:.*

HTAS-1          ASVFMAATLEYLTTELMEMSAIAANESKKSRVTPRHLHLAIYGDQETAQLLDKVTLPQGG
H2A             APVYLAAVLEYLAAEVLELAGNAARDNKKTRIAPRHLQLAVRNDEELNKLLAGVTIAQGG
HIS-35          APVYLAAVLEYLAAEVLELAGNAARDNKKTRIAPRHLQLAVRNDEELNKLLAGVTIAQGG
HTZ-1           AAVYSAAILEYLTAEVLELAGNASKDLKVKRITPRHLHLAIRGDEELDTLI-KATIAGGG
                *.*: ** ****::*::*::. *:.: * .*::****:**: .*:*   *:  .*:. **

HTAS-1          VTPMPIHPSLLPKKKAKEDDKENNS-----
H2A             VLP-NIQAVLLPKK--TGGDKE--------
HIS-35          VLP-NIQAVLLPKK--TAGDKE--------
HTZ-1           VIP-HIHRYLMNKKGAPVPGKPGAPGQGPQ
                * *  *:  *: **     .*
```

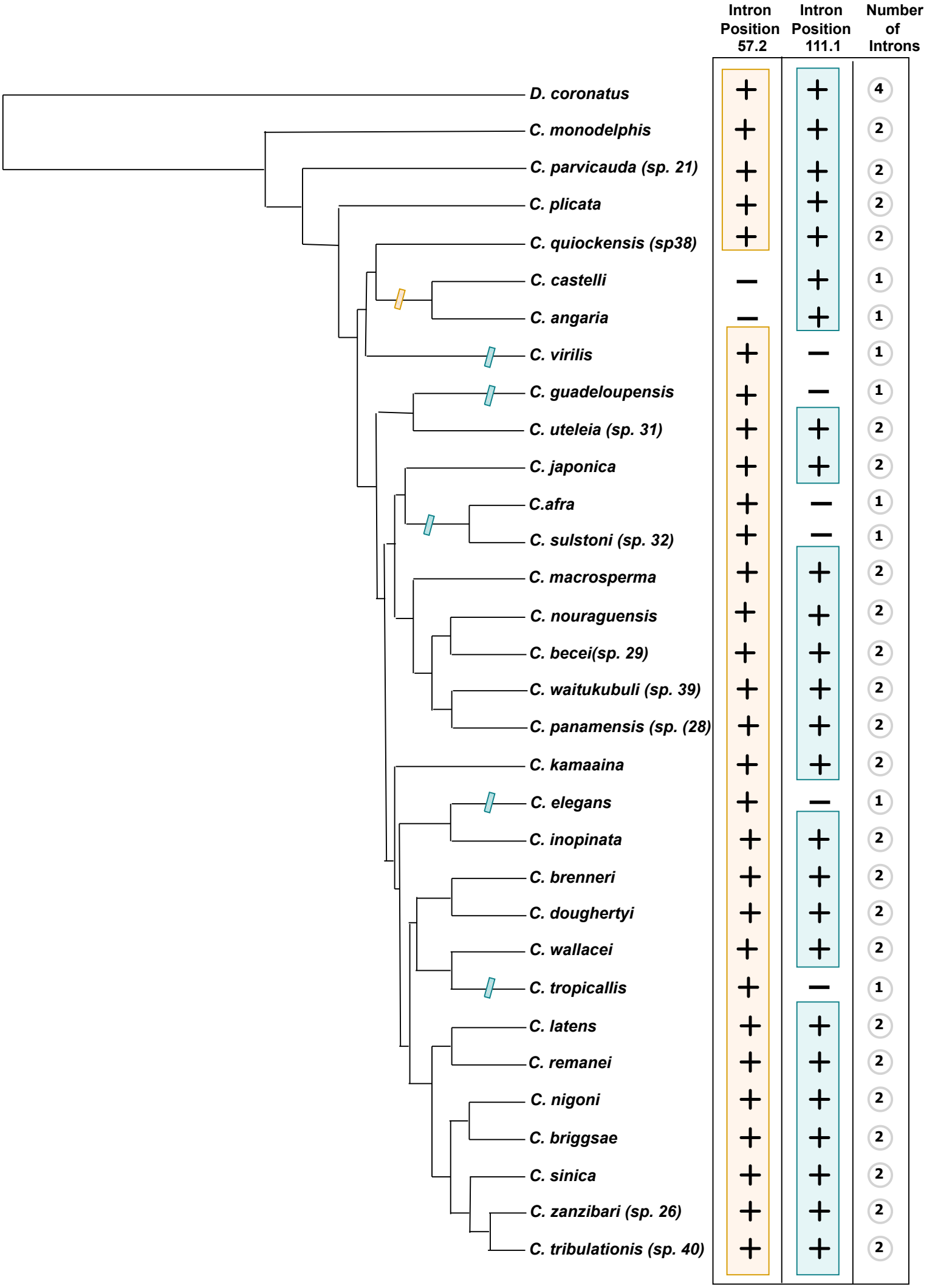

Intron Position 27.0
D. coronatus
C. monodelphis
C. parvicauda (sp. 21)
C. plicata
C. quiockensis (sp38)
C. castelli
C. angaria
C. virilis
C. guadeloupensis
C. uteleia (sp. 31)
C. japonica
C.afra
C. sulstoni (sp. 32)
C. macrosperma
C. nouraguensis
C. becei(sp. 29)
C. waitukubuli (sp. 39)
C. panamensis (sp. (28)
C. kamaaina
C. elegans
C. inopinata
C. brenneri
C. doughertyi
C. wallacei
C. tropicallis
C. latens
C. remanei
C. nigoni
C. briggsae
