Department of Computational Biology and Medical Sciences
Graduate School of Frontier Sciences
The University of Tokyo

2016

Master's Thesis

Centromeric sequence organization in medaka

Shingo Tomioka

Supervised by
Professor Shinichi Morishita

Submitted January 27, 2017

Introduction

The centromere

The centromere is a genomic region where kinetochore forms in mitosis and meiosis and plays a critical role for proper chromosome segregation. The centromere is characterized by the presence of centromere-specific histone H3 variant CENH3 (also known as CENP-A), and underlying DNA is commonly comprised of repetitive sequences [1]. Despite its fundamental biological importance, the mechanisms of centromere specification and formation are still insufficiently understood [2].

DNA sequences at the centromere

Repeat-based centromere is the most common structure in eukaryote species [1]. One of the major components of the repeat-based centromeres is a tandemly-repeated satellite DNA. Despite the conserved biological function of the centromere, DNA sequences at the centromere evolves rapidly (known as "centromere paradox") and the satellite sequences are generally species-specific [3]. A well-known example is a 171-bp AT-rich alpha-satellite observed in human and many other primate species [4]. In the human genome, tandemly-repeated alpha-satellites comprise hundred kilobase- to megabase-sized arrays in each chromosome [5]. The satellite DNA can be highly diverged within a species, whereas the inner centromeric regions tend to have highly homogenous repeat arrays. In many primate species including human, the core centromeres consist of higher-order repeats (HORs) where several alpha-satellite monomers comprise a repeat unit which itself iterates tandemly with extremely high identity (>95%) [5]. Another major component of the repeat-based centromeres is retrotransposons, which has been widely observed in plant species [6,7]. Satellites and retrotransposons are not mutually exclusive, rather intermingled structure of them are commonly observed [1]. These regional centromeres are flanked by heterochromatic pericentromeres. Centromeres and pericentromeres are characterized by distinct chromatin structures which are regulated by different epigenetic marks [2,8].

While repeat-based regional centromeres are the most common structure in eukaryotes, some species have 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 species in insects and plants, the spindle microtubules attach all along a chromosome and the entire chromosome functions as a centromere, called "holocentromere" [1, 9].

Centromere specification and role of underlying DNA sequences

Understanding the machinery of centromere formation is still ongoing challenge [1, 2, 9]. The chromosomal position where the centromere forms is considered to be specified by the cooperation of centromere-specific histone variant CENP-A, histone modifications and underlying DNA sequences [9]. However, specific DNA sequences are not considered to be indispensable for the centromere formation, which is inferred by the observations of several "atypical" centromeres.

Neocentromeres are the centromeres that form on ectopic sites distantly from original centromeres along with a loss or an inactivation of the original ones and have been widely observed in many species [10, 11]. Neocentromeres can form on loci that do not have repetitive sequences or specific sequence motifs. Dicentric chromosomes have also been widely observed, which possess two regions with centromeric sequences; one of them directs the centromere formation whereas the other one remains inactivated [12–14]. It is also known that some species have a few chromosomes with repeat-free centromeres whereas their other chromosomes have repeat-based centromeres [15–17].

The wide prevalence of repetitive DNA among eukaryote species despite these seeming dispensability of it for centromere formation suggest that repetitive DNA may stabilize the centromere, rather than specify its formation. Some researchers have proposed that the repetitive DNA may serve as a wide range of "safe" region with little genes thus provide plasticity for sliding of kinetochore formation [1, 9]. It has been hypothesized that tandem repeats occur spontaneously at any genomic position and amplify by unequal crossover between sister chromatids, homologous chromosomes and non-homologous chromosomes [4, 18, 19]; therefore the naturally-expanded repetitive regions may be selected as a suitable "safe" region for the centromere formation. Moreover, the neocentromeres and repeat-free centromeres may be premature states of newly-formed centromeres that will gradually acquire repetitive DNA [9]. On the other hand, the presence of a widely-conserved centromere protein binding motif (CENP-B box) in centromeric satellites [3] and several recent functional studies [20, 21] imply more direct contribution of the underlying DNA sequences to the centromere formation and function.

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 [22–24], the presence of chromosome-specific alpha-satellite HORs [5] and super-chromosomal subfamilies of alpha-satellites [25, 26]. 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 [4].

The genome projects era

Although the basic characteristics of centromeric sequences were revealed by the 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 its completion in 2003, large portion of centromeric sequences were missed as huge gaps [27, 28]. Whereas divergent monomeric portion around pericentromeres were assembled in many chromosomes, the assemblies reached more homogeneous HOR regions in a limited number of chromosomes. Nevertheless, subsequent analyses on these few chromosomes revealed sequence landscapes with never-seen resolution [29–32]. The unsatisfactory sequence assembly at the repetitive centromeric regions were the case in other contemporarily-assembled species [33, 34].

Second-generation sequencing-based studies

Although second-generation sequencers (SGSs) represented by Illumina and 454 dramatically lowered sequencing cost and accomplished an increased number of genome assemblies [35], they achieved virtually no improvement in centromeric sequence assembly because of their short read length. Nevertheless, their high throughput sequencing combined with chromatin immunoprecipitation (ChIP-seq) facilitated identification of centromere-associated sequences and characterization of functional regions in the assembled centromeric sequences [36].

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 [37, 38], whereas others identified novel HOR patterns from assembled sequences [39] or from WGS data [40].

Melters *et al.* [38] identified candidate centromeric satellite sequences of 282 species (204 animals and 78 plants) 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 the 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 other genomic regions [3].

Another remarkable computational study is from Miga *et al.* [41], 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 other downstream analyses. The generated centromeric sequences have been positioned in the centromeric regions in the latest human reference genome (GRCh38) [42].

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 [43] and Oxford Nanopore [44] are expected as promising tools for centromere studies [45, 46].

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 [47]. 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 [48]. In recent years, a number of studies have reported dramatic improvement in genome assemblies using PacBio sequencing, some of which improved centromere assembly as well [49–51]. 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 [49]. Some studies targeting centromere-associated repeat-rich regions have also been conducted with PacBio sequencing [52, 53], and a computational tool for detecting HOR patterns from long reads was also developed [54].

Oxford Nanopore sequencing yields even longer read length (e.g. MinION sequencer routinely yields >150 kb read) with accuracy of ~92% [44]. 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 [55–57], however only limited amount of them have been included in the assembled genomes [50, 57, 58]. Medaka, as well as zebrafish, has traditionally played an important role as a model organism of fish species [59]. Cloning-based assembly of its genome was conducted about ten years ago [60], 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* [38], 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 [61].

Genomic abundance of centromeric repeats

This study started with a candidate centromeric satellite sequence of medaka which was identified in a previous computational study [38]. 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 [62]. 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 [41]. 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 [38].

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. S2). 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.* [63] (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 [64, 65].

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. 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 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 [65], 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. S2), 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 the FISH experiments were carried out by a collaborator (see Methods).

The candidate satellite identified by Melters *et al.* [38] was first used as a hybridization probe and signals were observed only from 5~7 chromosome pairs (Fig. S1). In order to map the centromeres of the other chromosomes, additional probes were designed (see Methods). The additional probes successfully hybridized to some chromosomes that the first probe failed to hybridize, with approximately the same positions as expected by the *in silico* centromeric repeat distribution (Fig. S1, Table 1), although two additional probes hybridized to less chromosomes than expected by 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

The repetitive nature of centromeric sequences inevitably accompanies the possibility of misassembly. Although long-range assembly correctness on the assembled genomes was verified with BAC-end and fosmid-end reads and Hi-C contact frequency data, exact sequence ordering within long centromeric repeat arrays can be inaccurate. 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 [66] with a strict 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 (Fig. S3). Although the assembly validation using somewhat erroneous PacBio reads 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 [4]. 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 [67] 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 [4].

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. S2). This discordant classification may be because the assemblies of each strain captured different subregion of the corresponding repeat arrays or due to

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)	

Total amount of the centromeric repeats identified in the chromosomes are shown. The total amount in the contigs that were anchored to the chromosomes and in the unanchored contigs are also shown (fraction of the centromeric repeats in the unanchored contigs are shown in the brackets). The centromeric positions were determined by the repeat distribution on each chromosome, employing the nomenclature by Levan *et al.* [63]. 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).

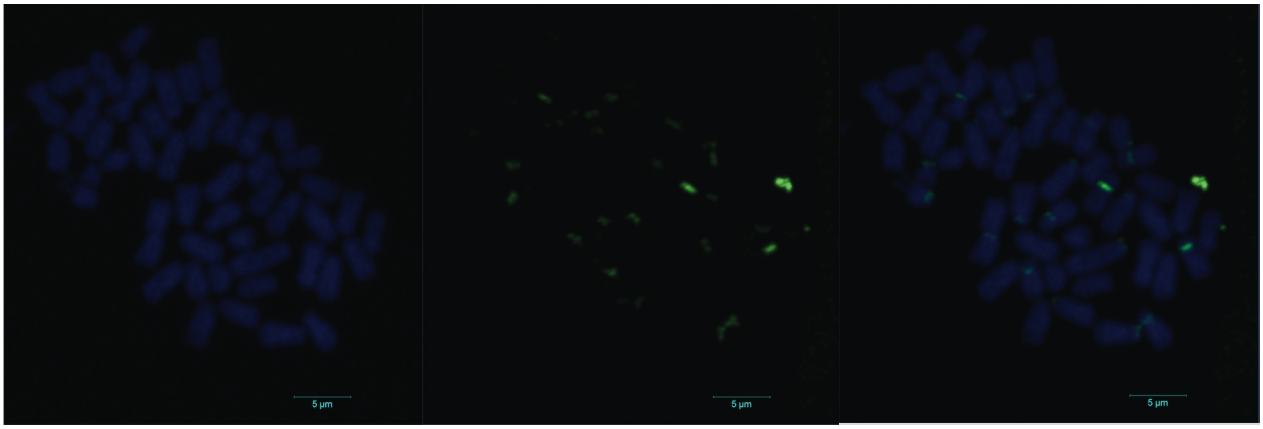

Figure 1: The candidate centromeric satellite sequence and three derivative sequences localized to the centromeres of ~13 pairs of chromosomes. (left) DNA is stained with DAPI. (center) probes are stained green. (right) two images are combined.

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. S2), 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 centromeric sequences were conserved among species within about 50 million years after separation [38].

Sequence organization at the centromeres

Sequence organization on the assembled centromeric sequences were analyzed. Self-alignment dot plots of the centromeric sequences on each chromosome are shown in Supplementary Figure S3.

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 were 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 smaller 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 nearly identical (Fig. 4). The assignment of these representative monomers to the different subfamilies was due to the definition of the distance used for the hierarchical clustering, which was calculated by alignment identity of two sequences and thus large indels cause substantial loss in the 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 [27]. 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 the 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 was 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 were comprised of SF 3 satellites. In order to examine if the sequences of these centromeric transition regions are conserved among the strains, dot plots were drawn (Fig. 6B, C). Whereas modest sequence conservation was observed in the surrounding unique regions, apparent conservation of the sequence structure, including linear ordering of the satellites, was not conserved. Similar unconservation of the sequence ordering in centromeric transition regions were observed in some other chromosomes (Fig. S4). These results imply that the centromeric repeat arrays have evolved rapidly compared to the other genomic region, in line with traditional observations [4].

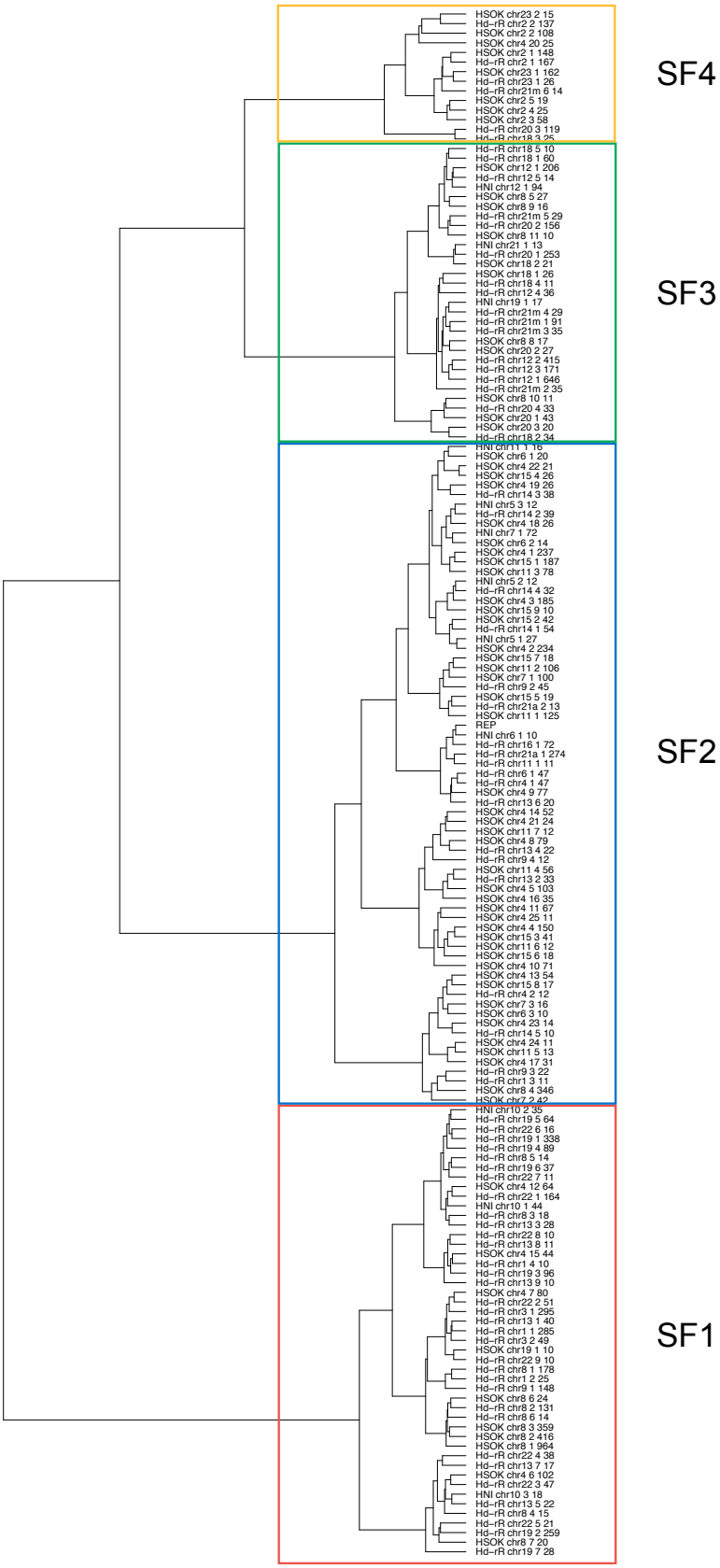

Figure 2: Hierarchical clustering of chromosome-representative monomers. Monomers are labeled as species, chromosome, cluster index, number of the cluster constituents. The clustering revealed four large subfamilies of satellite monomers.

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.

The self-alignment dot plots revealed the presence of higher-order repeats (HORs) in many chromosomes (Fig. S3). The HORs observed here seem somewhat diverged, unlike highly-identical HORs observed in the core centromeric regions of human and some other primates [5]. This is, however, reasonable because a stretch of highly-identical HOR array longer than the PacBio read length (~50 kb at maximum) would fail to be assembled. Therefore, the absence of highly-homogenized HORs in the assembled sequences does not exclude the possibility of the presence of them in inner centromeric regions in medaka. Moreover, the fact that all the contigs containing the centromeric repeats were truncated in the middle of the centromeric arrays implies the presence of highly-homogenized sequences.

Discussion

Due to the long-standing difficulty in assembling repetitive DNA, in-depth analyses of centromeric sequences have been limited [1]. Based on the high-quality genomes of the three medaka inbred strains which were recently assembled with PacBio long reads, this study has conducted a comprehensive analysis of the centromeric sequences in medaka. This is, to the author's knowledge, the first in-depth study of centromeric sequence organization in fish species.

This study has revealed the presence of four subfamilies of the centromeric satellites. The satellites of each subfamily are possessed by generally different subset of chromosomes. The presence of centromeric satellite subfamilies have been also observed in human, in which, similarly to the medaka satellite subfamilies, each subfamily belongs to generally different subsets of chromosomes [26]. Although 22 out of the 24 chromosomes were classified to one or two subfamilies, the classification could be imperfect. This is because many chromosomes have relatively short repeat arrays in the assemblies which locate at the ends of much larger original arrays and the peripheral regions tend to accumulate mutations [18, 29].

