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

Histone variants are paralogs that replace canonical histones in nucleosomes, often imparting novel functions. Despite their importance, 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 108 nematode genomes to trace the evolutionary histories of the three H2A variants found in *Caenorhabditis elegans*: the ancient H2A.Z^{HTZ-1}, the sperm-specific HTAS-1, and HIS-35, which differs from canonical H2A by a single glycine-to-alanine C-terminal change. We find disparate evolutionary histories. Although the H2A.Z^{HTZ-1} protein is highly conserved, its gene exhibits recurrent intron gain and loss. This pattern suggests that it is intron presence, rather than specific intron sequences or positions, that may be important to H2A.Z functionality. In contrast, for HTAS-1 and HIS-35, we find variant-specific intron positions that are 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. HIS-35 exhibits gene retention in some descendent lineages but also recurrent gene loss in others, suggesting that histone variant use or functionality is highly flexible in this case. We also find that the single amino acid differentiating HIS-35 from core H2A is ancestral and common across canonical *Caenorhabditis* H2A sequences and identify one nematode species that bear identical HIS-35 and canonical H2A proteins, findings that are not predicted from the hypothesis that HIS-35 has a distinct function. Instead, we speculate that HIS-35 enables H2A expression across the cell cycle or in distinct tissues; genes encoding such partially-redundant functions may be advantageous yet relatively replaceable over evolutionary times, consistent with the patchwork pattern of retention and loss of both genes. Our study shows the evolutionary trajectory for histone H2A variants with distinct functions 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, and which wrap around ~147 bps of DNA (1-4). In many species, genes encoding canonical core histones (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 their rapid, coordinated expression (4-7). The

45 expression of canonical histones is tightly coupled with the S-phase of the cell cycle
46 because of the critical need for the large bulk of histone proteins required to package and
47 compact the newly synthesized DNA (8-10).

48 Histones have evolved variant forms that further regulate chromatin compaction and
49 affect processes like transcription, DNA repair, and development (11-20). Such histone
50 variants are often considered to be a part of the 'histone code' and control distinct sets of
51 genes during specific times and tissues in both animals and plants (11, 12, 21-23). Among
52 all the core histones, H2A is the fastest evolving histone, showing the most diversity in the
53 variant types and expression patterns (24, 25). For example, the highly conserved H2A.Z
54 is present in every eukaryotic species and ubiquitously expressed. It has been shown in
55 different species to play a variety of roles including transcriptional regulation,
56 heterochromatin boundaries, DNA repair, DNA replication, and dosage compensation (26-
57 31). In contrast, H2A.X has arisen independently in several lineages, with functions in
58 DNA damage response and transcription (18, 31, 32), while macroH2A, which is involved
59 in X inactivation and stress response, is present only in some lineages, making its origins
60 unclear (3, 4, 24, 33-36). Finally, short histone variants, such as H2A.B, H2A.P, H2A.L,
61 and H2A.Q, are expressed in testis (except H2A.B which is also expressed in the brain
62 cells) (25, 37-40). In short, a great wealth of histone H2A variants are observed, ranging
63 from those shared across all eukaryotes to others that are species-specific (3, 4, 18, 19, 35).
64 H2A variants also have distinct functions in a broad range of processes and can be
65 ubiquitously expressed or expressed only in certain tissues (11, 25, 41, 42). However, with

66 the exception of the well-studied ancient variant H2A.Z, how histone H2A variants have
67 arisen and evolved remains unknown.

68 Clues to how histone variants differ may stem from the distinct gene structures and
69 expression patterns compared to canonical histones. Unlike canonical histone genes,
70 variant histone gene expression is not restricted to S phase expression and can be tissue-
71 specific. Most histone variants are typically found in a single copy in the genome and
72 contain introns in their pre-mRNAs (5-7, 43, 44). Notably, the roles of these introns in
73 variant function remain poorly studied, including whether the introns themselves impart
74 novel functionality. For instance, sequences found within introns of other genes can
75 regulate gene expression (43, 45-48). Introns can also allow for alternative splicing, which
76 is common in animals and plays important role in development and disease (49-52). Introns
77 have been used in other studies to determine the evolutionary trajectory of other protein
78 families (53-57). Thus, the presence of introns in histone variants may provide useful clues
79 to exploring the histone evolution and function across closely related species.

80 Core histone functions are expected to be highly conserved across eukaryotes, given
81 their central roles in ensuring DNA packaging and protection (4, 61). Thus, observed
82 protein changes in core histones are expected to be largely neutral with respect to protein
83 function. On the other hand, amino acid differences between variant histones and core
84 histones are thought to generally lead to functional differences (3, 4, 62). Indeed, protein
85 sequence differences between well-studied variants and their core homologs have been
86 shown to affect chromatin structure and function. The histone H3.3 variants, for example,
87 which differ from each other by 4-5 amino acids, play distinct roles in transcriptional

88 activation, chromatin remodeling, heterochromatin formation, and development (14, 16,
89 63-66). On the other hand, the functional significance of protein differences between
90 variant histones and their core counterparts in some cases is unclear. For instance, the
91 protein sequence of *C. elegans* HIS-35 differs by a single 'A' instead of a 'G' at the 124th
92 position from the S-phase *C. elegans* H2A, the functional significance of which is
93 unknown.

94 In this study, we focus on the evolutionary histories of the three very different H2A
95 variants found in *Caenorhabditis elegans:* H2A.Z^{HTZ-1}, HIS-35, and HTAS-1 (Figure 1).
96 H2A.Z^{HTZ-1} is an ortholog of the evolutionarily conserved variant, H2A.Z. HIS-35
97 intriguingly differs from core H2A by a single amino acid difference. HTAS-1 shows a
98 greater divergence from core H2A, particularly at the highly divergent C and N termini,
99 and appears to be expressed only in sperm and has only been reported in *C. elegans* to date
100 (67). The presence of a discrete number of H2A variants with distinct functions and/or
101 expression patterns in combination with the availability of sequence data across a large
102 number of nematode species allows a unique opportunity to track the evolutionary
103 trajectory of this histone variant family.

