Article

Chinese Science Bulletin

SPECIAL ISSUE
Adaptive Evolution and Conservation Ecology of Wild Animals

June 2013 Vol.58 No.18: 2174–2178
doi: 10.1007/s11434-012-5623-z

Multiple intron gain and loss events occurred during the evolution of *Cenp-A* gene

FAN XinYu[1], YU Li[2,3], XU HuaiLiang[1*] & LI Ying[4*]

[1] *College of Animal Science and Technology, Sichuan Agricultural University, Ya'an 625014, China;*
[2] *Laboratory for Conservation and Utilization of Bio-resource, Yunnan University, Kunming 650091, China;*
[3] *Key Laboratory for Animal Genetic Diversity and Evolution of High Education in Yunnan Province, Yunnan University, Kunming 650091, China;*
[4] *Institute of Animal Breeding and Genetics, Sichuan Agricultural University, Ya'an 625014, China*

Received July 27, 2012; accepted September 3, 2012; published online January 14, 2013

Centromere protein A (CENP-A) is a histone H3 like protein, and it plays a very important role in chromosomal segregation during mitosis and meiosis. The analyses on the exon-intron organization of the *Cenp-A* gene in representative genomes revealed that multiple intron gain and loss events have occurred during the evolution of *Cenp-A* gene in opisthokonta (common ancestor of fungi and animals). Moreover, our results revealed that at least two positions were conserved in the intron gain and loss events during the evolution of the *Cenp-A* gene.

CENP-A, evolution, intron gain and loss

Citation: Fan X Y, Yu L, Xu H L, et al. Multiple intron gain and loss events occurred during the evolution of *Cenp-A* gene. Chin Sci Bull, 2013, 58: 2174–2178, doi: 10.1007/s11434-012-5623-z

The centromere is composed of a tandemly repetitive satellite sequence and a protein complex [1]. More and more evidences have shown that centromere protein A (CENP-A) binding onto the centormeric region is an early step in kinetochore formation, although the process of the assembly of the complex is not clear [2]. CENP-A is a variant protein of histone H3 [3] containing a highly variant N-terminal tail, which diverges greatly in both the lengths and the amino acid compositions, while the C-terminal domain shares an average of 57% amino acid identities with histone H3 [4]. Histone H3 was replaced by CENP-A at the centromeric region of the necleosome [5]. The nearly invariant histone H3 has been maintained by a strong purifying selection during eukaryote evolution [6]. In contrast, *Cenp-A* evolved rapidly, especially in *Drosophila* [7,8] and *Arabidopsis* [9], where the rapid evolution is associated with positive selection [8].

Cenp-A gene has been detected in all examined eukaryotes [10]. In *Drosophila,* only one exon is identified, but in all mammals, birds and frogs, a 4 exon-3 intron organization (or gene structure) have been observed [11]. However, the gene structural evolution of *Cenp-A* is unclear. Here we compared the structure of *Cenp-A* gene in representative species from fungi to mammals and observed multiple intron gain and loss events in the *Cenp-A* gene during the eukaryotic evolution.

1 Materials and methods

The cDNA sequences of *Cenp-A* gene from mammals (human, rhesus monkey, cattle, dog, mouse and rat), amphibians (frog), fishes (zebrafish, fugu, tetraodon), echinodermata (sea urchin), nematodes (*C. elegant*), insects (*A. gambiae*, *D. meglanogaster*, *S. aegypti*) and fungi (*S. cerevisiae*) were downloaded from GenBank (the accession numbers are shown in Table 1). Here we only selected those species for which both the cDNA of *Cenp-A* and the genomic sequences

*Corresponding authors (email: endlesssnow2012@gmail.com; huailxu@yahoo.com)

Table 1 The cDNA sequences used and their chromosomal distribution of *Cenp-A* gene

