Introduction

The centromere

The centromere is a chromosomal region where kinetochore forms and plays a critical role for proper chromosome segregation in mitosis and meiosis. The centromere is characterized by the presence of centromere-specific histone H3 variant CENH3 (also known as CENP-A). In the majority of species studied so far, the centromere is comprised of repetitive DNA [1]. Despite its fundamental biological importance, the mechanism how the position of the centromere is specified are still insufficiently understood [2].

Sequence composition of the centromere

Repeat-based regional centromere is the common structure in eukaryote species. Many species possess satellite DNA specific to the species. A well-known example is 171-bp AT-rich alpha-satellite observed in human and many other primate species. In the human genome, tandemly-repeated alpha-satellites comprise hundred kilobase- to megabase-sized arrays in each chromosome. The alpha-satellite monomer sequence exhibits high divergence up to 40% within a species. In core centromeric regions, multiple monomers comprise a higher-order repeat units which themselves iterate tandemly with extremely high identity (>95%); this structure is called higher-order repeats (HORs). Another major component of eukaryote centromeres is retrotransposons. Retrotransposon-based centromeres were widely observed in plant species. Satellites and retrotransposons are not mutually exclusive, rather intermingled structure of them are commonly observed. These regional centromeres are flanked by heterochromatic pericentromeres. Centromeres and pericentromeres are characterized by distinct chromatin structures which are regulated by different sets of epigenetic marks [].

While repeat-based regional centromeres are the most common structure in eukaryotes, some species possess different types of centromeres. Budding yeast *Saccharomyces cerevisiae* has a "point centromere" where ~125-bp specific sequences form centromeres. In nematode *C. elegans* and some insects and plants, the spindle microtubules attach all along a chromosome and the entire chromosome functions as a centromere, called "holocentromere".

Early studies

Fundamental understanding of centromeric sequence characteristics was established by a number of early studies in 1980's and 1990's, mainly targeting human centromeres. These studies based on experimental methods such as genomic fragmentation by restriction enzymes, pulsed field gel electrophoresis and DNA hybridization.

The findings include approximate estimation of repeat array size and its divergence among individuals [3], the presence of chromosome-specific alpha-satellite HOR patterns [4] and super-chromosomal subfamilies of alpha-satellites [5]. It was also revealed that alpha-satellite is widely shared within primate species and even HOR patterns are shared with closely-related species including chimpanzee, gorilla and orangutan [6].

The genome projects era

Although the basic characteristics of centromeric sequences were revealed by these early studies, understanding detailed sequence organization of the centromere has been challenging in many species, due to the difficulty of assembling its highly-repetitive sequences. In the human genome project, which declared completion in 2003, large portion of centromeric sequences were missed as huge gaps. Whereas divergent monomeric portion around pericentromeres were assembled in many chromosomes, the assemblies reached more homogeneous HOR regions in a limited number of chromosomes [7, 8]. Nevertheless, subsequent analyses on these few chromosomes revealed sequence landscapes with never-seen resolution [9–12].

Second-generation sequencing-based studies

Although second-generation sequencers (SGSs) represented by Illumina and 454 accomplished a number of genome assemblies [13], they achieved virtually no improvement in centromeric sequence assembly because of their short read length, emphasizing that centromere studies need specific approaches considering its sequence characteristics. Nonetheless, their high throughput sequencing combined with chromatin immunoprecipitation (ChIP-seq) facilitated identification of centromere-associated sequences in many species [] and characterization of functional regions in the assembled centromeric sequences [14].

Computational studies

A number of computational studies on centromeric sequences were conducted using Sanger and Illumina whole genome shotgun (WGS) sequencing data, some of which made remarkable achievement. Some studies identified candidate centromeric satellite sequences from WGS data [15, 16], whereas others identified novel HOR patterns from assembled sequences [17] or from WGS data [18].

Melters *et al.* [16] identified candidate centromeric satellite sequences of 282 species (204 animal and 78 plant species) using WGS data from various sequencing platforms, mainly from Sanger and Illumina. They based on the assumption that the most abundant tandem repeat in a genome derives from centromeric sequences, which is true for most species whose centromeric sequences has been previously characterized. This study revealed that centromeric satellites from various eukaryotic species do not share common properties such as repeat unit length, GC content or genomic abundance and that centromeric satellite sequences are conserved among only closely-related species of within 50 million years after separation. These results confirmed a traditional view that centromeric sequences evolves rapidly, independently of the rest of genomic sequences [19].

Another remarkable computational study is from Miga *et al.* [20], in which they generated centromeric array sequences

of each human chromosome, using graph-based probabilistic models constructed from Sanger WGS reads. Although the generated sequences do not guarantee long-range ordering of the satellite sequences, they adequately represent local ordering, thus provide useful scaffolds for mapping sequencing reads and/or other downstream analyses. The generated centromeric sequences, moreover, has been included in the latest human reference genome (GRCh38) [21].

Feasibility of long-read sequencing

Despite these development in sequencing technologies and computational methods, long-range organization of centromeric sequences could not be resolved mainly due to the short read length of Sanger or SGS technologies. However, recently-emerging long-read technologies of PacBio [22] and Oxford Nanopore [23] are expected as promising tools for centromere studies [24, 25].

PacBio single-molecule real-time (SMRT) sequencing yields average read length of ~15kb and longest of ~50kb with P6-C4 chemistry. This long read length enables to capture long-range structure such as HORs directly and provides more opportunity to anchor repetitive sequences to adjacent unique regions. Although error rate of PacBio sequencing is relatively high (~15%), the error pattern is believed to be completely random, thus can be successfully corrected with enough sequencing coverage [26]. Furthermore, in contrast to Sanger or SGS technologies which suffer from sequencing bias deriving from PCR amplification and/or vector cloning steps, amplification-free PacBio sequencing involves no apparent sequencing bias [27]. In recent years, a number of studies have reported dramatic improvement in genome assemblies using PacBio sequencing (TODO: add citation!), some of which improved centromere assembly as well [28–30]. Among these, a *de novo* assembly of a grass genome covered three of the nine centromeric regions which are comprised mainly of 155-bp satellite sequences and spans ~400 kb [28]. Some studies targeting centromere-associated repeat-rich regions have also been conducted with PacBio sequencing [31,32], and a computational tool for detecting HOR patterns from long reads was also developed [33].

Oxford Nanopore sequencing yields even longer read length (e.g. MinION sequencer routinely yields >150 kb read) with accuracy of ~92% [23]. However, some early-adopting laboratories of it including the author's laboratory observe unstable read length and much higher error rate than the officially announced rate (Kin Fai Au, personal communication), thus the community awaits improvement and sophistication of its sequencing technology and protocols.

Limited number of in-depth studies, especially in fish species

Due to the long-standing difficulty in assembling centromeric regions, in-depth analysis of centromeric sequences have been conducted in a limited number of species, including human, some other primates, mouse, some plants, *Drosophila* and yeasts, and scarce in fish species. Centromere-associated satellites have been identified or confirmed in zebrafish, seabass and stickleback by CENP-A-targeted ChIP and/or FISH experiments [34–36], however only limited amount of them have been included in the assembled genomes [29, 36, 37]. Medaka, as well as zebrafish, has traditionally played an important role as a model organism of fish species [38]. Cloning-based assembly of its genome was conducted about ten years ago [39], however centromeric regions were largely missed in gaps, as was the case in many contemporary assembly projects. A 156-bp candidate centromeric satellite sequence of medaka was identified in a computational study by Melters *et al* [16], though whether this sequence truly derives from the centromere has not been confirmed by experimental methods such as ChIP-seq or FISH.