104 **RESULTS**

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

106 Taking advantage of vast sequencing (68-70) efforts across nematode species, we
107 used BLAST searches across 108 reported nematode genomes to identify all annotated

108 copies of H2A and H2A-related gene variants across the diversity of nematodes. After
109 filtering and collapsing identical proteins, we were left with 408 unique sequences. We
110 then used standard phylogenetic methods to reconstruct the evolutionary history of these
111 sequences (Figure S1, intronless (thus putative core or canonical) H2As are highlighted in
112 yellow, HTZ-1 homologs are highlighted in pink, putative HIS-35 homologs are
113 highlighted in blue, and putative HTAS-1 homologs are highlighted in gray). However,
114 scrutiny of the recovered phylogenetic tree revealed several bizarre findings. For instance,
115 core H2A proteins were grouped as separate clades that included very deeply-diverged
116 nematode sequences; on the other hand, species-specific variants were often found grouped
117 far from core proteins from the same or related species. Perhaps the clearest case can be
118 shown by performing phylogenetic reconstruction on core H2A orthologs and on the HIS-
119 35 orthologs identified based on intron positions (described in detail below). Here, we
120 expect distinct clades of H2A and HIS-35 sequences, yet we recover no such separate
121 clades (Figure S2, HIS-35 is highlighted).

122 Some of these anomalies are as expected by errors in phylogenetic reconstruction
123 due to model misspecification (the phenomenon in which differences between the assumed
124 model of sequence evolution and the actual evolutionary process led to errors in
125 phylogenetic reconstruction) (71). The possibility of model misspecification is increased
126 in the case of histone genes by several factors: extreme differences in rates across sites
127 (some sites show conservation across eukaryotes, others are variable within
128 *Caenorhabditis*); large relative differences in rates across gene lineages (the extremely

129 slow evolution of core proteins, but substantially more change in some variants); and the
130 generally small number of total sites given the short length of histone proteins.

131 **Intron position and phase distinguish the three H2A variants of *C. elegans*.**

132 As an alternative approach, we used another potential source of phylogenetic
133 information: the position of the spliceosomal introns that disrupt nuclear genes, including
134 variant histones. Intron positions can be conserved over very long times in orthologous
135 genes (53-56, 58, 59). Whereas early work considered the possibility of intron 'sliding',
136 in which an intron would migrate a few base pairs along a gene, recent work has shown
137 that intron sliding is a very rare occurrence, and thus that intron positions are very often
138 conserved over very long times (54, 59). For example, in *Theileria* apicomplexans, 99.7%
139 of intron positions are conserved between *T. parva* and *T. annulata*, diverging roughly 82
140 million years ago (53). In mammals, 99.9% of intron positions are conserved between
141 humans and dogs, diverging around 100 million years ago (55).

142 We began our study of intron-exon structures in H2A variants by obtaining intron-
143 exon structures for all H2A gene family members and performing alignments to determine
144 intron position sharing across genes. We first aligned the three main *C. elegans* H2A
145 variants (Figure 1), HIS-35, HTAS-1, and H2A.Z^{HTZ-1}, each of which has a single intron
146 position (highlighted boxes in Figure 1). Scrutiny of the intron positions showed that the
147 three genes contain introns at different positions, differing in the codon that they interrupt
148 or and/or in the phase at which they disrupt the codon. Interestingly, the intron position in

149 HIS-35 falls very near to that found in H2A.Z^{HTZ-1}. However, these introns are unlikely to
150 represent a shared intron, for two reasons. First, intron positions rarely slide between
151 phases (54, 59). Second, HIS-35's near identity to H2A (see below) strongly suggests that
152 it more likely evolved from intronless H2A by gene duplication and intron gain, and not
153 from H2A.Z^{HTZ-1}. Thus, the differences in the intron positions between the three *C. elegans*
154 variants suggest that the three variants evolved independently as different duplicates of
155 core H2A, rather than from each other.

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

158 The *C. elegans* sperm-specific H2A variant, HTAS-1, contains a single intron between
159 codons 26 and 27 ("phase 0"; Figure 1, highlighted with yellow). Alignment across all the
160 H2A variants of 108 nematodes revealed 16 genes that share an intron at the exact
161 homologous position as *C. elegans* HTAS-1 (Figure 2) from species falling within a single
162 subclade of 22 species (represented by the red hash mark on the tree branch in Figure 2)
163 within the *Caenorhabditis* genus, suggesting an origin of this intron position within the
164 common ancestor of these species. Scrutiny of the sequence gene tree (Figure S1,
165 highlighted in gray, IP_152) revealed that this same set of genes (i.e., those sharing the
166 intron) appears together as a clade, reinforcing the notion that the genes sharing the HTAS-
167 1 intron position represent a set of orthologs. We next sought to identify potential HTAS-
168 1 orthologs in additional *Caenorhabditis* species, both the 6/22 within the HTAS-1-
169 containing subclade that lack intron-containing HTAS-1 candidates (marked by purple

170 hash on the tree branch in Figure 2) as well as the 10 species outside this subclade. No
171 genes from any such species were found grouping with the putative HTAS-1 ortholog
172 clade; thus, overall, we couldn't find any other genes that are candidates for HTAS-1
173 orthologs. In total, the data is consistent with the origin of HTAS-1 and its gene-specific
174 intron occurring in the same ancestor of a subset of *Caenorhabditis* species.

175 As noted above, in six species within the HTAS-1-containing clade of
176 *Caenorhabditis* (*C. wallacei*, *C. brenneri* and *C.inopinata* (*sp34*), *C. kamaaina*, *C.*
177 *panamensis* (*sp28*), *and C. japonica*), we could not find annotated genes with either an
178 intron position at the HTAS-1-specific position (Figure 2, marked with purple hashes on
179 the tree branch) or other genes that grouped with the putative HTAS-1 orthologs (Figure
180 S1). However, we were unable to distinguish whether these absences represent true
181 biological loss or failure of gene annotation, which is particularly possible given that
182 histone genes' short lengths increase the possibility they would be missed by standard
183 annotation methods. Indeed, we identified one species harboring a likely HTAS-1
184 sequence for which only the second exon was annotated (Figure S1), however, TBLASTN
185 searches against the genome failed to confidently detect this gene. This is likely because,
186 given the much higher rate of evolution of HTAS-1, HTAS-1 from one species often shows
187 a greater degree of core H2A from another species than to its HTAS-1 ortholog;
188 consequently, TBLASTN searches using HTAS-1 from one species typically give many
189 hits, with the likely HTAS-1 not being within the top dozen hits, making it difficult to
190 distinguish gene absence from the failure of annotation. In total, then, the data is consistent

191 with a single origin of HTAS-1 and its characteristic intron position within the ancestor of
192 a subset of studied *Caenorhabditis* species, but we cannot determine with certainty whether
193 HTAS-1 has been lost in up to six independent lineages (Figure 2).

194 **Intron position conservation suggests HIS-35 originated in the *Caenorhabditis-***
195 ***Diploscapter* ancestor and has undergone subsequent gene retention and loss**

196 The variant HIS-35 of *C. elegans* has a phase zero intron between the 50th and the 51st
197 codon (Figure 1). Alignment across all H2A variants revealed 20 species with genes that
198 share this intron position (Figure 3, marked with a plus sign). In all cases, these genes
199 show very high sequence similarity with H2A at the protein level. These genes are from
200 species falling in the clades of *Caenorhabditis* and its sister genus *Diploscapter*,
201 representing 20/32 species within this clade, suggesting an origin of this intron position
202 within the common ancestor of *Diploscapter* and *Caenorhabditis* (Figure 3).

