Results

Genome assembly of three medaka strains

Three medaka inbred strains were recently sequenced with PacBio single-molecule real-time (SMRT) sequencing and were assembled by the author's laboratory (Ichikawa *et al.*, unpublished; see Methods for an overview of the assembly procedure). Two strains (Hd-rR and HNI) were established from northern and southern Japanese populations, respectively and the other one (HSOK) was established from eastern Korean population. The northern and southern Japanese populations are estimated to have separated 18 million years ago (MYA), whereas the ancestor of the two Japanese populations and that of the eastern Korean population are estimated to have separated 25 MYA [**?**].

Genomic abundance of centromeric repeats

Melters *et al.* (2013) estimated that the medaka candidate centromeric satellite comprise 0.32% of the medaka genome. However this estimation can underestimate the true genomic abundance due to its identification strategy. In order to better infer the genomic abundance of the centromeric satellite, PacBio raw reads were searched for the centromeric satellite sequence.

Genomic fraction of the centromeric repeat was estimated by searching PacBio subreads for the representative monomer sequence. The genomic fraction in Hd-rR and HNI genomes were estimated to be ~1%, while that in the HSOK genome was ~2% (Table 1). This difference is consistent with the previous observations that centromeric repeat array size in a chromosome can vary up to 20-fold within a species [**?**]. Assuming the genome size to be 800 Mb, the centromeric satellite comprise 8–16 Mb of the genome, which implies each chromosome has around 500 kb of centromeric satellite on average. This is concordant with the observations that the centromere of many higher eukaryotes studied to date are characterized by hundreds to thousands of kilobases of satellite sequences [**?**]. Although quantifying the centromeric satellite in erroneous PacBio reads can lead to slight underestimation, it provides much more reliable estimation than estimating by short Sanger sequencing reads.

Validation of centromeric sequence assembly

Repetitive nature of centromeric sequences often accompanies the possibility of misassembly []. In order to validate the assembly at the centromeric regions, PacBio raw subreads were mapped to the assembled genomes and read coverage over centromeric regions was visually inspected on a genomic browser.

PacBio subread were mapped to the medaka genomes by BLASR [] with a stringent mapping parameters (see Methods). The assembly validity was then visually inspected on the genomic browser by confirming enough number of subreads covered the centromeric repeat arrays without breaks. Most part of the centromeric sequences were covered by enough number of subreads, although a small number of exceptions were observed in chromosomes 9, 13 and 20 in the Hd-rR genome, which contained one or two breaking points that were not spanned by subreads (Supplementary Fig. S1). Although PacBio read-based assembly validation cannot completely exclude the possibility of mis-assembly, indeed long-range ordering over the centromeric repeat arrays can be inaccurate, nevertheless relatively narrow range of assembly can be ascertained and that is surely informative for observing sequence composition of a specific chromosome or inter-chromosomal sequence similarity.

Centromeric repeat distribution

The distribution of centromeric repeats in the three medaka strain genomes were revealed by searching their genomes using RepeatMasker (Table S2). For those chromosomes that have >1 kb centromeric repeat, positions of the centromeres in chromosomes were classified, employing the nomenclature defined by Levan *et al.* (1964) (Table S2). Although the nomenclature was originally based on microscopic inspection of the centromeres in chromosomes rather than repeat distribution in the DNA sequence level, nevertheless the sequence-based classification conducted here is informative for inferring evolutionary relationship between the chromosomes. The composition of positional types were consistent with a previous karyotype study []. Centromeric positions of the same chromosome were mostly conserved among the strains, confirmed by observing the corresponding pair of genetic markers flanked the repeat arrays, with only two exceptions in chromosomes 4 and 6 (Supplementary Fig. S2). For chromosome 4, Hd-rR had an acrocentric repeat array whereas HSOK had a metacentric array. For chromosome 6, all the three strains had acrocentric repeat arrays but those of Hd-rR and HSOK and that of HNI were on the opposite side of the chromosome. As the karyotype study has revealed that the three strains possess slightly different sets of centromeric positions [], the difference of chromosomes 4 and 6 may be derived from the *bona fide* karyotype difference. Of note, Hd-rR chromosome 21 possessed metacentric and acrocentric arrays of nearly the same length (41.6 kb and 45.5 kb, respectively; Supplementary Fig. S2), thus it may be a dicentric chromosome where one of the arrays forms the functional centromere whereas the other is silenced.