This study

The author's laboratory recently assembled three medaka inbred strain genomes using PacBio long reads and achieved dramatic improvement in the assembly quality (Ichikawa *et al.*, unpublished). Based on these high quality medaka genomes, this study conducted in-depth analysis of medaka centromeric sequences. This study revealed the presence of inter-chromosomal relationship of the satellite sequences and its conservation among the strains. Also the evidence of higher-order structure (HORs) was captured.

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 from eastern Korean population. The two 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 [40].

Genomic abundance of centromeric repeats

This study started with a candidate centromeric satellite sequence of medaka which was identified in a previous computational study [16]. In that study, Melters *et al.* estimated that the 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 with RepeatMasker (version 4.0.6; A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org). The genomic fraction in Hd-rR and HNI genomes were estimated to be ~1%, while that in the HSOK genome was ~2% (Table S2). This difference is consistent with the previous observations that centromeric repeat array size in a chromosome can vary up to 20-fold among individuals within a species [20]. 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 [1]. Although quantifying the centromeric satellite in erroneous PacBio reads can lead to slight underestimation, the estimation should be more reliable than the clustering-based estimation using short Sanger reads in the previous study [16].

Centromeric repeat distribution

The distribution of the centromeric satellites in the three medaka genomes was investigated. The three assembled genomes were searched for the candidate centromeric satellite sequences using RepeatMasker (Table 1, Fig. S1). The results revealed that all the identified centromeric satellite arrays were truncated by contig gaps at either or both ends, suggesting none of the centromeric regions was spanned by a single contig. In the Hd-rR and HSOK genomes, ~1-Mb centromeric satellites were identified in total, respectively, whereas only ~80 kb was identified in the HNI genome. This substantial difference in the amounts of identified centromeric satellite is presumably due to the difference in read length. The HSOK genome was sequenced with the newest P6-C4 chemistry and the average read length was 11 kb; Hd-rR was sequenced with the combination of P6-C4 and older P5-C3 and P4-C2 chemistries and the average read length was 6.5 kb; HNI was sequenced with P5-C3 and P4-C2 with the average read length of 3.6 kb (Table S1). In addition, substantial amount of the centromeric satellite was identified in the contigs that failed to anchor to the chromosomes. The enrichment of identified centromeric satellite in unanchored contigs to in anchored contigs was as big as 12-fold in HSOK and 27-fold in HNI, in contrast to relatively small 3-fold enrichment in Hd-rR (Table 1). In the Hd-rR genome assembly, contigs were scaffolded using BAC-/Fosmid-end sequencing reads and Hi-C contact frequency data, which successfully anchored a number of contigs containing centromeric satellites, emphasizing the effectiveness of complementing the long-read sequencing with other methods that capture even longer-range information.

For those chromosomes that have >1 kb centromeric repeat, positions of the centromeres in chromosomes were classified into metacentric, submetacentric, subtelocentric and acrocentric, employing the nomenclature by Levan *et al.* [41] (Table 1). Although this nomenclature originally based on karyotype observation rather than DNA sequence level and the positions induced from the two levels can slightly differ, the sequence-based classification conducted here is nevertheless informative for interpreting subsequent analyses. The number of chromosomes classified to each type was in line with previous karyotype studies [].

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 (Fig. S1). 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 located on the opposite end 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 *bona fide* karyotype difference. Notably, Hd-rR chromosome 21 possessed metacentric and acrocentric arrays of nearly the same length (41.6 kb and 45.5 kb, respectively; Fig. S1), thus this chromosome may dicentric where one of the arrays forms the functional centromere whereas the other is silenced.

Centromeric sequence mapping by FISH

To confirm that the candidate centromeric satellite sequence truly localizes to the centromeres, FISH experiment was conducted. Probe sequences were designed by the author and FISH experiments were carried out by a collaborator (see Methods).

The candidate satellite was first used as a hybridization probe and signals were observed only from 5~7 pairs of chro-

Table 1: Centromeric repeat distribution

	Hd-rR		HNI		HSOK	
chromosome	total repeat (bp)	position	total repeat (bp)	position	total repeat (bp)	position
1	48,805	SM	0	-	0	-
2	54,844	M	3,831	M	64,213	M
3	52,681	ST	0	-	0	-
4	10,513	A	0	-	305,521	M
5	0	-	10,605	A	0	-
6	8,226	A	1,635	A	7,020	A
7	0	-	12,911	A	25,917	A
8	59,863	SM	0	-	324,346	SM
9	40,159	SM	0	-	0	-
10	0	-	14,685	ST	0	-
11	4,755	A	4,513	A	66,412	A
12	232,280	SM	25,683	SM	40,516	SM
13	35,778	A	0	-	0	-
14	33,284	A	0	-	0	-
15	0	-	0	-	63,112	A
16	12,804	A	0	-	0	-
17	1,588	A	0	-	0	-
18	23,853	SM	0	-	9,236	SM
19	131,040	SM	4,830	SM	4,757	SM
20	96,309	ST	0	-	17,574	ST
21	87,124	M/A	2,131	A	0	-
22	61,066	A	0	-	4,942	A
23	6,580	M	0	-	25,847	SM
24	0	-	0	-	0	-
anchored total	1,001,552		80,824		959,413	
unanchored total	3,279,256	(5.89%)	2,254,882	(3.16%)	11,273,168	(17.5%)
total	4,280,808		2,335,706		12,232,581	
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* [**?**]. 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).

mosomes. Additional probes were designed to complement hybridization to other chromosomes. For each chromosome, satellite monomers were collected from satellite arrays and the collected monomers were aligned back to the original array with BLASTN [42]. Then a monomer with the highest score was chosen as the representative monomer of the chromosome, where the score was defined as:

$$score = \sum_{\text{hits}} \text{alignment identity} \cdot \frac{\text{alignment length}}{\text{query length}}.$$

Representative monomers obtained from each chromosome were then aligned to the Hd-rR genome with BLASTN and three monomers that exhibited high identity to different subsets of chromosomes were chosen as additional probe sequences (Fig. **??**). The additional probes successfully hybridized to some chromosomes that the first probe failed to hybridize, although two additional probes hybridized to less number of chromosomes than expected from the *in silico* alignment results. When all the probes combined, signals were observed at the centromeres of ~13 pairs of chromosomes (Fig. 1). This result confirmed that the candidate centromeric satellite truly derives from the centromeres. Moreover, the number of the chromosomes having each centromeric positions were largely consistent with the sequence-based results in the previous section.

Validation of centromeric sequence assembly

Repetitive nature of centromeric sequences inevitably accompanies the possibility of misassembly. In order to validate the centromeric sequence assembly, PacBio raw subreads were mapped to the assembled genomes and read coverage over centromeric regions was visualized for manual inspection.

PacBio subread were mapped to the medaka genomes by BLASR [43] with a stringent mapping parameters (see Methods). The assembly validity was then manually inspected on the genomic browser by confirming that 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. S2). Although PacBio read-based assembly validation cannot completely exclude the possibility of misassembly, indeed long-range ordering over the centromeric repeat arrays can be inaccurate, nevertheless relatively narrow range of assembly can be ascertained and that is adequately informative for observing sequence composition of a specific chromosome or inter-chromosomal sequence similarity.

Inter-chromosomal centromeric sequence conservation

It is widely known that in some species centromeric sequences exhibit inter-chromosomal conservation that are considered to derive from evolutionary rearrangements of chromosomes and/or frequent sequence exchange as a result of co-localization in the nucleus [6]. In order to reveal the presence of inter-chromosomal relationship of centromeric repeats in the medaka genomes, satellite sequences from each chromosome were compared.