203 We sought to determine whether the 12 descendent species without a clear candidate
204 HIS-35 gene represent true losses or merely failures of gene annotation. TBLASTN and
205 BLASTN searches failed to recover likely HIS-35 orthologs absent from the annotation.
206 In contrast to the case of HTAS-1 above, we are confident in the ability of these methods
207 to detect HIS-35 orthologs, since we were able to detect the annotated HIS-35 ortholog by
208 identical methods (i.e., the positive control). For instance, using *C. japonica* HIS-35 as a
209 query, BLASTN searches revealed the (annotated) putative ortholog in *C. macrosperma*,
210 but did not reveal unannotated candidate orthologs in *C. nouroguensis*, *C. waitukubuli, C.*

211 *panamensis*, or *C. becei*, which are equally closely related to *C. japonica*. In addition,
212 whereas species lacking HTAS-1 appear to be randomly scattered across the HTAS-1-
213 containing clade, as expected by random gene annotation failures, species lacking HIS-35
214 within the HIS-35-containing clade group in subclades. This is as expected by true
215 biological loss and not by random annotation failures. Thus, we conclude that HIS-35 has
216 been lost some 5 times independently in different *Caenorhabditis* lineages.

217 **Conservative protein evolution of HIS-35 and evidence for occasional gene conversion**
218 **with core H2A**

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

225 However, when we looked at position 124 the canonical H2A sequences of all the
226 *Caenorhabditis* species we actually found out that the "A" is ancestral to canonical H2A
227 (Figure 3). This conservation of the "A" at position 124 suggests that HIS-35 likely has not
228 diverged in function from the ancestral H2A despite this variant being retained across many
229 *Caenorhabditis* species. We also found that the predicted protein sequences of HIS-35 and
230 H2A of species *C. kamaaina* are exactly the same (Figure 3). These findings are not as

231 expected if canonical H2A and HIS-35 proteins have different functions, but are as
232 expected if the proteins are functionally redundant.

233 The multiple copies of core histone genes are known to undergo so-called concerted
234 evolution, with sequences being transferred between gene copies by recombination (72-
235 75). We, therefore, wondered whether concerted evolution plays a role in the identical
236 protein sequence changes observed in the H2A and HIS-35 paralogs of some species.
237 Under concerted evolution, the two sequences undergoing concerted evolution are
238 homogenized (one overwrites the other). Consequently, the prediction is that such events
239 should lead the interconverting partners to the group together on a phylogenetic tree. To
240 search for evidence of concerted evolution, we reconstructed separate phylogenetic trees
241 of exon-1 and exon-2 for all H2A and HIS-35 sequences. Most of the reconstructed tree
242 largely reflected the species tree, suggesting against the possibility of widespread concerted
243 evolution (Figures S3 and S4). However, we did observe the grouping of the two gene
244 sequences for *Caenorhabditis sp21*, consistent with the concerted evolution of HIS-35 and
245 H2A in only this species. Concerted evolution of these genes is consistent with a lack of
246 functional differentiation, though admittedly the low rate of such events weakens the
247 strength of this argument.

248 **The intron loss and gain in HTZ-1 show a dynamic evolutionary history**

249 The ubiquitously expressed *C. elegans* H2A variant H2A.Z^{HTZ-1} is the ortholog of
250 H2A.Z, which is evolutionary conserved across all the eukaryotes (3, 76). *C. elegans*

251 H2A.Z^{HTZ-1} has an intron that splits the 57th codon at position 2 (Figure 1). Within the
252 alignment across all H2A variants, we searched for genes that share the single intron
253 position of *C. elegans* H2A.Z^{HTZ-1}, revealing 30 genes share this position (Figure 4, marked
254 with a plus sign; Figure S1). These are putative H2A.Z^{HTZ-1} orthologs; consistent with this
255 hypothesis, they grouped together on the tree. Unexpectedly, we found that 22 of these 30
256 putative H2A.Z^{HTZ-1} genes have two (or more in *Diploscapter*) introns in their genes, one
257 at position 57.2 and the other at position 111.1. Both introns have been repeatedly
258 individually lost in different lineages (including in the lineage leading to the single-intron
259 *C. elegans* H2A.Z^{HTZ-1} gene). Species including *C. elegans*, *C. tropicallis*, *C. sp32*, *C. afra*,
260 *C. guadeloupensis*, *C. virilis* have lost the second H2A.Z intron which is at position 111.1,
261 whereas *C. casteli* and *C. angaria* have lost their first intron. These results are consistent
262 with general results in protein-coding genes, wherein intron loss is common across the
263 *Caenorhabditis* phylogeny (77, 78). Nonetheless, the finding that general trends of intron
264 loss may equally apply to histone variant genes is important in understanding the functional
265 implications of variant introns.

266 Intron presence is a conspicuous difference between core and variant histones,
267 raising the question of the functional significance of variant histone introns: specifically,
268 are variant histone introns important for differentiating the specific functions of those
269 variants from their core paralogs? The finding of recurrent loss of introns from a variant
270 gene suggests that the specific positions or sequences of introns in a particular histone
271 variant gene may not be of particular importance. Nonetheless, it is of note that all observed

272 H2A.Z orthologs contain at least one intron, suggesting that the presence of at least one
273 intron, at whatever position, could be important for efficient expression of variant histone
274 genes, consistent with their expression through canonical gene expression pathways, in
275 which intron presence often promotes gene expression.