Species	Gene	Sequences origination	Chromosome	Strand	Start	End	Span	No. Exon/ Intron
Human	*Cenp-A*	AAH02703	2	+	26862569	26869651	7083	4/3
Chimpanzee	*Cenp-A*	this study	2a	+	27378536	27385614	7079	4/3
Rehsus	*Cenp-A*	XP_001087306	13	+	26733155	26739887	6733	4/3
Bovine	*Cenp-A*	XM_869623	11	+	55208380	55212532	4153	4/3
Bovine	*Cenp-A*-L-1	Li and Huang (2008)	scaffold24854	–	748	1676	929	2/1
Bovine	*Cenp-A*-L-2	Li and Huang (2008)	27	–	399914	400839	926	2/1
Bovine	*Cenp-A*-L-3	Li and Huang (2008)	4	+	5923139	5923554	416	1/0
Bovine	*Cenp-A*-L-4	Li and Huang (2008)	4	–	5898445	5899361	929	2/1
Bovine	*Cenp-A*-L-5	Li and Huang (2008)	scaffold522	–	103014	103370	406	1/0
Bovine	*Cenp-A*-L-6	Li and Huang (2008)	scaffold1160	+	31104	31459	356	1/0
Bovine	*Cenp-A*-L-7	Li and Huang (2008)	scaffold8622	+	23762	24431	670	2/1
Bovine	*Cenp-A*-L-8	Li and Huang (2008)	scaffold83	–	816862	817277	416	1/0
Bovine	*Cenp-A*-L-9	Li and Huang (2008)	scaffold15295	+	1482	2159	678	2/1
Bovine	*Cenp-A*-L-10	Li and Huang (2008)	13	+	4794302	4794878	577	2/1
Dog	*Cenp-A*	XP_859713	17	+	23798396	23799742	1347	3/2
Mouse	*Cenp-A*	AAH11038	5	+	30943610	30950005	6396	4/3
Rat	*Cenp-A*	XP_001069485	6	–	25688066	25693824	5759	4/3
Frog	*Cenp-A*	NM_001016585	1026	–	175365	183745	8381	4/3
Fugu	*Cenp-A*	Régnier et al. (2003)	Un	–	238884614	238885770	1157	4/3
Tetraodon	*Cenp-A*	Régnier et al. (2003)	Un_random	–	47265831	47267213	1383	4/3
Zebrafish	*Cenp-A*	AAH44483	8	–	2482889	2483326	438	1/0
A. aegypti	*Cenp-A*	EAT38856	supercont1.387	+	861256	861906	615	1/0
A. gambiae	*Cenp-A*	EAL39661	2L	–	46425886	46426644	759	1/0
D. melanogaster	*Cid*	AY126932	2R	+	9001598	9002275	678	1/0
Sea urchins	*His-69*	XP_788572	scaffold76804	+	11432	17114	5683	4/3
C. elegans	*Hcp-3*	NM_066727	III	–	9615328	9616345	1018	4/3
S. cerevisiae	*Cse4*	AAB60309	11	–	345716	346405	690	1/0

are available, because of the great divergence in both the amino acid composition and the length of the N-terminal region among different species, which makes it difficult to obtain the *Cenp-A* gene from the genome sequence even by using a cDNA sequence from a slightly remote species as a query. We had no problem in extracting chimpanzee *Cenp-A* coding sequences from its genomic sequence by using the human *Cenp-A* as a query. The *Cenp-A* gene sequences were extracted by mining their genome database (http://genome.cse.ucsc.edu/cgi-bin/hgBlat). And the obtained genomic DNA sequences and the cDNA sequences were used to conduct cDNA-to-genomic sequence alignment on Spidey (http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/), which provided the exon-intron structures. Repetitive elements in *Cenp-A* gene sequences were identified by the RepeatMasker program (http://www.repeatmasker.org/).

2 Results

2.1 Gene structure evolution of *Cenp-A*

The results from database searches show that some of *Cenp-A* genes are intronless. The sizes of these genes and their locations in chromosomes are shown in Table 1, and the *Cenp-A* gene structural evolution is shown in Figure 1. *Cenp-A* genes in one fungus (*S. cerevisiae*) and in three insects (*A. gambiae*, *D. meglanogaster* and *S. aegypti*) are intronless, which is also observed in all the 11 published Drosophila genomes. However, *Cenp-A* gene is interrupted by three introns in *C. elegans* and sea urchin (Figure 1). Most interestingly, different exon-intron structures are identified