Centromeric repeat arrays in each chromosome were decomposed into satellite monomers by RepeatMasker and the monomers were clustered by DNACLUST [44] 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 pairwise distance was calculated. Based on this distance, hierarchical clustering of the chromosome-representative monomers were performed. The chromosome-representative monomers were clustered into four groups, revealing the presence of super-chromosomal subfamilies (Fig. 2, Table 2). 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 the others, thus they are assigned to the dominant subfamily. Chromosomes 9 and 21 were classified into two subfamilies with no significant preference. Chromosomes 17 and 24 could not 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 (SF) 2 for acrocentric, SF 1 and 3 for submetacentric and subtelocentric and SF 4 for metacentric, respectively (Table 2). This tendency is analogous to the traditional observation that human acrocentric chromosomes share highly identical alpha-satellite sequences [6].

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 SF 1 while that of HNI into SF 3, 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 (Fig. S1). This discordant classification may be because the assemblies of each strain captured different subregion of the corresponding repeat arrays or due to 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 SF 2, those from the metacentric array of Hd-rR and from the acrocentric array of HNI into SF 3. The two acrocentric arrays from Hd-rR and HNI were located at close but distinct positions in the chromosome (Fig. S1), thus it may well contain different repeat sequence profiles and be classified into different subfamilies. The overall conservation of centromeric satellites among the three strains which separated 18 and 25 million years ago is in line with the previous observation that

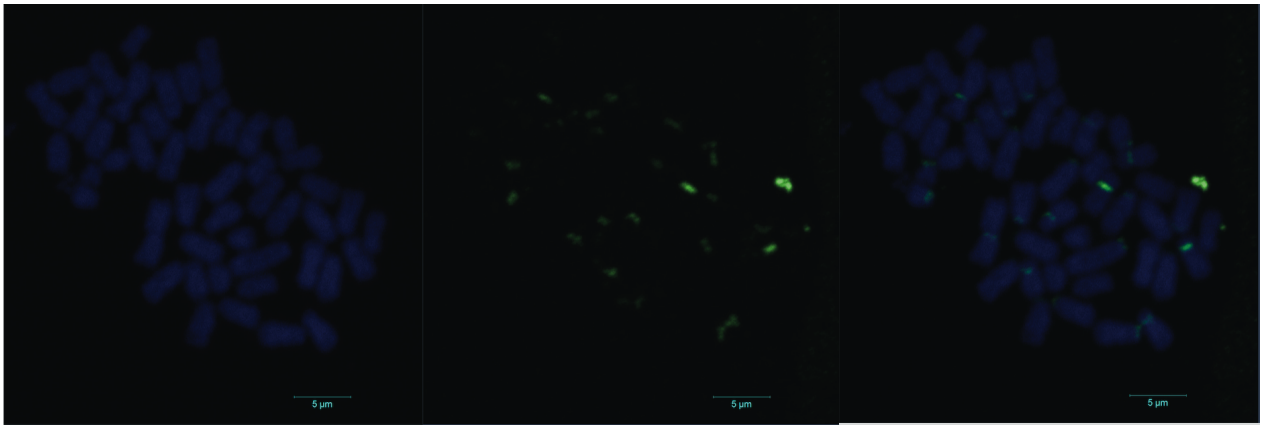


Figure 1: The candidate centromeric satellite sequence and three derivative sequences localized to the centromeres of ~13 pairs of chromosomes. (left) DNA is stained with DAPI. (center) probes are stained green. (right) two images are combined.

centromeric sequences were conserved among species within about 50 million years after separation [16].

Sequence organization at the centromeres

Sequence organization on the assembled centromeric sequences were analyzed. Dot plot of centromeric sequences of each chromosome are shown in Supplementary Figure S2.

HSOK chromosome 8 captured the longest centromeric arrays, namely two arrays of 250 kb and 95 kb flanking an assembly gap (Fig. 3A). These two arrays comprised of the satellites from three subfamilies (SF 1, SF 2, SF 3). SF 1 satellites comprise large inner portion of the arrays, interspersed by SF 2 satellites; these sequences were flanked by much less amount of SF 3 satellites. Multiple alignment of the chromosome-representative monomers revealed that the representative monomer of the forth largest cluster which belongs to SF 2 possessed ~10-bp insertion compared to the representative monomers belonging to SF 1, yet otherwise looks virtually identical (Fig. 4). The assignment of these representative monomers to different clusters was due to the alignment-identity-based definition of the distance between satellite sequences in which large insertion leads to substantial loss of alignment identity. On the other hand, the representative monomers belonging to SF 3 exhibit distinct sequence composition from the monomers in SF 1 and SF 2. Interestingly, the orientation of the satellite sequences switched at the boundaries of SF 1 and SF 3 arrays (Fig. 3A). This suggests the scenario that the SF 1 array inserted into the SF 3 array as a result of a sequence conversion, unequal crossover or other chromosome rearrangement events. Switches of sequence orientation in satellite arrays have also been observed in the pericentromeric regions of human chromosomes [7]. Overall similar sequence organization was observed in the same chromosome of Hd-rR, which had 20-kb and 40-kb SF 1 arrays flanking an assembly gap and a 1-kb SF 3 array at outside of the 20-kb SF 1 array (Fig. 3B).

Another interesting example was HSOK chromosome 4 which captured a over 300-kb nearly continuous array (Fig. 5). This array comprised mainly of SF 2 satellites, interspersed with shorter SF 1 satellite arrays. Also small amount of SF 4 satellites were observed in downstream portion. Furthermore, frequent switches of sequence orientation were observed, some of which correspond to the SF boundaries whereas others do not.

Chromosome 12 was the only chromosome that all the three strain genomes captured >10-kb centromeric arrays. The Hd-rR assembly reached the centromeric region from the both sides; HNI reached from the p-arm side; HSOK reached from the q-arm side (Fig. 6A). All the arrays comprised of SF 3 satellites. In order to examine if the sequences of these centromeric transitional regions are conserved among the strains, dot plots were drawn (Fig. 6B, C). Whereas modest sequence conservation was observed in the surrounding unique regions, the sequence structure in the centromeric arrays were not conserved. Similar unconservation of sequence structures in centromeric transitional regions were observed in some other chromosomes (Fig. S3). These results suggest that the repetitive sequences at the centromeres have evolved independently of surrounding unique regions, in line with traditional observations [6].

HOR structures

Dot plot of the centromeric sequences revealed the presence of higher-order repeats (HORs) in many chromosomes.

