Introduction

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].

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-satellite comprise hundred kilobase- to megabase-sized arrays in each chromosome. The alpha-satellite monomer 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 (discussed in detail below).

While repeat-based regional centromeres are the most common structure in eukaryotes, some species possess different types of centromeres. Budding yeast *Saccharomyces cerevisiae* has specific ~125-bp sequences at the centromere; this is called a "point centromere". 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".

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 understandings include approximate estimation of repeat array size and its divergence among individuals [?], the presence of chromosome-specific alpha-satellite HOR patterns [3] and super-chromosomal subfamilies of alpha-satellites [4]. 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 [5].

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 remained as huge gaps. Whereas divergent monomeric portion around pericentromeres were assembled in many chromosomes, only a limited number of chromosomes reached more homogeneous HOR regions [6,7]. Nevertheless, subsequent analyses on a few chromosomes that did assemble till the core centromeric regions offered sequence landscapes with never-seen resolution [8–11].

Although second-generation sequencers (SGSs) represented by Illumina and 454 accomplished a number of genome assemblies [12], 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 [13].

A number of computational studies 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 [14, 15], whereas others identified novel HOR patterns from assembled sequences [16] or from WGS data [17]. Melters *et al.* [15] 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 [18]. Another remarkable computational study is from Miga *et al.* [19], 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 downstream analyses like sequencing read mapping. The generated centromeric sequences, moreover, has been included in the latest human reference genome (GRCh38).

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 of SGS technologies. However, recently-emerging long-read technologies of PacBio [?] and Oxford Nanopore [20] are expected as promising tools for centromere studies [21, 22]. For example, 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 chance 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 [?]. 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 [23]. 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 [24–26]. 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 [24]. Some studies targeting centromere-associated repeat-rich regions have also been conducted [?, 27]. In addition, a computational tool for detecting HOR patterns from long reads was developed [28]. Oxford Nanopore sequencers yield even longer read length (e.g. MinION sequencer routinely yields >150 kb read) with accuracy of ~92% [20]. However, some early-adopting laboratories 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 sequencing technology and protocols.

Results

Genome assembly of three medaka strains

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

Genomic abundance of centromeric repeats

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

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

Validation of centromeric sequence assembly

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

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

Centromeric repeat distribution

The distribution of centromeric repeats in the three medaka strain genomes were revealed by searching their genomes using RepeatMasker (Table 2). For those chromosomes that have >1 kb centromeric repeat, positions of the centromeres in chromosomes were classified, employing the nomenclature defined by Levan *et al.* (1964) (Table 2). Although the nomenclature was originally based on microscopic inspection of the centromeres in chromosomes rather than repeat distribution in the DNA sequence level, nevertheless the sequence-based classification conducted here is informative for inferring evolutionary relationship between the chromosomes. The composition of positional types were consistent with a previous karyotype study []. Centromeric positions of the same chromosome were mostly conserved among the strains, confirmed by observing the corresponding pair of genetic markers flanked the repeat arrays, with only two exceptions in chromosomes 4 and 6 (Supplementary Fig. S2). For chro-

Table 1: Centromeric repeat genomic abundance

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

mosome 4, Hd-rR had an acrocentric repeat array whereas HSOK had a metacentric array. For chromosome 6, all the three strains had acrocentric repeat arrays but those of Hd-rR and HSOK and that of HNI were on the opposite side of the chromosome. As the karyotype study has revealed that the three strains possess slightly different sets of centromeric positions [], the difference of chromosomes 4 and 6 may be derived from the *bona fide* karyotype difference. Of note, Hd-rR chromosome 21 possessed metacentric and acrocentric arrays of nearly the same length (41.6 kb and 45.5 kb, respectively; Supplementary Fig. S2), thus it may be a dicentric chromosome where one of the arrays forms the functional centromere whereas the other is silenced.

Inter-chromosomal centromeric sequence conservation

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

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

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

Methods

Sequencing and genome assembly

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

The genomes were sequenced with PacBio SMRT sequencing and were assembled into contigs using FALCON assembler []. The contigs were then polished with PacBio reads using Quiver [] and with Illumina reads using Pilon []. A number of contigs that contained long centromeric repeat arrays were not polished with Pilon because it was observed that extremely more bases were corrected on centromeric regions than other genomic regions presumably due to mismap-

Table 2: Centromeric repeat distribution

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

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

Table 3: Super-chromosomal subfamilies of centromeric repeats

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

Chromosomes were classified into four subfamilies (SF). Chromosomes in brackets are the ones that have significantly more amount of repeats classified into another subfamily. Hd-rR chromosome 21 possessed two distantly-positioned arrays, thus is notated as 21m (metacentric) and 21a (acrocentric; see Table 2 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 2.

