BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 17 2010, pages 2101–2108
doi:10.1093/bioinformatics/btq343

Sequence analysis

Advance Access publication July 8, 2010

Global sequence characterization of rice centromeric satellite based on oligomer frequency analysis in large-scale sequencing data

Jiří Macas[1,*], Pavel Neumann[1], Petr Novák[1] and Jiming Jiang[2]

[1]Institute of Plant Molecular Biology, Biology Centre ASCR, Branisovska 31, CZ-37005, Ceske Budejovice, Czech Republic and [2]Department of Horticulture, University of Wisconsin-Madison, WI 53706, USA

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Satellite DNA makes up significant portion of many eukaryotic genomes, yet it is relatively poorly characterized even in extensively sequenced species. This is, in part, due to methodological limitations of traditional methods of satellite repeat analysis, which are based on multiple alignments of monomer sequences. Therefore, we employed an alternative, alignment-free, approach utilizing k-mer frequency statistics, which is in principle more suitable for analyzing large sets of satellite repeat data, including sequence reads from next generation sequencing technologies.

Results: k-mer frequency spectra were determined for two sets of rice centromeric satellite CentO sequences, including 454 reads from ChIP-sequencing of CENH3-bound DNA (7.6 Mb) and the whole genome Sanger sequencing reads (5.8 Mb). k-mer frequencies were used to identify the most conserved sequence regions and to reconstruct consensus sequences of complete monomers. Reconstructed consensus sequences as well as the assessment of overall divergence of k-mer spectra revealed high similarity of the two datasets, suggesting that CentO sequences associated with functional centromeres (CENH3-bound) do not significantly differ from the total population of CentO, which includes both centromeric and pericentromeric repeat arrays. On the other hand, considerable differences were revealed when these methods were used for comparison of CentO populations between individual chromosomes of the rice genome assembly, demonstrating preferential sequence homogenization of the clusters within the same chromosome. k-mer frequencies were also successfully used to identify and characterize smRNAs derived from CentO repeats.

Contact: macas@umbr.cas.cz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 24, 2010; revised on June 8, 2010; accepted on June 23, 2010

1 INTRODUCTION

Highly abundant, tandemly arranged DNA repeats, referred to as satellite DNA, are among the major constituents of complex eukaryotic genomes. They are composed of basic repeated units, or monomers, which are usually only tens to hundreds of nucleotides long (Macas *et al.*, 2002). However, they can be amplified to millions of copies, making up to 20% of plant nuclear DNA (Ingham *et al.*, 1993) and even higher proportions in some insect or rodent genomes (Hacch and Mazrimas, 1974; Pons *et al.*, 1997). Up to megabase-sized arrays of head-to-tail arranged monomers constitute specific genomic regions appearing as heterochromatic bands or spots on mitotic chromosomes or interphase nuclei. Satellite repeats are among the most dynamic components of eukaryotic genomes, undergoing rapid changes in their sequences and abundance, which often result in emergence of genus-specific or even species-specific repeats (Macas *et al.*, 2000; Tek *et al.*, 2005). In contrast to this rapid inter-specific diversification, monomers of a given satellite are usually highly homogenized within a species as a result of their concerted evolution (Elder and Turner, 1995). This uniformity, together with the genomic organization of satellite repeats into long contiguous arrays represent serious obstacles to their investigation using conventional cloning and sequencing approaches (Lee *et al.*, 2006; Song *et al.*, 2001). Consequently, large proportions of satellite sequences are left out from genome assemblies even in extensively studied model species. As the number of genomes investigated using large-scale sequencing increases, there is a mounting amount of unexplored satellite sequence data available in the repositories of unassembled reads. This increase has recently been accelerated by the development of next generation sequencing technologies, which have made whole genome shotgun sequencing possible in a wide range of species. All these sequencing projects generate a wealth of data, which, if properly analyzed, could be used to investigate the patterns of intra- and inter-specific evolution of satellite repeats.

Sequence analysis of satellite DNA is traditionally based on a definition of basic repeated units (monomers or higher order repeats), which are subjected to pairwise or multiple sequence alignments to assess their similarities. While this approach is intuitive and in many respects convenient, it also includes several drawbacks: (i) it is prone to bias introduced by subjective decisions when defining individual repeat units within a longer sequence, especially in the case of less conserved repeats; (ii) processing large amounts of sequence data in this way is laborious; and (iii) multiple sequence alignments are feasible only with up to hundreds of monomers, as computational and visualization constraints are

*To whom correspondence should be addressed.

becoming limiting concomitant with the increasing number of analyzed sequences. These problems hinder efficient utilization of sequence data from large-scale sequencing of model genomes based on Sanger sequencing and become even more obvious when handling the sequence data generated by the next generation, high-throughput sequencing technologies including Roche/454 Life Sciences, Solexa/Illumina, ABI/SOLiD and Helicos systems. High-throughput sequencing is based on cloning-free, massively parallel processing of millions of individual templates, generating up to gigabases of sequence data in a single run (Shendure and Ji, 2008). The absence of the cloning step and production of large quantities of sequence reads make these techniques ideal for in-depth studies of satellite repeats in complex genomes. Indeed, it has been demonstrated that 454 sequencing facilitates highly efficient identification of novel families of satellite DNA in plants (Macas *et al.*, 2007). However, massively parallel sequencing data cannot be utilized for satellite sequence analysis using the traditional, monomer-based approaches, as the short length of the sequence reads hampers extraction of the full-length monomer sequences and the extremely large volume of the produced sequencing data prevents its analysis based on multiple sequence alignments.

Here, we explore an alternative way to globally characterize satellite repeats based on alignment-free sequence analysis using *k*-mer frequency statistics (Vinga and Almeida, 2003). The analysis starts with the decomposition of each sequence into a set of overlapping *k*-mers (oligomers of fixed length *k*, also called *L*-tuples or words) and determination of a spectrum of *k*-mer frequencies for the whole collection of sequences. This information is then used to identify the most conserved sequence fragments within the analyzed set, to infer the prevailing monomer length and to reconstruct its sequence. It is also employed to assess global similarities between various sequence sets. As the analysis is centered on short (10–17 nt) *k*-mers, it can be applied to analyzing and comparing sets of sequence fragments of various lengths, ranging from long Sanger reads down to the short sequences (~35 nt) generated by some of the next generation sequencing technologies.