C. sinica
C. zanzibari (sp. 26)
C. tribulationis (sp. 40)
MARSIVPRHDNT-----ARVPKTSPQKLKKAARGAQVSKSAKIPKR-KRTTRAKRAGLTFPVGRVDSKFRKSAVGVNRRIGATAAVFLTAALEYMCSELLELSGIAAHQNKKKKITPRFLMLAVRGDEELDQLFAKTIIAQAGVQPKPIHESLLPKNARK----------
MARTVVPRQDNT-----ARVPKT-LQEPKKAAHGATNP---KSSKV-KRTTRTKQAGLIFPVGRIDSKLRKSPAAVNKRISGAASVYLAAALEYLCSEMVELSGLAAQQHKKKMITPRFVQLAVRGDEELNKMFSKVIIAQGGVMPKPIHESLLPKK----------
MAKTKQKVASR-------EIPRVT-------NSKRG-----VSKSQTKKVSRSERAGLNFPVGRVDRKLRQT--IVNKRISAGASIFVAAIIEYMSTELIDLALSAATANKKHRITPRFLQLAARGDEDFAKLFTRVVIAQGGVTPSPTHQALLPKKKM----------
MARTKQPSTSATRVA--NAVPKV--------NSKKS----ISKNQVKKKKSRSDRAGLTFPVGRIDRKLRKT--ILNKRVGAGASVFLAAIMEFMSAELIELANDAAVAHKKRRITPRFLQLAIRGDDELAKMFSRVMIAQGGVLPKPIHESLLPKK----------
MARTKQTGTSANKIKGINVVPKV--------NSKKS----ISKNQVKK-TSRSERAGLNFPVGRIDRKLRKT--IVNKRVGAGAAVFLAAIMEYMSAELIDLANIAAVAHKKLRITPRFLQLAIRGDEELSKMFSRVMIAQGGVMPKPIHESLLPQKK----------
MARTKKPLALKNGSP--SQVPKTHGTVAVLKSKIV-----KNKSKSSKRVSRSQRAGIIFPVGRVDRKLRES--IIHKRISGVAAVFMAAIVEYLSTEVVDLALEVAKVHQKRRITPRFLQLAVRGDEDFAKMFSRIMIAQGGVLPTPIHESLLQKKK----------