276 **Discussion**

277 *Introns as sources of phylogenetic information*

278 This study is among the first to leverage the information present in intron positions
279 to decipher the evolutionary history of histone variants (60). Previous studies have shown
280 intron position conservation among widely diverged eukaryotic species (53-55, 57-60). For
281 instance, intron positions are highly conserved between humans, mice, and fish (57). Thus,
282 intron positions contain a record of evolutionary history that can facilitate insights into
283 gene history. The utility of introns here is different for the different variants and allows
284 different questions to be addressed. The clearest case comes from HIS-35, in which nearly
285 complete evolutionary conservation of H2A and HIS-35 sequences leads to a lack of
286 phylogenetic signal within proteins. At the other end of the spectrum lies HTAS-1, which
287 shows much more rapid sequence evolution. However, here, the discrepancy between rates
288 between HTAS-1 and H2A, and likely between different sites in the two proteins (i.e., the
289 N and C termini are highly constrained in H2A but fast-evolving for HTAS-1) makes it
290 impossible to define a single evolutionary model across the gene family. As expected by
291 general long branch attraction consideration, this leads to fast-evolving HTAS-1

292 incorrectly grouping far away from *Caenorhabditis* core H2A sequences. This generally
293 undermines our confidence in sequence-based phylogenetic reconstruction of HTAS-1
294 genes. Ironically, intron-defined HTAS-1 orthologs do group as a clade, indicating that
295 sequence-based phylogenetic reconstruction was likely successful for this group; however,
296 without the intron position information, our general lack of confidence in the methods'
297 success for HTAS-1 would lead us to question this finding. Thus, for this case, even though
298 the sequence-based phylogenetic methods apparently correctly identified the HTAS-1
299 clade, the orthologous information from intron positions was necessary for us to be
300 confident in the sequence-based phylogeny. Importantly, intron positions were
301 indispensable in distinguishing the origins of HTAS-1. Because HTAS-1 arose in an
302 ancestor containing two genes encoding nearly identical proteins (H2A and HIS-35), it is
303 very difficult to determine whether H2A evolved from the preexisting variant HIS-35 or
304 *de novo* from H2A. The fact that all candidate HTAS-1 genes lack the HIS-35 intron
305 strongly suggests *de novo* evolution from H2A, not HIS-35.

306 *Evidence from phylogenetic analysis for functional significance of recently-evolved histone*
307 *variants.*

308 In addition to tracing the origins and subsequent history of gene loss and retention,
309 our results provide insights into the possible functions of histone H2A variants. For
310 example, HIS-35 differs by a single amino acid from the S-phase H2A ('A' instead of 'G'
311 at the 124th position). Given that characterized histone variants are thought to largely
312 represent functionally distinct proteins, one hypothesis is that this single difference

313 functionally differentiates HIS-35 protein from H2A. One way that a single amino acid
314 change could have outsize effects is through altering the landscape of posttranslational
315 modifications, which are key to histone function. For instance, this is the case with H3
316 variants, H3.1 and H3.2 which differ from one another by single amino acid and show
317 distinct patterns of expression and post-translational modifications (64, 65).

318 Given it is possible for a small amino acid difference to change function, we
319 approached the single H2A/HIS-35 difference with the hypothesis that the single difference
320 could lead to functionally different proteins. However, when we looked at the H2A
321 sequences of all the *Caenorhabditis* species, we found an 'A' at position 124 to be
322 ancestral. Considering the presence of an A at position 124 in other canonical H2As,
323 suggests the variant HIS-35 might have the same function as the canonical H2A. Lack of
324 functional differentiation is also consistent with the similarity of A and G amino acids. This
325 hypothesis is also supported by the case of *C. kamaanina*, in which the encoded HIS-35
326 and H2A protein sequences are exactly the same. While these results do not disprove the
327 hypothesis that H2A and HIS-35 encode proteins with important functional differences
328 (except in the case of *C. kamaanina*), we propose instead that HIS-35 protein's functional
329 importance lies in allowing for the expression of a protein with overlapping or redundant
330 functions to canonical H2A that is not restricted to S-phase, as is the case with canonical
331 H2A. This could allow expression in differentiated cells that do not undergo mitosis and
332 enable tissue-specific expression. Such a potential semi-redundancy could help to explain
333 the ambivalent phylogenetic pattern, in which retention of HIS-35 in most species suggests

334 functional importance whereas loss in 5 independent lineages suggests conditional
335 expendability. Interestingly, a similar pattern of lineage-specific loss has been observed
336 for H2B variants, which has also been interpreted as encoding functionally important but
337 partially redundant functions (79).

338 *Function in reproduction for the sperm-specific variant HTAS-1*

339 Sperm-specific proteins show generally elevated rates of evolution, consistent with
340 strong selection on sperm functions because of sperm competition (67, 80). The current
341 data show that this is decisively the case for the sperm-specific variant HTAS-1. We show
342 that *C. elegans* HTAS-1's greater divergence from core H2A is not because of differences
343 in an evolutionary age, but very much despite it: HTAS-1 is most divergent variant protein
344 despite being the most recent to diverge from core H2A, having since evolved at a rate
345 many times higher than any other H2A paralog. This high rate of evolution strongly
346 suggests that HTAS-1 may be adapted to play roles important for reproduction in some
347 species.

348 However, our data also have somewhat ambiguous implications for HTAS-1
349 function. On the one hand, though HTAS-1 has been maintained for long periods of time
350 in many lineages it may have also been lost in multiple independent lineages, as found for
351 HIS-35. While rapid evolution of HTAS-1 made it impossible to exclude the possibility
352 that apparent HTAS-1 losses actually represent gene annotation failures, the fact that

353 careful scrutiny revealed no gene annotation failures for HIS-35 weighs against the
354 possibility of annotation failure, suggesting real loss of HTAS-1.

355 On the other hand, the much larger degree of protein sequence difference between
356 HTAS-1 and H2A would seem to decrease the probability that HTAS-1 protein is
357 functionally identical to H2A protein particularly given the extended C & N terminus of
358 HTAS-1 which has previously been reported to play a vital role in DNA compaction,
359 chromosome segregation, and fertility (67). Moreover, the particular chromatin constraints
360 of sperm production raise the possibility that HTAS-1 proteins could encode distinct
361 functions relative to H2A proteins, for instance by leading to greater sperm DNA
362 compaction; however, it is also possible that a distinct H2A paralog is simply necessary to
363 ensure expression of H2A proteins well after germline mitosis in later stages of
364 spermatogenesis that undergo transcription, DNA recombination and repair, and
365 division. Thus, more study of the functional significance of HTAS-1 homologs in different
366 species is clearly needed to distinguish between these possibilities.

367 *Concluding remarks*

368 To summarize, these results show exceptions to previously reported patterns,
369 challenging sometimes implicit assumptions about non-core histones. First, whereas
370 protein sequence differences between core and variant histone paralogs are often assumed
371 to reflect differences in protein function, here we show that the variant protein HIS-35 is
372 likely to have a redundant function with core H2A despite the sequence difference. Second,

373 while all *C. elegans* H2A variants have a single intron, our observation of multi-intron
374 variants and of recurrent intron loss, suggests that specific introns may not have crucial
375 roles in the expression of histone variants. Instead, the role of introns in variant histones
376 may simply lie in introns' general roles in promoting expression. Third, the combination
377 of conservation and loss of variant histones points to potentially lineage-specific, partially
378 redundant, or easily replaced roles of some histone variants. Future studies should explore
379 the generality of these patterns across other lineages of eukaryotes. In addition to our
380 specific findings for histone variant biology, these results highlight that introns can be
381 useful in the reconstruction of the histories of complex gene families.