In this work, we used the approach outlined above for global sequence characterization of the rice (*Oryza sativa*) satellite repeat CentO, which is one of the major components of the centromeric regions of this species' chromosomes. The prevalent monomer size of CentO repeats is 155 bp and the satellite was estimated to make up ~1.6% of the rice genome (Cheng *et al.*, 2002). The analysis was performed on several datasets acquired by different large-scale sequencing approaches. Data from the 454 sequencing of DNA fragments obtained by chromatin immunoprecipitation (ChIP-seq) with the centromeric histone CENH3 antibody (Yan *et al.*, 2008) were used to investigate the repeats associated with functional centromeres. The other analyzed datasets included CentO repeats extracted from whole genome shotgun Sanger sequencing reads and from CentO arrays on individual chromosomes of the rice genome assembly. We also demonstrated how *k*-mer spectra can be used for the identification of small RNAs (smRNAs) derived from CentO sequences, because this class of transcripts originating from centromeric repeats is supposed to play an important role in centromere formation and function (Alshire and Karpen, 2008; Carone *et al.,* 2009). This analyzes provided, for the first time, a global sequence characterization of various genomic populations of CentO repeats.

2 METHODS

All sequencing data used in this study were from japonica rice (*O.sativa* ssp. *japonica* var. Nipponbare). The 454 reads of rice sequences associated with centromeric histone H3 were generated previously using chromatin immunoprecipitation with CENH3 antibody followed by 454 sequencing using GS-20 instrument (Yan *et al.*, 2008). Whole genome shotgun Sanger sequencing reads (Goff *et al.*, 2002) were retrieved from the NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/) using a query selecting for 'shotgun' trace type and 'SBI' (Syngenta) sequencing center. It should be noted that contrary to the 454 sequences these reads were obtained by sequencing cloned fragments. Although in the case of CentO repeats we did not observe any discrepancies, caution should be taken in other studies when comparing these types of sequencing data as the cloning step may introduce some distortions in sequence representation. The assembled genome pseudomolecules (version 5) corresponding to individual rice chromosomes were downloaded from the Rice Genome Annotation project (http://rice.plantbiology.msu.edu/pseudomolecules/). Centromeric regions of the assembled pseudomolecules were screened using the dot-plot analysis program Gepard (Krumsiek *et al.*, 2007) to identify blocks of CentO repeats. These blocks were retrieved and kept separate for further analysis (see Supplementary Material S1 for coordinates of CentO blocks within assembled chromosomes). In addition, all identified CentO sequences (898 684 bp) were put in the same orientation and used to build a BLAST (Altschul *et al.*, 1997) database representing the genomic diversity of CentO. This database was then employed for similarity-based identification of CentO sequences within the sets of CENH3-ChIPed 454 and genomic shotgun reads. The reads were compared to the database using the blastn program (*e*-value cutoff 0.01, low-complexity filtering off) and the reads were considered to contain CentO when producing hits with bit score of at least 60. This bit score provided a relatively relaxed cutoff, corresponding for example to an 89% identity over a 54 bp region, which should ensure efficient capturing of all variants of CentO sequences.

The sequence analyzes were performed using custom-made BioPerl (http://www.bioperl.org) and R (R Development Core Team, 2009; http://www.R-project.org) scripts executed on a dual-processor server with 16 GB RAM running under the Debian Linux operating system (the scripts are available from the corresponding author). *k*-mer frequencies in the analyzed sequence sets were calculated by moving a sliding window of the length *k* in one nucleotide steps over each sequence within the set, extracting corresponding *k*-mer sequences and storing their counts in a hash table. The frequencies were then calculated by dividing the counts by a total number of extracted *k*-mers, and F_{100} values were obtained by multiplying the frequencies by 15 500. This value corresponds to the length of 100 CentO monomers, considering that the monomer length with the greatest frequency in rice genome is 155 bp (Lee *et al.*, 2006; Ma and Jackson, 2006). Quantification of differences in *k*-mer frequencies between two datasets were performed by calculating their Euclidean distance (Vinga and Almeida, 2003).

Reconstruction of the most conserved CentO regions was performed using a list of 17-mers sorted according to their frequencies, starting with the most frequent *k*-mer and using a threshold of 10% of the starting *k*-mer frequency for its extension (see also Fig. 4). The *k*-mers already used in the reconstruction were removed from the list and the next most frequent *k*-mer was used as a seed for a new fragment reconstruction, until no *k*-mers with $F_{100} \geq 2$ were available. Sequence logos (Schneider and Stephens, 1990) of the reconstructed fragments were generated using frequency matrices where each position was assigned a value of the most frequent *k*-mer overlapping this position. To build a consensus logo, the reconstructed fragments were aligned using SeaView (Galtier *et al.*, 1996) and their frequency matrices were merged according to their positions on

the alignment (Supplementary Material S2). Reconstruction of chromosome-specific CentO fragments was performed based on differences between *k*-mer frequencies calculated for CentO sequences from a single chromosome versus all the remaining chromosomes pooled together.

3 RESULTS

3.1 Preparing the datasets of CentO sequences

Genomic sequences located within rice centromeres were isolated based on their association with the centromere-specific histone variant CENH3 (Yan *et al.*, 2008). DNA fragments recovered from the ChIP assay employing CENH3-specific antibody were subjected to 454 sequencing, which yielded 325 298 reads with an average length of 106 nt. Using a sequence similarity search against a comprehensive compilation of known CentO sequences revealed that 115 036 (35.4%) of these reads were derived from CentO repeats. Compared to the estimated proportion of CentO in the rice genome (1.6%, Cheng *et al.*, 2002), the immunoprecipitated sample was enriched for CentO by about 22-fold, thus demonstrating the efficiency of the assay.