MARLK----------QRPNRILNTSTKTSSA------KKKKKR---ISRSTRSGLTFPVGRIHRKLRET-TRGKQRISAGASVFMAATLEYLTTELMEMSAIAANESKKSRVTPRHLHLAIYGDQETAQLLDKVTLPQGGVTPMPIHPSLLPKKKAKEDDKENNS--
MARRK----------THP-AVMRVVTPP----------KRKR---KTKSERAGLLLPVGRVHRKLRET-VHG-QRVSPGASVYLTAALEYLACEVLELAGNLAKDSKKKRISPRSLNLAIRNDLELDELFKEVTISEGGVAPKGILPQLLPKGKQGKDHNASK---
MARTKQVSA----------KGGIAKV-------STSSK---QTQSQRAGLTLPVARIHRHLRKETHEG-QRISAGAAVYLASVLEYLTTEIIELSGVAAKEHKKTRISPRFLNLAIRHDDELSEMFKDVTISQGGVKPQQLAALHPKPRG
MARTKNATP--------LPSP-PGTPQHTTKAKAQ-------SKKR---TTRSKRAGLTFPVGRIHRKLQVA-TNG-HRVSPGASVYLAAVLEYMTAEILEISGMVAEQSKRKTVTPRMMQLAVQQDNEFAEMFKDITISQGGVLPKPIHESLLPKKNASGQGTSAGGAP
MARTKNATP--------LPSP-PGTPQHTTKAKAQ-------SKKR---TTRSKRAGLTFPVGRIHRKLQVA-TNG-HRVSPGASVYLAAVLEYMTAEILEISGMVAEQSKRKTVTPRMMQLAVQQDNEFAEMFKDITISQGGVLPKPIHESLLPKKNASGQGTSAGGAP
MGRSSHSSS--------VAQLSSGSRPSLLKSKP--------IHLR---KSRSKRAGLTFSVSRVDRKLHEG-AHG-NRISSGASVYLAAVLQYLTSELVDMSVIESKKKNKKRITPRLITLAIKTDNDFNSLFKDITIAQGGVLPKPTKE----IKSNSNQN-------
MGRSSDSPSPT------VAQSISGSRPSLLKSKP--------IHLR---KSRSKRAGLTFSVSRVDRKLHEG-ANG-NRISSGASVYLAAVLEYLTSEIVGISGIESKKKNKKRVTPRHITLAIKMDNDFNSLFKDITIAQGGVLPKPKEE----IQSNSNQN-------