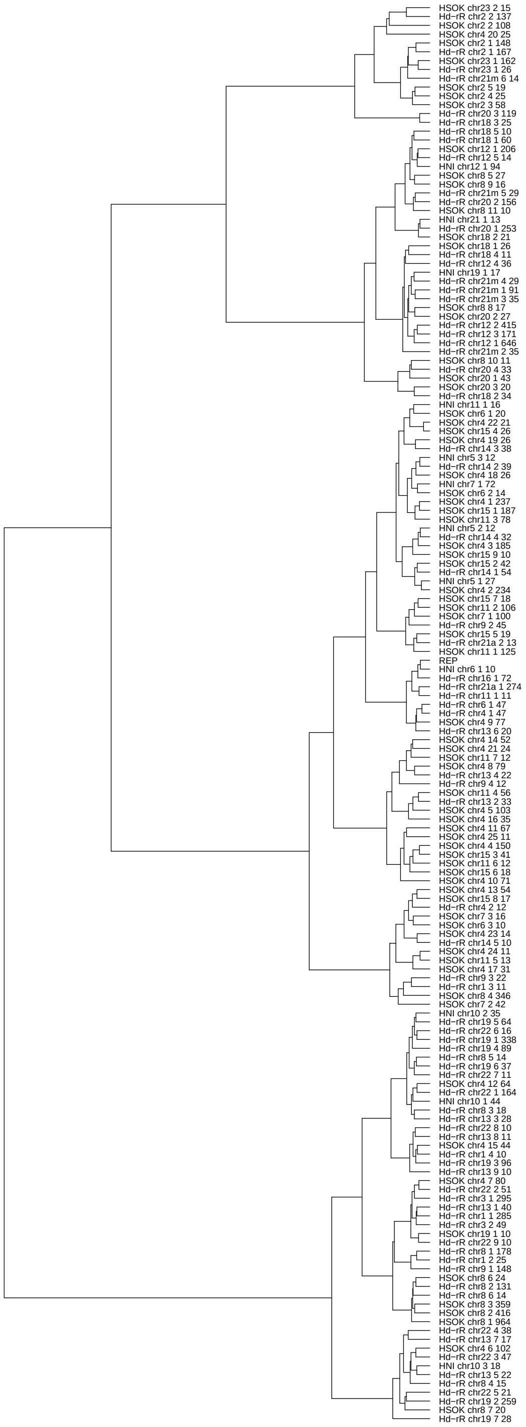

Figure 1: Hierarchical clustering of chromosome-representative monomers. Monomers are labeled as species, chromosome, cluster index, number of the cluster constituents.

ping of short reads. The polished contigs were mapped to the chromosomes using SNP genetic markers. Hd-rR contigs were further scaffolded using BAC- and fosmid-end pair reads and a number of unanchored contigs were positioned into the chromosomes using Hi-C contact frequency data.

Validating centromeric sequence assembly

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

Estimating genomic abundance of centromeric repeats

In order to minimize the effect of high error rate of PacBio sequencing on abundance estimation of the centromeric repeats, only high quality subreads were used for this step. Specifically, subreads were filtered with the criteria that average base quality over the all bases >10. Also, subreads shorter than 1 kb were excluded. The filtered subreads were then scanned by RepeatMasker with a sensitive setting using the medaka representative centromeric satellite monomer sequence as a custom library. Genomic fraction of the medaka centromeric satellite for each strain was estimated by the ratio of total amount of masked centromeric satellite in the total length of the filtered subreads (Table 1).

Revealing centromeric repeat distribution and centromeric positions

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

Inter-chromosomal centromeric sequence comparison

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

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

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

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

[4] 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, aug 2001.

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

[6] 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, nov 2004.

[7] Xinwei She, Julie E Horvath, Zhaoshi Jiang, Ge Liu, Terrence S Furey, Laurie Christ, Royden Clark, Tina Graves, Cassy L Gulden, Can Alkan, Jeff A Bailey, Cenk Sahinalp, Mariano Rocchi, David Haussler, Richard K Wilson, Webb Miller, Stuart Schwartz, and Evan E Eichler. The structure and evolution of centromeric transition regions within the human genome. *Nature*, 430(7002):857–64, aug 2004.

[8] 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.