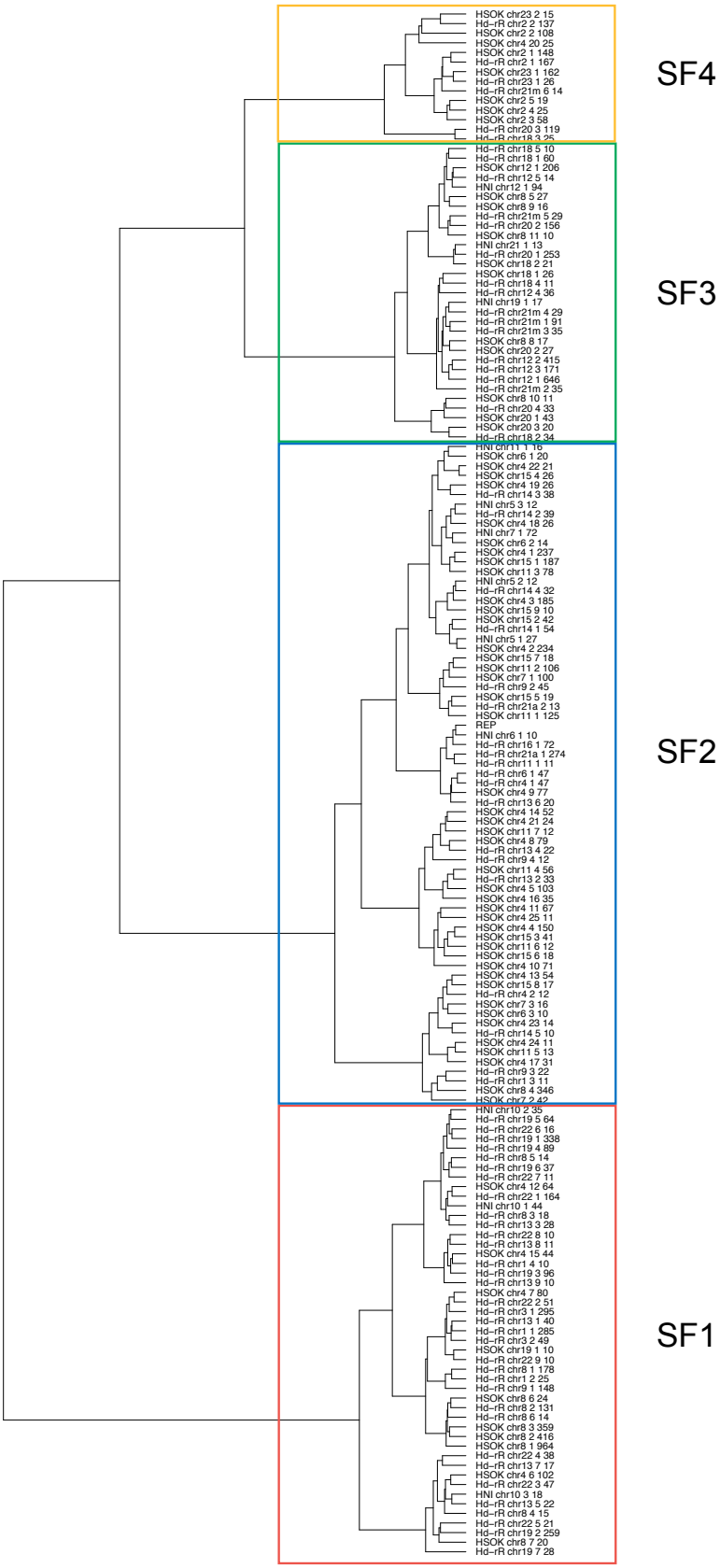


Figure 2: Hierarchical clustering of chromosome-representative monomers. Monomers are labeled as species, chromosome, cluster index, number of the cluster constituents. The clustering revealed four large subfamilies of satellite monomers.

Table 2: Super-chromosomal subfamilies of centromeric repeats

SF	Hd-rR	HNI	HSOK	combined	positions
1	1,3,8,9,13,19,22	10	8,19 (4)	1,3,8,9,10,13,19,22 (4)	6SM+2ST+2A (1M)
2	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)
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 1 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 1.

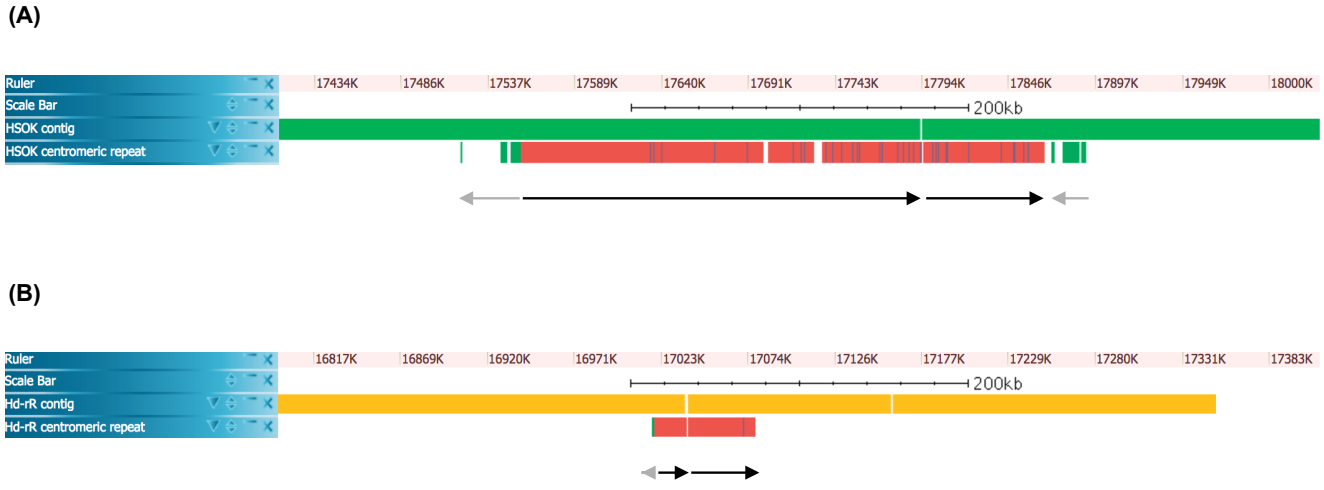


Figure 3: Sequence organization of chromosome 8 centromeric regions. (A) HSOK chromosome 8 had 250-kb and 95-kb repeat arrays flanking an assembly gap. SF 1 satellites (red) comprise large inner portion of the arrays, interspersed by SF 2 satellites (blue). These sequences were flanked by shorter SF 3 satellite arrays (green). The orientation of the satellite sequences switched at the boundaries of SF 1 and SF 3 arrays (indicated by black and grey arrows). (B) Hd-rR had similar sequence organization as HSOK.

```
6  (24)  SF1    ---AAATAAGACTAACTTTGACTTTTTAGTGCATTTTGTGCAAAAAATTCACTTT-----------TCATTCAAAGGTGTCAAAAAAGCG      76
1  (964) SF1    AACTGCAAATGAGAACTTTGACTTTTGAGTTCATTTTATGCTCAAGAATCAGTTTTTCA----------TTCAAAAGTGTCAAAAAAGCG      80
2  (416) SF1    AACTGCAAATGAGAACTTTGAATTTTGAGTGCTTTTTGTGTACTAACATCAGTTTT---------TTTTTTCAAAAGTGTCAGAAAAGCG      81
4  (346) SF2    AACTGCAAATGAGAACTTTGACTTTTGAGTGCATTTTATGCTCAAGAATCAGTTTTTCAAAAAACATTTTGCAAAAATGTCAAAAAAGCG      90
3  (359) SF1    AACTGCAAATGAGAACTTTGACTTTTGAGTGCATTTTGTGCAGAAAAATCAGTTTTTCA----------TTCAAAATTGTCAAAAAAGCG      80
7  (20)  SF1    -------------AACTTTGAGTTTTTAGTGCTTTTTGTGCATAAAAATAAGTTTTTCA----------TTCAAAAGTGTCAAAAAGGTG      67
10 (11)  SF3    -AACTACAATGAGATCTTTTCCTTTTAAGTGCGTTTTTTGCTCAAAAAACAATGTTGTCA----------CCAAAAGTTTAGATAAAGC      78
8  (17)  SF3    AATTACATATGAGATCTTTGCCTTTTGAG-AGCGACTGTGATCAAAAATCATTTTGTCACTCAAAAGTGTTAAAACTGTCAA--AAAAGC      87
11 (10)  SF3    AACTGCAAATGGGATTTTTGCCTTTTGAGTGCATTTTGTGCTCAAAAATCATTTTGTCACTCAAAAGTGCTAAAAGTGTCAAAAAAGATT      90
5  (27)  SF3    AAATATAAATGAGAATTTTGCCTTTTGGCTGCGTTTTGTGTTAAAAAAACATTTTGTCACTCAAAAGTGCCAAAAGTGTCAAAAAAAAGC      90
9  (16)  SF3    AATTATAAATAAGATCTTTGCCTTTTGAGTGCATTTTGTGCTCAAATATCATTTTTTAACTCAAAAGTGCTAAAAGTGTCAAAAAAAAGC      90
                           *  ***   ****         * **    *     * * *                 **      *   *

6  (24)  SF1    TTTTCAGTTCTAAATAGCACTGTTTTGACTTCTCAACTCACGTGACAAGAAATAA---CACTTTCTT--    140
1  (964) SF1    TTTTCAGCTCTAAATAGGAATAATTGGACTTCTCAACTCACAACGTTGCAAGTAATAAATCATTCTTT-    148
2  (416) SF1    TTTTCAGGTCTAAAAAGCACTGTTTGGACTTCTCAACTCACAACGTGACAACAAATAACACTTTCTT--    148
4  (346) SF2    TTTTCAGGCTAAAATAGCACTGTTTTGACTTCTCAACTCACAACGTGACAAGAAATAACACTTTCTT--    157
3  (359) SF1    TTGTCAGCTCTAAATAGCAATGTTTGGACTTTTAAGCTTATCACATGACAAGACATAACACTTTCTT--    147
7  (20)  SF1    TTGTCAGCTCTAAAAAGCATTGTTTGGACTTCTCATCTCACAACGTGACAAGAACTAACATT-------    129
10 (11)  SF3    ATTTTTAATCAAATTCTAACTGTTTTGACTTTTTAACTTGAAATGTGACCAAAAAGCAACACTTTTTT-    146
8  (17)  SF3    ATTTTGGCTCAATTTCTAAAAGTTTTGACATTTCAACTTGAAATGTGACCAAATAACAGCACTTTTATG    156
11 (10)  SF3    TTTTTTTCTGAAATTGTAACTGTTTTGACTTTTCAACTTGAGATGTGACCAGAAATCAACACTTTTATG    159
5  (27)  SF3    ATTTTGGCTCAAATTCTGACTGTTTTGACATTTCAACTTGAAATATGACCAAACAACAGCTCTTTTATG    159
9  (16)  SF3    ATTTTGGCTCAAATTCTGACTGTTTTGACTGTGCATCTTTAAACATGACCAAAAAACAACACTTTTTT-    158
                 * *       *      *    ** ***     * **           *
```