As the 454 reads ranged from 46 to 253 nt, they were trimmed to the same length prior to analysis in order to ensure equal contribution of each read to the global sequence statistics. This precaution should avoid a bias towards certain sequences because the read length produced during a given number of 454 sequencing cycles depends on template complexity (Rahmann, 2006). Thus, only the first 90 nucleotides were used from reads of this or longer length. Using this cutoff still allowed use of 95.4% of the reads, while the remaining 4.6% of the reads, which were shorter than 90 nt, were discarded. In addition, low-quality (2.6%) and duplicated reads representing sequencing artifacts (19.5%) were also removed, resulting in a final set of 84 271 trimmed reads, corresponding to a total of 7 584 390 nucleotides that were used for further analysis.

In addition to the CentO sequences obtained by CENH3-immunoprecipitation (further referred to as the 'ChIP' sample), we also investigated CentO repeats identified in the collection of unassembled whole genome shotgun reads produced by conventional (Sanger) sequencing in frame of the rice sequencing project (Goff *et al.*, 2002). These sequences (further referred to as the 'WGS' sample) were used for comparison with the ChIPed sample as they should represent a total genomic pool of CentO sequences (both centromeric and peri-centromeric), contrary to the ChIP sample, which should mostly include centromere-associated repeats. To get a dataset comparable to the ChIP data, WGS sequences were prepared for analysis by randomly extracting a single 90 bp segment from each shotgun read and selecting the fragments containing CentO repeats using the same procedure as for the ChIP sample. Out of 5 438 615 shotgun read segments there were 65 150 (1.2%) sequences that passed the selection criteria, providing 5 836 500 nt of sequence data.

3.2 Calculating and comparing *k*-mer spectra

As the first step in sequence analysis, the *k*-mer composition of each read was determined, providing a list of all *k*-mers occurring in the analyzed dataset (*k*-mer spectrum of the dataset). Summary counts of *k*-mers were calculated for the whole dataset and abundance of individual *k*-mer sequences was expressed as their frequency in the analyzed sequences (number of occurrences/total number of counted *k*-mers). These calculations were performed

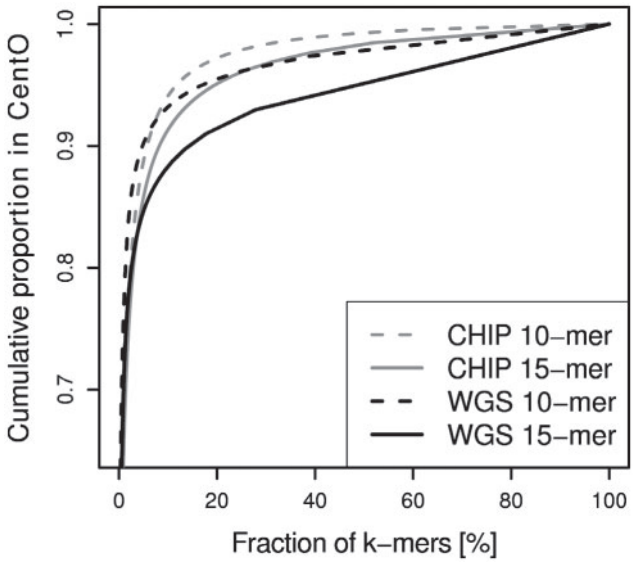

Fig. 1. Cumulative plots of *k*-mer proportions in ChIP and WGS sets. *k*-mers were sorted in descending order of their frequencies.

for *k*-mers with $k = 10, 13, 15$ and 17. The observed frequencies spanned over four orders of magnitude for each k in both ChIP and WGS datasets. Unique *k*-mer frequencies were quite low, on the order of 10^{-7}. Therefore, we expressed the *k*-mer abundance as their frequency multiplied by 15 500 to obtain values providing more comprehensible information about *k*-mer conservation within CentO repeated units. The value 15 500 corresponds to 100 CentO monomers of length 155 bp, which is their predominant size in the rice genome (Lee *et al.*, 2006; Ma and Jackson, 2006). Thus, this value, further referred to as F_{100}, could be used as an approximate indicator of the proportion of CentO monomers in which the *k*-mer occurs. For example, the frequency of the most abundant ChIP 10-mer 'ATGTCCAAAA' was 0.0051, which corresponds to a F_{100} of 79.2 and indicates that this sequence is present in ~79% of CentO monomers. In spite of the large number of different *k*-mer sequences encountered within CentO reads, there was only a small fraction of the most frequent *k*-mers, which accounted for majority of CentO sequences. For instance, 5.7% and 4.8% of the most frequent 10-mers represented 90% of all CentO sequences in the ChIP and WGS samples, respectively (Fig. 1).

A graphical comparison of *k*-mer frequencies between ChIP and WGS revealed high similarity between these two sets, especially with regards to the most conserved *k*-mers (Fig. 2). Frequencies of most *k*-mers with $F_{100} > 10$ differed less than 2-fold between the sets, and frequency values differing more than 3-fold were found only for less frequent *k*-mers ($F_{100} < 1$). The exception was a subset of *k*-mers that were about 10-fold more frequent in ChIP than in WGS (F_{100} of 1.0–10 in ChIP versus 0.1–1.0 in WGS, Fig. 2). However, as these *k*-mers mostly included homopolymer sequences (usually tracts of at least four 'T's or 'A's), the likely explanation for this observation is that their increased frequency in the ChIP fraction was a consequence of sequencing errors known to be generated by 454 sequencing of the homopolymer templates (Huse *et al.*, 2007).