[9] Mark T. Ross, Darren V. Grafham, Alison J. Coffey, Steven Scherer, Kirsten McLay, Donna Muzny, Matthias Platzer, Gareth R. Howell, Christine Burrows, Christine P. Bird, Adam Frankish, Frances L. Lovell, Kevin L. Howe, Jennifer L. Ashurst, Robert S. Fulton, Ralf Sudbrak, Gaiping Wen, Matthew C. Jones, Matthew E. Hurles, T. Daniel Andrews, Carol E. Scott, Stephen Searle, Juliane Ramser, Adam Whittaker, Rebecca Deadman, Nigel P. Carter, Sarah E. Hunt, Rui Chen, Andrew Cree, Preethi Gunaratne, Paul Havlak, Anne Hodgson, Michael L. Metzker, Stephen Richards, Graham

Scott, David Steffen, Erica Sodergren, David A. Wheeler, Kim C. Worley, Rachael Ainscough, Kerrie D. Ambrose, M. Ali Ansari-Lari, Swaroop Aradhya, Robert I. S. Ashwell, Anne K. Babbage, Claire L. Bagguley, Andrea Ballabio, Ruby Banerjee, Gary E. Barker, Karen F. Barlow, Ian P. Barrett, Karen N. Bates, David M. Beare, Helen Beasley, Oliver Beasley, Alfred Beck, Graeme Bethel, Karin Blechschmidt, Nicola Brady, Sarah Bray-Allen, Anne M. Bridgeman, Andrew J. Brown, Mary J. Brown, David Bonnin, Elspeth A. Bruford, Christian Buhay, Paula Burch, Deborah Burford, Joanne Burgess, Wayne Burrill, John Burton, Jackie M. Bye, Carol Carder, Laura Carrel, Joseph Chako, Joanne C. Chapman, Dean Chavez, Ellson Chen, Guan Chen, Yuan Chen, Zhijian Chen, Craig Chinault, Alfredo Ciccodicola, Sue Y. Clark, Graham Clarke, Chris M. Clee, Sheila Clegg, Kerstin Clerc-Blankenburg, Karen Clifford, Vicky Cobley, Charlotte G. Cole, Jen S. Conquer, Nicole Corby, Richard E. Connor, Robert David, Joy Davies, Clay Davis, John Davis, Oliver Delgado, Denise DeShazo, Pawandeep Dhami, Yan Ding, Huyen Dinh, Steve Dodsworth, Heather Draper, Shannon Dugan-Rocha, Andrew Dunham, Matthew Dunn, K. James Durbin, Ireena Dutta, Tamsin Eades, Matthew Ellwood, Alexandra Emery-Cohen, Helen Errington, Kathryn L. Evans, Louisa Faulkner, Fiona Francis, John Frankland, Audrey E. Fraser, Petra Galgoczy, James Gilbert, Rachel Gill, Gernot Glöckner, Simon G. Gregory, Susan Gribble, Coline Griffiths, Russell Grocock, Yanghong Gu, Rhian Gwilliam, Cerissa Hamilton, Elizabeth A. Hart, Alicia Hawes, Paul D. Heath, Katja Heitmann, Steffen Hennig, Judith Hernandez, Bernd Hinzmann, Sarah Ho, Michael Hoffs, Phillip J. Howden, Elizabeth J. Huckle, Jennifer Hume, Paul J. Hunt, Adrienne R. Hunt, Judith Isherwood, Leni Jacob, David Johnson, Sally Jones, Pieter J. de Jong, Shirin S. Joseph, Stephen Keenan, Susan Kelly, Joanne K. Kershaw, Ziad Khan, Petra Kioschis, Sven Klages, Andrew J. Knights, Anna Kosiura, Christie Kovar-Smith, Gavin K. Laird, Cordelia Langford, Stephanie Lawlor, Margaret Leversha, Lora Lewis, Wen Liu, Christine Lloyd, David M. Lloyd, Hermela Loulseged, Jane E. Loveland, Jamieson D. Lovell, Ryan Lozado, Jing Lu, Rachael Lyne, Jie Ma, Manjula Maheshwari, Lucy H. Matthews, Jennifer McDowall, Stuart McLaren, Amanda McMurray, Patrick Meidl, Thomas Meitinger, Sarah Milne, George Miner, Shailesh L. Mistry, Margaret Morgan, Sidney Morris, Ines Müller, James C. Mullikin, Ngoc Nguyen, Gabriele Nordsiek, Gerald Nyakatura, Christopher N. O'Dell, Geoffery Okwuonu, Sophie Palmer, Richard Pandian, David Parker, Julia Parrish, Shiran Pasternak, Dina Patel, Alex V. Pearce, Danita M. Pearson, Sarah E. Pelan, Lesette Perez, Keith M. Porter, Yvonne Ramsey, Kathrin Reichwald, Susan Rhodes, Kerry A. Ridler, David Schlessinger, Mary G. Schueler, Harminder K. Sehra, Charles Shaw-Smith, Hua Shen, Elizabeth M. Sheridan, Ratna Shownkeen, Carl D. Skuce, Michelle L. Smith, Elizabeth C. Sotheran, Helen E. Steingruber, Charles A. Steward, Roy Storey, R. Mark Swann, David Swarbreck, Paul E. Tabor, Stefan Taudien, Tineace Taylor, Brian Teague, Karen Thomas, Andrea Thorpe, Kirsten Timms, Alan Tracey, Steve Trevanion, Anthony C. Tromans, Michele D'Urso, Daniel Verduzco, Donna Villasana, Lenee Waldron, Melanie Wall, Qiaoyan Wang, James Warren, Georgina L. Warry, Xuehong Wei, Anthony West, Siobhan L. Whitehead, Mathew N. Whiteley, Jane E. Wilkinson, David L. Willey, Gabrielle Williams, Leanne Williams, Angela Williamson, Helen Williamson, Laurens Wilming, Rebecca L. Woodmansey, Paul W. Wray, Jennifer Yen, Jingkun Zhang, Jianling Zhou, Huda Zoghbi, Sara Zorilla, David Buck, Richard Reinhardt, Annemarie Poustka, André Rosenthal, Hans Lehrach, Alfons Meindl, Patrick J. Minx, LaDeana W. Hillier, Huntington F. Willard, Richard K. Wilson, Robert H. Waterston, Catherine M. Rice, Mark Vaudin, Alan Coulson, David L. Nelson, George Weinstock, John E. Sulston, Richard Durbin, Tim Hubbard, Richard A. Gibbs, Stephan Beck, Jane Rogers, and David R. Bentley. The DNA sequence of the human X chromosome. *Nature*, 434(7031):325–337, mar 2005.