Figure 4: Multiple sequence alignment of HSOK chromosome 8 representative monomers. 11 representative monomers of HSOK chromosome 8 were aligned using Clustal Omega (version 1.2.3) [45]. The labels of each sequence represent cluster index (as a descending order of cluster size), number of monomers belonging to the cluster (in brackets) and belonging subfamilies. Asterisks ("*") indicate the nucleotides shared in all the representative monomers. Representative monomer 4 which belongs to SF 2 has ~10-bp insertion compared to SF 1 representative monomers, yet otherwise shares virtually the same sequence composition. SF 3 representative monomers have distinct sequence composition from SF, and SF representative monomers.

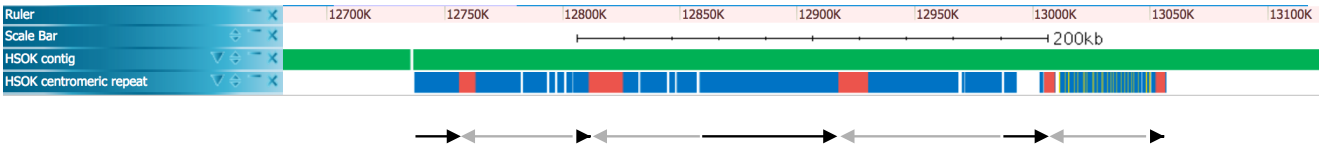


Figure 5: Sequence organization of HSOK chromosome 4 centromeric region. The ~300-kb nearly continuous array was truncated by the contig end at the left end. The array comprised mainly of SF 2 satellites (blue) and these are interspersed by shorter SF 1 satellite arrays (red). Also small amount of SF 4 satellites (yellow) were observed in the right portion. Frequent switches of sequence orientation were observed (indicated by black and grey arrows).

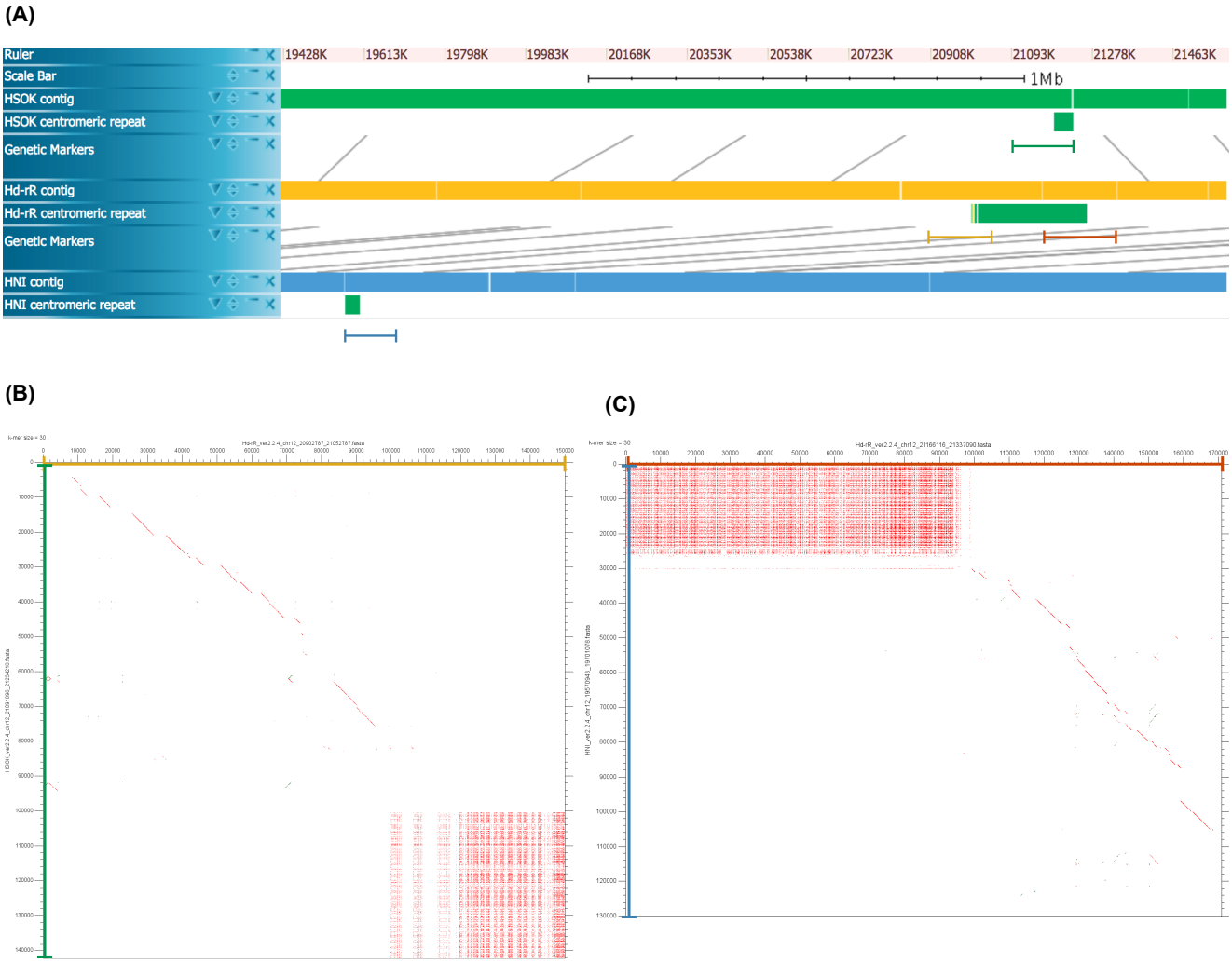


Figure 6: Comparison of the centromeric transitional regions of chromosome 12. (A) The Hd-rR assembly reached the centromeric region from the both sides (separated by a contig gap); HNI reached from the p-arm side; HSOK reached from the q-arm side. The grey lines indicate the positions of corresponding genetic markers. (B) Sequences of the q-arm transitional regions of Hd-rR and HSOK was compared. (C) Sequences of the p-arm transitional regions of Hd-rR and HNI was compared. Dots represent 30-bp exact matches between two sequences. Whereas modest conservation was observed in the surrounding unique sequences (indicated by the chained diagonal lines), no clear conservation was observed within the centromeric array sequences.