It was also observed that the subfamilies had different preference for the centromeric positions in chromosomes. For example, the medaka SF 2 showed clear preference for acrocentric chromosomes. This is analogous to the human acrocentric chromosomes (chr. 13, 14, 21 and 22) which share highly identical alpha-satellites [4]. The human acrocentric chromosomes possess nucleolus organizers region (NOR) at their short arms and are present in close proximity in the nucleolus, where frequent sequence exchange are believed to occur [4]. Although the presence of NORs have not been investigated in this study, the same mechanism may work in medaka and produce the high centromeric sequence similarity in the acrocentric chromosomes. It is uncertain if any biological mechanism works for the concerted centromeric positions in the other subfamilies; or they might just be the "remnants" that are left after collecting the acrocentric chromosomes.

This study has also revealed the conservation of centromeric satellites among the three strains which are estimated to have separated 18 (Hd-rR and HNI) and 25 (HSOK and the previous two) million years ago (MYA) [61]. This conservation is consistent with a previous comparative study of 282 animal and plant species that observed the centromeric satellites are conserved among species within 50 million years after separation [38], whereas unconserved centromeric repeat as a result of rapid diversification has also been observed in some closely-related species [69]. It has been known that the teleost lineage underwent a whole-genome duplication (WGD) 336–404 MYA and since that medaka has largely retained the genomic structure without major chromosome rearrangements [60]. The inter-chromosomal centromeric sequence comparison in this study have shown that the chromosome pairs that derived from the WGD do not necessarily possess similar centromeric satellite sequences. Combined with the Melters *et al.*'s observation, it can be speculated that medaka centromeric sequences have massively diverged since the WGD. Furthermore, the sequence organization including linear ordering of satellites were not conserved among the strains, suggesting that the centromeric sequences have evolved independently after the strain separations.

The medaka genome assemblies using PacBio long reads captured centromeric arrays in many chromosomes, including some long arrays over 100 kb. However the amount of centromeric repeats identified in the contigs that were anchored to the chromosomes was substantially below the estimated genomic abundance of the centromeric repeats in all the strains. Indeed large amount of the repeats were found

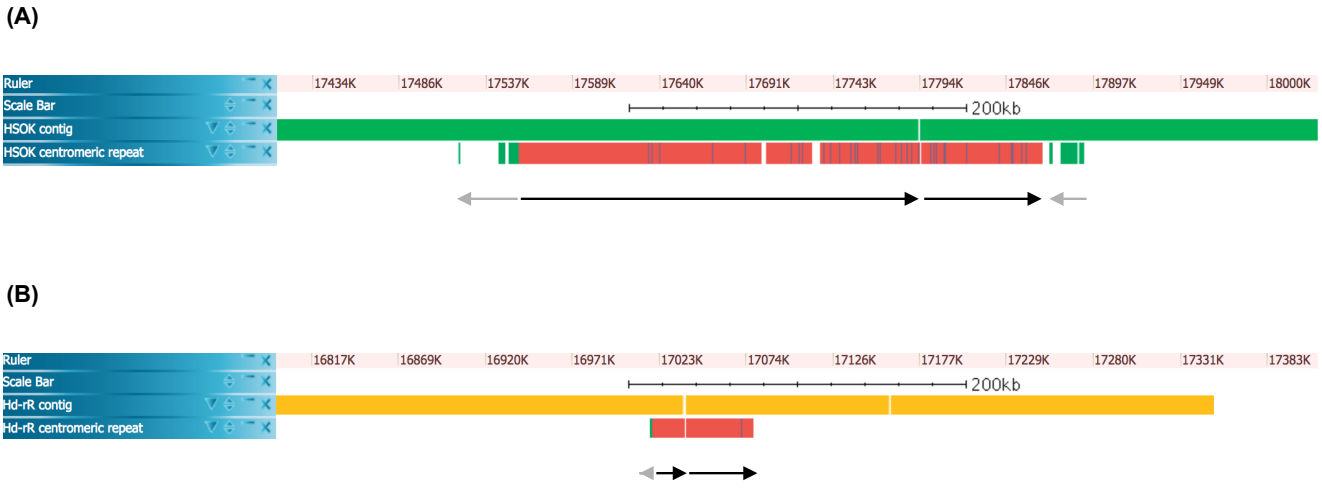

Figure 3: Sequence organization of chromosome 8 centromeric regions. (A) HSOK chromosome 8 had 250-kb and 95-kb repeat arrays flanking an assembly gap. SF 1 satellites (red) comprise large inner portion of the arrays, interspersed by SF 2 satellites (blue). These sequences were flanked by shorter SF 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) [68]. 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.

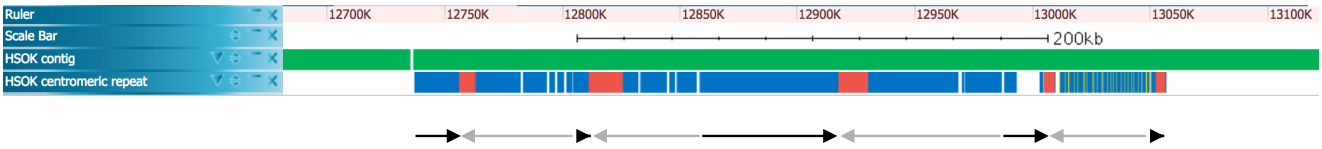

Figure 5: Sequence organization of HSOK chromosome 4 centromeric region. The ~300-kb nearly continuous array was truncated by the contig end at the left end. The array was comprised mainly of SF 2 satellites (blue) and these are interspersed by shorter SF 1 satellite arrays (red). Also small amount of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

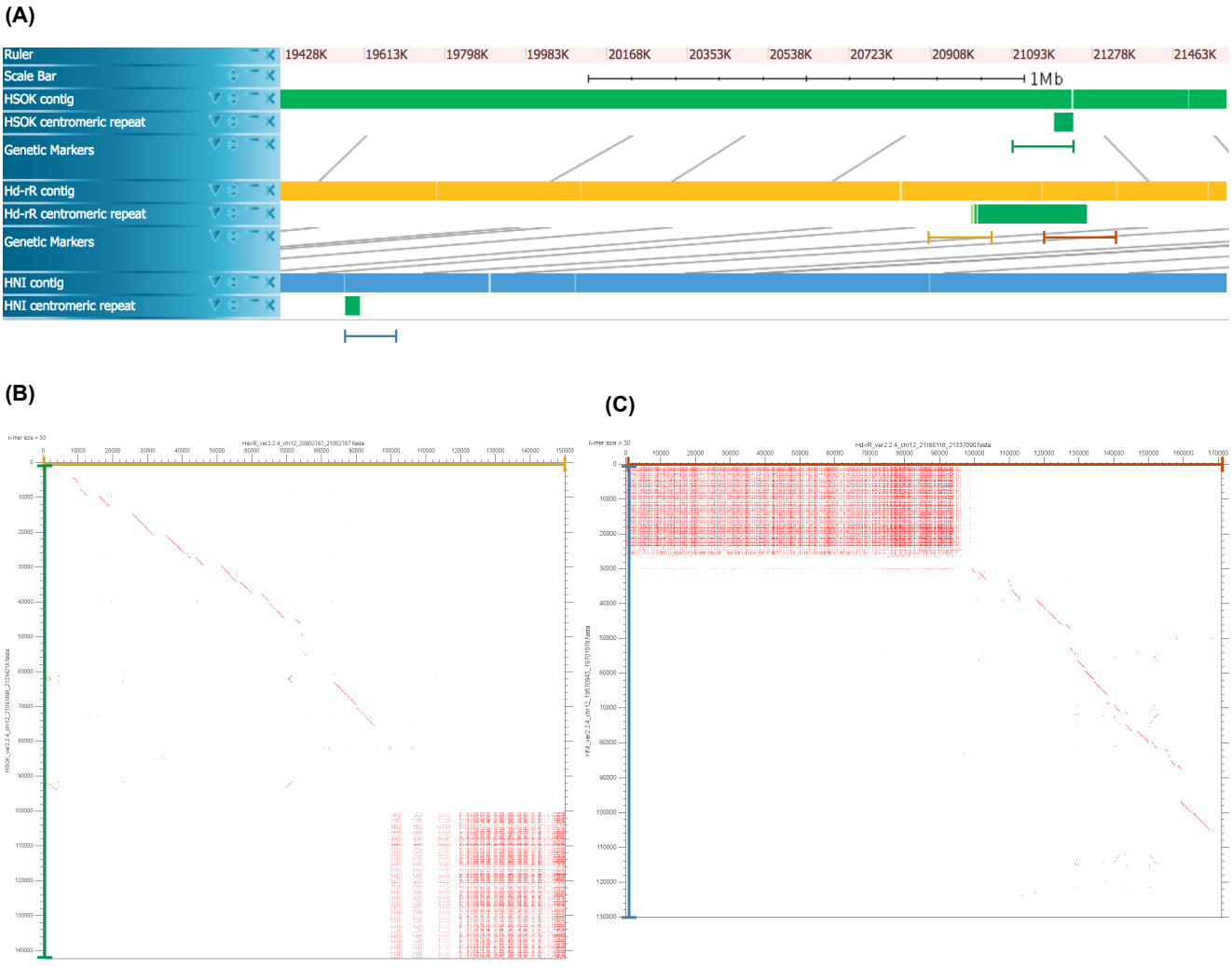

Figure 6: Comparison of the centromeric transition regions of chromosome 12. (A) The Hd-rR assembly reached the centromeric region from the both sides (separated by a contig gap); HNI reached from the p-arm side; HSOK reached from the q-arm side. The grey lines indicate the positions of corresponding genetic markers. (B) Sequences of the q-arm transition regions of Hd-rR and HSOK was compared. (C) Sequences of the p-arm transition 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.

in unanchored contigs. BAC/fosmid-end reads and Hi-C contact frequency data were used for the Hd-rR genome assembly, which successfully anchored a number of contigs containing centromeric repeats and resulted in higher portion of centromeric repeats to be anchored despite shorter read length than HSOK. This result emphasizes the complementary power of other long-range information methods in addition to the long read sequencing. The sum of the centromeric repeats in anchored and unanchored contigs did not reach the estimated genomic abundance, suggesting uncertain amount of the repeats were lost in the assembly process. Highly-homogenized HORs which might exist in the inner centromeric regions may have caused the under-represtation of centromeric repeats in the assembled contigs.

The centromeric sequence organization of three medaka strains revealed in this study provides insights into centromeric sequence evolution. Also, the characterized centromeric sequences provides a important basis for understanding the possible contribution of DNA sequences to the centromere specification and function.

Methods

Genome assembly

Assembly of 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 described.

The genomes were sequenced with PacBio RSII sequencer and were assembled into contigs with FALCON assembler [70]. The contigs were then polished with PacBio reads using Quiver [71] and with Illumina reads using Pilon [72]. 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. Also a number of unanchored contigs were positioned into the chromosomes using Hi-C contact frequency data.

Validation of centromeric sequence assembly

PacBio raw subreads were mapped to the assembled genomes by BLASR (version 5.2.6fa6cc2) [66] with parameters of "--bestn 10 --minReadLength 10000 --minSubreadLength 10000 --minAlnLength 5000 --minPctSimilarity 80". Then the mapped subreads that have >85% sequence identity in the 1-kb sequences at the both alignment ends were selected and visualized on the genomic browser. The centromeric repeat regions were manually inspected and confirmed that they were covered by enough number of overlapping subreads (at least 5 subreads and typically much more reads at every position) without breaks (Fig. S3).

Estimation of 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) [62] with 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).

Identification of centromeric repeats in the genomes and classification of 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.* [63]. 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.

Designing FISH probe sequences

The candidate centromeric satelilte identified by Melters *et al.* [38] was first used as a FISH probe but hybridized to only 5~7 pairs of chromosomes. Subsequently, additional probe sequences were designed as follows. For each chromosome, satellite monomers were collected from satellite arrays and the collected monomers were aligned back to the original array with BLASTN (in BLAST+ suite, version 2.3.0) [73]. 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. S1).

FISH experiments

The FISH experiments were 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).

Selection of chromosome-representative monomers and hierarchical clustering

Centromeric repeat arrays in each chromosome of the three strains were decomposed into satellite monomers by RepeatMasker. The monomer sequences within each chromosome were then clustered into groups of >85% sequence similarity by DNACLUST [67] (parameters: "-d -l -s 0.85 --no-k-mer-filter"). 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.* [38] was performed by needle program in EMBOSS suite (version 6.5.7) [74]. 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 (version 3.2.4) with "ward.D2" method.

Acknowledgements

The author would like to thank Professor Shinichi Morishita for guiding my research work, Yuta Suzuki, Yuichi Motai, Yoshihiko Suzuki, Kazuki Ichikawa, Jun Yoshimura and other laboratory members, Hiroyuki Takeda, Ryohei Nakamura, Masahiko Kumagai and Yusuke Inoue (in the Takeda laboratory at the Department of Biological Sciences, Graduate School of Science, The University of Tokyo) for valuable discussion, advice and supports.

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

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

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

[6] Guixiang Wang, Xueyong Zhang, and Weiwei Jin. An overview of plant centromeres. *Journal of Genetics and Genomics*, 36(9):529–537, 2009.

[7] Inna Lermontova, Michael Sandmann, Martin Mascher, Anne Catherine Schmit, and Marie Edith Chabouté. Centromeric chromatin and its dynamics in plants. *Plant Journal*, 83(1):4–17, 2015.

[8] Jolien S Verdaasdonk and Kerry Bloom. Centromeres: unique chromatin structures that drive chromosome segregation. *Nature Reviews Molecular Cell Biology*, 12(5):320–332, 2011.

[9] Tatsuo Fukagawa and William C. Earnshaw. The Centromere: Chromatin Foundation for the Kinetochore Machinery. *Developmental Cell*, 30(5):496–508, 2014.

[10] Owen J. Marshall, Anderly C. Chueh, Lee H. Wong, and K. H Andy Choo. Neocentromeres: New Insights into Centromere Structure, Disease Development, and Karyotype Evolution. *American Journal of Human Genetics*, 82(2):261–282, 2008.

[11] Kristin C. Scott and Beth a. Sullivan. Neocentromeres: A place for everything and everything in its place. *Trends in Genetics*, 30(2):66–74, 2014.

[12] William C. Earnshaw and Barbara R. Migeon. Three related centromere proteins are absent from the inactive centromere of a stable isodicentric chromosome. *Chromosoma*, 92(4):290–296, 1985.

[13] Noemi C. Steiner and Louise Clarke. A novel epigenetic effect can alter centromere function in fission yeast. *Cell*, 79(5):865–874, 1994.

[14] Fangpu Han, Jonathan C Lamb, and James A Birchler. High frequency of centromere inactivation resulting in stable dicentric chromosomes of maize. *Proceedings of the National Academy of Sciences of the United States of America*, 103(9):3238–43, 2006.

[15] Francesca M. Piras, Solomon G. Nergadze, Elisa Magnani, et al. Uncoupling of Satellite DNA and Centromeric Function in the Genus Equus. *PLoS Genetics*, 6(2):e1000845, 2010.

[16] Wei Hao Shang, Tetsuya Hori, Atsushi Toyoda, et al. Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. *Genome Research*, 20(9):1219–1228, 2010.

[17] Devin P Locke, LaDeana W Hillier, Wesley C Warren, et al. Comparative and demographic analysis of orang-utan genomes. *Nature*, 469(7331):529–33, 2011.

[18] G. Smith. Evolution of repeated DNA sequences by unequal crossover. *Science*, 191(4227):528–535, 1976.

[19] Brian Charlesworth, Paul Sniegowski, and Wolfgang Stephan. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, 371(6494):215–220, 1994.

[20] Jorja G Henikoff, Jitendra Thakur, Sivakanthan Kasinathan, and Steven Henikoff. A unique chromatin complex occupies young α-satellite arrays of human centromeres. 1(1):1–26, 2015.

[21] Megan E Aldrup-MacDonald, Molly E Kuo, Lori L. Sullivan, Kimberline Chew, and Beth A Sullivan. Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Research*, 26(10):1301–1311, 2016.

[22] 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.

[23] Melanie M. Mahtani and Huntington F. Willard. Pulsed-field gel analysis of alpha-satellite DNA at the human X chromosome centromere: High-frequency polymorphisms and array size estimate. *Genomics*, 7(4):607–613, 1990.

[24] G M Greig, S Parikh, J George, V E Powers, and H F Willard. Molecular cytogenetics of alpha satellite DNA from chromosome 12: fluorescence in situ hybridization and description of DNA and array length polymorphisms. *Cytogenetics and cell genetics*, 56(3-4):144–8, 1991.

[25] I. A. Alexandrov, S. P. Mitkevich, and Y. B. Yurov. The phylogeny of human chromosome specific alpha satellites. *Chromosoma*, 96(6):443–453, 1988.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

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

[33] Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[34] R. A. Hoskins, J. W. Carlson, C. Kennedy, et al. Sequence Finishing and Mapping of Drosophila melanogaster Heterochromatin. *Science*, 316(5831):1625–1628, 2007.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

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

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

[44] 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.

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

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

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

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

[49] 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.

[50] 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.

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

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

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

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

[61] 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.