382 **MATERIAL AND METHODS**

383 **Data source**

384 Genomic sequences and gene feature format files of 108 nematode species were obtained
385 from WormBase and the *Caenorhabditis* database (68, 69).

386 **Data mining and processing**

387 All the known genes of 108 nematode species with characterized exon-intron
388 structures were fetched from their genomes using their respective 'gene feature format'
389 file. We then noted the positions of the introns in the header of their respective genes and
390 translated the *Caenorhabditis* gene sequences.

391 To identify the homologs of H2A and their variants, BLASTP, version 2.9.0+, was
392 performed using standard parameters while treating the translated gene sequences (of 108
393 nematode species) as the database and H2A and variant (H2A.Z^{HTZ-1}, HTAS-1, HIS-35)
394 protein sequences as the query (81). Using a maximum e-value of 1e-10, 8003 hits were
395 retrieved which were the homologs of H2A and H2A-variant genes. We then removed
396 dubious genes encoding proteins more than 200 amino acids long, because histone proteins
397 are generally shorter. We collapsed the genes whose introns align at the position and have
398 an identical protein sequence. After filtering off the genes we were left with 408 distinct
399 entries.

400 Previous studies have shown the intron position conservation among widely diverged
401 eukaryotic species (53-55, 57-60). Therefore, to assess the intron position conservation
402 among the putative H2A variant genes, we performed a Multiple Sequence Alignment
403 (MSA) using the default parameters of CLUSTALW (82). We mapped the intron positions
404 of each gene onto the corresponding protein CLUSTALW alignment, allowing us to
405 identify as potential H2A.Z^{HTZ-1}, HTAS-1, HIS-35 orthologs those genes with intron
406 positions matching *C. elegans* intron positions.

407 **Phylogenetic Analysis**

408 Multiple sequence alignment of 408 H2A variants homologs was performed using
409 default parameters of MUSCLE to generate a phylogenetic tree (Figure S1) using IQtree,
410 which does an automatic selection of the model by doing a model fit test and likelihood

411 scoring (83, 84). VT+R9, the variable time method, was selected by IQtree for our
412 data. The tree didn't yield a clear phylogenetic signal for HIS-35 or H2A.Z^{HTZ-1}, with
413 homologs exhibiting the *C. elegans* HIS-35 or H2A.Z^{HTZ-1} intron positions are scattered
414 over the tree (Figure S1; IP 234 and IP 233 respectively). However, when we took a closer
415 look at the HIS-35 characteristic intron-containing genes, we see that genes containing an
416 intron at the *C. elegans* HIS-35 intron position (Figure S1, IP 234) are restricted to most
417 species of *Caenorhabditis* and its sister genus *Diploscapter*. A clear clade of species was
418 seen which had HTAS-1 characteristic intron position (Figure S1, IP 152). For consistency
419 in Figures 2-4, we show the portion of the phylogenetic tree containing *Caenorhabditis*
420 species and *Disploscapter*.

421 **Confirmation of H2A variant losses**

422 We found a loss of H2A.Z^{HTZ-1}, HTAS-1, HIS-35 characteristic introns in a few
423 lineages (marked by a minus sign in Figures 2, 3, and 4). To know whether this is a real
424 loss or reflected errors in gene annotation, tblastn searches were performed across the
425 genome of these species. This manual curation led to the variant's characteristic intron
426 splice sites being identified by eye in a few species due to alignment gaps at the exact intron
427 position, indicating that these species truly contain the variant and that failure to initially
428 identify the variant is due to a failure of the annotation to include these genes. We included
429 a material and method flowchart as the Figure S4.

430

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 NICHD R03HD093990.

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.

450 7. Romeo V, Schümperli D. Cycling in the nucleus: regulation of RNA 3′ processing and
451 nuclear organization of replication-dependent histone genes. Curr Opin Cell Biol.
452 2016;40:23-31.

453 8. Biterge B, Schneider R. Histone variants: key players of chromatin. Cell Tissue Res.
454 2014;356(3):457-66.

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

457 10. Talbert PB, Ahmad K, Almouzni G, Ausió J, Berger F, Bhalla PL, et al. A unified
458 phylogeny-based nomenclature for histone variants. Epigenetics & chromatin.
459 2012;5(1):1-19.

460 11. Martire S, Banaszynski LA. The roles of histone variants in fine-tuning chromatin
461 organization and function. Nature reviews. Molecular cell biology. 2020 Sep;21(9):522-
462 41.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

491 26. Faast R, Thonglairoam V, Schulz TC, Beall J, Wells JR, Taylor H, et al. Histone variant
492 H2A. Z is required for early mammalian development. Current Biology.
493 2001;11(15):1183-7.

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

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

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

500 30. Petty EL, Collette KS, Cohen AJ, Snyder MJ, Csankovszki G. Restricting dosage
501 compensation complex binding to the X chromosomes by H2A. Z/HTZ-1. PLoS genetics.
502 2009;5(10):e1000699.

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

506 32. Rogakou EP, Pilch DR, Orr AH, Ivanova VS, Bonner WM. DNA double-stranded
507 breaks induce histone H2AX phosphorylation on serine 139. J Biol Chem.
508 1998;273(10):5858-68.

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

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

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

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

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

522 38. Shaytan AK, Landsman D, Panchenko AR. Nucleosome adaptability conferred by
523 sequence and structural variations in histone H2A–H2B dimers. Curr Opin Struct Biol.
524 2015;32:48-57.

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

527 40. Hoghoughi N, Barral S, Vargas A, Rousseaux S, Khochbin S. Histone variants:
528 essential actors in male genome programming. The Journal of Biochemistry.
529 2018;163(2):97-103.

530 41. Millar CB. Organizing the genome with H2A histone variants. Biochem J.
531 2013;449(3):567-79.

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

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

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

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

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

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

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

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

549 50. Armstrong JF, Pritchard-Jones K, Bickmore WA, Hastie ND, Bard JB. The expression
550 of the Wilms' tumour gene, WT1, in the developing mammalian embryo. Mech Dev.
551 1993;40(1-2):85-97.

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

554 52. Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes Dev.
555 2003;17(4):419-37.

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

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

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

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

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

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

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

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

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

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

581 63. Ahmad K, Henikoff S. Histone H3 variants specify modes of chromatin assembly.
582 Proceedings of the National Academy of Sciences. 2002;99(suppl 4):16477-84.

583 64. Hake SB, Allis CD. Histone H3 variants and their potential role in indexing mammalian
584 genomes: the “H3 barcode hypothesis”. Proceedings of the National Academy of Sciences.
585 2006;103(17):6428-35.