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 [46]. The contigs were then polished with PacBio reads using Quiver [47] and with Illumina reads using Pilon [48]. 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 [43]. 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. S2).

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 (version 4.0.6; A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org) 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 S2).

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.

Centromere FISH

FISH experiment was carried out by Yusuke Inoue at the Department of Biological Sciences, Graduate School of Science, The University of Tokyo.

Centromeric satellite DNA were synthesized by annealing and extension of two DNA oligos using TaKaRa ExTaq (TaKaRa), followed by subcloning into pCR™II-TOPO®vector (Thermo). DNA probes were prepared by cutting and labeling the plasmid DNA with biotin, using Nick Translation Kit (Roche). Medaka fibroblast cells were treated with 0.05 μg/ml of corcemid (for probe1,2) or 1 μM of nocodazole (for probe3, 4, all) for 4–5 hours. After trypsinization, cells were hypotonically swollen in 75mM KCl for 20 minutes, fixed with ice-cold Carnoy's solution (1:3 acetic acid: methanol), then spread onto slides. After RNase treatment and denaturation of chromosomal DNA, hybridization was carried out by dropping probe DNA solution onto slides and incubating at 37 ℃ for overnight. After washing, chromosomal DNA was incubated with avidin-FITC (Vector Laboratories) for 1 hour. After final wash, slides were mounted with Vectashield Plus DAPI (Vector Laboratories). Images were acquired using a fluorescence microscope (LSM710, Zeiss).

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 [44]. 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 needle program in EMBOSS suite [49]. 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.

References

[1] Miroslav Plohl, Nevenka Meštrović, and Brankica Mravinac. Centromere identity from the DNA point of view. *Chromosoma*, 123(4):313–325, 2014.

[2] Kara L McKinley and Iain M Cheeseman. The molecular basis for centromere identity and function. *Nat Rev Mol Cell Biol*, 17(1):16–29, 2016.

[3] Rebecca Oakey and Chris Tyler-Smith. Y chromosome DNA haplotyping suggests that most European and Asian men are descended from one of two males. *Genomics*, 7(3):325–330, 1990.

[4] Huntington F Willard and John S Waye. Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends in genetics : TIG*, 3(7), 1987.

[5] Ivan Alexandrov, Alexei Kazakov, Irina Tumeneva, Valery Shepelev, and Yuri Yurov. Alpha-satellite DNA of primates: old and new families. *Chromosoma*, 110(4):253–266, 2001.

[6] Huntington F Willard. Evolution of alpha satellite. *Current opinion in genetics & development*, 1:509–514, 1991.

[7] M Katharine Rudd and Huntington F Willard. Analysis of the centromeric regions of the human genome assembly. *Trends in genetics : TIG*, 20(11):529–33, 2004.

[8] Xinwei She, Julie E Horvath, Zhaoshi Jiang, et al. The structure and evolution of centromeric transition regions within the human genome. *Nature*, 430(7002):857–64, 2004.

[9] Mary G Schueler, Anne W Higgins, M Katharine Rudd, Karen Gustashaw, and Huntington F Willard. Genomic and Genetic Definition of a Functional Human Centromere. *Science*, 294(October):109–115, 2001.

[10] Mark T. Ross, Darren V. Grafham, Alison J. Coffey, et al. The DNA sequence of the human X chromosome. *Nature*, 434(7031):325–337, 2005.

[11] Chad Nusbaum, Tarjei S Mikkelsen, Michael C Zody, et al. DNA sequence and analysis of human chromosome 8. *Nature*, 439(7074):331–5, 2006.

[12] M Katharine Rudd, Gregory a Wray, and Huntington F Willard. The evolutionary dynamics of alpha-satellite. *Genome Research*, 16:88–96, 2006.

[13] Michael C. Schatz, Arthur L. Delcher, and Steven L. Salzberg. Assembly of large genomes using second-generation sequencing. *Genome Research*, 20(9):1165–1173, 2010.

[14] Karen E Hayden, Erin D Strome, Stephanie L Merrett, et al. Sequences associated with centromere competency in the human genome. *Molecular and cellular biology*, 33(4):763–72, 2013.

[15] Can Alkan, Maria Francesca Cardone, Claudia Rita Catacchio, et al. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. *Genome Research*, 21(1):137–145, 2011.

[16] Daniël P Melters, Keith R Bradnam, Hugh a Young, et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, 14(1):R10, 2013.

[17] Marija Rosandić, Vladimir Paar, Matko Gluncić, Ivan Basar, and Nenad Pavin. Key-string algorithm–novel approach to computational analysis of repetitive sequences in human centromeric DNA. *Croatian medical journal*, 44(4):386–406, 2003.

[18] Can Alkan, Mario Ventura, Nicoletta Archidiacono, et al. Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. *PLoS computational biology*, 3(9):1807–18, 2007.

[19] Steven Henikoff, Kami Ahmad, and Harmit S Malik. The Centromere Paradox : Stable Inheritance with Rapidly Evolving DNA. *Science*, 293(August):1098–1103, 2001.

[20] Karen H Miga, Yulia Newton, Miten Jain, et al. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Research*, 24(4):697–707, 2014.

[21] GenomeRef. Announcing GRCh38, http://genomeref.blogspot.jp/2013/12/announcing-grch38.html, last accessed 2017/01/27, 2013.

[22] John Eid, Adrian Fehr, Jeremy Gray, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–8, 2009.

[23] Miten Jain, Hugh E. Olsen, Benedict Paten, et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1):239, 2016.

[24] Megan Aldrup-MacDonald and Beth Sullivan. The Past, Present, and Future of Human Centromere Genomics. *Genes*, 5(1):33–50, 2014.

[25] Karen H. Miga. Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Research*, pages 421–426, 2015.

[26] Gene Myers. Efficient Local Alignment Discovery amongst Noisy Long Reads. In *WABI*, pages 52–67. 2014.

[27] Michael G Ross, Carsten Russ, Maura Costello, et al. Characterizing and measuring bias in sequence data. *Genome biology*, 14(5):R51, 2013.

[28] Robert VanBuren, Doug Bryant, Patrick P. Edger, et al. Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature*, 527(7579):508–11, 2015.

[29] Shubha Vij, Heiner Kuhl, Inna S Kuznetsova, et al. Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. *PLoS genetics*, 12(4):e1005954, 2016.

[30] Yinping Jiao, Paul Peluso, Jinghua Shi, et al. The complex sequence landscape of maize revealed by single molecule technologies. *bioRxiv*, pages 1–19, 2016.

[31] Thomas K. Wolfgruber, Megan M. Nakashima, Kevin L. Schneider, et al. High Quality Maize Centromere 10 Sequence Reveals Evidence of Frequent Recombination Events. *Frontiers in Plant Science*, 7(308):1–14, 2016.

