Centromeric sequence organization in medaka

Shingo Tomioka

(Morishita laboratory: Laboratory of Omics)

Background: The centromere is a chromosomal region where kinetochore forms in mitosis and meiosis and plays a critical role for proper chromosome segregation. Despite its fundamental biological importance, the mechanism of the centromere specification and formation, and the contribution of underlying DNA sequences on them are not well understood. In many eukaryote species, the centromeres comprised mainly of repetitive DNA, namely satellite DNA and/or retrotransposons. The repetitive sequences can span several mega-bases at the centromeres, making it difficult to assemble these regions with traditional Sanger sequencing or second-generation sequencers such as Illumina and 454 because of their relatively-short read length. Consequently, in-depth analyses of the repetitive centromeric sequences have been limited. Recently-developing long-read sequencing technologies of PacBio and Oxford Nanopore have great potential to capture longer-range sequence structure and dramatically improve the understanding of centromeric sequences. In the recent few years, several studies using PacBio sequencing that revealed centromeric sequence organization with largely-improved resolution have been reported.

The author's laboratory has recently assembled the genomes of three medaka inbred strains (Hd-rR, HNI and HSOK established from southern and northern Japanese populations, and east Korean population, respectively), which are estimated to have separated around 20 million years ago, using PacBio sequencing reads. Based on these newly-assembled medaka genomes, this study has conducted an in-depth analysis of centromeric sequences in medaka.

Results: The centromeric satellite was estimated to comprise 1–2% (8–16 Mb) of the medaka genome, which is in the same order as other eukaryote species whose centromeric sequences have been characterized to date. The centromeric repeats were identified in 19 out of 24 chromosomes in the Hd-rR genome, 9 in HNI and 13 in HSOK, reaching over 300 kb in a single chromosome at maximum. All the centromeric repeat arrays were truncated at their contig ends, suggesting that highly-homogenized repeat sequences comprise the core centromeric regions.

Inter-chromosomal comparison of the centromeric satellite sequences revealed the presence of four super-chromosomal subfamilies in the satellites. Each satellite subfamily belongs to generally different subsets of the chromosomes. Similar subfamilies have been observed in human centromeric alpha-satellite. Interestingly, the subfamilies exhibit distinct preference for the centromeric positions in chromosomes. It has been known that the teleost lineage underwent a whole-genome duplication (WGD) more than 300 million years ago and since then medaka has largely retained the genomic structure without major chromosome rearrangements. The result of the inter-chromosomal sequence comparison showed that the chromosome pairs that derived by the WGD do not necessarily share similar satellite sequences. This observation of inter-chromosomal sequence similarity that is irrelevant from the evolutionary rearrangement of the chromosomes suggest that centromeric sequences have been exchanged between the chromosomes and homogenized, the process known as concerted evolution. The distinct preference of the centromeric positions between the satellite subfamilies imply that the centromeric positions may affect the effectiveness of inter-chromosomal sequence exchange.

It was also observed that the three strains conserve similar centromeric satellite sequences at the corresponding chromosomes. On the other hand, long-range sequence organization such as linear ordering of the satellites were not conserved among the strains, suggesting that the centromeric sequence

arrays have undergone independent structural rearrangements within each species after their separation.

Conclusion: This is, to the best of the author's knowledge, the first in-depth analysis of the centromeric sequence organization in fish species. This study revealed the conservation of centromeric satellite sequences among the three medaka strains which separated around 20 million years ago. The inter-chromosomal sequence similarity irrelevant of genomic rearrangements supports the process of the concerted evolution in medaka centromeric sequences. The centromeric sequence organization characterized in this study provides insights on centromeric sequence evolution and serves as a basis for understanding the possible role of DNA sequences in centromere specification and function.