[62] A.F.A. Smit, R. Hubley, and P. Green. RepeatMasker, http://repeatmasker.org.

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

[64] Hiroshi Uwa and Yoshio Ojima. Detailed and banding karyotype analyses of the medaka, Oryzias latipes in cultured cells. *Proceedings of the Japan Academy. Ser. B: Physical and Biological Sciences*, 57(2):39–43, 1981.

[65] 宇和紘. 核型と進化, メダカの生物学. 東京大学出版会, 162-182, 1990.

[66] 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.

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

[68] 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.

[69] Hye-Ran Lee, Wenli Zhang, Tim Langdon, et al. Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in Oryza species. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11793–8, 2005.

[70] 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.

[71] 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.

[72] 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.

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

[74] 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

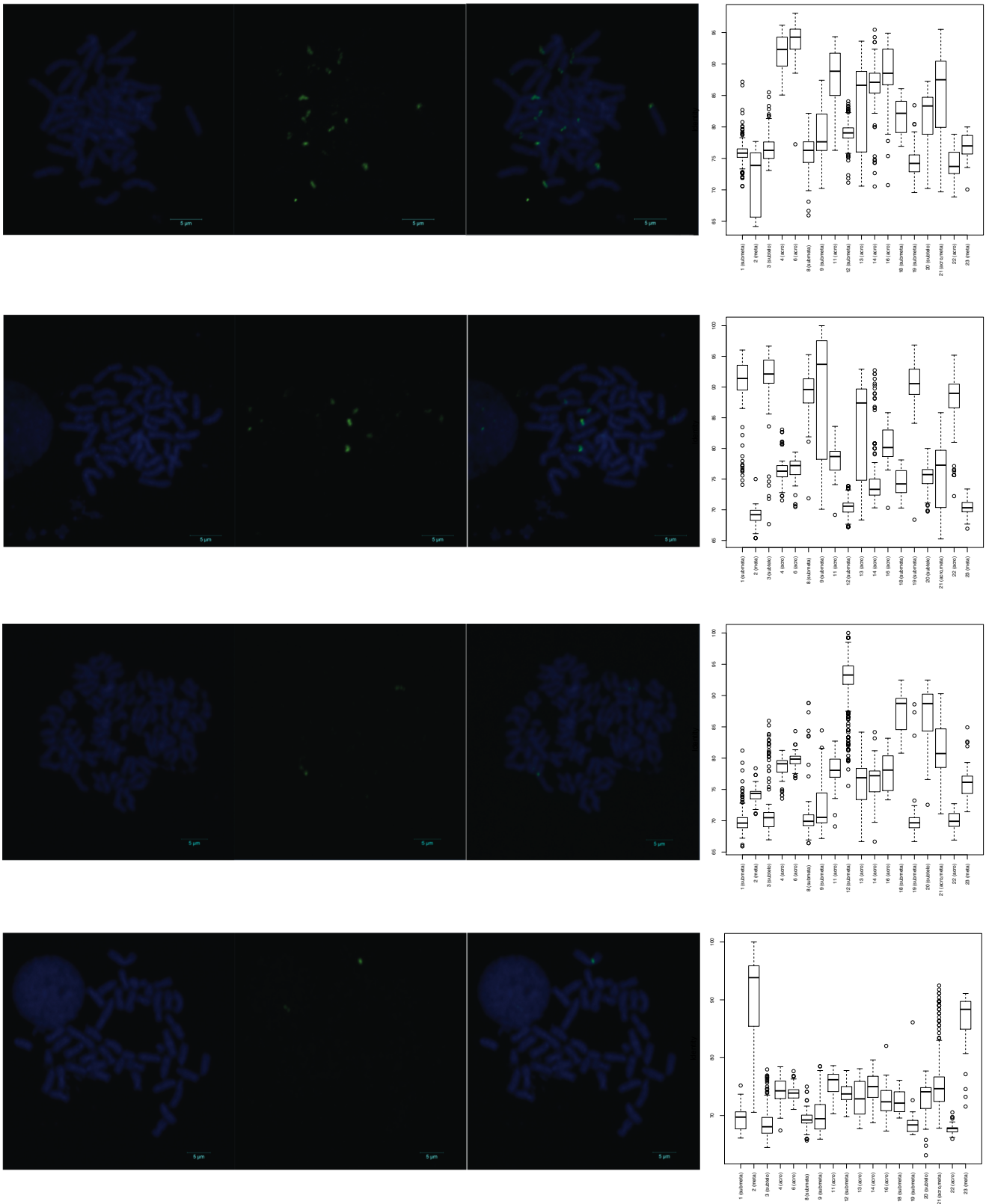

Figure S1: FISH with four probes. (Left) FISH; DNA is stained with DAPI (left); probes are stained green (center); combined (right). (Right) similarity of the probe sequence to satellite sequences in each chromosomes are plotted *in silico*. The top probe used the candidate centromeric satellite sequence identified by Melters *et al.* [38] and hybridized to 5~7 pairs of chromosomes. The other probes were additionally designed and hybridized to other ~6, 1 and 1 pairs of chromosomes, respectively.

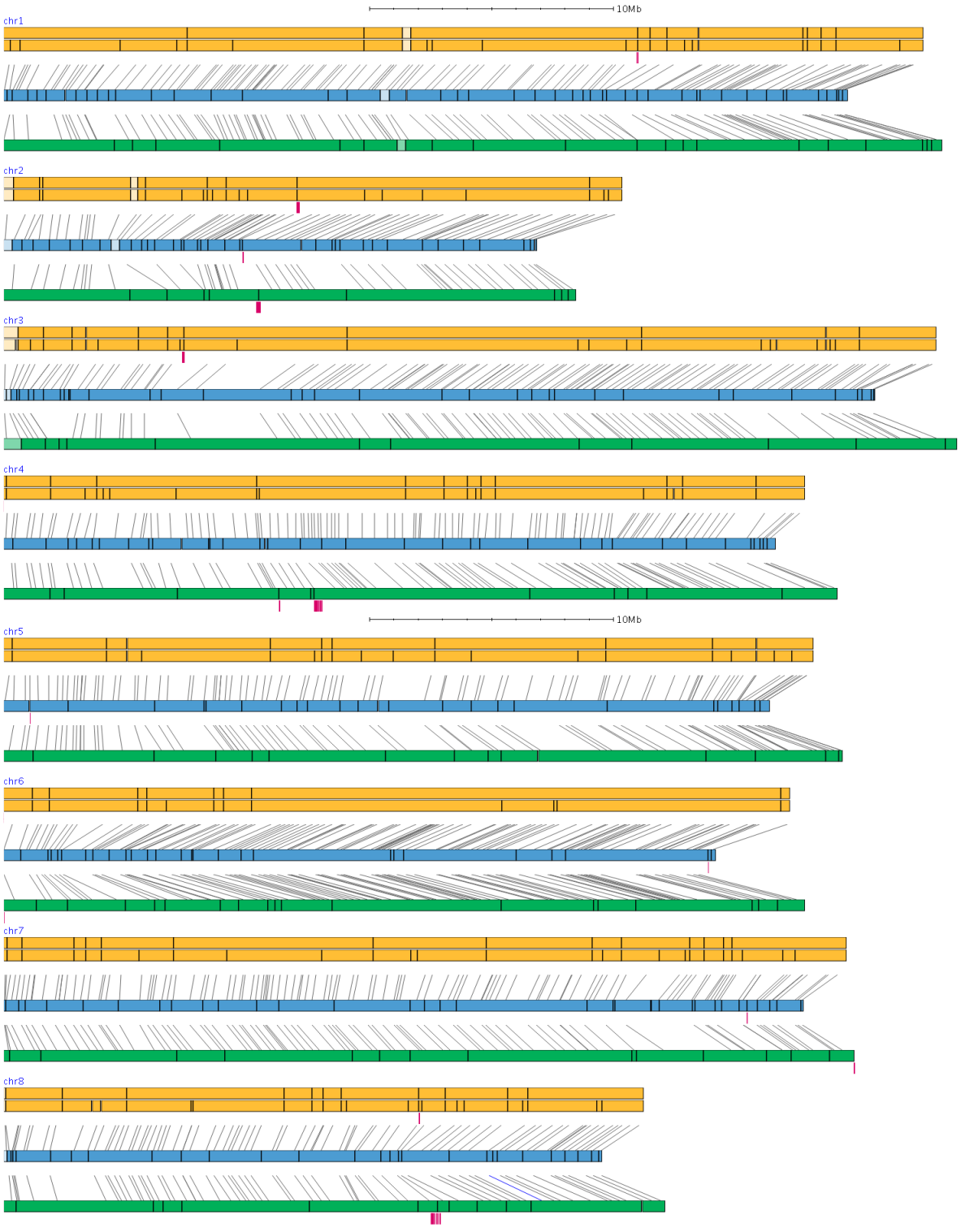

Figure S2: Centromeric repeat distribution. Yellow, Hd-rR scaffolds (upper) and contigs (lower); blue, HNI contigs; green, HSOK contigs; red, centromeric repeats; grey, corresponding genetic markers.

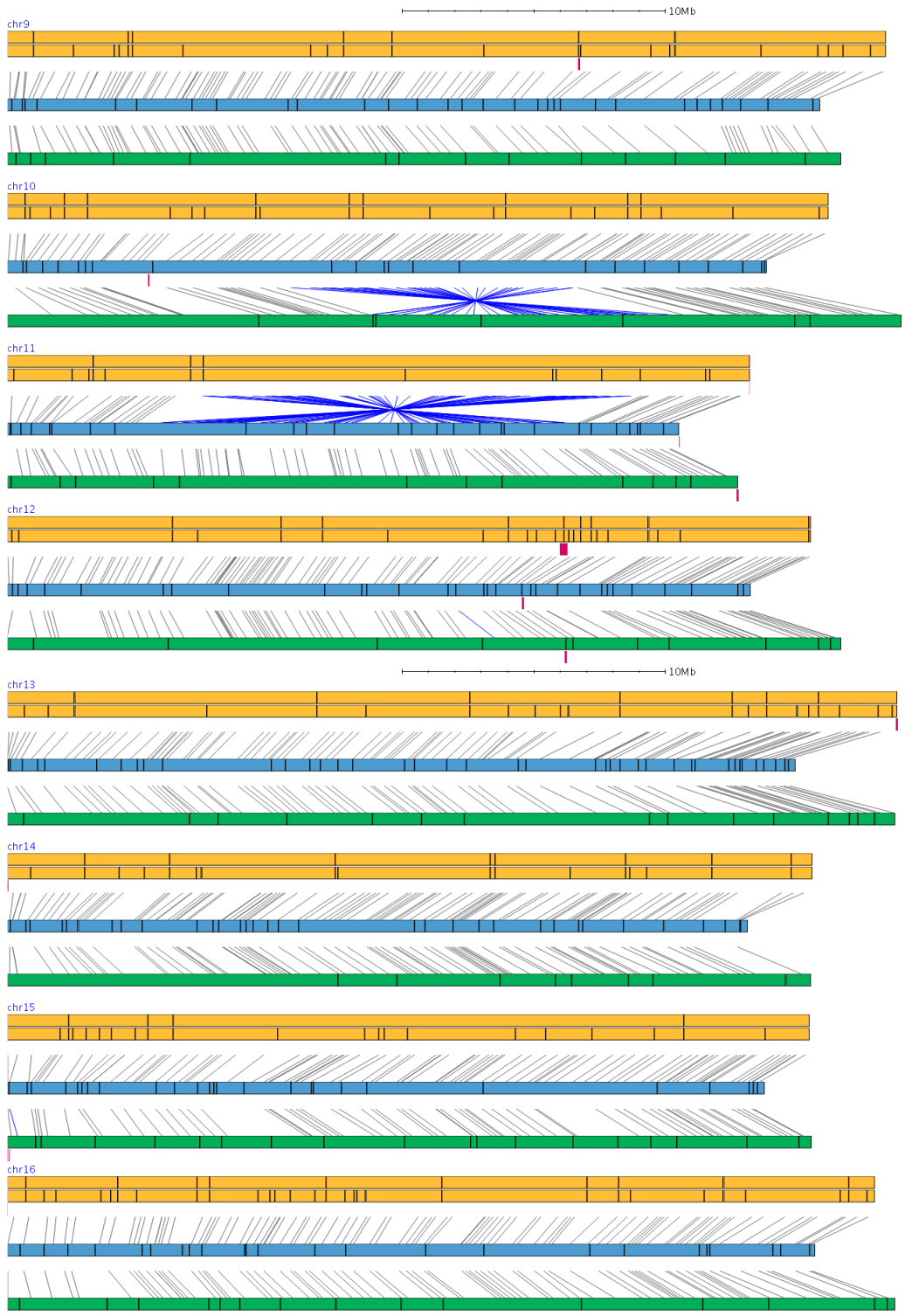

Figure S2: Centromeric repeat distribution. Yellow, Hd-rR scaffolds (upper) and contigs (lower); blue, HNI contigs; green, HSOK contigs; red, centromeric repeats; grey, corresponding genetic markers.

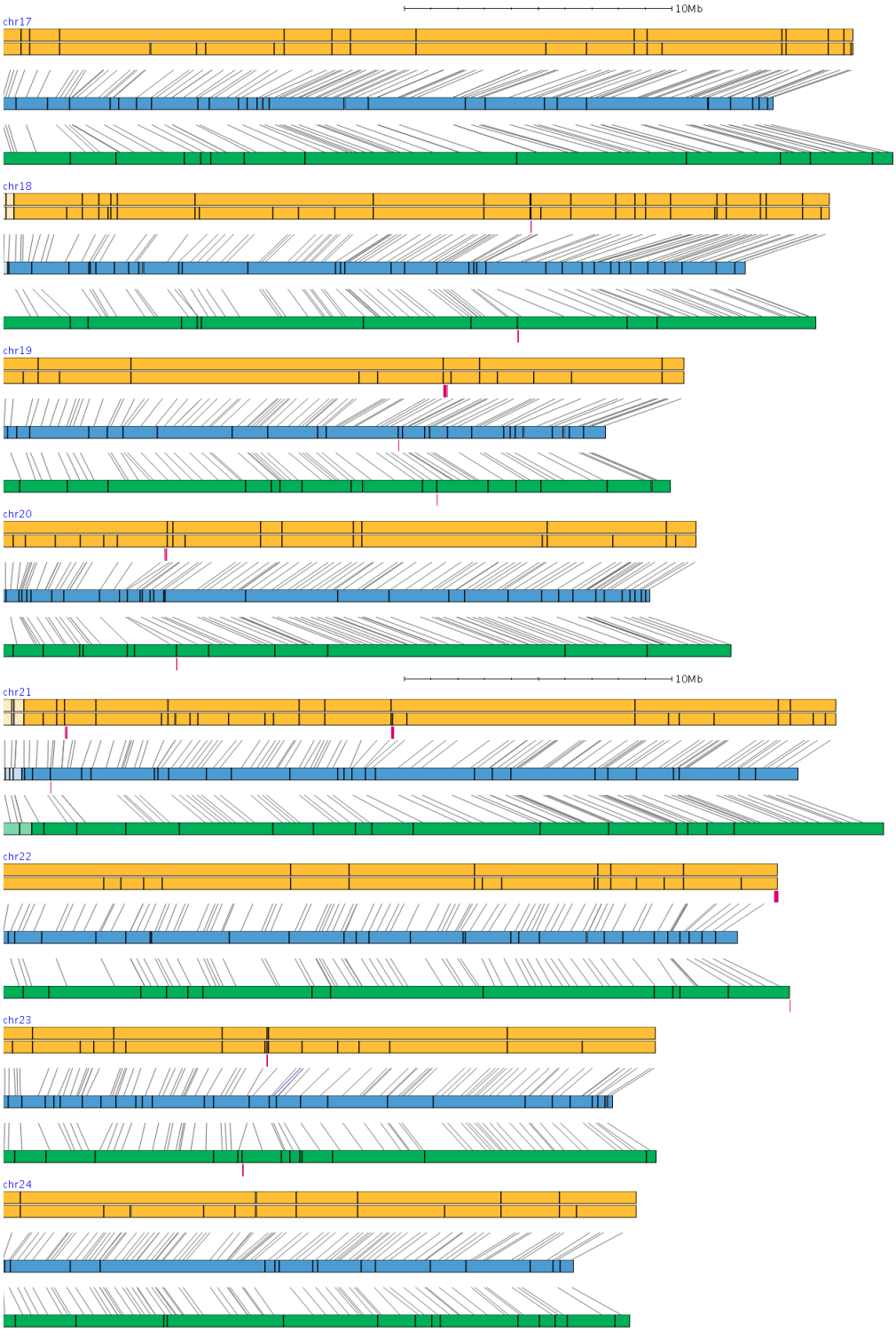

Figure S2: Centromeric repeat distribution. Yellow, Hd-rR scaffolds (upper) and contigs (lower); blue, HNI contigs; green, HSOK contigs; red, centromeric repeats; grey, corresponding genetic markers.