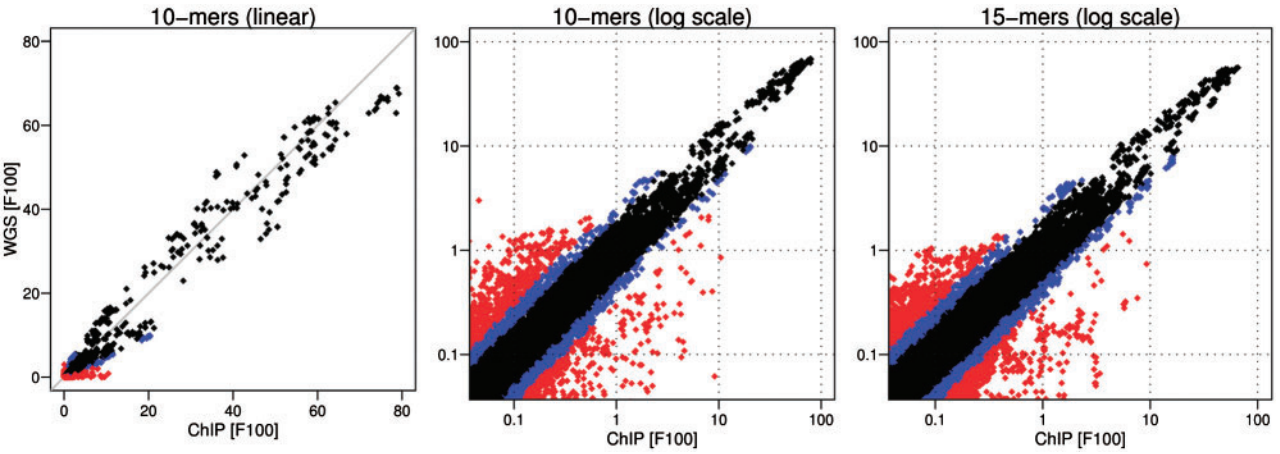

Fig. 2. Comparison of *k*-mers frequencies between ChIP and WGS sets. Each *k*-mer is represented by a dot and colored according to ratio of frequencies observed in the two sets (black, <2-fold; blue, 2-3-fold; red, >3-fold difference).

A global measure of dissimilarity between two sequence sets can be obtained by evaluating the differences between frequencies at which individual *k*-mers occur in the two sets. Based on this principle, we compared ChIP and WGS sets by calculating their Euclidean distance (d_E). We also extended our comparison to the set of CentO repeats extracted from version 5 of the japonica genome assembly (marked as 'JAT') and then to the sub-populations present on different chromosomes (pseudomolecules) of the assembly. These sub-populations were represented by individual clusters (contiguous arrays of CentO repeats separated by other sequences) identified in the (peri-)centromeric regions of the assembled chromosomes. The Euclidean distance values shown in Figure 3 are based on 15-mer frequencies; however, we obtained principally identical results when employing other *k*-mer lengths. This analysis confirmed the similarity of ChIP and WGS sets ($d_E = 104$), suggesting there are no significant differences between centromere-associated and the whole genome population of CentO sequences. Comparison of the ChIP and WGS sets to the sequences extracted from the genome assembly (JAT) resulted in slightly higher values ($d_E = 144$ and 126, respectively), most likely reflecting the fact that the genome assembly does not completely cover centromeric regions of most chromosomes, and thus a substantial part of CentO repeats is missing from the assembly. However, much higher dissimilarities were observed between CentO sets from individual chromosomes, which differed to various extents. For example, the highest d_E values were obtained when comparing chromosomes 6, 3 and 8 to others but not in their mutual comparisons, revealing that these three chromosomes share similar CentO sequences (Fig. 3). As CentO repeats in most centromeres are arranged into several clusters made up of contiguous arrays of the repeat monomers separated by non-CentO sequences, we also performed the comparison on this level in order to get more detailed information about CentO variability. This analysis revealed that the degree of dissimilarity is generally considerably lower between clusters from the same chromosome than between different chromosomes, suggesting intra-chromosomal homogenization of CentO repeats. This was especially obvious for chromosomes 6, 7 and 9.

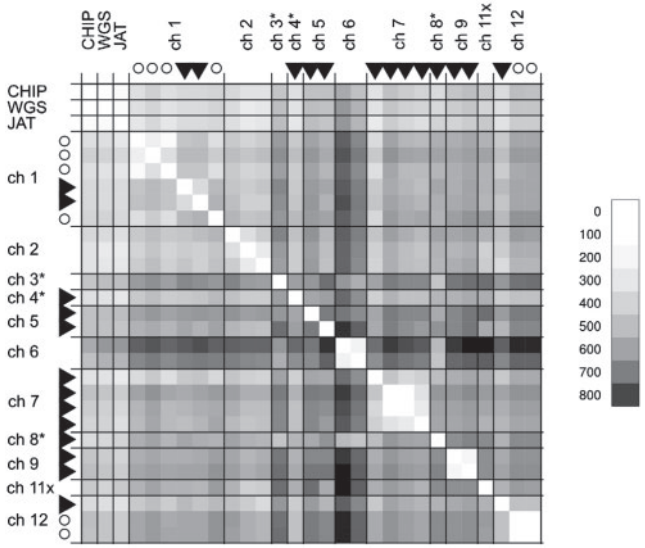

Fig. 3. Sequence divergence of various CentO sets measured by their Euclidean distance. Squares on the plot represent pairwise comparisons of the sets of shotgun reads (ChIP and WGS) and CentO sequences extracted from the rice genome assembly, which were either pooled for the whole genome (JAT) or examined separately for each chromosome (ch1–ch12). CentO clusters at least 10 kb in length were further distinguished within the chromosomes; if not present, all CentO sequences were pooled (chromosomes marked with *). Chromosome 10 is omitted due to very short CentO array and chromosome 11 is represented by the BAC AC146908. CentO clusters located within and outside CenH3-binding domains are marked with black triangles and open circles, respectively (according to Yan *et al.*, 2008). For size and position of CentO clusters on chromosomes see Supplementary Material S1. Plot shading corresponds to Euclidean distances (based on 15-mer frequencies) according to the scale, with similar sequences appearing lighter and divergent ones darker.