MARTTGASSP-------ASSPNRVSGPSVMTPKRK-SSVMKAKKER---KSRSKRAGLTFPVSRIHRNLRES-TTV-QRISAGASVYLSAVLEYMASEIADIGGDAAKQHKKSRIIPRHLQLAIRGDDEFDVFFKDITIAQGGVLPKPTN----------
MARTTKASSP-------ASSPSRGRGPSVSTPKPK-SSEMKAKKEN---KSRSKRAGLTFPVSRVYRNLRKK-VPG-QRISAGASTYLAAVLEYLATEIVDISGEAAKGNKKSRITPRHMQMAIRGDDEFDVFFKDITIAQGGVQPNPTN----------
MARTTAASSP-------VPSPSRATESPVVAPEPK-SSKKKADKKR---KSRSKRAGLTFPVSRVHRNLRKF-NPG-QRVSAGASTYLTAVLEYLITELVDIGGEAAEQHKKLRVTPRHLQLAIRGDDEFDVFFKNITIAQGGVPPKQAKL----------

H2A


D. coronatus
C. monodelphis
C. parvicauda (sp. 21)
C. plicata
C. quiockensis (sp38)
C. castelli
C. angaria
C. virilis
C. guadeloupensis
C. uteleia (sp. 31)
C. japonica
C.afra
C. sulstoni (sp. 32)
C. macrosperma
C. nouraguensis
C. becei(sp. 29)
C. waitukubuli (sp. 39)
C. panamensis (sp. (28)
C. kamaaina
C. elegans
C. inopinata
C. brenneri
C. doughertyi
C. wallacei
C. tropicallis
C. latens
C. remanei
C. nigoni
C. briggsae
C. sinica
C. zanzibari (sp. 26)
C. tribulationis (sp. 40)

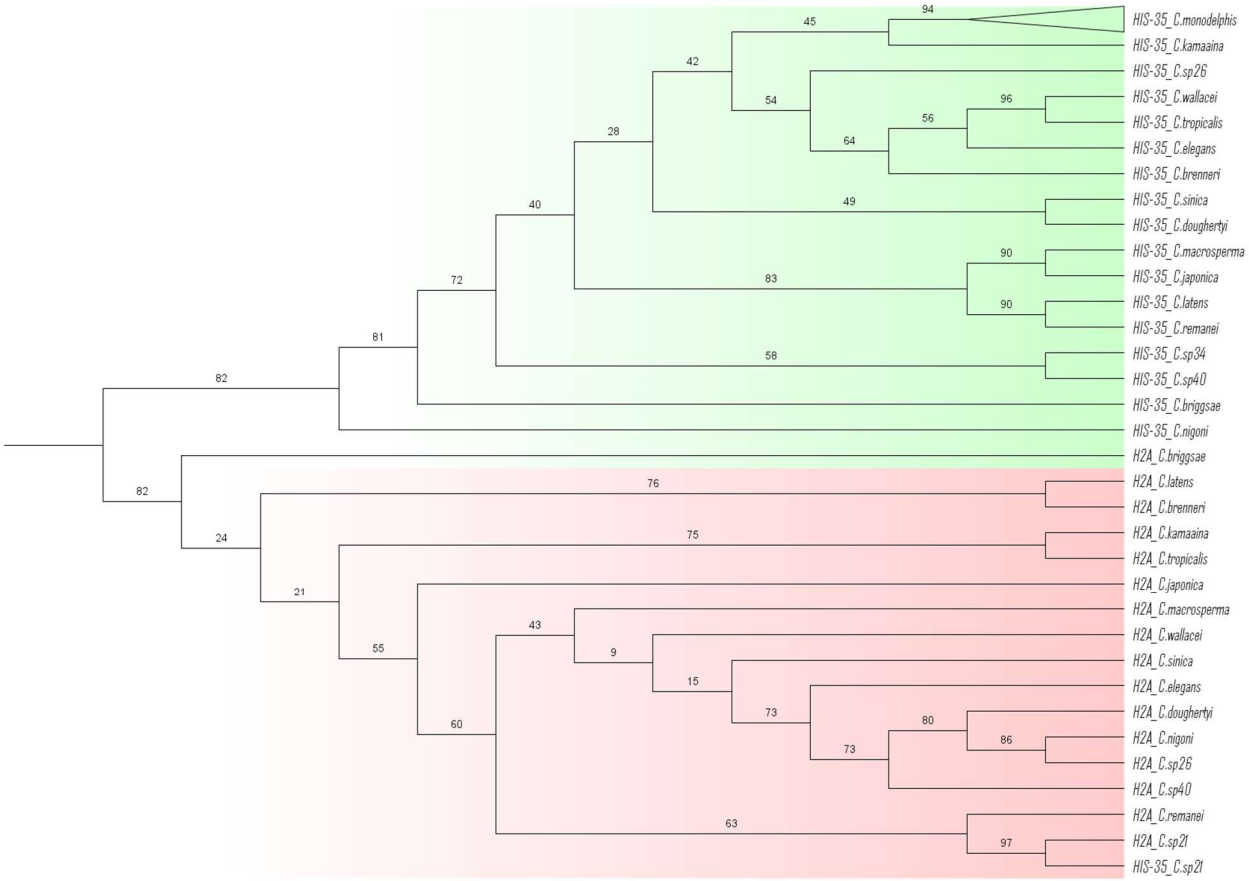
HIS-35


Intron Position 51.0
D. coronatus
C. monodelphis