Hd-rR chr.1

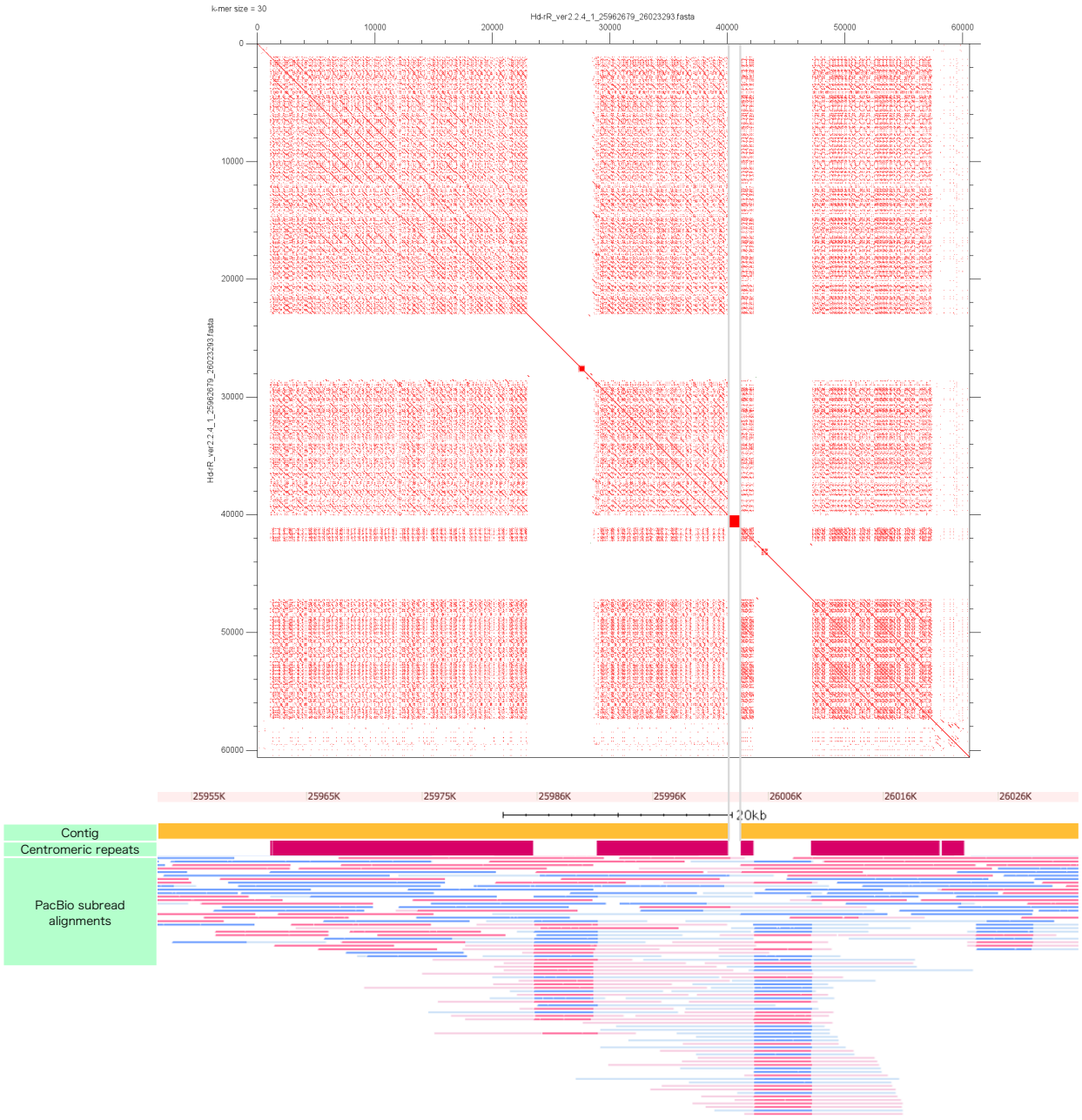

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

Hd-rR chr.2

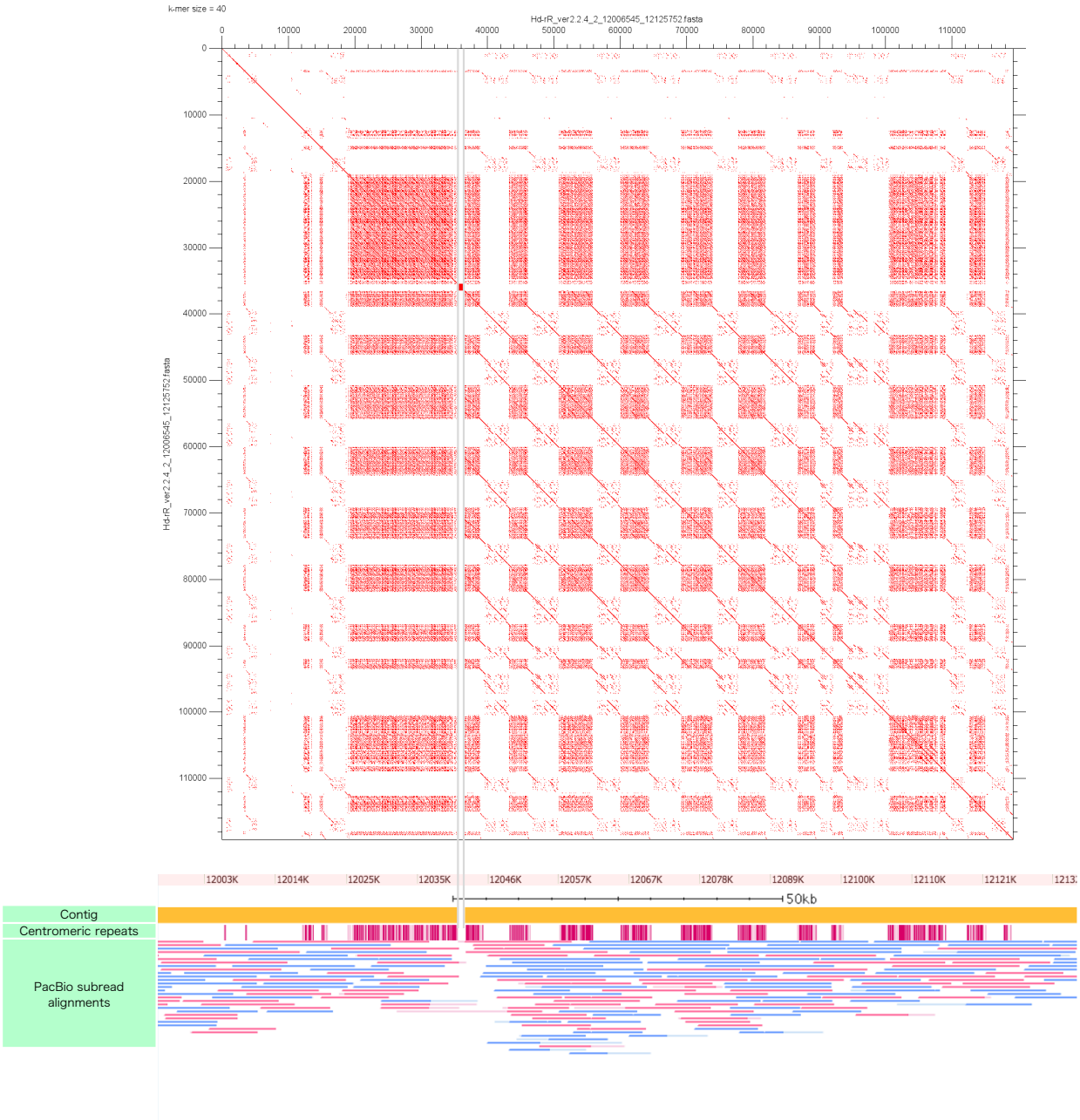

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

Hd-rR chr.3

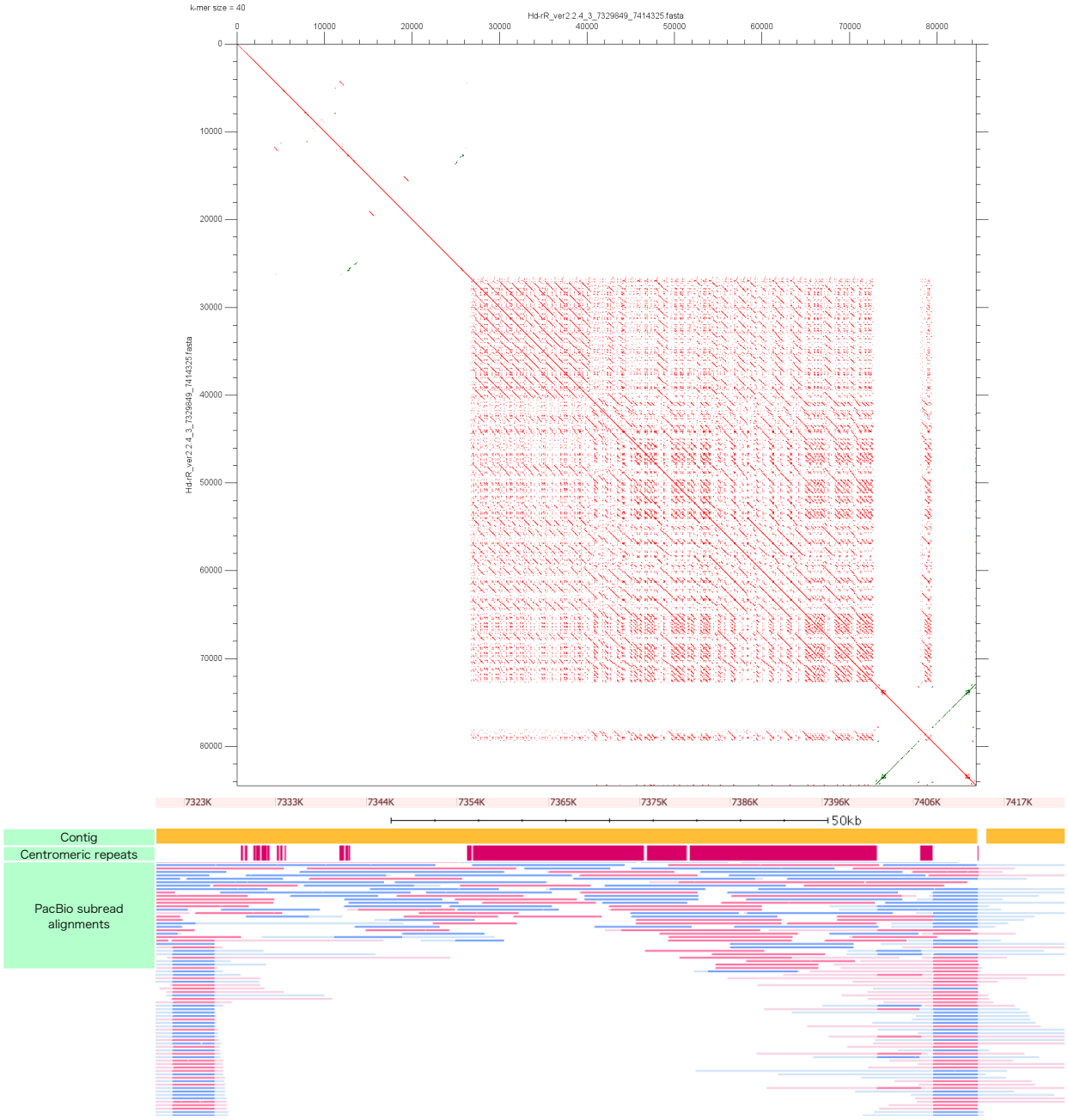

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

Hd-rR chr.8

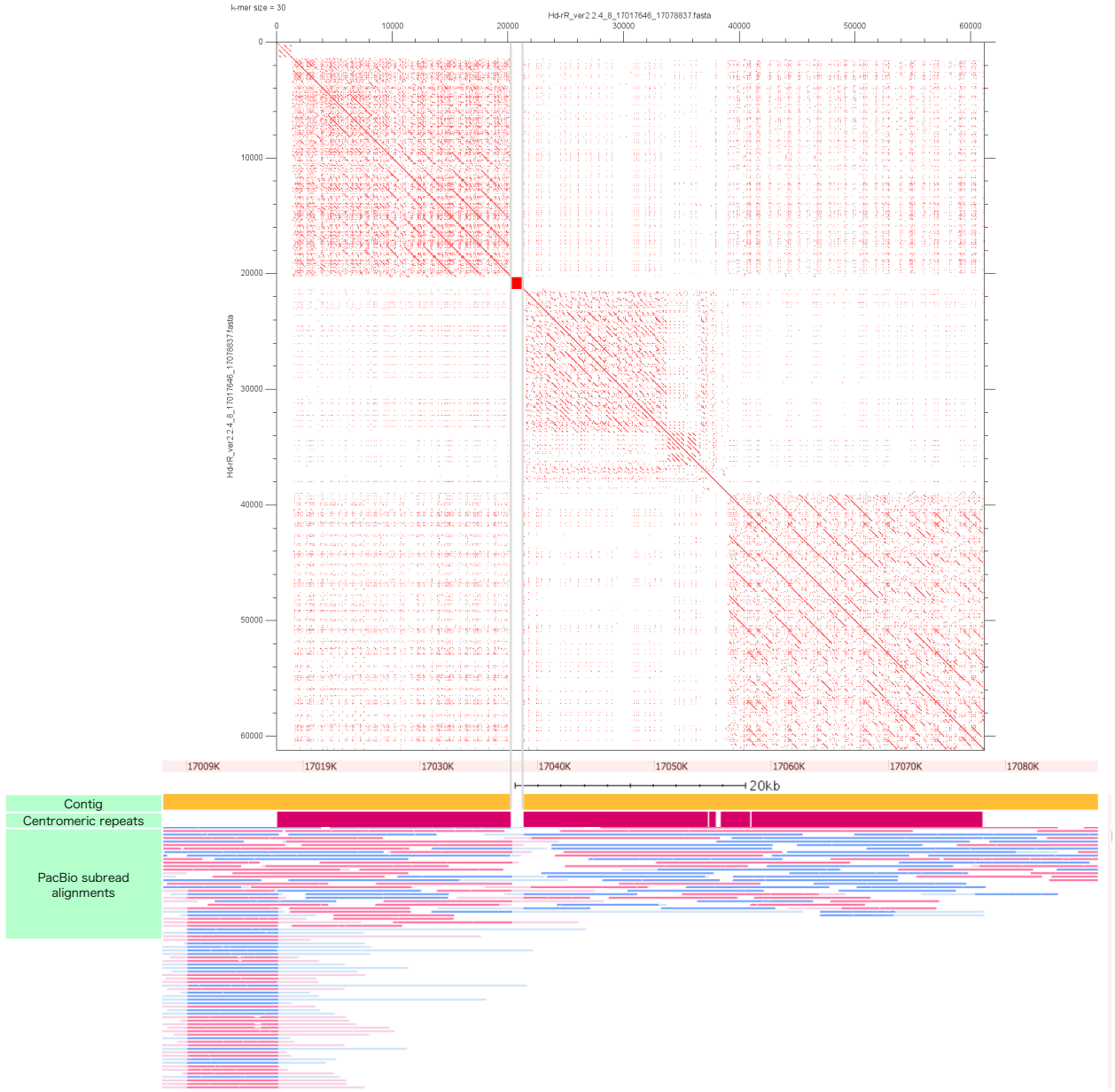

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

Hd-rR chr.9

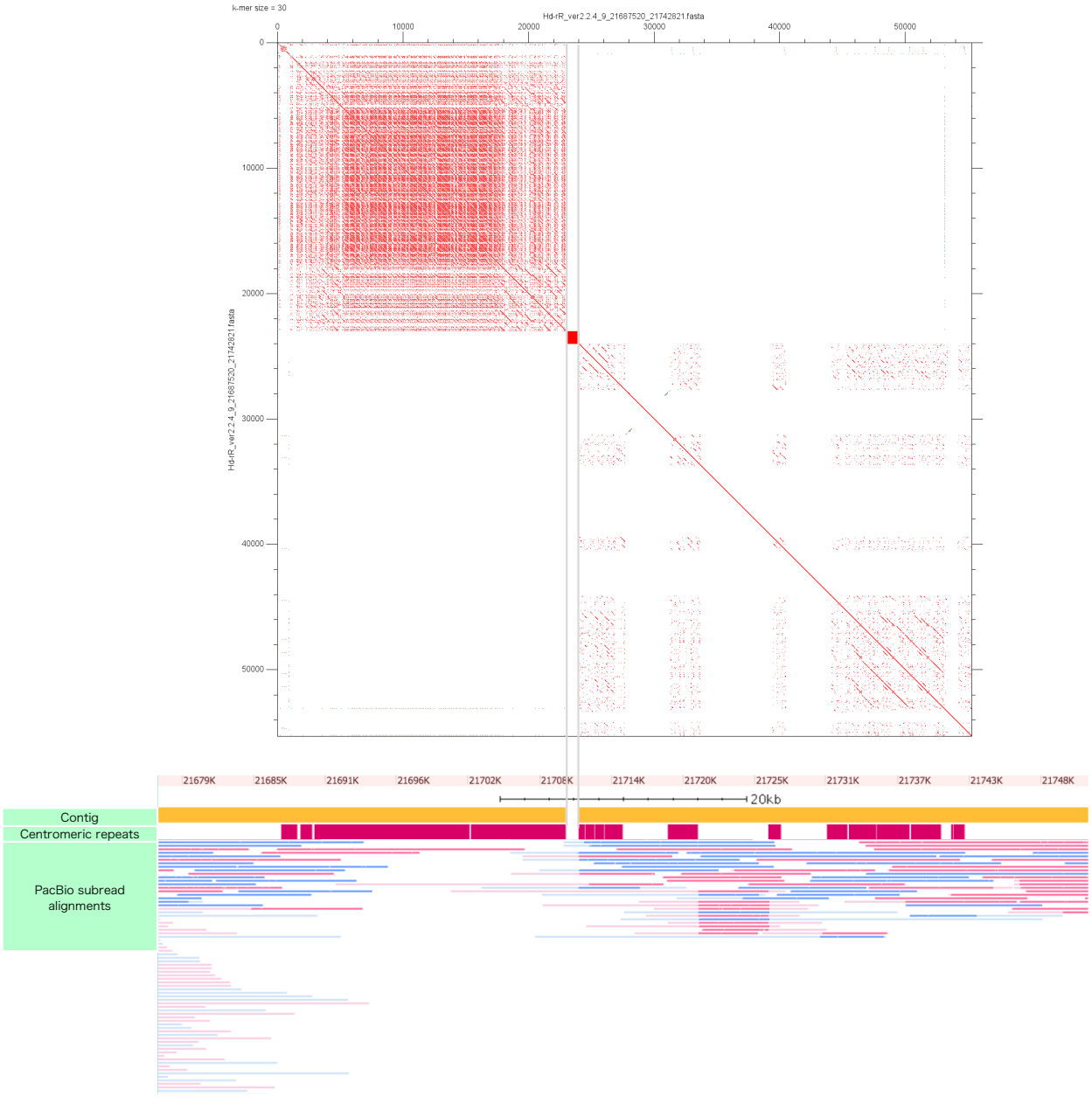

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

Hd-rR chr.12