[32] Daniel E Khost, Danna G Eickbush, and Amanda M Larracuente. Single molecule long read sequencing resolves the detailed structure of complex satellite DNA loci in. *Bioarchiv*, 2016.

[33] Volkan Sevim, Ali Bashir, Chen-shan Chin, and Karen H Miga. Alpha-CENTAURI: assessing novel centromeric repeat sequence variation with long read sequencing. *Bioinformatics*, 32(13):1921–4, 2016.

[34] Ruth B. Phillips and Kent M. Reed. Localization of repetitive DNAs to zebrafish (Danio rerio) chromosomes by fluorescence in situ hybridization (FISH). *Chromosome Research*, 8(1):27–35, 2000.

[35] Inna S Kuznetsova, Natascha M Thevasagayam, Prakki S R Sridatta, et al. Primary analysis of repeat elements of the Asian seabass (Lates calcarifer) transcriptome and genome. *Frontiers in Genetics*, 5(July):1–14, 2014.

[36] Jennifer N. Cech and Catherine L. Peichel. Identification of the centromeric repeat in the threespine stickleback fish (Gasterosteus aculeatus). *Chromosome Research*, 23(4):767–779, 2015.

[37] Kerstin Howe, Matthew D Clark, Carlos F Torroja, et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, 496(7446):498–503, 2013.

[38] Joachim Wittbrodt, Akihiro Shima, and Manfred Schartl. Medaka - A Model Organism From the Far East. *Nature Reviews Genetics*, 3(1):53–64, 2002.

[39] Masahiro Kasahara, Kiyoshi Naruse, Shin Sasaki, et al. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447(7145):714–719, 2007.

[40] Davin H E Setiamarga, Masaki Miya, Yusuke Yamanoue, et al. Divergence time of the two regional medaka populations in Japan as a new time scale for comparative genomics of vertebrates. *Biology letters*, 5(6):812–6, 2009.

[41] Albert Levan, Karl Fredga, and Avery A. Sandberg. Nomenclature for centromeric position on chromosomes. *Hereditas*, 52:201–220, 1964.

[42] C Camacho, G Coulouris, V Avagyan, et al. BLAST plus: architecture and applications. *BMC Bioinformatics*, 10(421):1, 2009.

[43] Mark J Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, 13(1):238, 2012.

[44] Mohammadreza Ghodsi, Bo Liu, and Mihai Pop. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC bioinformatics*, 12:271, 2011.

[45] Fabian Sievers, Andreas Wilm, David Dineen, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7(1):539, 2011.

[46] Chen-Shan Chin, Paul Peluso, Fritz J. Sedlazeck, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13(12):1050–1054, 2016.

[47] Chen-Shan Chin, David H Alexander, Patrick Marks, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10(6):563–569, 2013.

[48] Bruce J. Walker, Thomas Abeel, Terrance Shea, et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, 9(11), 2014.

[49] Peter Rice, Ian Longden, and Alan Bleasby. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(1):276–277, 2000.

Supplements

Table S1: PacBio sequencing statistics

	Hd-rR	HNI	HSOK
Number of cells	38 (P6-C4) + 35 (P5-C3) + 78 (P4-C2)	24 (P5-C3) + 144 (P4-C2)	97 (P6-C4)
Number of filtered subreads	13,359,879	14,777,797	5,527,528
Total bases (bp)	87,095,247,396	52,830,178,508	60,649,832,062
Average read length (bp)	6,519	3,575	10,972