Table 1: Centromeric repeat genomic abundance

strain	total subreads	passed subreads	passed subreads	repeats in passed subreads	estimated genomic abundance
Hd-rR	13,359,879	4,586,550 (34.33%)	34,933,754,979 bp	354,930,731 bp (1.02%)	8.13 Mb
HNI	14,777,797	7,265,969 (49.17%)	28,478,925,597 bp	338,807,989 bp (1.19%)	9.52 Mb
HSOK	5,527,528	1,955,979 (35.39%)	23,106,352,588 bp	460,716,149 bp (1.99%)	15.95 Mb

Table 2: Centromeric repeat distribution

	Hd-rR		HNI		HSOK	
chromosome	total repeat (bp)	position	total repeat (bp)	position	total repeat (bp)	position
1	48805	SM	0	-	0	-
2	54844	M	3831	M	64213	M
3	52681	ST	0	-	0	-
4	10513	A	39	-	305521	M
5	0	-	10605	A	0	-
6	8226	A	1635	A	7020	A
7	0	-	12911	A	25917	A
8	59863	SM	0	-	324346	SM
9	40159	SM	141	-	137	-
10	0	-	14685	ST	0	-
11	4755	A	4513	A	66412	A
12	232280	SM	25683	SM	40516	SM
13	35778	A	608	-	901	-
14	33284	A	532	-	0	-
15	0	-	51	-	63112	A
16	12804	A	1241	-	0	-
17	1588	A	311	-	559	-
18	23853	SM	0	-	9236	SM
19	131040	SM	4830	SM	4757	SM
20	96309	ST	181	-	17574	ST
21	87124	M/A	2131	A	0	-
22	61066	A	0	-	4942	A
23	6580	M	0	-	25847	SM
24	0	-	0	-	0	-
anchored total	1,001,552		83,928		961,010	
unanchored total	3,279,256	(5.89%)	2,254,882	(3.16%)	11,273,168	(17.5%)
total	4,280,808		2,338,810		12,234,178	
positions summary	2M+6SM+2ST+8A (6U)		1M+2SM+1ST+5A (15U)		2M+5SM+1ST+5A (11U)	

RepeatMasker hits against the medaka centromeric satellite were collected over each chromosome. The centromeric positions were determined by repeat distribution on chromosomes employing the nomenclature by Levan *et al* (1964). Note that Hd-rR chromosome 21 possessed centromeric repeat arrays of nearly the same length (41.6 kb and 45.5 kb) at the positions corresponding to metacentric and acrocentric, thus described as 'M/A'. M, metacentric; SM, submetacentric; ST, subtelocentric; A, acrocentric; U, unknown (due to the lack of centromeric repeats).

Inter-chromosomal centromeric sequence conservation

Previous studies have revealed that centromeric sequences exhibit inter-chromosomal conservation that are considered to be derived from evolutionary process of chromosome formation []. In order to reveal the presence of inter-chromosomal relationship of centromeric repeats in medaka genomes, chromosomal-representative satellite monomers were collected and clustered.

Specifically, centromeric repeat arrays in each chromosome were decomposed into satellite monomers by RepeatMasker and the monomers were clustered by DNACLUST [] with >85% sequence similarity threshold. For those clusters that have ≥10 members, the monomer with the longest sequence in the cluster was chosen as the representative monomer of the cluster. All-vs-all pairwise alignment of the representative monomers from each chromosome along with the representative monomer identified by Melters *et al.* was performed and the distance between the monomers were calculated. Based on this distance, hierarchical clustering of the chromosome-representative monomers were performed (Fig. 1). The chromosome-representative monomers were clustered into four groups, revealing the presence of super-chromosomal subfamilies (Table 3). Many (15 out of 24) chromosomes (chr. 2, 3, 5, 6, 7, 10, 11, 12, 14, 15, 16, 18, 20, 22 and 23) were assigned exclusively to one of the four subfamilies. Five chromosomes (chr. 1, 4, 8, 13 and 19) were clustered into two or three subfamilies but significantly more monomers were classified to one subfamily over others, thus assigned to the dominant subfamily. Chromosomes 9 and 21 were classified into two subfamilies with no significant preference. Chromosome 17 and 24 were not able to be classified due to the lack or insufficient amount of centromeric repeats in either of the three assembled genomes. Overall, 22 out of 24 chromosomes were assigned to one or two subfamilies. Intriguingly, each subfamily exhibited distinct preference of centromeric positions in chromosomes; namely subfamily 1 for acrocentric, subfamily 2 and 3 for submetacentric and subtelocentric and subfamily 4 for metacentric, respectively (Table 3).