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

Hd-rR chr.13

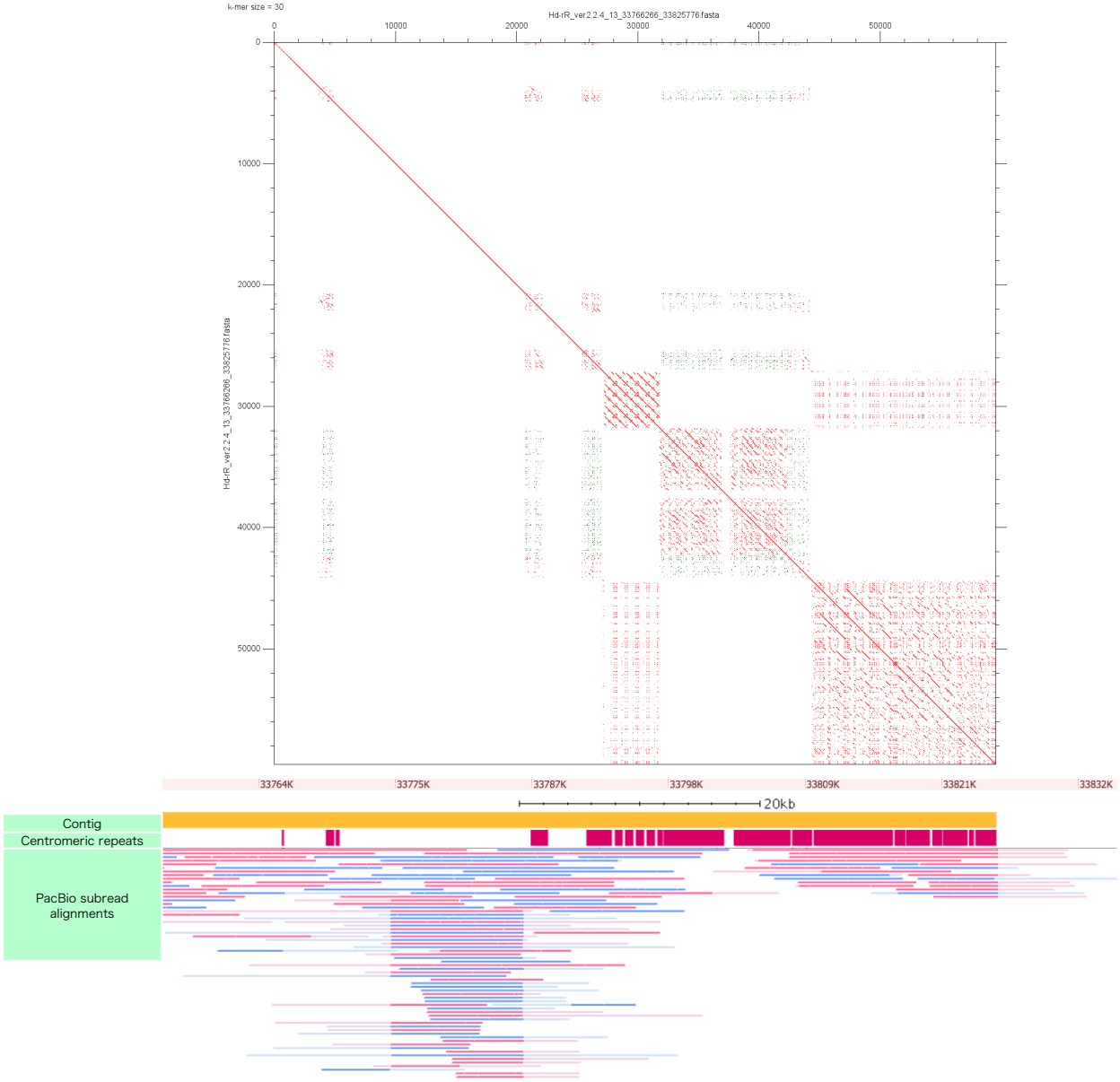

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

Hd-rR chr.14

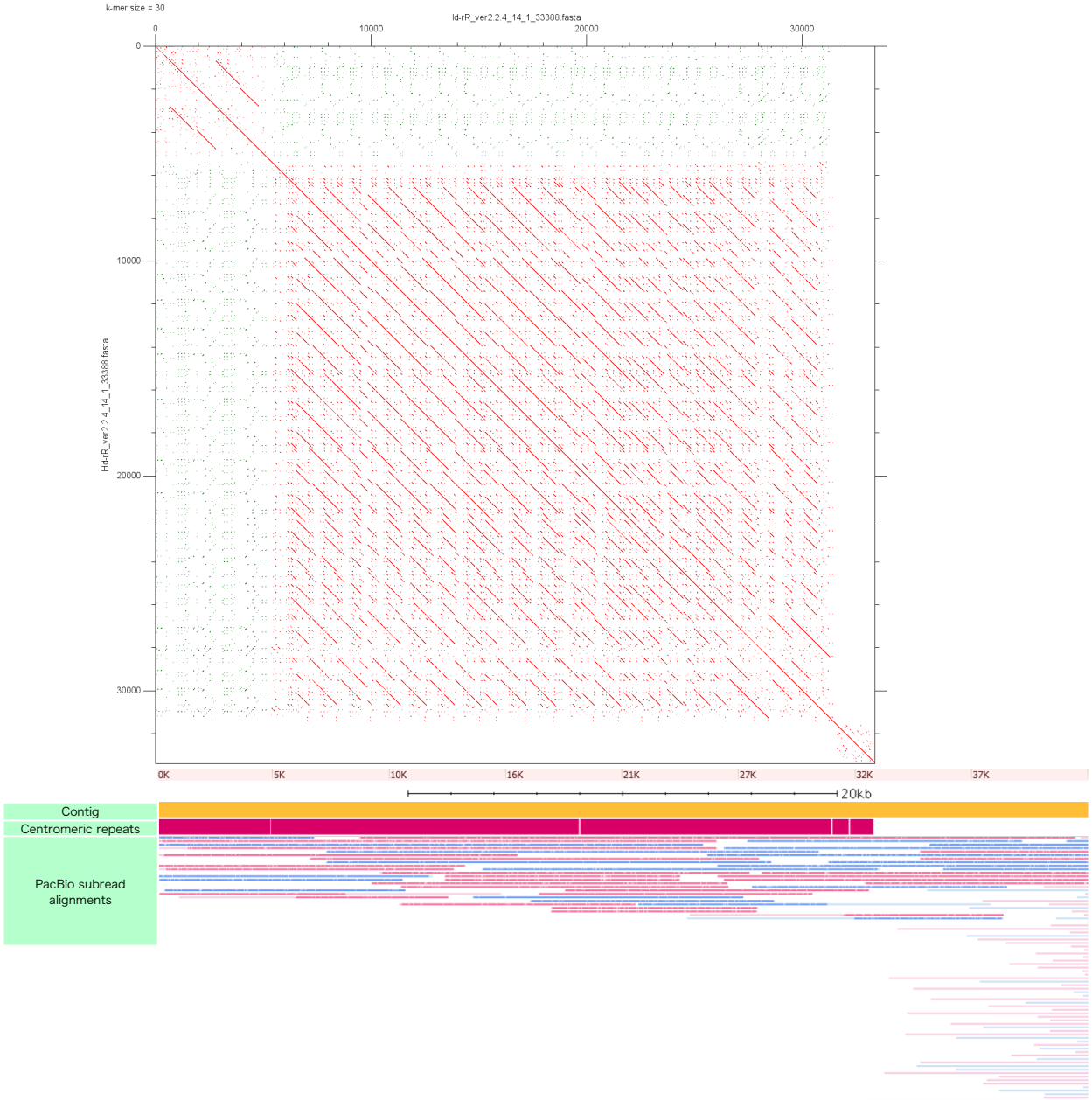

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

Hd-rR chr.18

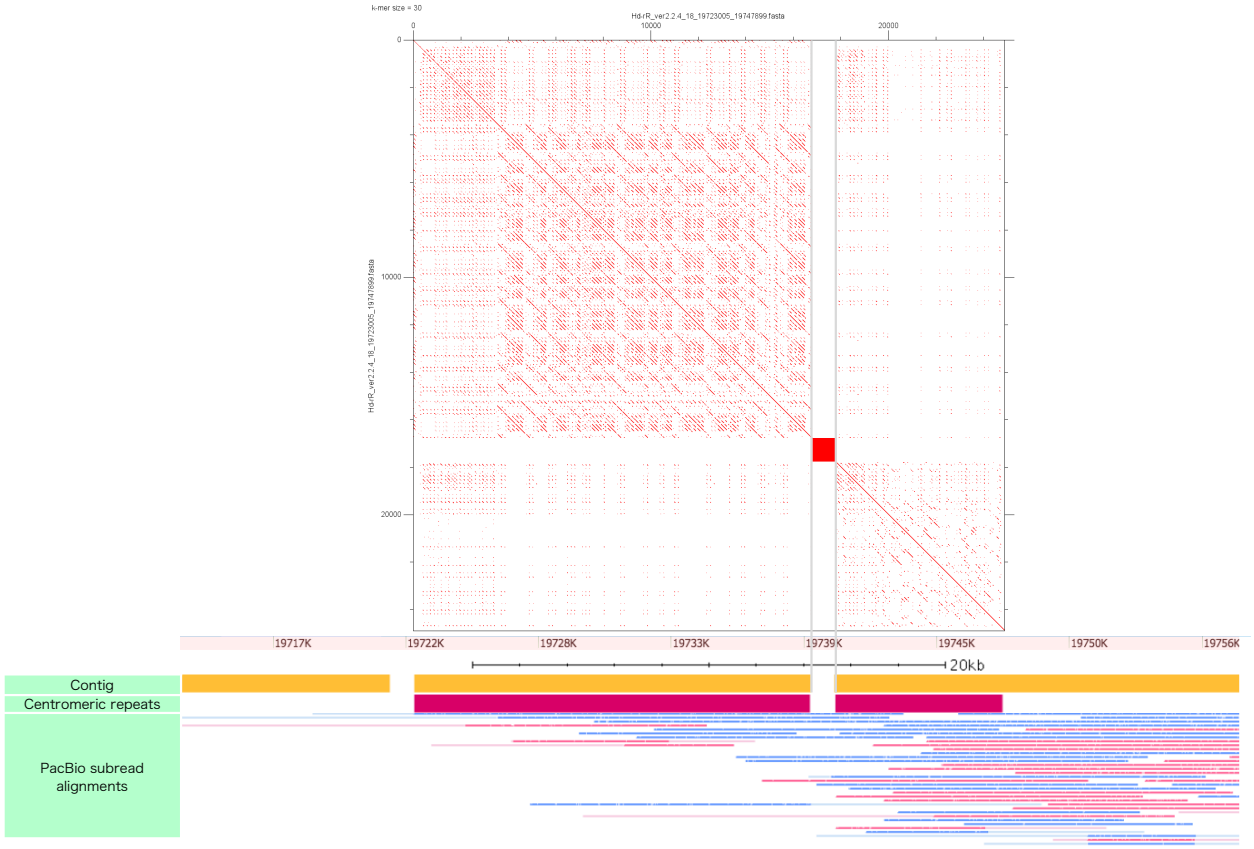

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

Hd-rR chr.19

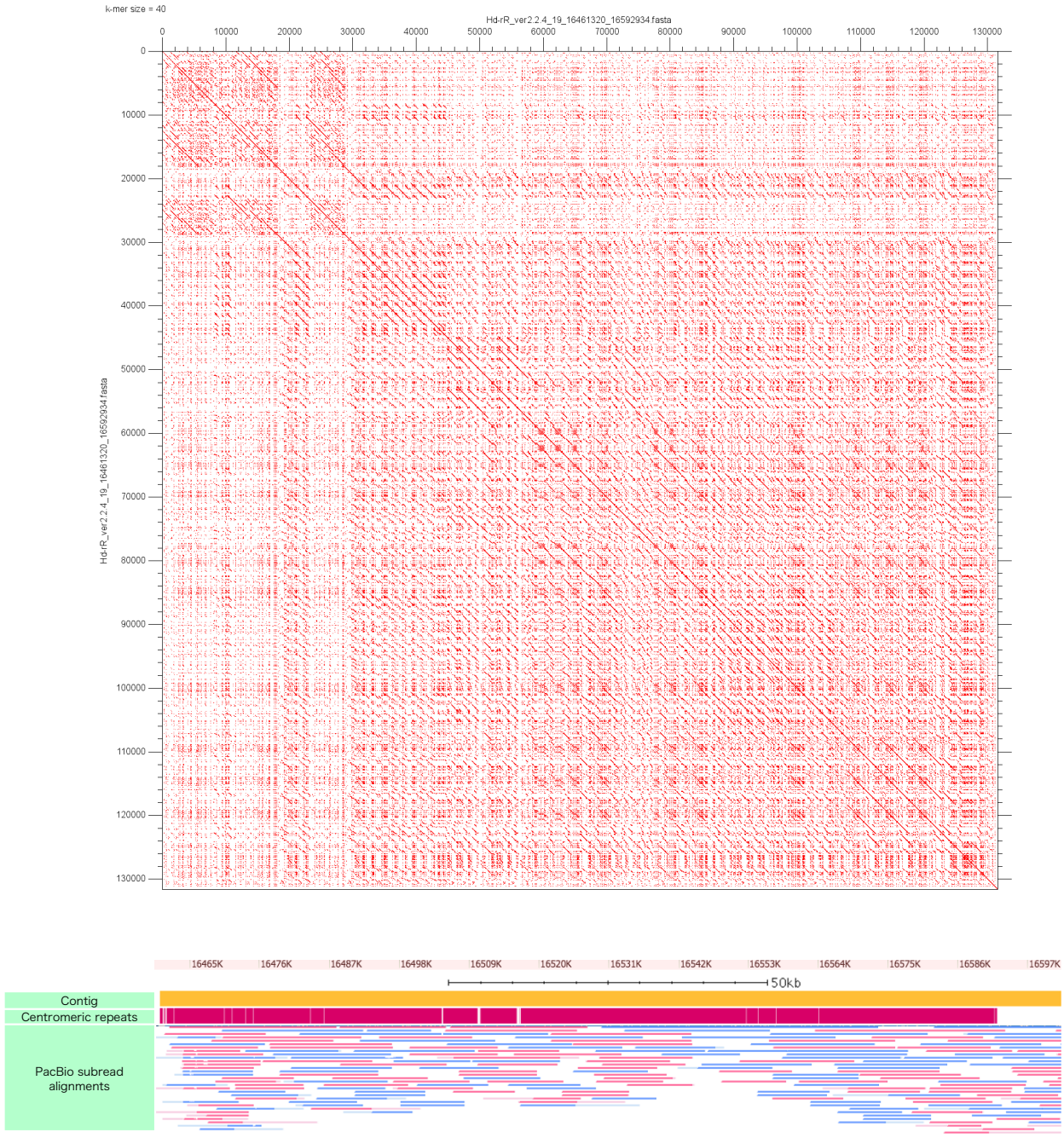

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

Hd-rR chr.20

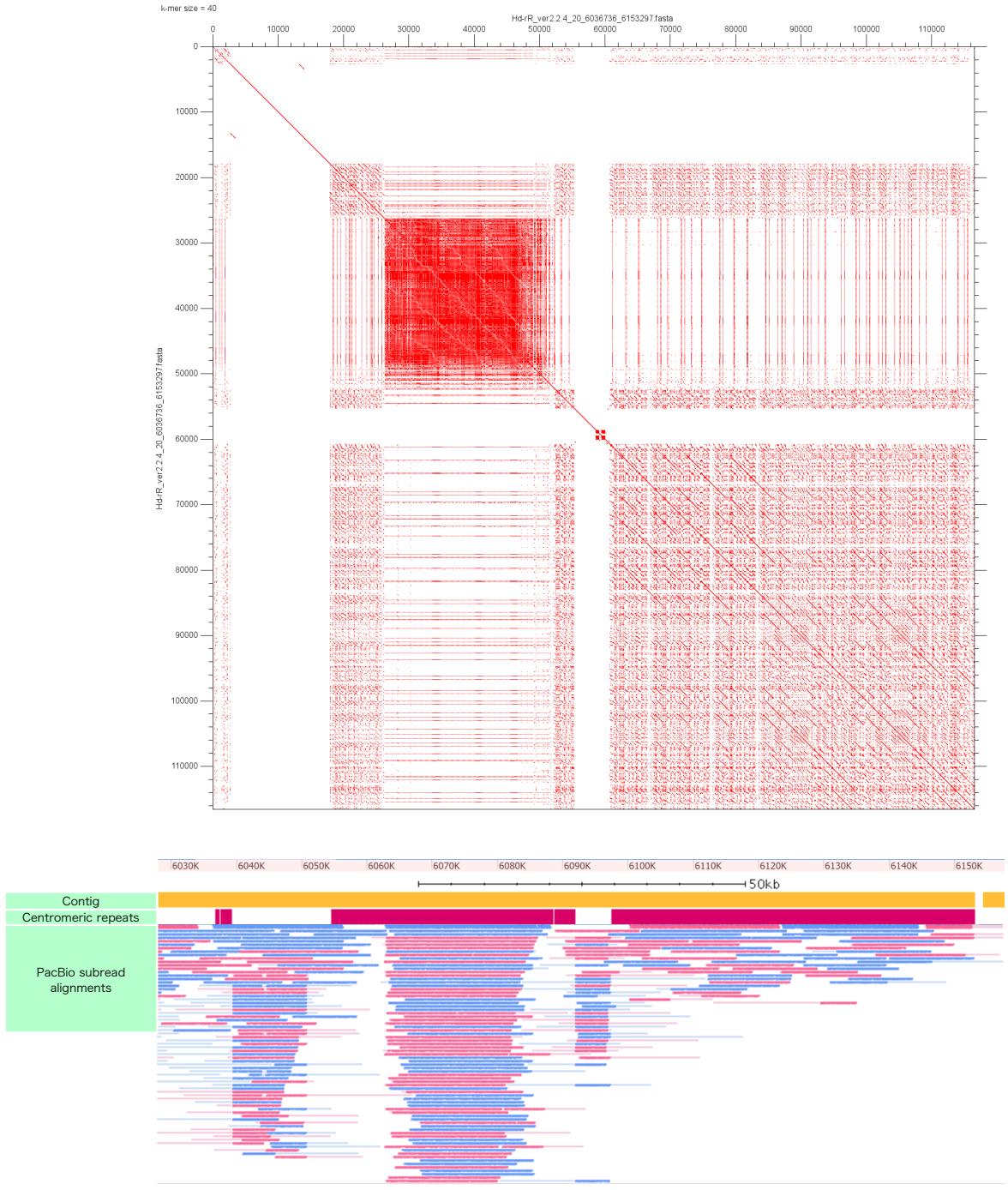