586 65. Hake SB, Garcia BA, Duncan EM, Kauer M, Dellaire G, Shabanowitz J, et al.
587 Expression patterns and post-translational modifications associated with mammalian
588 histone H3 variants. J Biol Chem. 2006;281(1):559-68.

589 66. Santenard A, Ziegler-Birling C, Koch M, Tora L, Bannister AJ, Torres-Padilla M.
590 Heterochromatin formation in the mouse embryo requires critical residues of the histone
591 variant H3. 3. Nat Cell Biol. 2010;12(9):853-62.

592 67. Chu DS, Liu H, Wu TF, Ralston EJ, Nix P, Meyer BJ, et al. Sperm chromatin
593 proteomics identifies evolutionarily conserved fertility factors. Nature (London). 2006 Sep
594 07,;443(7107):101-5.

595 68. Davis P, Zarowiecki M, Arnaboldi V, Becerra A, Cain S, Chan J, et al. WormBase in
596 2022—data, processes, and tools for analyzing Caenorhabditis elegans. Genetics.
597 2022;220(4):iyac003.

598 69. Stevens L, Félix M, Beltran T, Braendle C, Caurcel C, Fausett S, et al. Comparative
599 genomics of 10 new Caenorhabditis species. Evolution Letters. 2019;3(2):217-36.

600 70. Carlton PM, Davis RE, Ahmed S. Nematode chromosomes. Genetics.
601 2022;221(1):iyac014.

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

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

606 73. Liao D. Concerted evolution. Nature Encyclopedia of the human genome. 2003;1:938-
607 42.

608 74. Baldo AM, Les DH, Strausbaugh LD. Potentials and limitations of histone repeat
609 sequences for phylogenetic reconstruction of Sophophora. Mol Biol Evol.
610 1999;16(11):1511-20.

611 75. Fitch DH, Strausbaugh LD. Low codon bias and high rates of synonymous substitution
612 in Drosophila hydei and D. melanogaster histone genes. Mol Biol Evol. 1993;10(2):397-
613 413.

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

616 77. Irimia M, Rukov JL, Penny D, Garcia-Fernandez J, Vinther J, Roy SW. Widespread
617 evolutionary conservation of alternatively spliced exons in Caenorhabditis. Mol Biol Evol.
618 2008;25(2):375-82.

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

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

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

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

630 82. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of
631 progressive multiple sequence alignment through sequence weighting, position-specific
632 gap penalties and weight matrix choice. Nucleic acids research. 1994 Nov 11,;22(22):4673-
633 80.

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

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

```
HTAS-1      MARLKQRPNRILNTS--TKTSSAKKKKKRISRSTRSGLTFPVGRIHRKLRETTRGKQRIS        58
HTZ-1       ----------MAGGKGKAGKDSGKSKSKVVSRSARAGLQFPVGRIHRFLKQRTTSSGRVG        50
HIS-35      -----------MSGRGKGGKA--KTGGKAKSRSSRAGLQFPVGRLHRILRKGN-YAQRVG        46
H2A         -----------MSGRGKGGKA--KTGGKAKSRSSRAGLQFPVGRLHRILRKGN-YAQRVG        46
                       .      .   *.  *  ***:*:** *****:** *:: .    *:.

HTAS-1      AGASVFMAATLEYLTTELMEMSAIAANESKKSRVTPRHLHLAIYGDQETAQLLDKVTLPQ        118
HTZ-1       ATAAVYSAAILEYLTAEVLELAGNASKDLKVKRITPRHLHLAIRGDEELDTLIK-ATIAG        109
HIS-35      AGAPVYLAAVLEYLAAEVLELAGNAARDNKKTRIAPRHLQLAVRNDEELNKLLAGVTIAQ        106
H2A         AGAPVYLAAVLEYLAAEVLELAGNAARDNKKTRIAPRHLQLAVRNDEELNKLLAGVTIAQ        106
            * * *: ** ****::*::*::. *:.: * .*::****:**: .*:*   *:  .*:

HTAS-1      GGVTPMPIHPSLLPKKKAKEDDKENNS-----    145
HTZ-1       GGVIPH-IHRYLMNKKGAPVPGKPGAPGQGPQ    140
HIS-35      GGVLPN-IQAVLLPKKTAGDKE----------    127
H2A         GGVLPN-IQAVLLPKKTGGDKE----------    127
            *** *  *:  *: ** .
```

Figure 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. HTAS-1 (highlighted with a yellow) has a phase 0 intron between the 26th and 27th amino acid; HIS-35 (highlighted with pink) has a phase 0 intron between 50th and 51st amino acid. HTZ-1 (highlighted with blue) has a phase 2 intron splitting the 57th amino acid.