In those chromosomes that had sufficient amount of centromeric repeats in multiple strains, most (7 out of 9) chromosomes were classified into the same subfamilies among strains. One of the exceptions was chromosome 19, where representative monomers from Hd-rR and HSOK were classified into SF2 while that of HNI into SF3, although the repeats from each strain were confirmed to locate in close position of the chromosome as they were flanked by a corresponding pair of genetic markers (Supplementary Fig. S2). This discordant classification may be due to assembly of different subregion of the corresponding repeat array among strains or may have been caused by misassembly in one or more strains. The other exception was chromosome 21, where the representative monomers from the acrocentric array of Hd-rR were classified into SF1, those from the metacentric array of Hd-rR and from the acrocentric array of HNI into SF3. The two acrocentric arrays from Hd-rR and HNI were located at close but distinct positions in the chromosome (Supplementary Fig. S2), thus it may as well contain different repeat sequence profiles and be classified into different subfamilies.

Methods

Sequencing and genome assembly

Sequencing and assembling the three medaka strain genomes were carried out by Kazuki Ichikawa and Jun Yoshimura in the same laboratory. The detail of the methods will be described in Ichikawa *et al.* (unpublished). Here a brief overview of the methods is given.

The genomes were sequenced with PacBio SMRT sequencing and were assembled into contigs using FALCON assembler []. The contigs were then polished with PacBio reads using Quiver [] and with Illumina reads using Pilon []. A number of contigs that contained long centromeric repeat arrays were not polished with Pilon because it was observed that extremely more bases were corrected on centromeric regions than other genomic regions presumably due to mismapping of short reads. The polished contigs were mapped to the chromosomes using SNP genetic markers. Hd-rR contigs were further scaffolded using BAC- and fosmid-end pair reads and a number of unanchored contigs were positioned into the chromosomes using Hi-C contact frequency data.

Validating centromeric sequence assembly

PacBio raw subreads were mapped to the assembled genomes by BLASR []. Those mapped subreads that had i) >5 kb alignment length, ii) >80% sequence identity over the entire alignment and iii) >85% sequence identity on both the 1-kb ends of the alignment were selected for visualization on the genomic browser. The centromeric repeat regions were inspected and confirmed on the genome browser that they were covered by enough number of overlapping subreads (at least 5 subreads and typically way more reads at every position) without breaks (Supplementary Fig. S1).

Estimating genomic abundance of centromeric repeats

In order to minimize the effect of high error rate of PacBio sequencing on abundance estimation of the centromeric repeats, only high quality subreads were used for this step. Specifically, subreads were filtered with the criteria that average base quality over the all bases >10.