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

Hd-rR chr.21

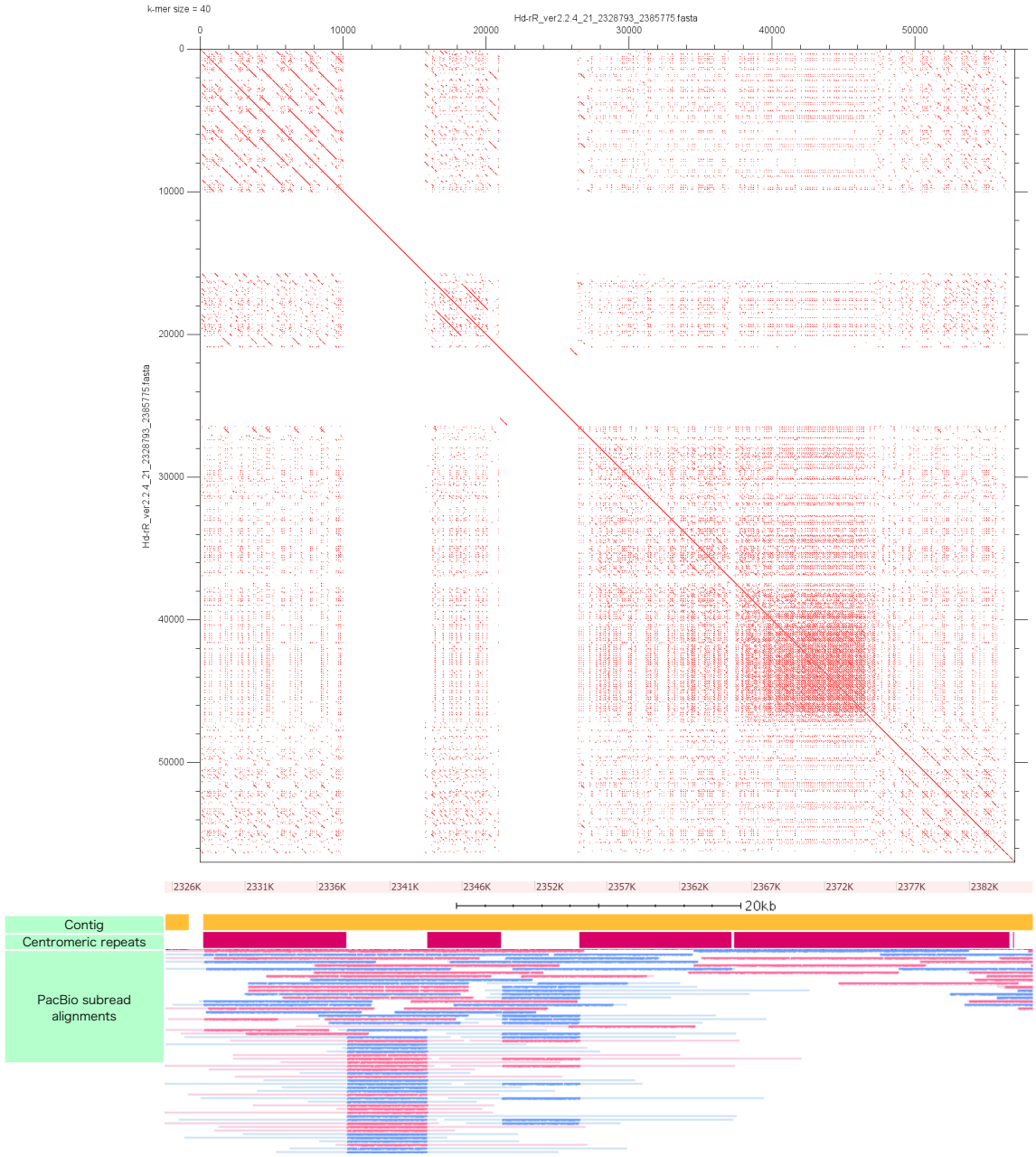

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

Hd-rR chr.22

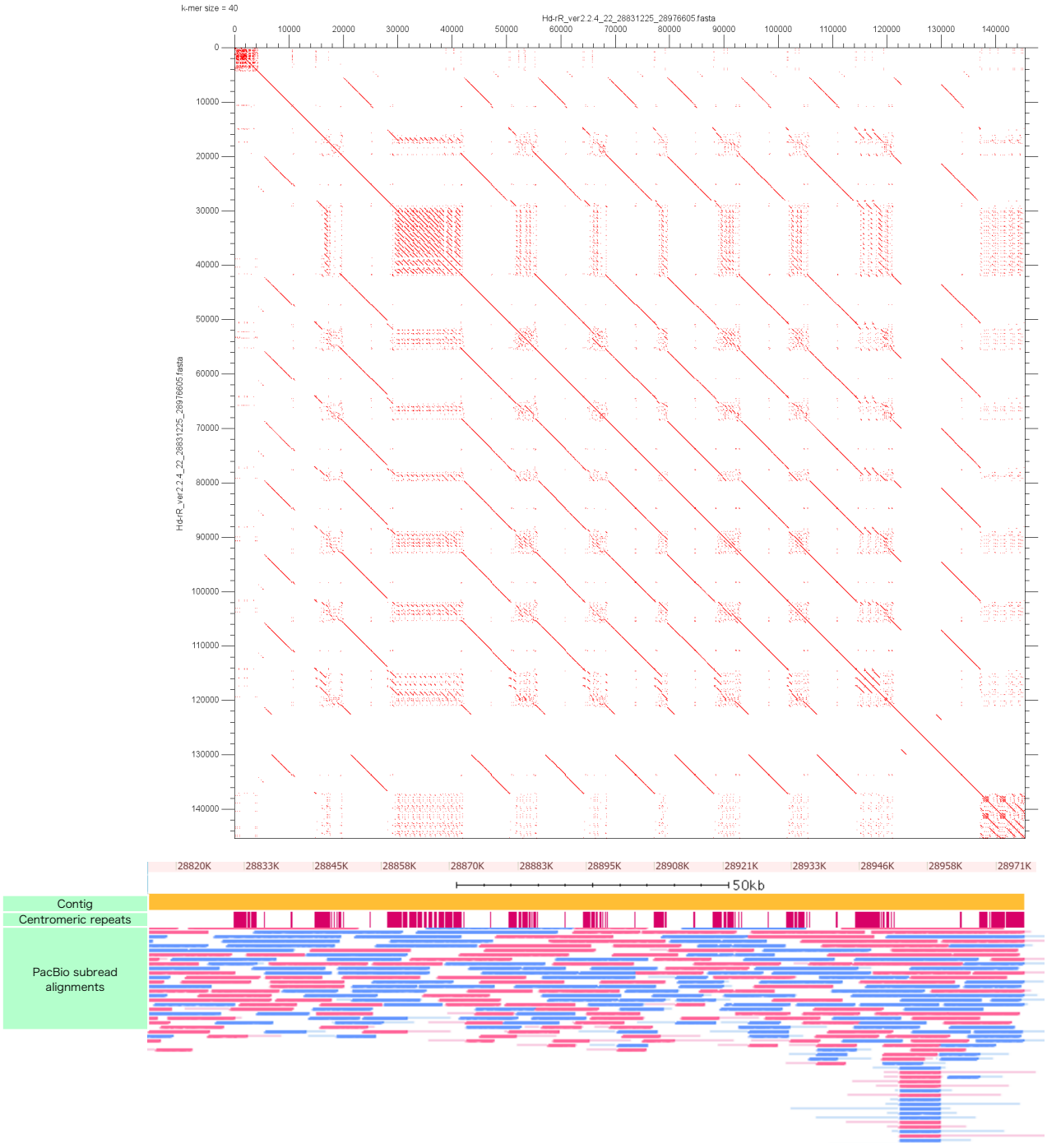

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

HSOK chr.2

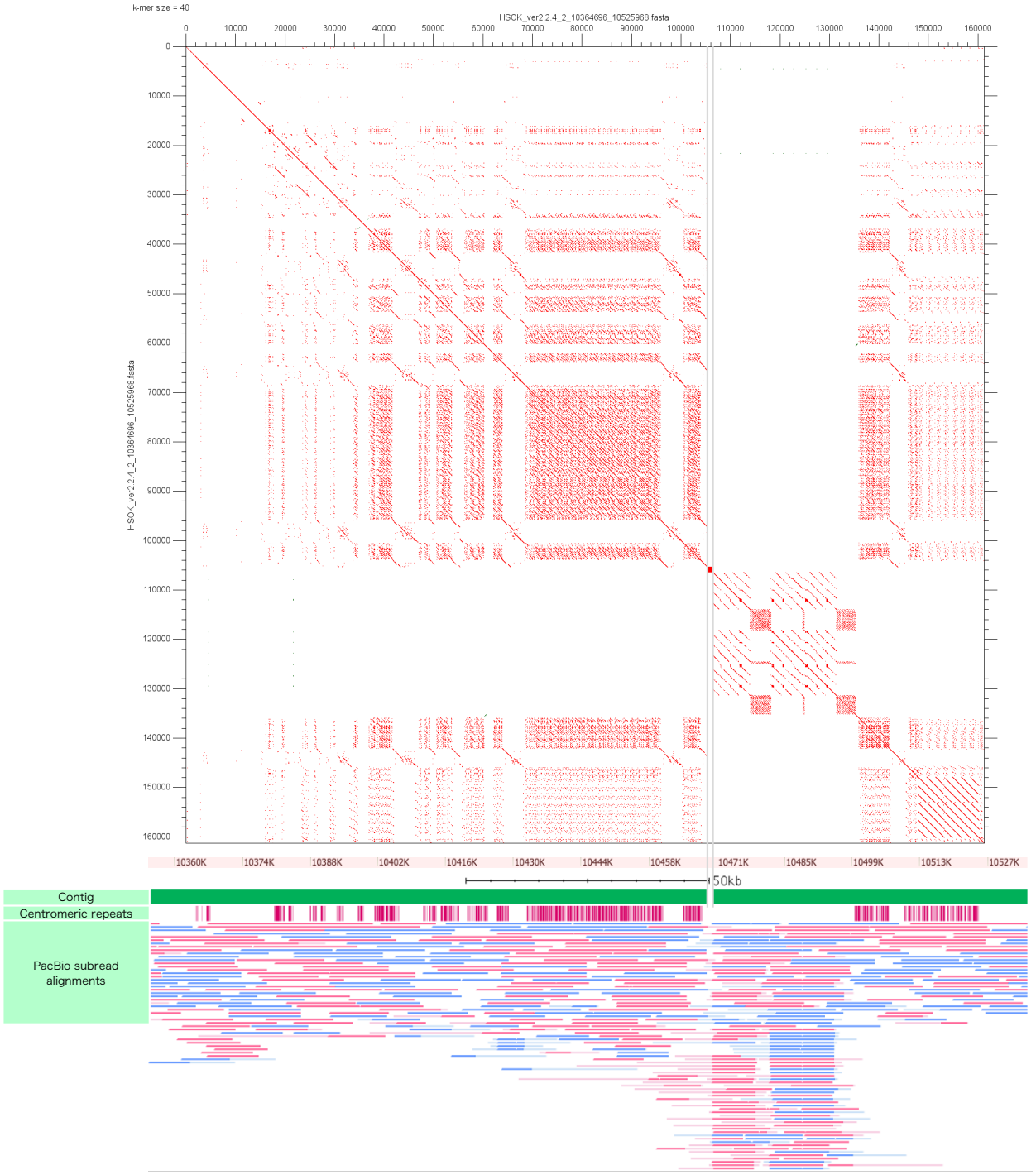

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

HSOK chr.4

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

HSOK chr.7

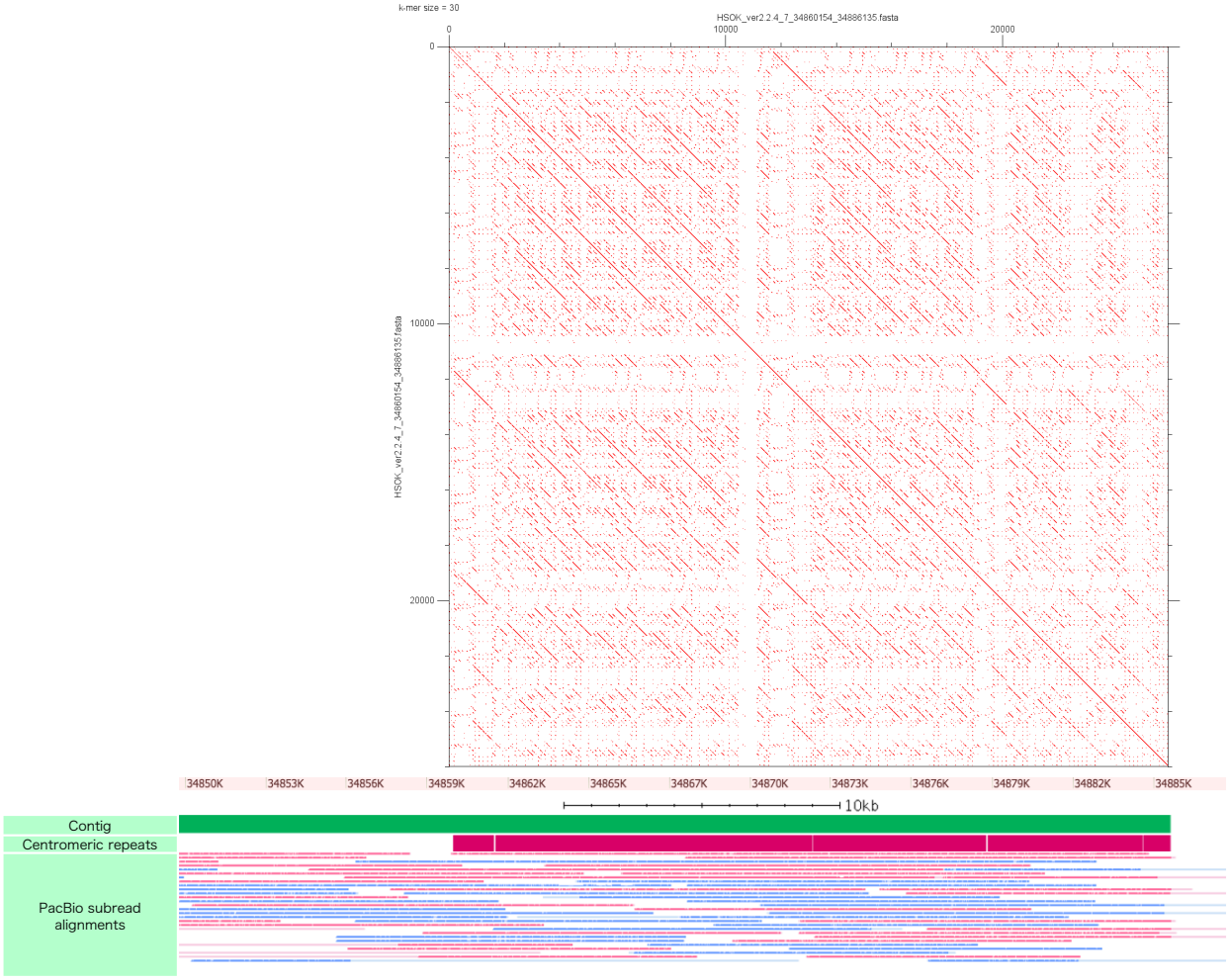

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