C. parvicauda (sp. 21)
C. plicata
C. quiockensis (sp38)
C. castelli
C. angaria
C. virilis
C. guadeloupensis
C. uteleia (sp. 31)
C. japonica
C.afra
C. sulstoni (sp. 32)
C. macrosperma
C. nouraguensis
C. becei(sp. 29)
C. waitukubuli (sp. 39)
C. panamensis (sp. (28)
C. kamaaina
C. elegans
C. inopinata
C. brenneri
C. doughertyi
C. wallacei
C. tropicallis
C. latens
C. remanei
C. nigoni
C. briggsae
C. sinica
C. zanzibari (sp. 26)
C. tribulationis (sp. 40)


Intron Position 57.2
Intron Position 111.1
Number of Introns
D. coronatus
C. monodelphis
C. parvicauda (sp. 21)
C. plicata
C. quiockensis (sp38)
C. castelli
C. angaria
C. virilis
C. guadeloupensis
C. uteleia (sp. 31)
C. japonica
C.afra
C. sulstoni (sp. 32)
C. macrosperma
C. nouraguensis
C. becei(sp. 29)
C. waitukubuli (sp. 39)
C. panamensis (sp. (28)
C. kamaaina
C. elegans
C. inopinata
C. brenneri
C. doughertyi
C. wallacei
C. tropicallis
C. latens
C. remanei
C. nigoni
C. briggsae
C. sinica
C. zanzibari (sp. 26)
C. tribulationis (sp. 40)

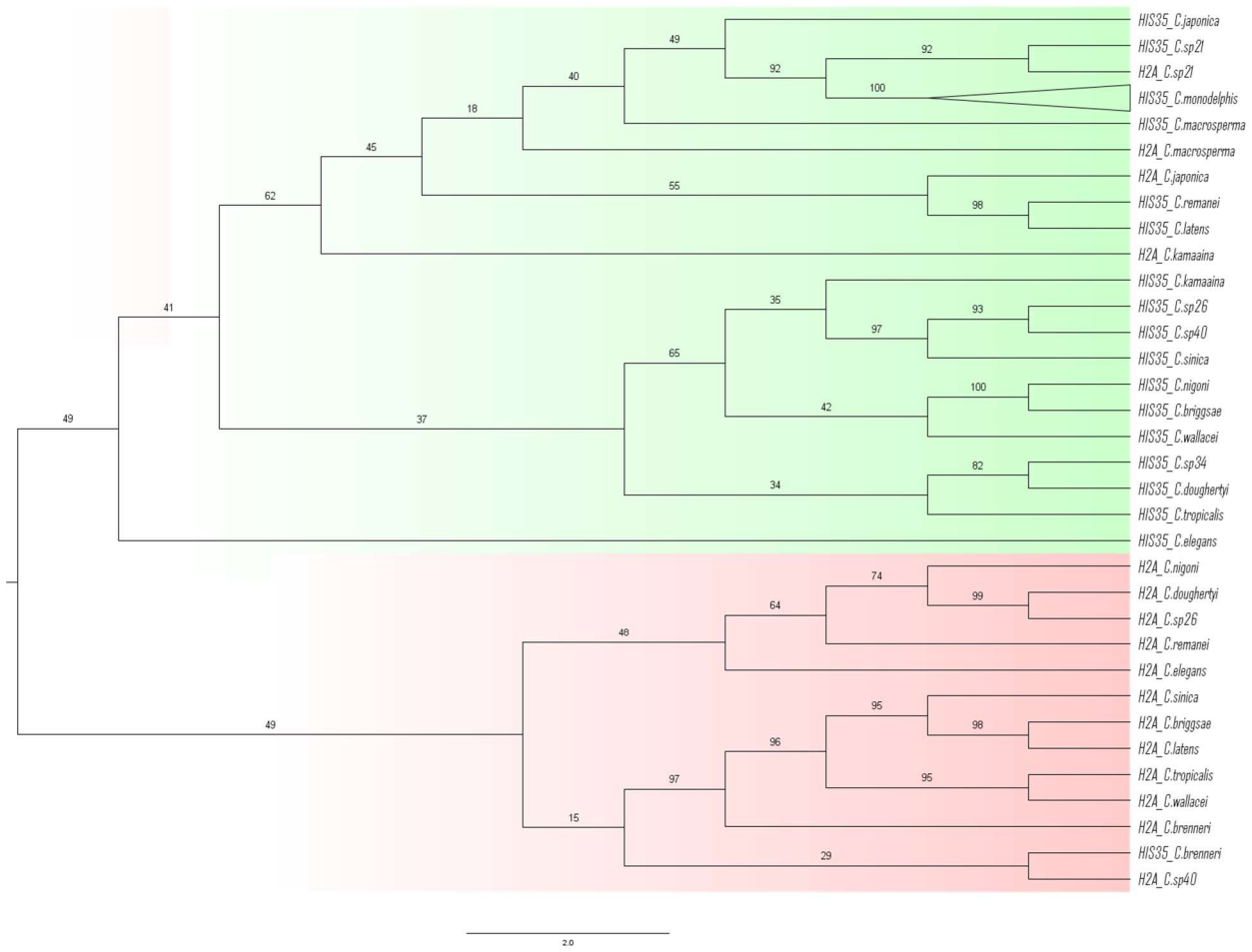
Supplementary Fig. 1

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”.