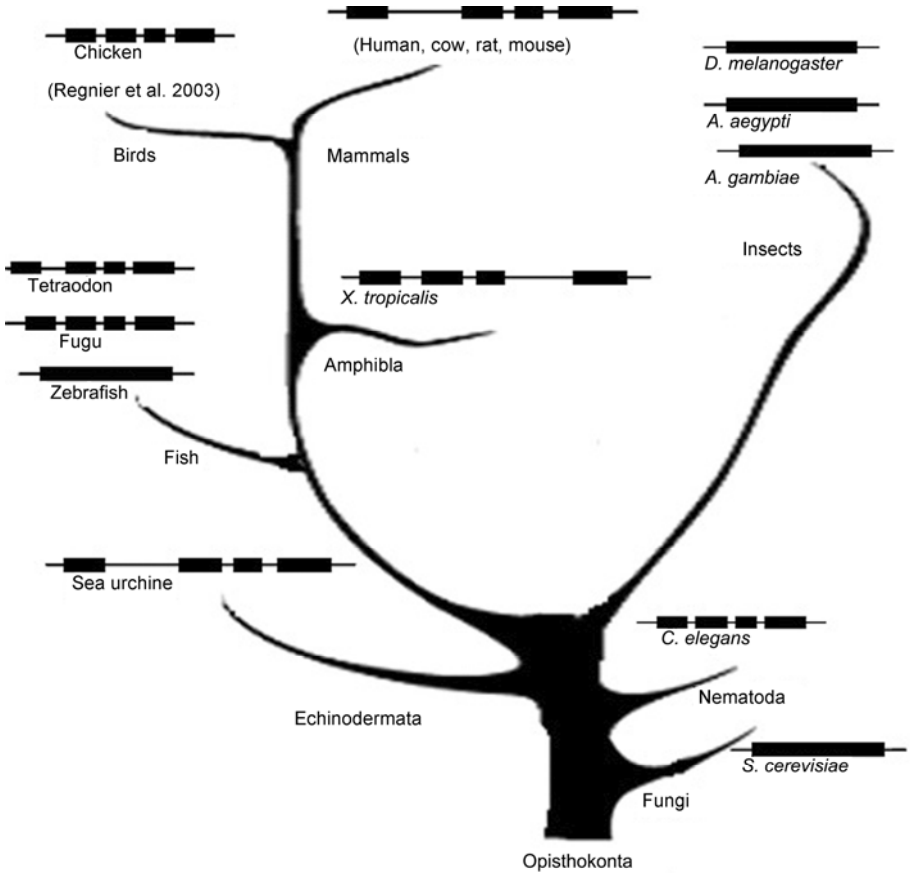

Figure 1 Scheme of the gene structure evolution of *Cenp-A* in opisthokonta. The tree is a simplified "tree of life" adapted from Eirin-Lopez et al. 2004, the intron-exon organizations are marked for those representative species, and the size of the lines (represent introns) and the boxes (represent exons) are not proportional to the lengths of the exon and the intron.

in fish, and three introns in fugu and in tetraodon are detected, but no intron is observed in zebrafish *Cenp-A* gene. All *Cenp-A* genes from amphibians, birds and mammals contain 4 exons and 3 introns except for that in cow, where a gene family with different exon-intron organizations among family members was observed [12]. We also proposed a most possible evolutionary relationship among different exon-intron forms (Figure 2 in [12]), in which we shown that only intron 2 was retained during the loss of introns form the original 4/3 to 1/0 structures [12]. Interestingly, the intron gain events have occurred at least twice: the first one occurred before the emergence of fungi and after the emergence of Nematoda (Figure 1); the second one occurred after the separation of the superorders Ostariophysi (including zebrafish, goldfish, and carp) and Acanthopterygii (including medaka, fugu, cichlid, etc.) [13]. And after the intron gain events, intron loss event has occurred at least once before the emergence of the insect. The positions and the phases (all phase 0) of intron 2 and intron 3 are conserved from sea urchin to mammals, in which the insertion site of the intron 2 in human is behind the 70th codon, and the insertion is behind the 96th codon for intron 3, and no intron sliding (change of the intron position) is observed, thus, suggesting the *Cenp-A* gene contains, at least, two hotspots of intron gain and loss.