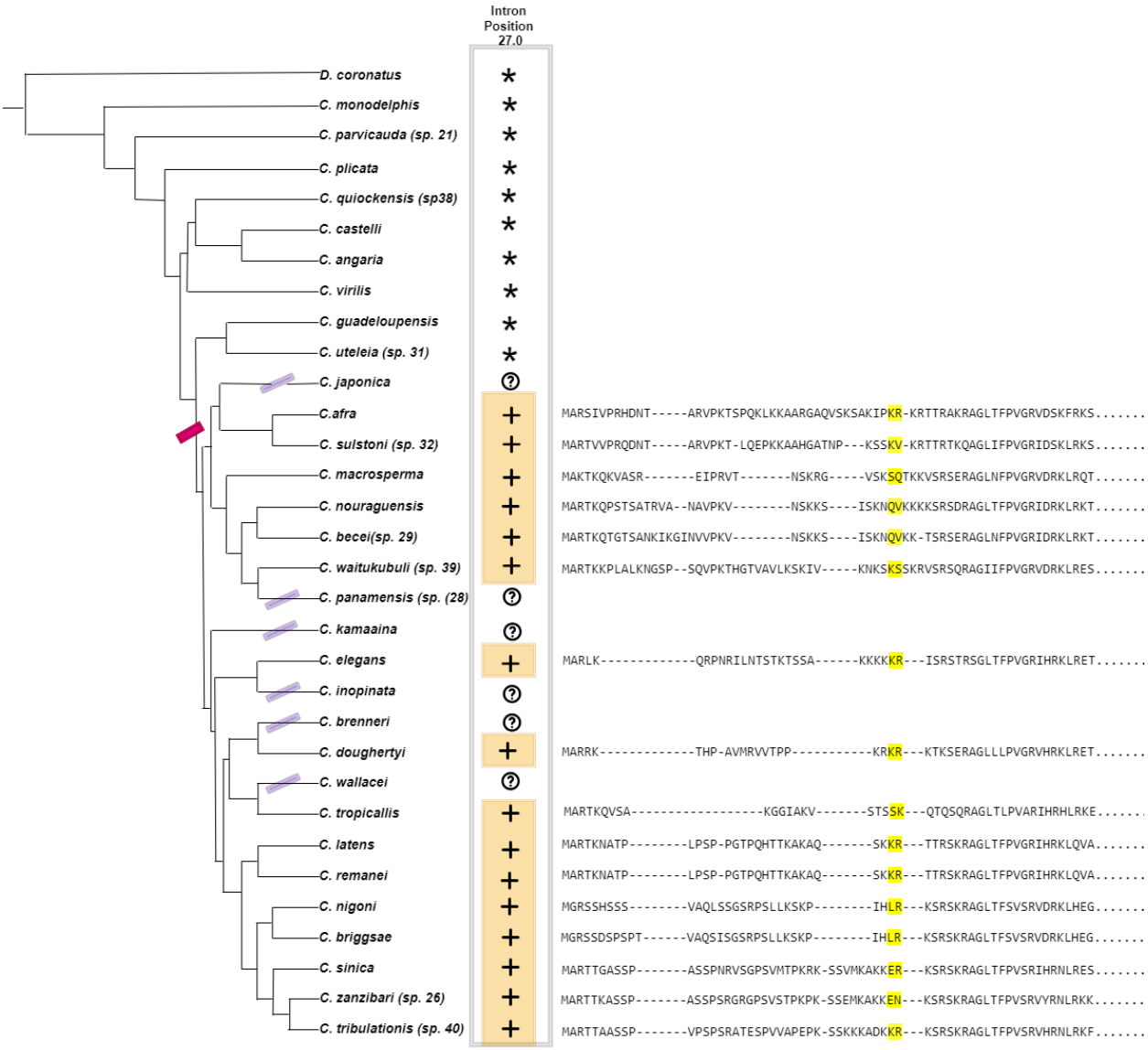
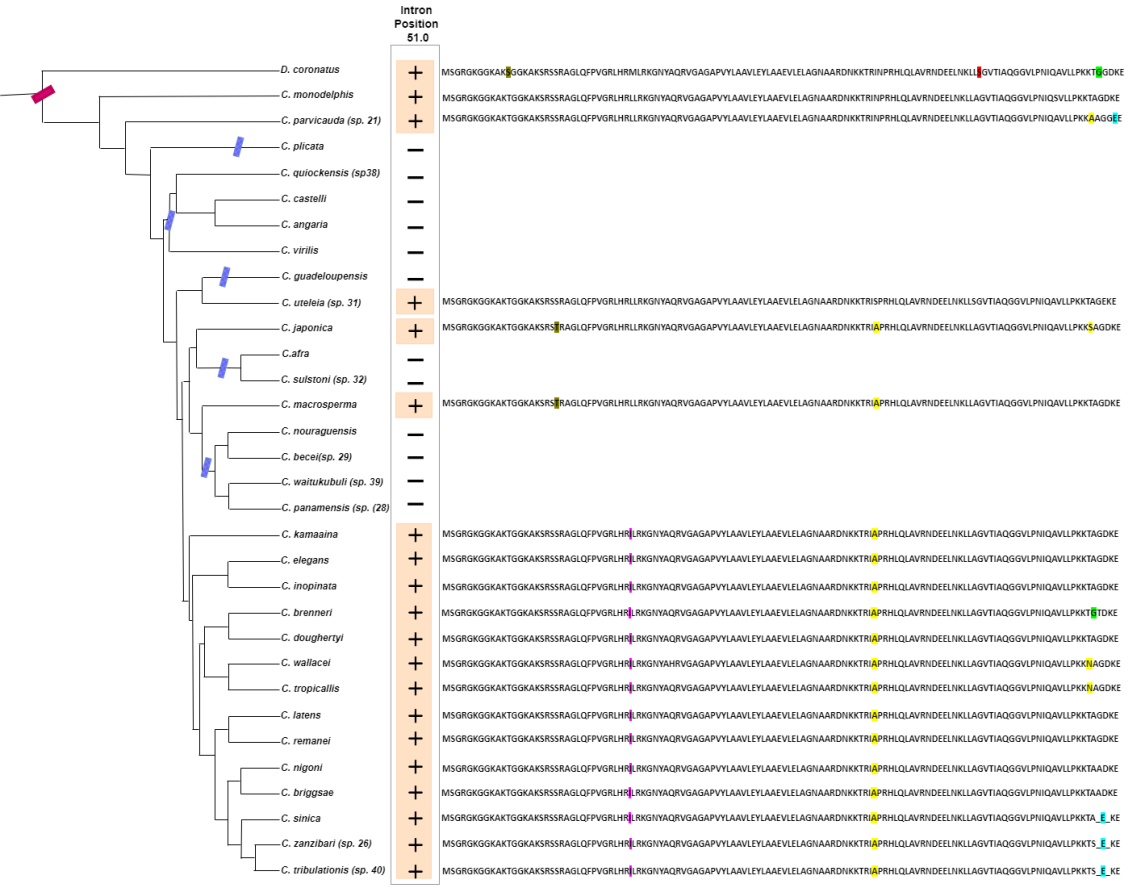


Figure 2: Intron position and sequence evidence indicate the origin of HTAS-1 within *Caenorhabditis* and subsequent loss. On the left is the previously reconstructed species tree of *Caenorhabditis* and *Diploscapter*. The likely origin of HTAS-1 is indicated with a dark pink bar. HTAS-1 presence in a species is indicated by the plus sign. The question mark denotes species where HTAS-1 homologs could not be identified.. 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). For consistency, we show the portion of the phylogenetic tree containing the *Caenorhabditis* species and *Disploscapter.*

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)

