Results

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.

Centromeric repeat distribution

The distribution of centromeric repeats in the three medaka strain genomes were revealed by searching their genomes using RepeatMasker (Table 2). 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 2). 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 figure S**??**). 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 figure S**??**), thus it may be a dicentric chromosome where one of the arrays forms the functional centromere whereas the other is silenced.

Methods

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. 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.

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).