2.2 Distributions of the repetitive elements in *Cenp-A* gene

We observed that even for those species with the same 4 exon-3 intron organizations of the *Cenp-A* gene, the lengths of the gene diverged greatly from 1018 in *C. elegans* to 8381 in frog. What makes this great difference? Is repetitive sequences insertion a possible reason? To test this possibility, we checked the distribution of repetitive elements in those 4 exon-3 intron *Cenp-A* genes. The results indicate that those short sequences (shorter than 2000 bp) contain no (or only one) matching repeats, while long sequences harbor more repetitive insertions, which differed both in the repeat types and their numbers. Table 2 shows the location and diversity of all repetitive elements found in the *Cenp-A* gene. We note that the divergence of the gene length is much less when those repeats are deleted; this is especially the case in mammals. The *Cenp-A* gene length in mammals divergent from 2307 to 3864 when the repetitive elements were deleted, indicating that the great divergence of the gene length of *Cenp-A* due mainly to the insertion of the repetitive elements.

3 Discussion

The study in the exon-intron organization of *Cenp-A* suggests that multiple intron gain and loss events occurred during the evolution of *Cenp-A* gene. These repeatedly occurred gain and loss of all the three introns (except for those in cow) is unexpected and erratic. More and more evidences have shown that introns are not "Junk DNA" as believed before. They may have functions, such as expression regulation [14], alternative splicing [15] and exon shuffling [16]. Although we cannot find any study to show direct evidence on whether or not any intron of *Cenp-A* has a functional effect, a research done by Osborn and Miller [17] showed that a intronless yeast CSE-4 (homologous gene of *Cenp-A*) can rescue a *Cenp-A* knocked down human cell, which suggest that the functional effect of *Cenp-A* introns if has, is not vital. This may further suggest that the multiple gain and loss of introns in *Cenp-A* might has just happened by chance and is evolutionarily neutral.

The intron density (number of introns per gene) is different among different genomes, so different tendencies of intron gain and loss are possible. The intron density from some representative species show that early branches are intron poor, while late branches are intron rich in the phylogeny of eukaryotic [18]. For example, birds and mammals have the values over 7, that for *S. cerevisiae* is only 0.053, but it is not always the case. For example, although *D. melganogaster* is higher than *C. elegans* in the phylogeny, its intron numbers per gene value is smaller than that in *C. elegans*. Most interestingly, the gene structure evolution of *Cenp-A* seems consistent with the intron density described by Jeffares et al. [18]. In *S. cerevisiae* and *D. molanogaster*, the *Cenp-A* gene contains no intron, but 3 introns are observed in other species, and the intron density is relatively low. The *Cenp-A* gene in zebrafish is also intronless, but its intron density is unclear yet. Thus whether the intron loss is also related to lower intron density in this species is unclear. Because of the great variances of the N-terminal of the *Cenp-A* both in the amino acid composition and in its length, the attempt to acquire a *Cenp-A* gene by searching its genome using a cDNA or amino acid sequences of *Cenp-A* from other taxon has failed. Thus the results in this paper may not reflect the whole gene structure evolution of *Cenp-A*. Even though, the results are still helpful for future studies to clarify how this gene origination changes occurred, and what are the factors that caused these changes.

The analyses on the distribution of the repetitive elements in the intron region of the *Cenp-A* gene show that the great gene length diversity of the *Cenp-A* genes was due mainly to the length differences of the repetitive elements in different species; this is especially the case in mammals. In several independent insertions, the most obvious example is occurred in amphibians and mammals, where the repetitive elements locate in intron 1 and intron 3 respectively (Table 2), so the insertion events is independent origin. And some linage specific repetitive elements have inserted into the *Cenp-A* genes after mammals diverged from other vertebrates, for example, one ELVR insertion is primate specific, and the insertions of ALU/B1, B2-B6, ID3, MIRS and LTR are rodent specific. This independent insertion even occurred after the divergence of mouse and rat (Table 2), and thus suggesting species specific insertions. The insertions of repetitive elements in genome are evolutionary neutral in general, however, we believe it is also the case in *Cenp-A* introns, because (1) the intron gain and loss of *Cenp-A* among different species is consistent with the intron density differences among corresponding species; and (2) the gain and loss of introns itself might be neutral as we discussed before.