[10] Chad Nusbaum, Tarjei S Mikkelsen, Michael C Zody, Shuichi Asakawa, Stefan Taudien, Manuel Garber, Chinnappa D Kodira, Mary G Schueler, Atsushi Shimizu, Charles a Whittaker, Jean L Chang, Christina a Cuomo, Ken Dewar, Michael G FitzGerald, Xiaoping Yang, Nicole R Allen, Scott Anderson, Teruyo Asakawa, Karin Blechschmidt, Toby Bloom, Mark L Borowsky, Jonathan Butler, April Cook, Benjamin Corum, Kurt DeArellano, David DeCaprio, Kathleen T Dooley, Lester Dorris, Reinhard Engels, Gernot Glöckner, Nabil Hafez, Daniel S Hagopian, Jennifer L Hall, Sabine K Ishikawa, David B Jaffe, Asha Kamat, Jun Kudoh, Rüdiger Lehmann, Tashi Lokitsang, Pendexter Macdonald, John E Major, Charles D Matthews, Evan Mauceli, Uwe Menzel, Atanas H Mihalev, Shinsei Minoshima, Yuji Murayama, Jerome W Naylor, Robert Nicol, Cindy Nguyen, Sinéad B O'Leary, Keith O'Neill, Stephen C J Parker, Andreas Polley, Christina K Raymond, Kathrin Reichwald, Joseph Rodriguez, Takashi Sasaki, Markus Schilhabel, Roman Siddiqui, Cherylyn L Smith, Tam P Sneddon, Jessica a Talamas, Pema Tenzin, Kerri Topham, Vijay Venkataraman, Gaiping Wen, Satoru Yamazaki, Sarah K Young, Qiandong Zeng, Andrew R Zimmer, Andre Rosenthal, Bruce W Birren, Matthias Platzer, Nobuyoshi Shimizu, and Eric S Lander. DNA sequence and analysis of human chromosome 8. *Nature*, 439(7074):331–5, jan 2006.

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

[12] M. C. Schatz, A. L. Delcher, and S. L. Salzberg. Assembly of large genomes using second-generation sequencing. *Genome Research*, 20(9):1165–1173, sep 2010.