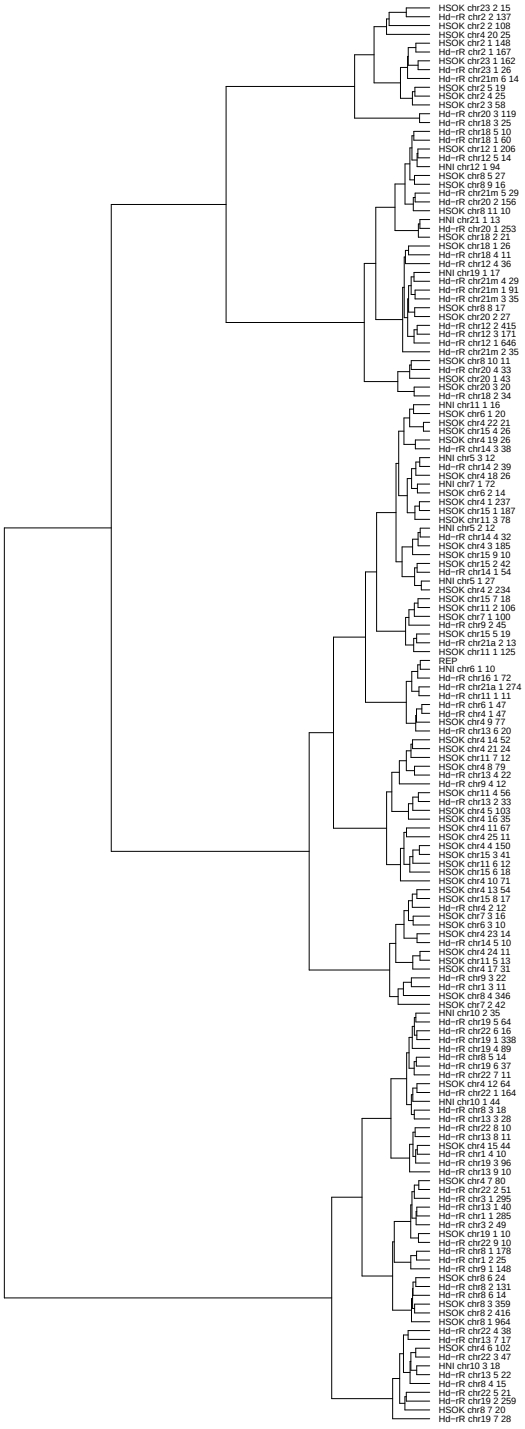

Figure 1: Hierarchical clustering of chromosome-representative monomers. Monomers are labeled as species, chromosome, cluster index, number of the cluster constituents.

Table 3: Super-chromosomal subfamilies of centromeric repeats

SF	Hd-rR	HNI	HSOK	combined	positions
1	4,6,9,11,14,16,21a (1,13)	5,6,7,11	4,6,7,11,15 (8)	4,5,6,7,9,11,14,15,16,21a (1,8,13)	1M+1SM+14A (2SM+1A)
2	1,3,8,9,13,19,22	10	8,19 (4)	1,3,8,9,10,13,19,22 (4)	6SM+2ST+2A (1M)
3	12,18,20,21m (8)	12,19,21a	12,18,20 (8)	12,18,20,21m,21a (8,19)	1M+8SM+2ST+1A (2SM)
4	2,23 (21m)		2,23 (4)	2,23 (4,21m)	3M+1SM (2M)

Chromosomes were classified into four subfamilies (SF). Chromosomes in brackets are the ones that have significantly more amount of repeats classified into another subfamily. Hd-rR chromosome 21 possessed two distantly-positioned arrays, thus is notated as 21m (metacentric) and 21a (acrocentric; see Table S2 for detail). Summarizing the chromosomes from the three strains, 22 out of the 24 chromosomes were assigned to one or two subfamilies. Notation of the centromeric positions are the same as Table S2.

Also, subreads shorter than 1 kb were excluded. The filtered subreads were then scanned by RepeatMasker with a sensitive setting using the medaka representative centromeric satellite monomer sequence as a custom library. Genomic fraction of the medaka centromeric satellite for each strain was estimated by the ratio of total amount of masked centromeric satellite in the total length of the filtered subreads (Table 1).

Revealing centromeric repeat distribution and centromeric positions

The three medaka strain genomes were searched for the medaka centromeric satellite by RepeatMasker with sensitive setting. For those chromosomes that have >1 kb centromeric repeat, positions of the centromeres were classified employing the nomenclature defined in Levan *et al.* (1964). The nomenclature divides a chromosome equally into eight portions and classify the chromosome by the position of the centromere from the two most inners to the two most outers as metacentric, submetacentric, subtelocentric and acrocentric. In this study, chromosomes were classified into a portion that contains the largest amount of centromeric repeats.

Inter-chromosomal centromeric sequence comparison

Centromeric repeat arrays in each chromosome of the three strains were decomposed into satellite monomers by RepeatMasker with sensitive setting. The monomer sequences within each chromosome were then clustered into groups of >85% sequence similarity by DNACLUST []. For those clusters that have ≥10 members, the monomer with the longest sequence in the cluster was chosen as the representative monomer of the cluster. All-vs-all pairwise alignment of the chromosome-representative monomers along with the representative monomer identified by Melters *et al.* was performed by EMBOSS needle program. The distance between a pair of two monomers was calculated as below:

$$\text{distance} = 1 - \frac{\text{number of matched bases}}{\text{length of shorter monomer}}$$

Based on this distance, hierarchical clustering of the chromosome-representative monomers were performed by "hclust" function in R with "ward.D2" method.

Supplements

Figure S1: Centromere landscapes in the medaka genomes

Figure S2: Centromeric repeat distribution