HSOK chr.8

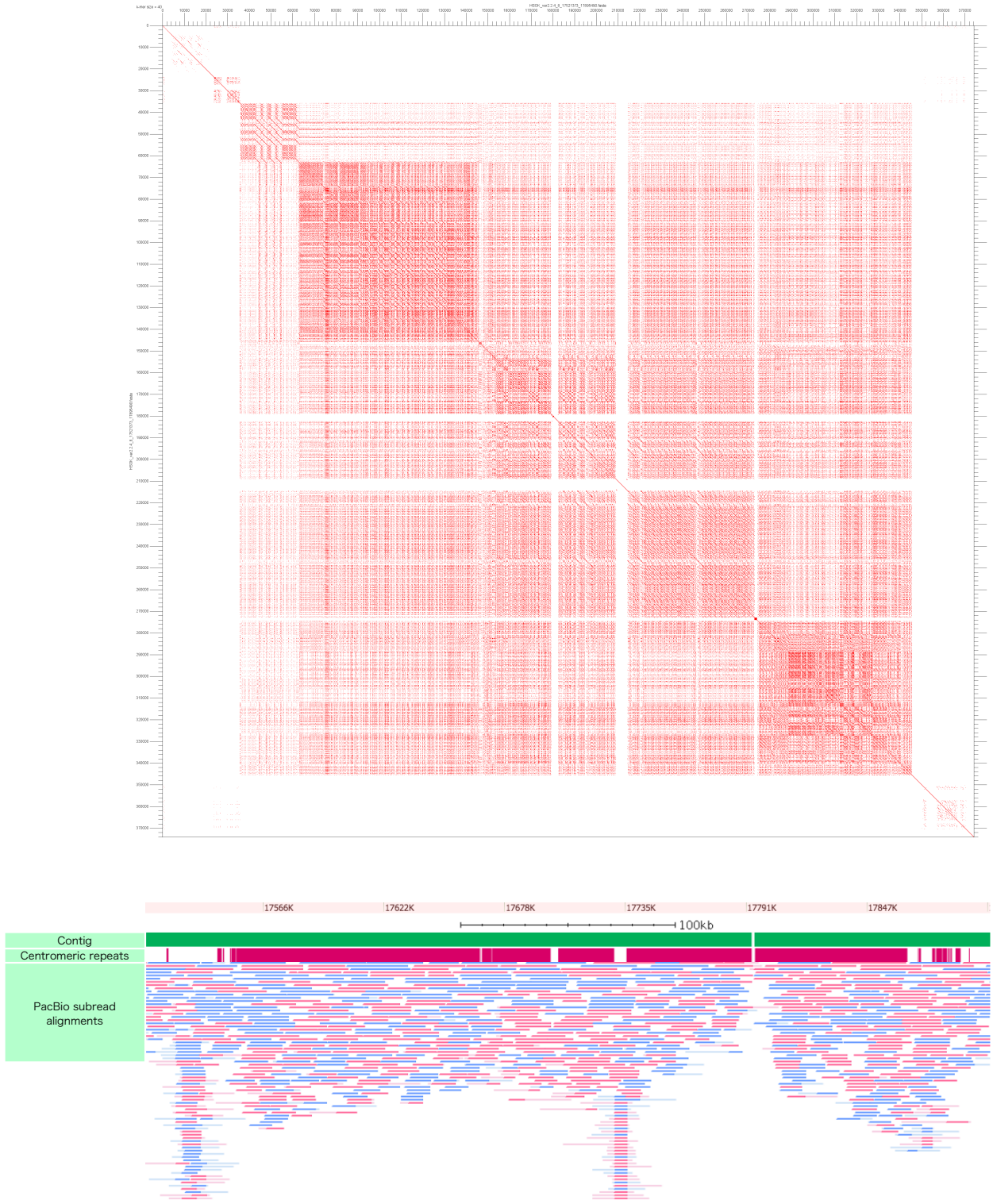

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

HSOK chr.11

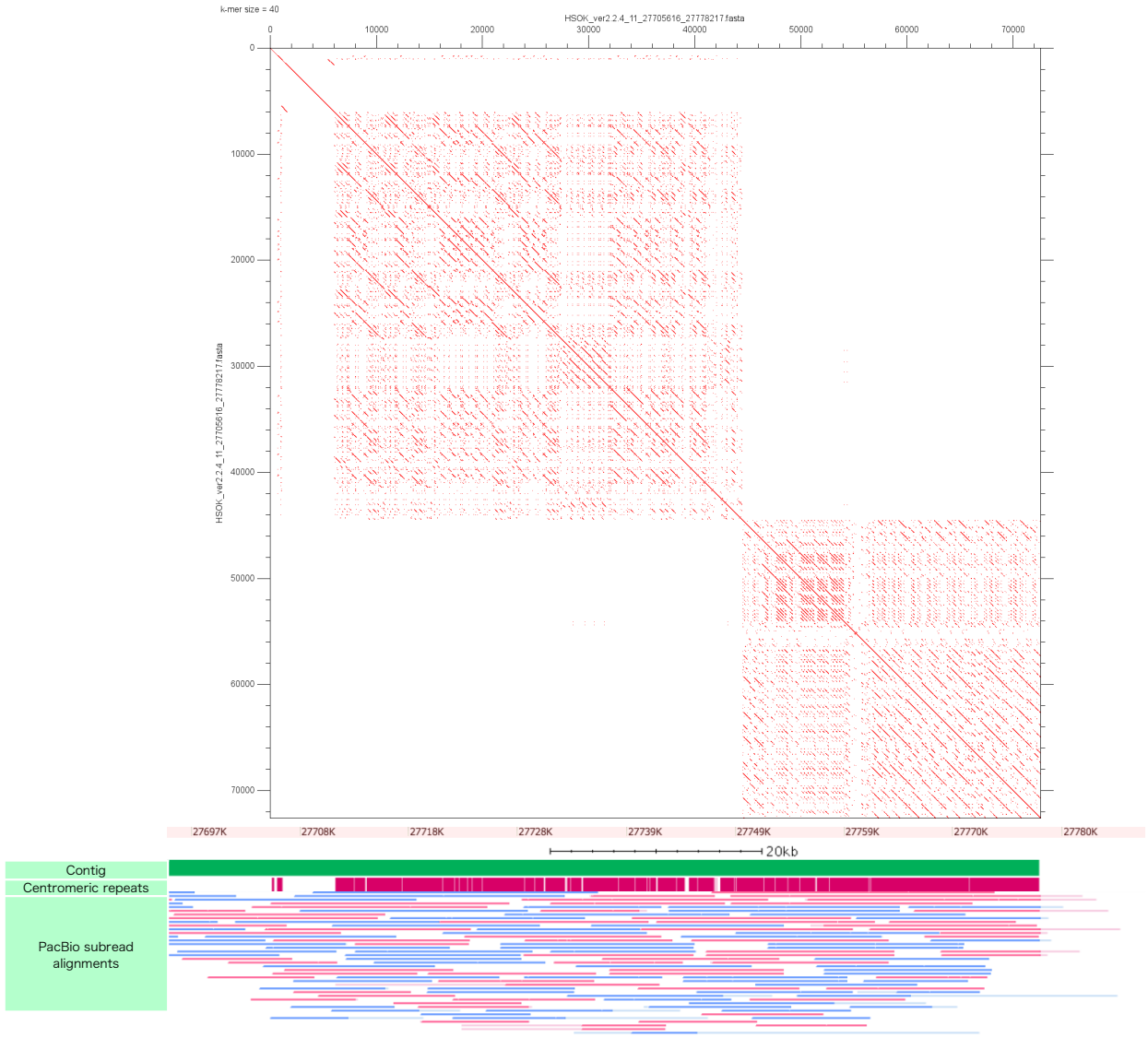

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

HSOK chr.12

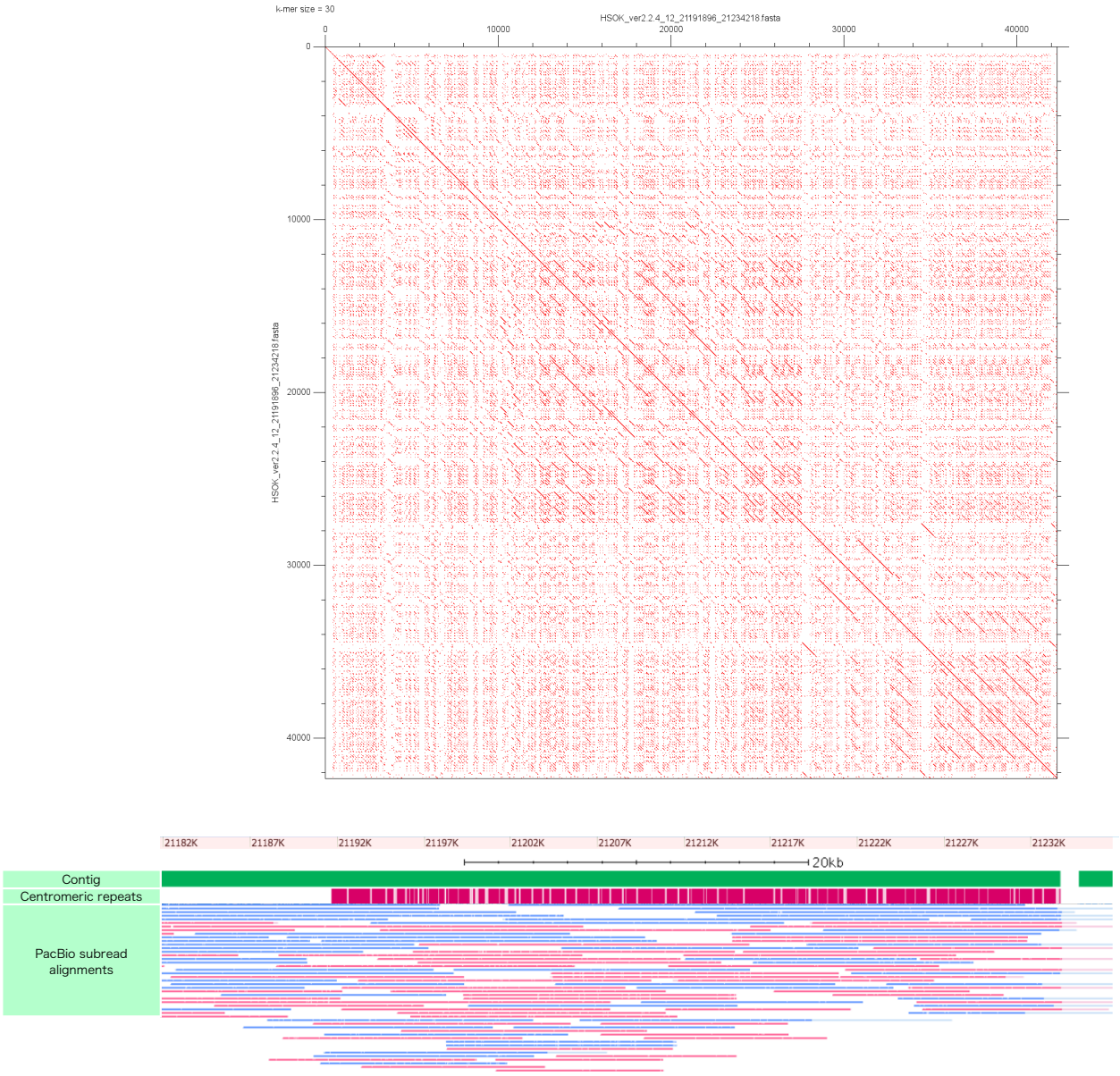

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

HSOK chr.15

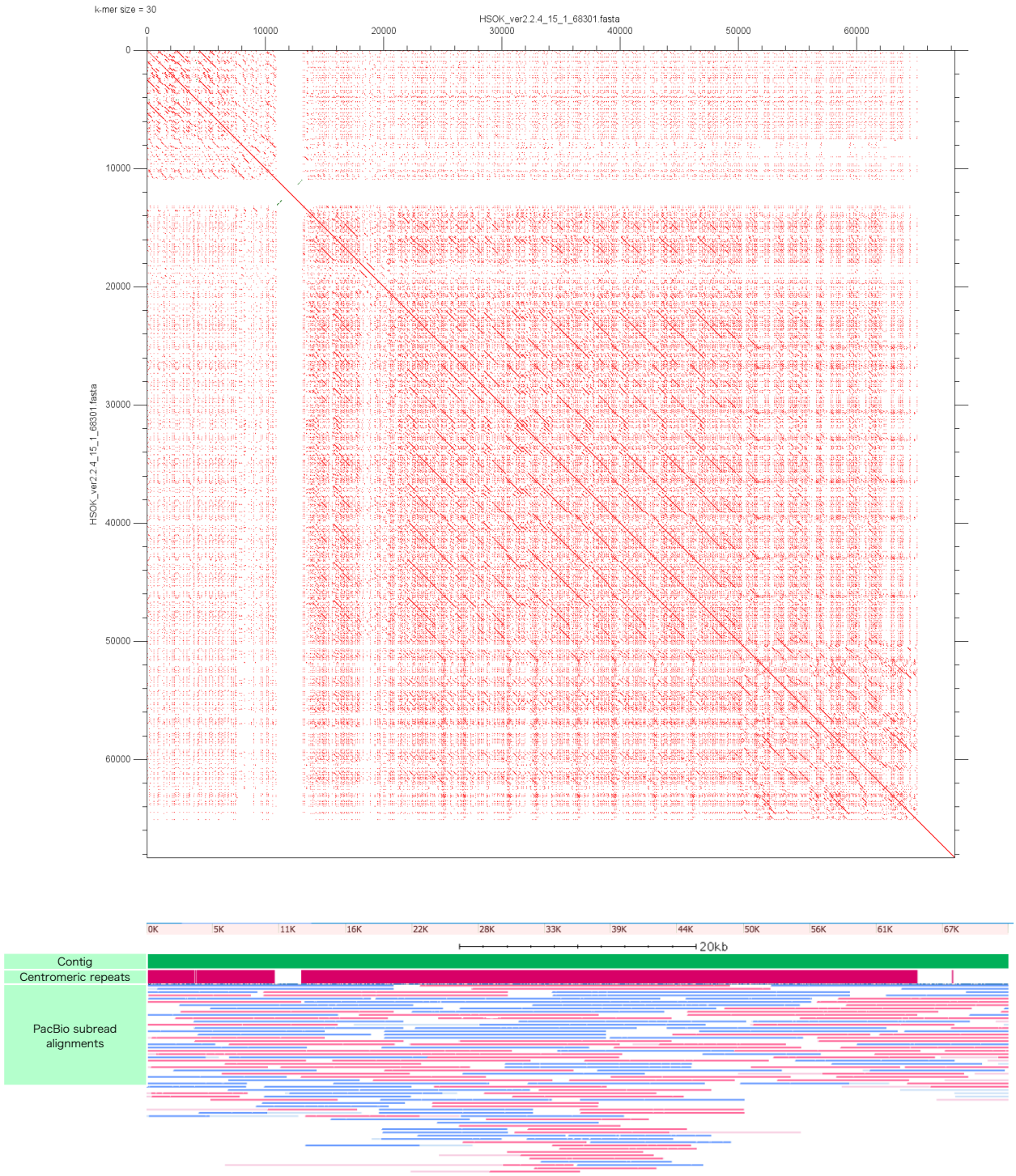

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

HSOK chr.20

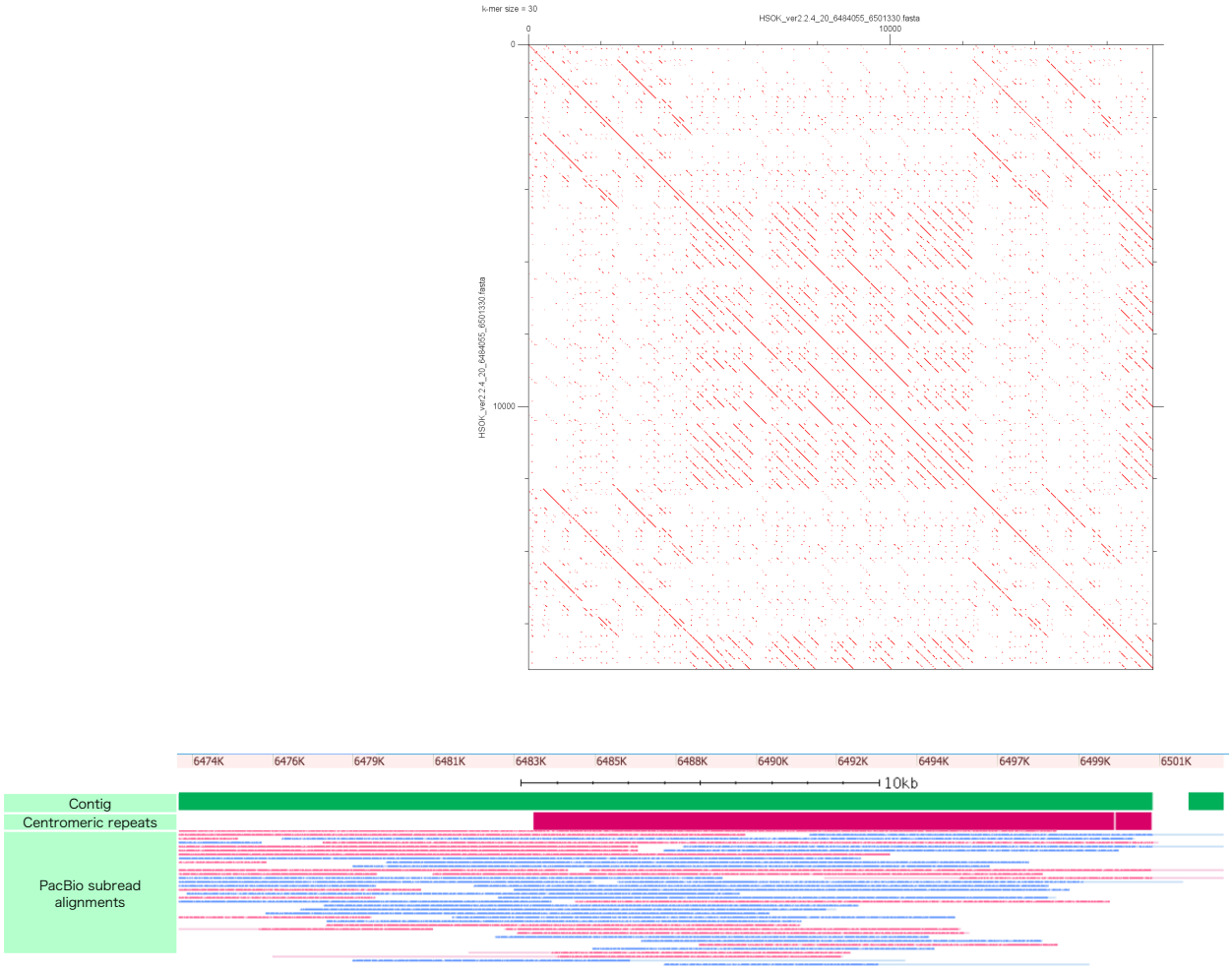

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

