Results

Genomic abundance of centromeric repeats

Melters *et al.* (2013) estimated that the medaka candidate centromeric satellite comprise 0.32% of the medaka genome. However this estimaeion 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.

PacBio subreads were first filtered with the criteria that read length >1 kb and base quality average over the all bases >10. The filtered subreads were then scanned by RepeatMasker 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). The genomic fraction in Hd-rR and HNI genomes were estimated to be ∼1%, while that in the HSOK genome was ∼2%. This difference is concordant with the previous observations that length of a centromeric array in a chromosome can vary up to 20-fold within a population [**?**]. Assuming the genome size to be 800 Mb, the centromeric satellite comprise 8–16 Mb of the genome, which implies each chromosome has 340–670 kb of centromeric sallite 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 reliable estimation than estimating by short Sanger sequencing reads.

Methods

centromeric repeat genomic abundance estimation

PacBio subreads were first filtered with the criteria that read length >1 kb and base quality average over the all bases >10. 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.