3.3 Sequence reconstruction based on *k*-mer frequencies

As the analyzed *k*-mers are, in principle, far shorter than the monomer length of most satellites (including CentO), we attempted

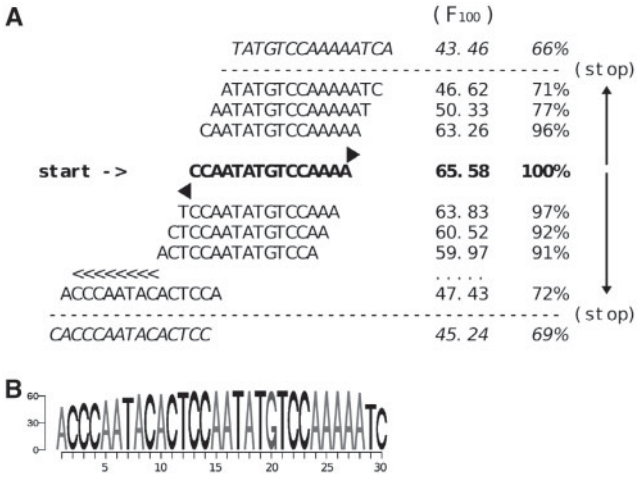

Fig. 4. Principle of sequence reconstruction. (**A**) The procedure starts with the most frequent *k*-mer (F_{100} = 65.58), which is repeatedly extended to both directions by finding the most frequent overlapping *k*-mers until their frequency drops below the specified threshold (set to 70% of the starting *k*-mer frequency in this example). (**B**) Sequence logo (Schneider and Stephens, 1990) is then constructed where height of the letters corresponds to *k*-mer frequencies.

to reconstruct longer sequence regions corresponding to the most conserved variants of the analyzed sequences. The principle of the reconstruction procedure is depicted in Figure 4. It employs a greedy algorithm starting with the most frequent *k*-mer, which is extended into both directions by repeatedly searching for one base extensions by finding *k*-mers overlapping with *k*-1 nucleotides. Only the *k*-mer with the highest frequency is used to determine the one base extension, and it is then, in turn, used as a query in the next round of extension until a specified threshold is reached. Figure 5A shows an example of reconstructed fragments from ChIP *k*-mers corresponding to the first 55 nt of the CentO monomer. These fragments represent the most abundant variants of the CentO repeats, which can be merged together to produce a sequence logo (Schneider and Stephens, 1990), giving an overview of the degree of conservation and composition of the major sequence variants (Fig. 5B and Supplementary Material S2). Although the prevailing CentO monomer size is 155 bp, the total length of the reconstructed monomer was 165 bp due to the occurrence of the longer CentO subfamily characterized by a 10 bp insertion. This insertion is located in positions 140–149 of the consensus, and the small height of the sequence logo at this region reflects the smaller abundance of this sequence variant in the genome. Performing the same reconstruction procedure for WGS *k*-mers and comparison of the reconstructed logo to the ChIP one confirmed the low divergence of these two samples as the two reconstructed sequences were almost identical, considering the prevalent bases at each position. However, there were minor differences in both the degree of conservation of various sequence regions and the presence of additional bases at certain positions (Fig. 5B).

The principle of sequence reconstruction depicted in Figure 4 can also be employed for identification of sequence variants enriched in one sample relative to the other, using differences between *k*-mer frequencies in the compared sets as the input for the reconstruction. In this case, the *k*-mers are assigned

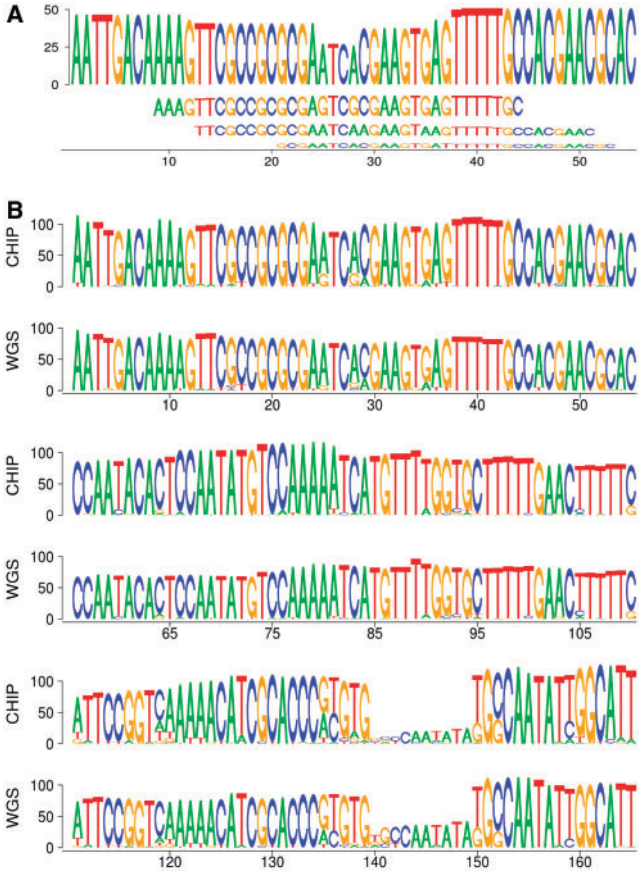

Fig. 5. (**A**) Examples of the reconstructed CentO sequence fragments corresponding to the first 55 nucleotides of the ChIP monomer. Consensus logo is constructed by merging these fragments while preserving information about proportions of individual sequence variants [the first line in (**B**)]. (B) Consensus sequences of ChIP and WGS monomers reconstructed from the most frequent 17-mers. Detailed lists and mutual positions of all reconstructed fragments used to build these logos are provided as Supplementary Material S2.

values calculated by subtraction of their observed frequencies in the two sets, and the reconstruction starts with those having the highest positive or negative values, depending on their enrichment in the first or the second set. This method was used for reconstruction of CentO fragments with unequal distribution among individual chromosomes, using data from the assembled rice chromosome pseudomolecules. Chromosome-specific fragments were identified for all chromosomes tested (Table 1), along with a number of other sequences that showed preferential localization to a subset of chromosomes (not shown). These data confirmed partial sequence diversification of CentO populations present on different chromosomes, supporting the hypothesis that CentO homogenization occurs primarily on intra-chromosomal scale.