[13] Karen E Hayden, Erin D Strome, Stephanie L Merrett, Hye-Ran Lee, M Katharine Rudd, and Huntington F

Willard. Sequences associated with centromere competency in the human genome. *Molecular and cellular biology*, 33(4):763–72, feb 2013.

[14] Can Alkan, Maria Francesca Cardone, Claudia Rita Catacchio, Francesca Antonacci, Stephen J. O'Brien, Oliver A. Ryder, Stefania Purgato, Monica Zoli, Giuliano Della Valle, Evan E. Eichler, and Mario Ventura. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. *Genome Research*, 21(1):137–145, 2011.

[15] Daniël P Melters, Keith R Bradnam, Hugh a Young, Natalie Telis, Michael R May, J Graham Ruby, Robert Sebra, Paul Peluso, John Eid, David Rank, José Fernando Garcia, Joseph L Derisi, Timothy Smith, Christian Tobias, Jeffrey Ross-Ibarra, Ian Korf, and Simon Wl Chan. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome biology*, 14(1):R10, jan 2013.

[16] 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, aug 2003.

[17] Can Alkan, Mario Ventura, Nicoletta Archidiacono, Mariano Rocchi, S Cenk Sahinalp, and Evan E Eichler. Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. *PLoS computational biology*, 3(9):1807–18, sep 2007.

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

[19] Karen H Miga, Yulia Newton, Miten Jain, Nicolas Altemose, Huntington F Willard, and W James Kent. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Research*, 24(4):697–707, apr 2014.

[20] Miten Jain, Hugh E. Olsen, Benedict Paten, Mark Akeson, D Branton, B Daniel, DW Deamer, M Andre, B Hagan, SA Benner, D Deamer, M Akeson, D Branton, JJ Kasianowicz, E Brandin, D Branton, DW Deamer, GM Cherf, KR Lieberman, R Hytham, CE Lam, K Kevin, A Mark, M Ayub, H Bayley, EA Manrao, IM Derrington, AH Laszlo, KW Langford, MK Hopper, G Nathaniel, PM Ashton, S Nair, T Dallman, S Rubino, W Rabsch, S Mwaigwisya, MT Bolisetty, G Rajadinakaran, BR Graveley, M Jain, IT Fiddes, KH Miga, HE Olsen, B Paten, M Akeson, J Quick, A Quinlan, N Loman, A Kilianski, JL Haas, EJ Corriveau, AT Liem, KL Willis, DR Kadavy, I Sović, M Šikić, A Wilm, SN Fenlon, S Chen, N Nagarajan, S Goodwin, J Gurtowski, S Ethe-Sayers, P Deshpande, MC Schatz, WR McCombie, AL Greninger, SN Naccache, S Federman, G Yu, P Mbala, V Bres, AD Hargreaves, JF Mulley, MD Cao, D Ganesamoorthy, A Elliott, H Zhang, MA Cooper, LJM Coin, K Judge, SR Harris, S Reuter, J Parkhill, SJ Peacock, E Karlsson, A Lärkeryd, A Sjödin, M Forsman, P Stenberg, RM Leggett, D Heavens, M Caccamo, MD Clark, RP Davey, NJ Loman, MJ Pallen, NJ Loman, AR Quinlan, M-A Madoui, S Engelen, C Cruaud, C Belser, L Bertrand, A Alberti, AS Mikheyev, MMY Tin, G Miles, J Hoisington-Lopez, E Duncavage, RR Miller, V Montoya, JL Gardy, DM Patrick, P Tang, MJ Pallen, J Quick, P Ashton, S Calus, C Chatt, S Gossain, J Hawker, J Quick, NJ Loman, S Duraffour, JT Simpson, E Severi, L Cowley, J Quick, AR Quinlan, NJ Loman, AC Ramgren, HS Newhall, KE James, J Risse, M Thomson, S Patrick, G Blakely, G Koutsovoulos, M Blaxter, J Wang, NE Moore, Y-M Deng, DA Eccles, RJ Hall, AC Ward, W Kim, M Watson, M Thomson, J Risse, R Talbot, J Santoyo-Lopez, K Gharbi, S Wei, Z Williams, J Schreiber, ZL Wescoe, R Abu-Shumays, JT Vivian, B Baatar, K Karplus, ZL Wescoe, J Schreiber, M Akeson, M Loose, S Malla, M Stout, CLC Ip, M Loose, JR Tyson, M Cesare, BL Brown, M Jain, AL Norris, RE Workman, Y Fan, JR Eshleman, W Timp, MC Frith, M Hamada, P Horton, NJ Loman, J Quick, JT Simpson, J Eid, A Fehr, J Gray, K Luong, J Lyle, G Otto, C Lee, C Grasso, MF Sharlow, T Szalay, JA Golovchenko, C Li, KR Chng, JHE Boey, HQA Ng, A Wilm, N Nagarajan, T Hoenen, A Groseth, K Rosenke, RJ Fischer, A Hoenen, SD Judson, S Zaaijer, S Chen, S Li, W Xie, X Li, C Zhang, H Jiang, AM Smith, R Abu-Shumays, M Akeson, DL Bernick, RY Henley, BA Ashcroft, I Farrell, BS Cooperman, SM Lindsay, M Wanunu, RP Horgan, LC Kenny, J Nivala, DB Marks, M Akeson, LE Hood, GS Omenn, RL Moritz, R Aebersold, KR Yamamoto, M Amos, Y-T Chen, C Iseli, CA Venditti, LJ Old, AJG Simpson, CV Jongeneel, K Berlin, S Koren, C-S Chin, JP Drake, JM Landolin, AM Phillippy, BD Ondov, TJ Treangen, P Melsted, AB Mallonee, NH Bergman, and S Koren. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1):239, 2016.