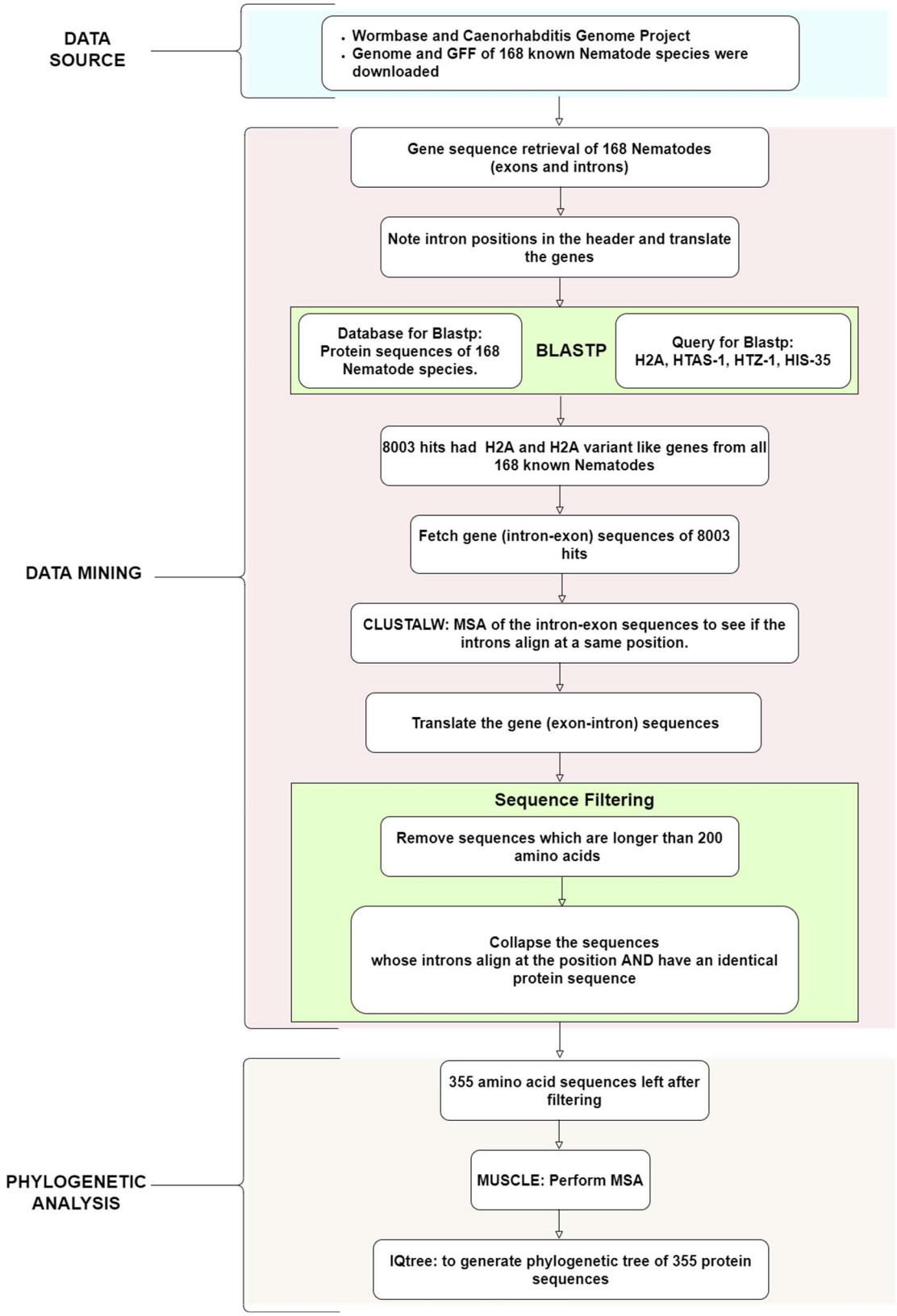
Supplementary Fig. 2

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