H2A

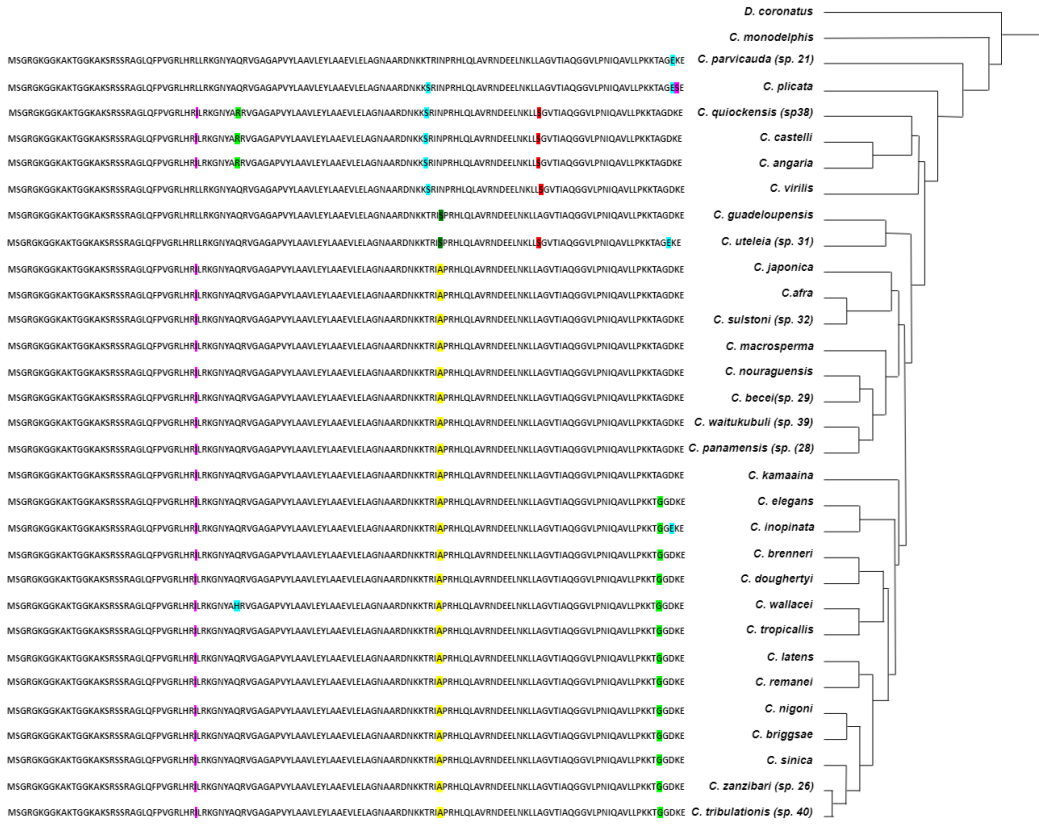

D. coronatus
C. monodelphis
C. parvicauda (sp. 21)
C. plicata
C. quiockensis (sp38)
C. castelli
C. angaria
C. virilis
C. guadeloupensis
C. uteleia (sp. 31)
C. japonica
C.afra
C. sulstoni (sp. 32)
C. macrosperma
C. nouraguensis
C. becei(sp. 29)
C. waitukubuli (sp. 39)
C. panamensis (sp. (28)
C. kamaaina
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)

Figure 3: Comparison of the variant HIS-35 (left) with canonical H2A (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).

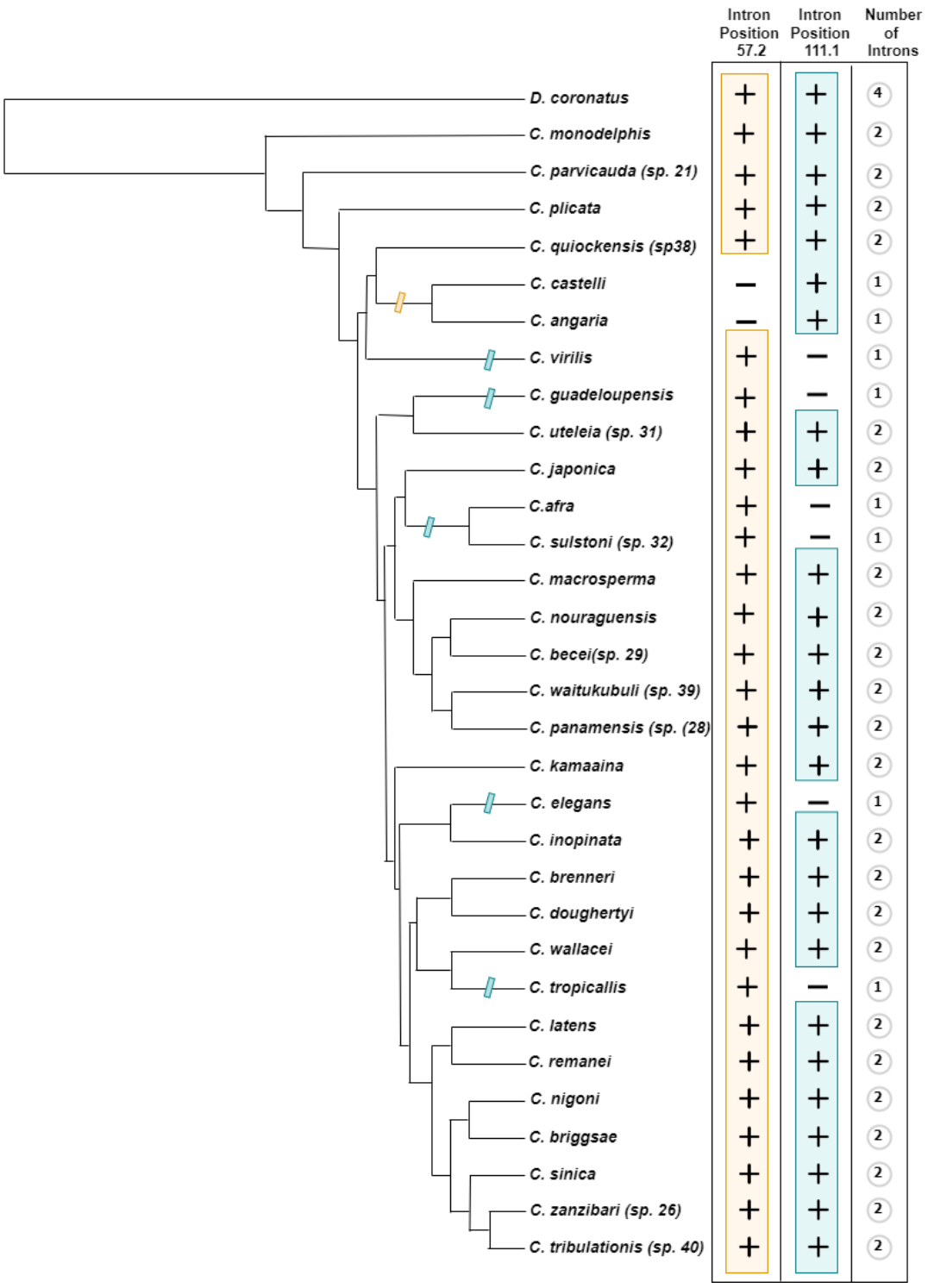


Figure 4: The dynamic history of intron loss and gain in HTZ-1. On the left is the previously reconstructed species tree of *Caenorhabditis*. HTZ-1 characteristic intron presence and absence in a species is indicated by the plus and minus sign. Yellow hash marks on the tree branch depict the loss of intron-1 whereas the blue hash mark on the branch suggests the loss of intron 2 in those lineages.