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

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

[23] Michael G Ross, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum, and David B Jaffe. Characterizing and measuring bias in sequence data. *Genome biology*, 14(5):R51, may 2013.

[24] Robert VanBuren, Doug Bryant, Patrick P. Edger, Haibao Tang, Diane Burgess, Dinakar Challabathula, Kristi Spittle, Richard Hall, Jenny Gu, Eric Lyons, Michael Freeling, Dorothea Bartels, Boudewijn Ten Hallers, Alex Hastie, Todd P. Michael, and Todd C. Mockler. Single-

molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature*, 527(7579):508–11, 2015.

[25] Shubha Vij, Heiner Kuhl, Inna S Kuznetsova, Aleksey Komissarov, Andrey A Yurchenko, Peter Van Heusden, Siddharth Singh, Natascha M Thevasagayam, Sai Rama Sridatta Prakki, Kathiresan Purushothaman, Jolly M Saju, Junhui Jiang, Stanley Kimbung Mbandi, Mario Jonas, Amy Hin Yan Tong, Sarah Mwangi, Doreen Lau, Si Yan Ngoh, Woei Chang Liew, Xueyan Shen, Lawrence S Hon, James P Drake, Matthew Boitano, Richard Hall, Chen-Shan Chin, Ramkumar Lachumanan, Jonas Korlach, Vladimir Trifonov, Marsel Kabilov, Alexey Tupikin, Darrell Green, Simon Moxon, Tyler Garvin, Fritz J Sedlazeck, Gregory W Vurture, Gopikrishna Gopalapillai, Vinaya Kumar Katneni, Tansyn H Noble, Vinod Scaria, Sridhar Sivasubbu, Dean R Jerry, Stephen J O'Brien, Michael C Schatz, Tamás Dalmay, Stephen W Turner, Si Lok, Alan Christoffels, and László Orbán. Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. *PLoS genetics*, 12(4):e1005954, apr 2016.

[26] Yinping Jiao, Paul Peluso, Jinghua Shi, Tiffany Liang, Michelle C Stitzer, Michael S Campbell, Joshua C Stein, Xuehong Wei, Chen-Shan Chin, Katherine Guill, Michael Regulski, Sunita Kumari, Andrew Olson, Jonathan Gent, Kevin L Schneider, Thomas K Wolfgruber, Michael R May, Nathan M Springer, Eric Antoniou, Richard Mccombie, Gernot G Presting, Michael Mcmullen, Kelly Dawe, Alex Hastie, David R Rank, and Doreen Ware. The complex sequence landscape of maize revealed by single molecule technologies. *bioRxiv*, pages 1–19, 2016.

[27] Thomas K. Wolfgruber, Megan M. Nakashima, Kevin L. Schneider, Anupma Sharma, Zidian Xie, Patrice S. Albert, Ronghui Xu, Paul Bilinski, R. Kelly Dawe, Jeffrey Ross-Ibarra, James A. Birchler, and Gernot G. Presting. High Quality Maize Centromere 10 Sequence Reveals Evidence of Frequent Recombination Events. *Frontiers in Plant Science*, 7(308):1–14, 2016.

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

Supplements

Figure S1: Centromere landscapes in the medaka genomes

Figure S2: Centromeric repeat distribution