3.4 Utilization of *k*-mer spectra for identification of smRNAs derived from CentO repeats

As the *k*-mer spectrum captures the whole range of sequence variability of the corresponding sequence set, it can be efficiently used for sequence similarity searches. The frequency associated with

Table 1. Examples of chromosome-specific CentO fragments reconstructed from *k*-mers differing in chromosome distribution

Target chrom.	Reconstructed fragments	Frequency in CentO arrays from individual chromosomes[a]									
		chr 01	chr 02	chr 03	chr 04	chr 05	chr 06	chr 07	chr 08	chr 09	chr 12
1	gggctcaatatatgccaat	9.5	0	0	0	0	0	0	0	0	0
2	tgtcggccaatattggcat	0	14.9	0	0	0	0	0	0	0	0
3	aaagttcgccgcgcatgccgcgaagtgagttt	0	0	20.3	0	0	0	0	0	0	0
4	cggtgaaaaacttcacacccacgtgtgcc	0	0	0	5.4	0	0	0	0	0	0
5	aaaaatcatgtttttggtgcttttc	0	0	0	0	10.0	0	0	0	0	0
6	cacgagtggcaataccggcattaattgac	0	0	0	0	0	15.0	0	0	0	0
7	tttttcacttcggtcgaaaacatcgc	0	0	0	0	0	0	12.8	0	0	0
8	attccggtatataacttcgcacccacgtg	0	0	0	0	0	0	0	11.6	0	0
9	tgccaattatgccaatattggcat	0	0	0	0	0	0	0	0	22.7	0
12	atcgcacccacctgtgccaatattgg	0	0	0	0	0	0	0	0	0	8.5

[a]Frequency of the reconstructed fragments was multiplied by 15 500 (analogous to the F_{100} frequency of *k*-mers).

each *k*-mer can then be used to determine whether the similarity hit corresponds to the sequence motif conserved in the query set. Using these principles, we searched for CentO-derived transcripts in the collection of 296 201 rice smRNA sequences published by Nobuta *et al.* (2007). This dataset is composed of 17-nt tags obtained by end-sequencing of smRNAs extracted from various tissues, and also includes information about the abundance of each sequence in the studied tissues. The smRNAs derived from CentO repeats were detected by their identity to the 17-mers present in the WGS set representing all genomic CentO sequences. There was a total of 213 CentO smRNAs identified (0.07% of the whole smRNA set), mostly corresponding to relatively rare transcripts with frequency of only 1–95 TPQ (Trancsripts Per Quarter million; Nobuta *et al.*, 2007). The abundance of individual smRNA sequences was not proportional to the frequency of corresponding *k*-mers in the genomic CentO sequences, suggesting that their precursors were not transcribed from the most conserved CentO repeats (Fig. 6A). Comparison of data from smRNA libraries constructed from six different tissues revealed considerable differences in both CentO smRNA diversity and abundance. Of 213 sequences, 194 (91%) were specific for a single library and there was none occurring in all libraries. The largest group was identified in immature panicles comprising 54 distinct smRNA sequences, while the smallest set was uncovered in ABA-treated seedlings containing 25 CentO smRNAs. The total abundance of all CentO smRNAs found in a single library ranged between 118 and 341 TPQ, being lowest in germinating seedlings and highest in immature panicles. Mapping CentO smRNA sequences (regardless of the library of origin) to the japonica rice assembly revealed their dispersed distribution along CentO clusters, having average density of 84 smRNA sites per 1 kb, which corresponds to about 13 sites per monomer. There was no CentO monomer found completely covered with smRNAs, as the longest stretch of CentO continuously overlaid by smRNA sequences was 138-bp long, potentially giving rise to 34 distinct smRNAs (data not shown). Interestingly, of 213 CentO smRNA sequences identified in this study, 137 (64%) originated from the bottom strand, while only 76 (36%) from the top strand (relative to the sequence orientation shown on Fig. 5B). The prevalence of the small RNAs originated from the bottom strand was most evident in immature panicles comprising five times more smRNA sequences

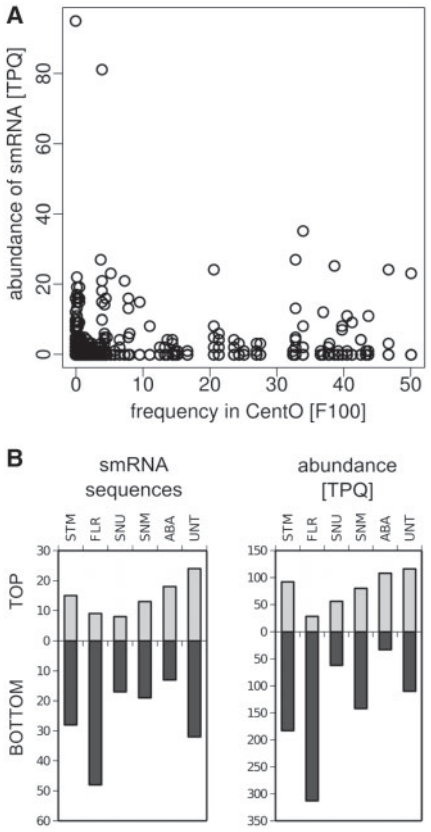

Fig. 6. Characteristics of CentO-derived smRNAs. (**A**) Comparison of the abunance of individual smRNA sequences with the frequency of corresponding *k*-mers in the genomic CentO sequences. (**B**) Abundance of CentO-derived smRNAs in different tissues expressed as the number of different smRNA sequences and as the total abundance of smRNA transcripts. The transcripts corresponding to top and bottom CentO strands are distinguished. The analysis was based on the smRNA databases of Nobuta *et al.* (2007): STM, stem; FLR, immature panicles; SNU, germinating seedlings; SNM, seedlings infected with *Magnaporthe grisea*; ABA, seedlings treated with ABA; UNT, untreated control to ABA.

derived from the bottom CentO strand than from the top strand, totally accounting for 11-fold higher proportion of CentO smRNA transcripts. In other libraries, however, this trend was less apparent or even reversed as in the case of ABA-treated seedlings (Fig. 6B).