Table S2: 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

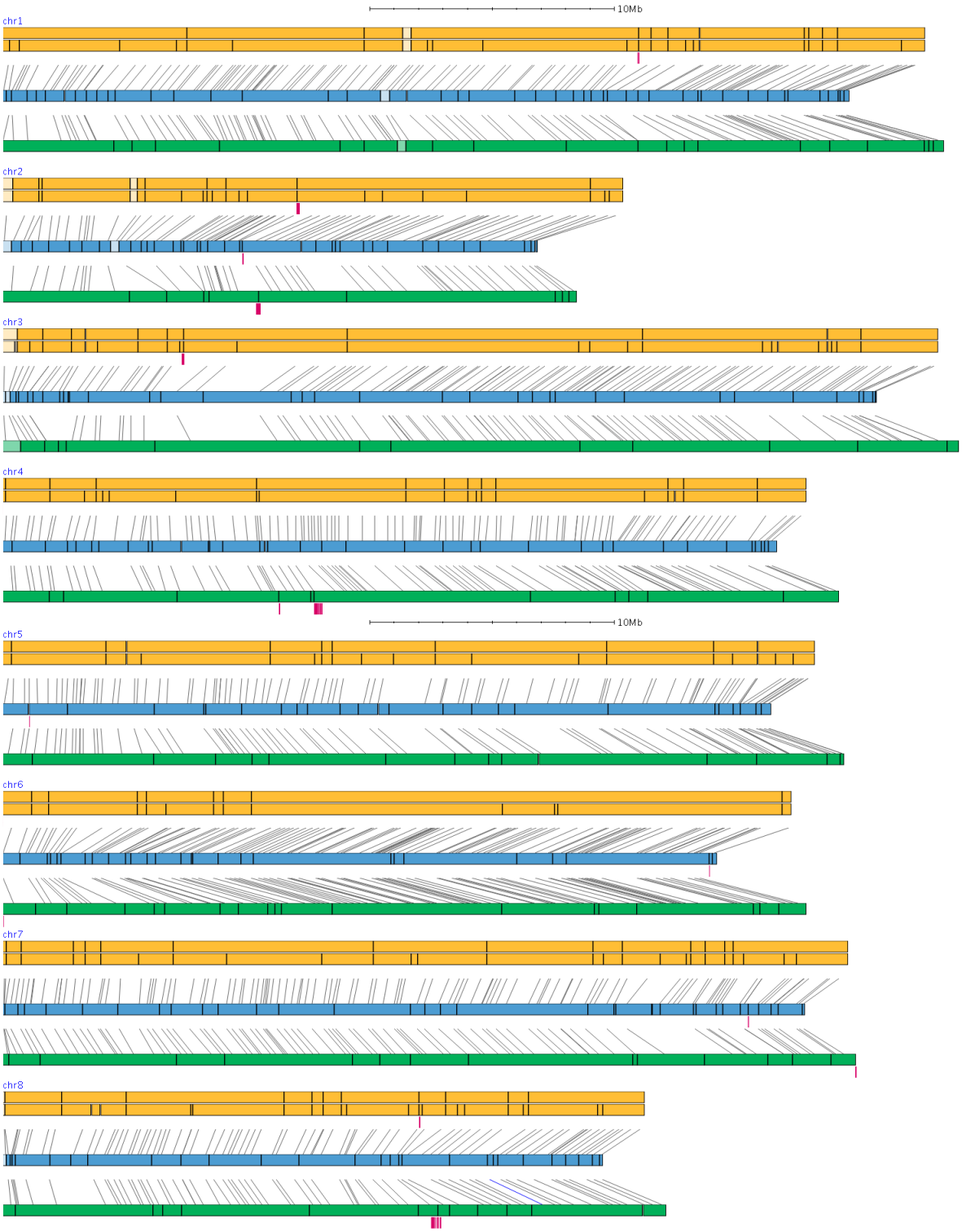


Figure S1: Centromeric repeat distribution

Figure S2: Centromere landscapes in the medaka genomes

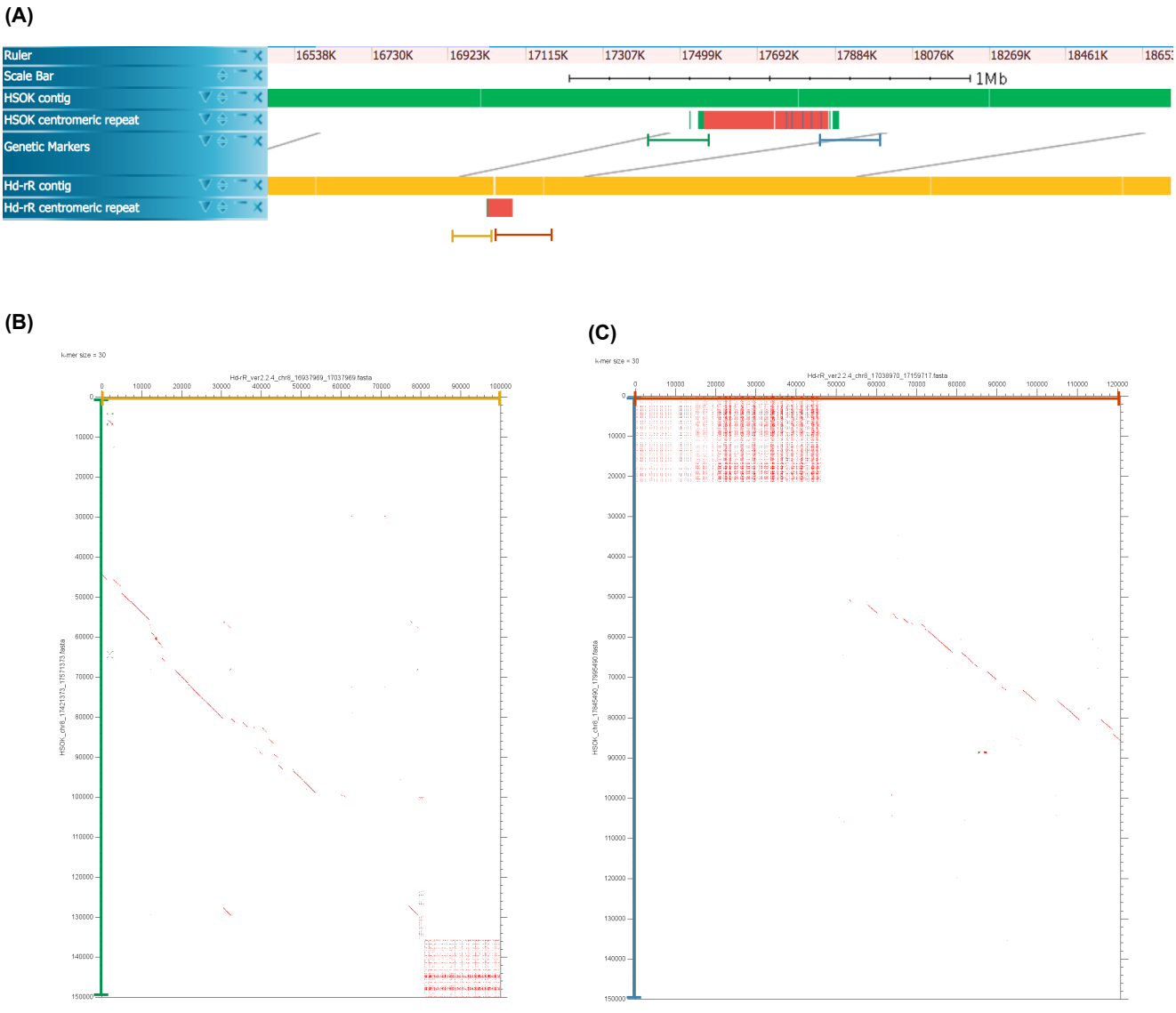
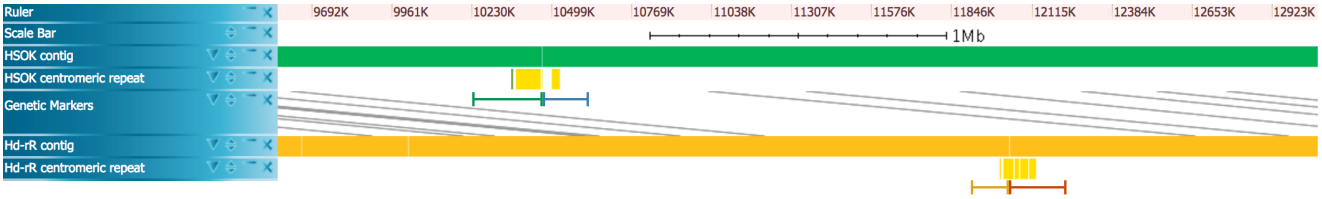
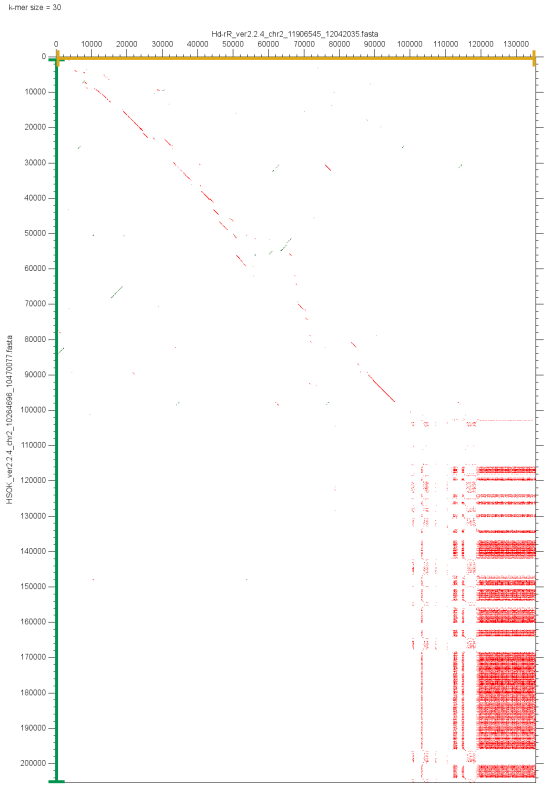


Figure S3: (A, B, C) Comparison of the centromeric transitional regions of chromosome 8 of Hd-rR and HSOK.

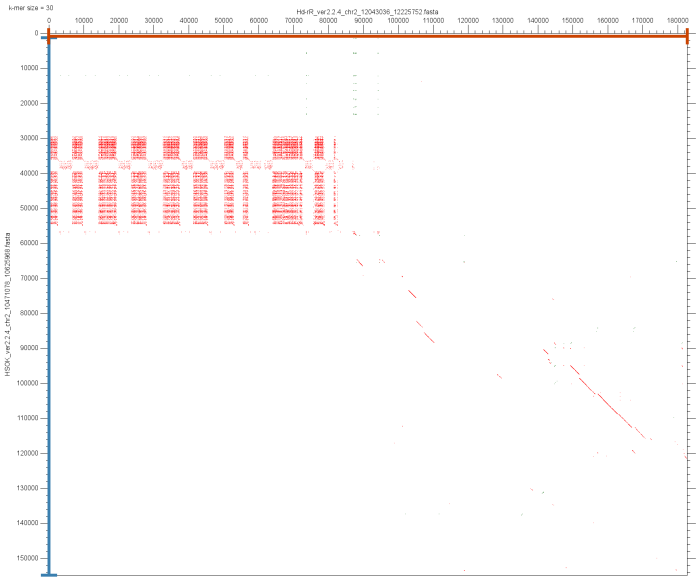
(D)



(E)



(F)



(D, E, F) Comparison of the centromeric transitional regions of chromosome 2 of Hd-rR and HSOK.