Table 2 Repetitive elements distribution in the *Cenp-A* genes

Gene	Repeat elements	Location	Total repeats	Gene span	Gene span after deleted repeats
Human *Cenp-A*	9 SINE 6 LINE and 1ERVL	intron 1	4184	7083	2899
Chimpanzee *Cenp-A*	9 SINE 6 LINE and 1ERVL	intron 1	4185	7079	2894
Rehsus *Cenp-A*	8 SINE 6 LINE and 1ERVL	intron 1	3811	6733	2922
Bovine *Cenp-A*	4 SINE 4 LINE	intron 1	1846	4153	2307
Mouse *Cenp-A*	14 SINE 4ALU/B1 6 B2-B6 3ID3 1 MIRS 2 LTR	intron 1	2640	6396	3756
Rat *Cenp-A*	12 SINE 1ALU/B1 7 B2-B6 3ID3 1 MIRS 1 LTR	intron 1	1895	5759	3864
Frog *Cenp-A*	4 DNA transposons and 1 satellites	intron 3	1579	8381	6802
Fugu *Cenp-A*	None	–	–	1157	–
Tetraodon *Cenp-A*	1 LINE	intron3	238	1383	1145
Sea urchins *His-69*	None	–	–	5683	–
C. elegans Hcp-3	None	–	–	1018	–

This work was supported by the National Natural Science Foundation of China (30970383), Program for New Century Excellent Talents in University (NCET).

1 Henikoff S, Ahmad K, Malik H S. The centromere paradox: Stable inheritance with rapidly evolving DNA. Science, 2001, 293: 1098–1102
2 Howman E V, Fowler K J, Newson A J, et al. Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice. Proc Natl Acad Sci USA, 2000, 97: 1148–1153
3 Palmer D K, O'Day K, Trong H L, et al. Purification of the centromere-specific protein *Cenp-A* and demonstration that it is a distinctive histone. Proc Natl Acad Sci USA, 1991, 88: 3734–3738
4 Sullivan K F. A solid foundation: Functional specialization of centromeric chromatin. Curr Opin Genet Dev, 2001, 11: 182–188
5 Warburton P E, Cooke C A, Bourassa S, et al. Immunolocalization of *Cenp-A* suggests a distinct nucleosome structure at the inner kinetochore plate of active centromeres. Curr Biol, 1997, 7: 901–904
6 Piontkivska H, Rooney A P, Nei M. Purifying selection and birth-and-death evolution in the histone H4 gene family. Mol Biol Evol, 2002, 19: 689–697
7 Malik H S, Henikoff S. Adaptive evolution of Cid, a centromere-specific histone in Drosophila. Genetics, 2001, 157: 1293–1298
8 Malik H S, Vermaak D, Henikoff S. Recurrent evolution of DNA-binding motifs in the Drosophila centromeric histone. Proc Natl Acad Sci USA, 2002, 99: 1449–1454
9 Talbert P B, Masuelli R, Tyagi A P, et al. Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. Plant Cell, 2002, 14: 1053–1066
10 Warburton P E. Epigenetic analysis of kinetochore assembly on variant human centromeres. Trends Genet, 2001, 17: 243–247
11 Figueroa J, Pendon C, Valdivia M M. Molecular cloning and sequence analysis of hamster *Cenp-A* cDNA. BMC Genomics, 2002, 3: 11
12 Li Y, Huang J F. Identification and molecular evolution of cow *Cenp-A* gene family. Mamm Genome, 2008, 19: 139–143
13 Chen W J, Orti G, Meyer A. Novel evolutionary relationship among four fish model systems. Trends Genet, 2004, 20: 424–431
14 Chartier F L, Bossu J P, Vu-Dac N, et al. Involvement of intronic sequences in the transcriptional regulation of apolipoprotein B, E and A-II genes. Z Gastroenterol, 1996, 34 Suppl 3: 44–45
15 Amy C M, Williams-Ahlf B, Naggert J, et al. Intron-exon organization of the gene for the multifunctional animal fatty acid synthase. Proc Natl Acad Sci USA, 1992, 89: 1105–1108
16 Bryk M, Belfort M. Spontaneous shuffling of domains between introns of phage T4. Nature, 1990, 346: 394–396
17 Osborn M J, Miller J R. Rescuing yeast mutants with human genes. Brief Funct Genomic Proteomic, 2007, 6: 104–111
18 Jeffares D C, Mourier T, Penny D. The biology of intron gain and loss. Trends Genet, 2006, 22: 16–22