4 DISCUSSION

Although the algorithms utilizing *k*-mer statistics has long been employed in some widely used programs including BLAST (Altschul *et al.*, 1997), they have recently drawn more attention due to their suitability for analyzing large amounts of sequence data. For example, *k*-mer frequencies derived from whole-genome Sanger and Solexa/Illumina reads provided an efficient means for the identification of repetitive sequences in genomic clones of maize and barley (Kurtz *et al*., 2008; Wicker *et al*., 2008), and evaluation of differences in *k*-mer composition was found superior to alignment-based methods in phylogenetic tree reconstruction including large datasets (Yang and Zhang, 2008). In this study, we demonstrated that *k*-mer frequency statistics can also be successfully adapted to the characterization and comparative analysis of large sets of satellite sequences. Our approach does not require *a priori* knowledge of repeat monomer length and is independent of the size of sequencing reads, provided it is equal or larger than the size of analyzed *k*-mers. It has been demonstrated that 17–20-mers provide sufficient specificity for repeat investigation in rice and barley genomes (Li *et al*., 2005; Liu *et al*., 2006; Wicker *et al*., 2008), and our analysis showed that due to relatively low sequence complexity of satellite repeats, the *k*-mer length can be decreased to 10 nt for this type of sequences. This is well below the 35 nt produced by the next generation sequencing technologies with the shortest read lengths. Thus, *k*-mer analysis provides an opportunity to explore the growing archives of unassembled sequencing reads produced by various technologies ranging from Sanger to Solexa/Illumina or SOLiD sequencing. This represents a significant advance in satellite repeat analysis, which up to now has been mostly based on multiple alignments of monomers extracted from assembled sequences, even in species where extensive shotgun sequencing data are available (Hall *et al*., 2003; Lee *et al.*, 2006). An exception was the approach developed for using whole-genome shotgun sequence data for global characterization of higher-order repeat structure of primate centromeric repeats (Alkan *et al*., 2007). However, this method required paired end Sanger sequencing data and relied on previous knowledge of higher order repeat sequences.

In addition to rich genomic resources available for the rice genome, the choice of the CentO satellite as a subject of this pilot study provided a possibility to compare our results to previous studies of this repeat using alignment-based analysis (Lee *et al.*, 2006; Ma and Jackson, 2006). This control proved the correct reconstruction of prevailing monomer size of 155 bp as well as the minor variant with 165 bp monomer (Fig. 5). To simplify the analysis, we took advantage of the availability of assembled CentO repeats for similarity-based selection of reads from shotgun sequencing data and for their arrangement in the same orientation. However, the same reconstruction procedure can also be used to reveal sequences of unknown satellites using whole genome reads. This is because the presence of high numbers of almost identical monomers in the genome results in a high frequency of *k*-mers from the corresponding satellite, which then usually appear on top of the list of the most frequent genomic *k*-mers. Consequently, the sequence reconstruction starting from the most frequent *k*-mers preferentially leads to assembly of fragments of satellite repeats. It should be noted that in this case all fragments will be reconstructed in duplicate, differing in orientation, due to the random orientation of sequencing reads. The forward and reverse duplicates of otherwise identical fragments can be merged, their overlaps identified and the monomer size determined by detecting 3′ and 5′ overlaps of the fragment assembly (Supplementary Material S2). These principles have been verified using 454 sequencing data from several species (our unpublished data) and Solexa/Illumina reads downloaded from the Arabidopsis 1001 genome project (Ossowski *et al*., 2008).

Comparison of *k*-mer spectra calculated from CentO sequences extracted from whole genome sequencing reads and from CENH3-associated DNA did not provide any evidence for selection of specific CentO variants in functional centromeres. This result can be explained by the fact that CentO repeats are predominantly located in the CENH3-associated chromatin in the rice genome. In 9 out of the 12 rice centromeres, the majority or entire CentO arrays were mapped in the CENH3-binding domains (Yan *et al.*, 2008). The CENH3-binding domains of three rice centromeres, Cen2, Cen6 and Cen11, were not determined (Yan *et al*., 2008). However, Cen2 and Cen6 contain ∼716 and 816 kb of CentO, respectively (Cheng *et al*., 2002). Thus, the majority of these repeats may also be associated with CENH3 in these three centromeres based on the 500–1000 kb average size of the CENH3-binding domains among the rice centromeres. Cen11 contains ∼2 Mb of CentO, representing the only centromere where large CentO arrays may be located outside of the CENH3-binding domain. However, it is likely that the centromeric and pericentromeric CentO arrays on Cen11 may be very similar in sequence, as observed for chromosomes 1 and 12 (Fig. 3).

On the other hand, sequence divergence was detected between CentO repeats from different chromosomes or their groups, revealing that CentO homogenization preferentially occurs on intra-chromosomal level and that spreading of sequence variants to other chromosomes is limited. A similar conclusion was drawn based on analysis of the CentO repeats associated with Cen1 and Cen8 using traditional sequence analysis methodologies (Lee *et al*., 2006). This is analogous to sequence evolution of centromeric satellites in human and *Arabidopsis* (Heslop-Harrison *et al*., 1999; Schindelhauer and Schwarz, 2002), as well as some non-centromeric repeats such as rDNA intergenic spacer-like satellites of *Vicia sativa* (Macas *et al*., 2003). However, this homogenization pattern is not universal as there exist satellite repeats in plants that are homogenized across the chromosomes and their subfamilies are then confined to specific chromosomal regions (Macas *et al*., 2006). The methods described here, allowing for easy detection of different sequence variants and for subsequent design of corresponding probes, should be instrumental in investigating this phenomenon further, especially with respect to using the next generation sequencing data recently generated from individual chromosomes (Mayer *et al.*, 2009). Yet another useful application is using *k*-mer spectra for detection of similarities in various distinct types of sequence data, as demonstrated by successful detection of rice smRNAs derived from CentO repeats. Compared to other similarity detection approaches, this method provides good sensitivity (the *k*-mer spectrum represents all sequence variants) as well as quantitative information about the *k*-mer frequency of the hits. The identification of CentO smRNAs is in agreement

with the study of Lee *et al.* (2006) who detected this type of CentO transcripts using the northern hybridization. The observed lack of correlation between the abundance of smRNAs and the corresponding CentO *k*-mers suggests that smRNA precursors are not transcribed from the most conserved CentO sequences. This could indicate that these transcripts represent a transcriptional noise driven by transcription of surrounding (non-CentO) sequences. On the other hand, we found multiple CentO regions (up to 138-bp long) identical to the detected smRNAs, which were mostly scattered within centromeric CentO arrays (data not shown). An interesting finding concerns the differences between various tissues and strands of smRNA transcription, which could point to tissue-specific transcription and/or processing of CentO transcripts by RNAi machinery. However, as the abundance of most CentO smRNAs was very low (42% of smRNAs had abundance of only 1 TPQ), these observations will have to be validated by deeper sequencing of small RNAs.

