DOI 10.18699/SBB-2022-076

Tracing the origin and evolution of the locus containing paralogous *CENH3* genes in cereals

Elisafenko E.A.[1, 2, 3], Evtushenko E.V.[2], Vershinin A.V.[2]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[2] *Institute of Molecular and Cellular Biology, SB RAS, Novosibirsk, Russia*

[3] *Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia*

* *antares@bionet.nsc.ru*

Key words: Centromere, *CENH3*, Gene duplication, Molecular evolution, Invasion of transposable elements, *Poaceae*

Motivation and Aim: The central component of centromere specification and function is the centromere-specific histone H3 (*CENH3*) [1]. The cereal family Poaceae is one of the largest and most diverse angiosperm families consisting of more than 12,000 species and more than 700 genera. Some cereal species (maize, rice) have one copy of the gene encoding this protein, while some (wheat, barley, rye) have two, *αCENH3* and *βCENH3* genes [2]. The discovery of *CENH3* duplications complicated the understanding of the evolution of the *CENH3* genes and raised the question about the mechanisms and causes of gains, losses and selective constraints. Using a homology-based approach, we trace back the mutual evolution of the structure of the *CENH3* genes and structure of the nearby regions in various tribes in the family Poaceae.

Methods and Algorithms: The sequences of *αCENH3 and βCENH3* genes were obtained for 23 species and the deduced protein sequences were used for phylogenetic analysis. The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model [3]. Evolutionary analyses were conducted in MEGA X [4]. To determine the exon-intron structure of the *αCENH3 and βCENH3* genes, RNA-seq reads from the SRA at NCBI were mapped onto the corresponding reference genome sequences or the assembled contigs of genome sequences (if the whole-genome reference sequences were not available) using HISAT2 [5]. The composition of repeated DNA sequences in the *CENH3* locus in different species was determined using the *Viridiplantae RepeatMasker* program and database [6]. The genes neighboring to *CENH3* genes were identified by similiarity with the sequences of *Triticum urartu* and *Oryza sativa* using TBLASTX at NCBI or, locally, BLASTN.

Results: We have established that the syntenic group or the *CENH3* locus with the *CENH3* gene and the boundaries defined by the *CDPK2* and *bZIP* genes first appeared about 50 Mya in a common ancestor of the Bambusoideae subfamilies, Oryzoideae and Pooideae. This locus came to Pooideae with one copy of *CENH3* in the most ancient tribes Nardeae and Meliceae (Fig. 1A). The *βCENH3* gene as a part of the locus appeared in the tribes Stipeae and Brachypodieae around 35–40 Mya. The duplication was accompanied by splicing abnormalities, leading to changes in the exon-intron structure and long deletions in the N-terminal domain of the protein *βCENH3*. Purifying selection acts mostly on *αCENH3s*, while *βCENH3*s and especially their NTT domains form more heterogeneous structures, in which clade-specific amino acid motifs are present. In barley species, the *βCENH3* gene assumed an inverted orientation relative to *αCENH3* and the left-border gene, *CDPK2*, was substituted with *LHCB-l* (Fig. 1). A head-to-head

orientation of the paralogs is a distinctive feature of all Triticeae species studied. The *βCENH3* inversion that we found in the *CENH3* locus of the barley species is, in our experience, a Triticeae-only feature and occurred in their common ancestor approximately 15–16 Mya. Invasion by mobile elements and concomitant rearrangements in the structure of the *CENH3* locus took place in an independent way during the formation of each species' genome. As a result of these processes, we observe significant variations in the size of the locus, the set of TE families and the amount of changes in their structure (Fig. 1). As the evolution and domestication of plant species went on, the locus was growing in size due to an increasing distance between *αCENH3* and *βCENH3* because of a massive insertion of the main LTR-containing retrotransposon superfamilies, *gypsy* and *copia*, without any evolutionary preference on either of them. A comparison of the molecular structure of the locus in the A, B and D subgenomes of the hexaploid wheat *Triticum aestivum* showed that invasion by mobile elements and concomitant rearrangements took place in an independent way even in evolutionarily close species.

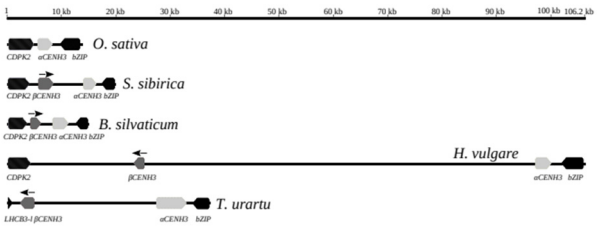

Figure 1. The structure and evolutionary changes of the *CENH3* locus.
Syntenic genes comprising the *CENH3* locus. The species with important events, such as the formation of a syntenic group (rice), the emergence of the beta paralog (*Stipa*, *Brachypodium*), the inversion of the beta paralog (barley) and replacement of the 5' marker gene (*T. urartu*) in their evolutionary history are given as examples.

Conclusion: The *CENH3* duplication in cereals was accompanied by changes in the exon-intron structure of the *βCENH3* paralog. The observed general tendency towards the expansion of the *CENH3* locus reveals an amazing diversity of ways in which different species implement the scenario described above.

Acknowledgements: The study is supported by grant of Russian Foundation for Basic Research (project No. 20–04-00699).

References

1. Postberg J., Forcob S., Chang W.-J., Lipps H.J. The evolutionary history of histone H3 suggests a deep eukaryotic root of chromatin modifying mechanisms. *BMC Evol Biol*. 2010;10:259.
2. Evtushenko E.V., Elisafenko E.A., Gatzkaya S.S., Lipikhina Y.A., Houben A., Vershinin A.V. Conserved molecular structure of the centromeric histone CENH3 in Secale and its phylogenetic relationships. *Sci Rep*. 2017;7:17628. doi: 10.1038/s41598-017-17932-8.
3. Jones D.T., Taylor W.R., Thornton J.M. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics.* 1992;8:275-282. doi: 10.1093/bioinformatics/8.3.275.
4. Kuma S., Stecher G., Li M., Knyaz C., Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol.* 2018;35:1547-1549.
5. Kim D., Langmead B., Salzberg S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12:357-360. doi: 10.1038/nmeth.3317.
6. Smit A., Hubley R., Green P. RepeatMasker Open-4.0.6 2013-2015. http://www.repeatmasker.org 2018.