HSOK chr.23

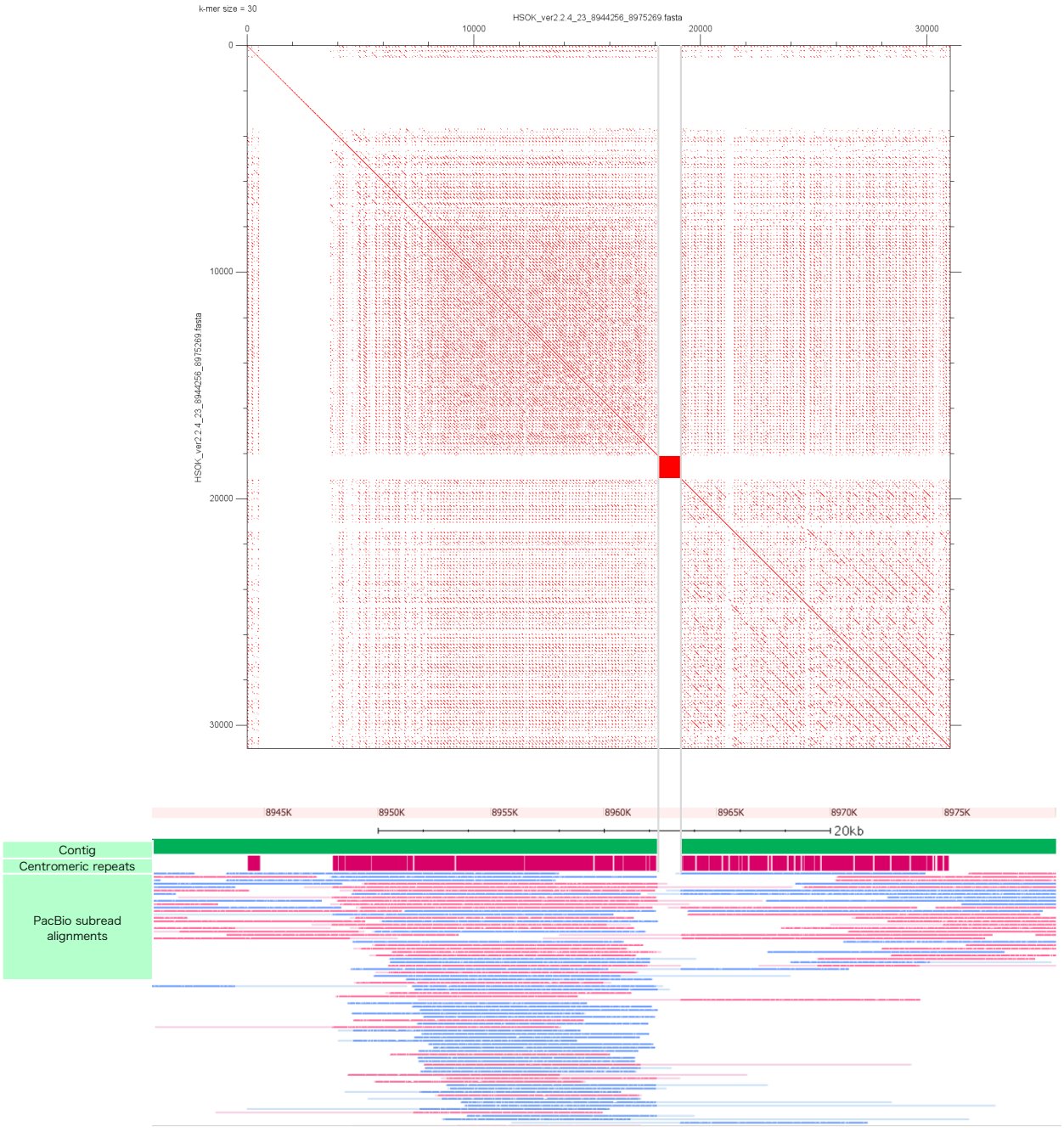

Figure S3: Centromere landscapes. (top) self-alignment dot plot. k-mer size of the matches is indicated at the top left of the figure. (bottom) distribution of centromeric repeats and PacBio subread mapping.

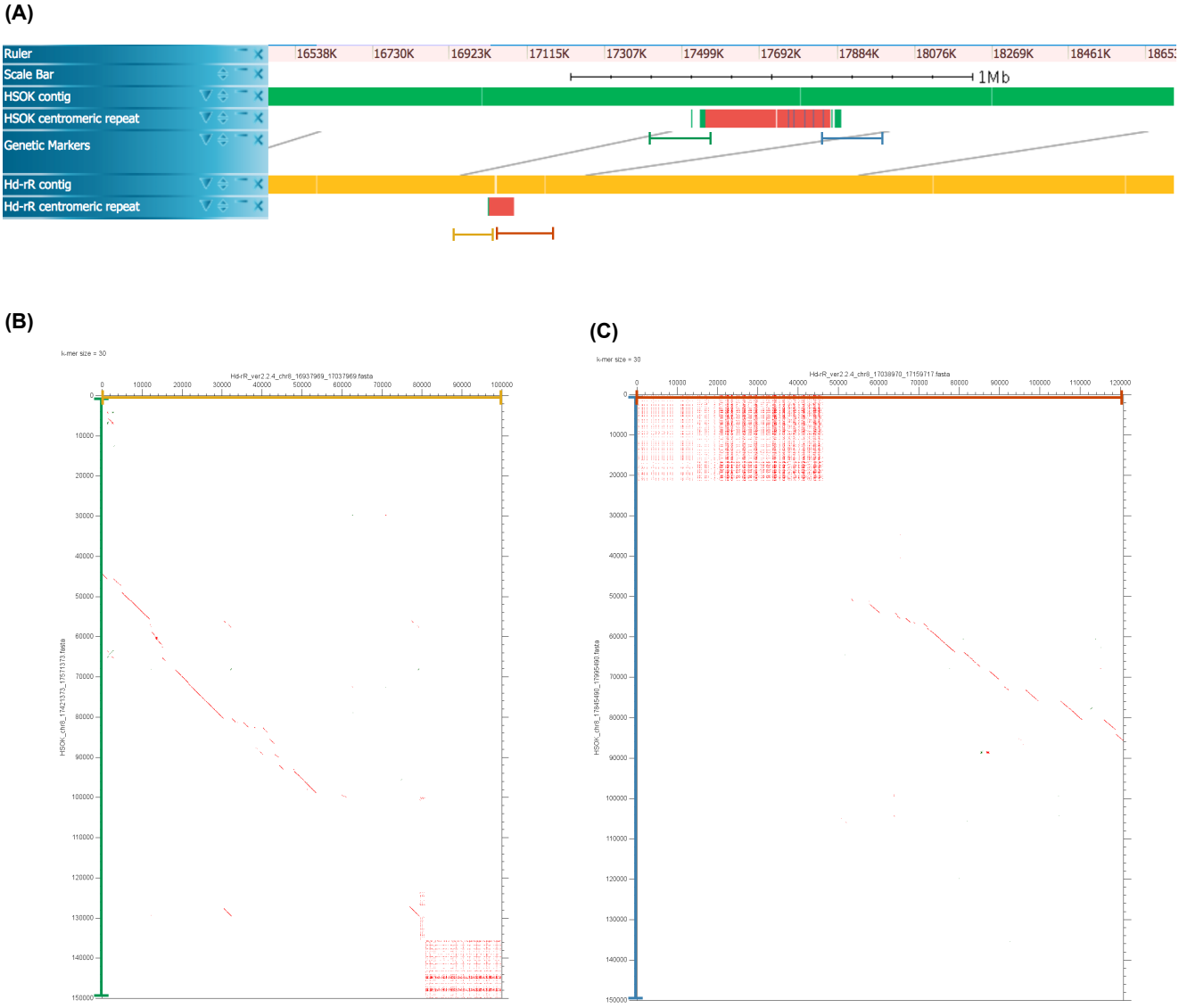

Figure S4: (A, B, C) Comparison of the centromeric transitional regions of chromosome 8 of Hd-rR and HSOK.

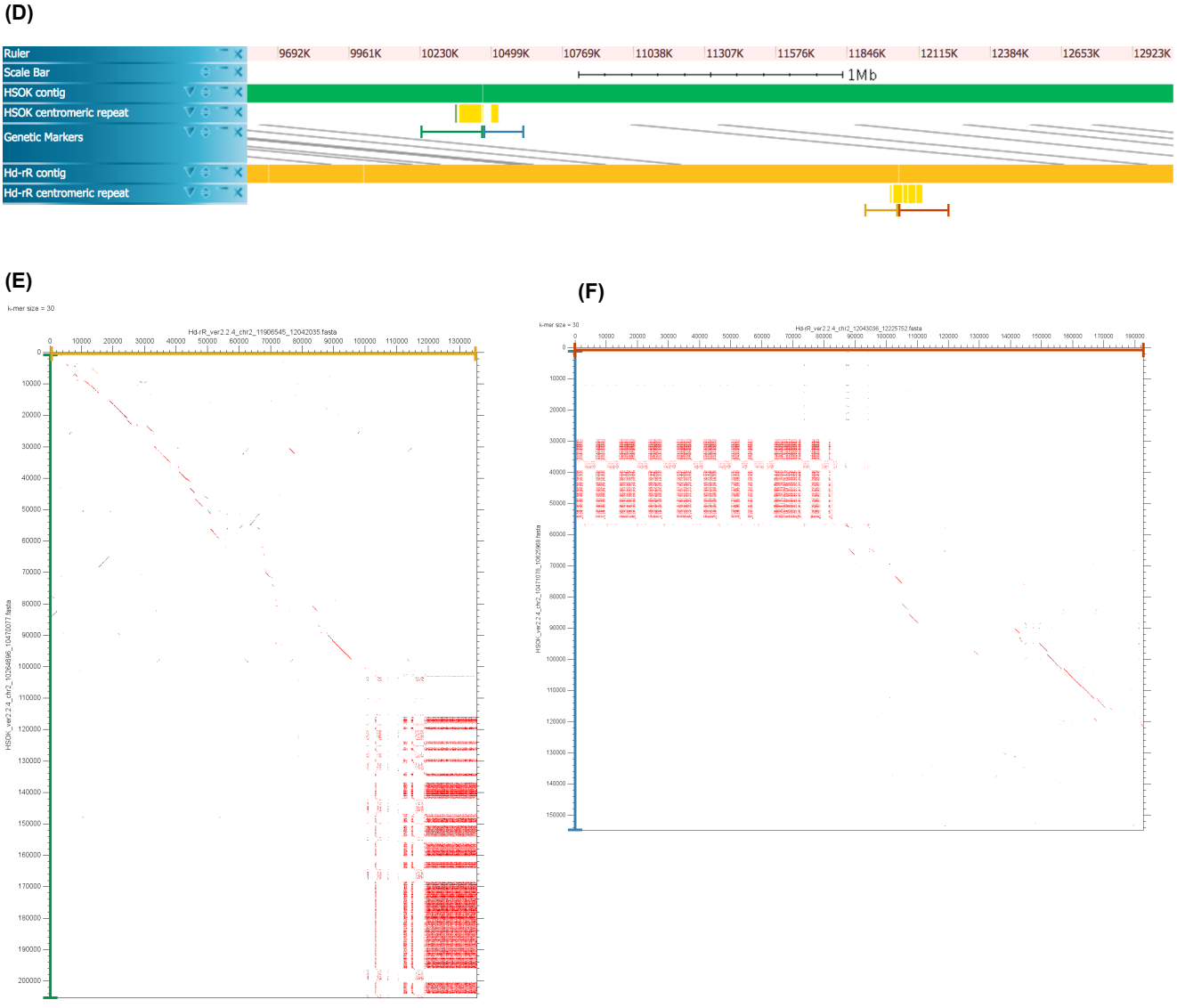

Figure S4: (D, E, F) Comparison of the centromeric transitional regions of chromosome 2 of Hd-rR and HSOK.