ACKNOWLEDGEMENTS

We thank S. M. Rafelski for critical reading of the manuscript.

Funding: Academy of Sciences of the Czech Republic (AVOZ50510513 to J.M.; KJB500960802 to P.N.); Ministry of Education, Youth and Sport of the Czech Republic (OC10037, LC06004 to J.M.); National Science Foundation (DBI-0603927 to J.J.).

Conflict of Interest: none declared.

REFERENCES

Alkan,C. *et al.* (2007) Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. *PLoS Comput. Biol.*, **3**, 1807–1818.

Allshire,R.C. and Karpen,G.H. (2008) Epigenetic regulation of centromeric chromatin: old dogs, new tricks? *Nat. Rev. Genet.*, **9**, 923–937.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Carone,D.M. *et al.* (2009) A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres. *Chromosoma*, **118**, 113–125.

Cheng,Z. *et al.* (2002) Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell*, **14**, 1691–1704.

Elder,J. and Turner,B. (1995) Concerted evolution of repetitive DNA-sequences in eukaryotes. *Q. Rev. Biol.,* **70**, 297–320.

Galtier,N. *et al.* (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.*, **12**, 543–548.

Goff,S.A. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.

Hacch,F. and Mazrimas,J. (1974) Fractionation and characterization of satellite DNAs of the kangaroo rat (*Dipodomys ordii*). *Nucleic Acids Res.*, **1**, 559–576.

Hall,S.E. *et al.* (2003) Centromere satellites from *Arabidopsis* populations: maintenance of conserved and variable domains. *Genome Res.*, **13**, 195–205.

Heslop-Harrison,J.S. *et al.* (1999) Polymorphisms and genomic organization of repetitive DNA from centromeric regions of *Arabidopsis* chromosomes. *Plant Cell*, **11**, 31–42.

Huse,S.M. *et al.* (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.*, **8**, R143.

Ingham,L.D. *et al.* (1993) Origin of the main class of repetitive DNA within selected *Pennisetum* species. *Mol. Gen. Genet.*, **238**, 350–356.

Krumsiek,J. *et al.* (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, **23**, 1026–1028.

Kurtz,S. *et al.* (2008) A new method to compute *K*-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics*, **9**, 517.

Lee,H. *et al.* (2006) Transcription and evolutionary dynamics of the centromeric satellite repeat CentO in rice. *Mol. Biol. Evol.*, **23**, 2505–2520.

Li,R.Q. *et al.* (2005) ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput. Biol.*, **1**, 313–321.

Liu,S. *et al.* (2006) Exact word matches in rice pseudomolecules. *Genome*, **49**, 1047–1051.

Ma,J. and Jackson,S.A. (2006) Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res.*, **16**, 251–259.

Macas,J. *et al.* (2000) Two new families of tandem repeats isolated from genus *Vicia* using genomic self-priming PCR. *Mol. Gen. Genet.*, **263**, 741–751.

Macas,J. *et al*. (2002) PlantSat: a specialized database for plant satellite repeats. *Bioinformatics*, **18**, 28–35.

Macas,J. *et al*. (2003) Sequence subfamilies of satellite repeats related to rDNA intergenic spacer are differentially amplified on *Vicia sativa* chromosomes. *Chromosoma*, **112**, 152–158.

Macas,J. *et al*. (2006) Sequence homogenization and chromosomal localization of VicTR-B satellites differ between closely related *Vicia* species. *Chromosoma*, **115**, 437–447.

Macas,J. *et al*. (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics*, **8**, 427.

Mayer,K.F.X. *et al.* (2009) Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol.*, **151**, 496–505.

Nobuta,K. *et al.* (2007) An expression atlas of rice mRNAs and small RNAs. *Nat. Biotechnol.*, **25**, 473–477.

Ossowski,S. *et al.* (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.*, **18**, 2024–2033.

Pons,J. *et al.* (1997) Conservation of satellite DNA in species of the genus *Pimelia* (Tenebrionidae, Coleoptera). *Gene*, **205**, 183–190.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing,Vienna, Austria.

Rahmann,S. (2006) Subsequence combinatorics and applications to microarray production, DNA sequencing and chaining algorithms. In Hutchison,D. *et al*. (eds) *Combinatorial Pattern Matching*. Springer, Berlin, pp.153–164.

Schindelhauer,D. and Schwarz,T. (2002) Evidence for a fast, intrachromosomal conversion mechanism from mapping of nucleotide variants within a homogeneous alpha-satellite DNA array. *Genome Res.*, **12**, 1815–1826.

Schneider,T.D. and Stephens,R. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

Song,J. *et al.* (2001) Instability of bacterial artificial chromosome (BAC) clones containing tandemly repeated DNA sequences. *Genome*, **44**, 463–469.

Tek,A.L. *et al.* (2005) Sobo, a recently amplified satellite repeat of potato, and its implications for the origin of tandemly repeated sequences. *Genetics*, **170**, 1231–1238.

Vinga,S. and Almeida,J. (2003) Alignment-free sequence comparison—a review. *Bioinformatics*, **19**, 513–523.

Wicker,T. *et al.* (2008) Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics*, **9**, 518.

Yan,H. *et al.* (2008) Intergenic locations of rice centromeric chromatin. *PLoS Biol.*, **6**, e286.

Yang,K. and Zhang,L. (2008) Performance comparison between *k*-tuple distance and four model-based distances in phylogenetic tree reconstruction. *Nucleic Acids Res.*, **36**, e33.