SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

Satellite DNA-like repeats are dispersed throughout the genome of the Pacific oyster *Crassostrea gigas* carried by *Helentron* non-autonomous mobile elements

Tanja Vojvoda Zeljko, Martina Pavlek, Nevenka Meštrović & Miroslav Plohl✉

Satellite DNAs (satDNAs) are long arrays of tandem repeats typically located in heterochromatin and span the centromeres of eukaryotic chromosomes. Despite the wealth of knowledge about satDNAs, little is known about a fraction of short, satDNA-like arrays dispersed throughout the genome. Our survey of the Pacific oyster *Crassostrea gigas* sequenced genome revealed genome assembly replete with satDNA-like tandem repeats. We focused on the most abundant arrays, grouped according to sequence similarity into 13 clusters, and explored their flanking sequences. Structural analysis showed that arrays of all 13 clusters represent central repeats of 11 non-autonomous elements named *Cg_HINE*, which are classified into the *Helentron* superfamily of DNA transposons. Each of the described elements is formed by a unique combination of flanking sequences and satDNA-like central repeats, coming from one, exceptionally two clusters in a consecutive order. While some of the detected *Cg_HINE* elements are related according to sequence similarities in flanking and repetitive modules, others evidently arose in independent events. In addition, some of the *Cg_HINE*'s central repeats are related to the classical *C. gigas* satDNA, interconnecting mobile elements and satDNAs. Genome-wide distribution of *Cg_HINE* implies non-autonomous *Helentrons* as a dynamic system prone to efficiently propagate tandem repeats in the *C. gigas* genome.

Satellite DNAs (satDNAs) and transposable elements (TEs) are two types of repetitive sequences that together represent the largest fraction of eukaryotic genomes[1]. SatDNAs are head-to-tail tandemly repeated non-coding DNA sequences primarily organized in long arrays associated with heterochromatin. Nevertheless, satDNA sequences can also be found dispersed in the euchromatic genome fraction as short arrays, single monomers or their fragments (reviewed in[2–4]). While satDNAs appear relatively static with respect to their localization on chromosomes, TEs are sequences able to move throughout the genome, ultimately forming interspersed repeats. TEs spread by a variety of mechanisms, RNA-mediated (Class I elements), and DNA-mediated (Class II elements), either autonomously or dependent on enzymes produced by the autonomous elements[5,6].

Regardless of the conceptual differences between satDNAs and TEs, numerous studies show that they can be interconnected in many different ways[7]. Thus, satDNA can be formed by tandem amplification of an entire TE or its part[8–11]. TEs themselves may have an internal region composed of sequences repeated in tandem. One example is *Tetris*, described in *Drosophila* as modularly structured non-autonomous foldback DNA transposon. It incorporates tandem repeats (TRs) that can act as building blocks in the formation of classical satDNA arrays[12].

Among TEs that may contain TRs in their structure one group stands out, *Helitrons*, a diverse superfamily of DNA transposons widespread in animals and plants[13]. These elements use rolling-circle replication (RCR) to spread through the genome, and they do not create target site duplications (TSD) upon insertion[14]. Because of the RCR mechanism, they are prone to capture and propagate diverse genomic sequences, including genes, contributing significantly to genome evolution[15–17]. Whole elements can also be repeated in tandem, in which case the inner copies are often truncated at the 3′ end[18,19]. One structural variant of the *Helitron* superfamily is *Helentron*, in its non-autonomous form known as *HINE* (*Helentron*-associated INterspersed Elements). *HINEs*

Division of Molecular Biology, Ruđer Bošković Institute, Bijenička 54, 10 000 Zagreb, Croatia. ✉email: plohl@irb.hr

are characterized by two modules which include subterminal inverted repeats and a short palindromic sequence at the 3′ end of the right module. A short array of satDNA-like TRs can be often embedded between these two modules[20,21].

Bivalve mollusks constitute a large class of marine and freshwater invertebrates carrying high ecological and commercial value[22]. The genome of the Pacific oyster *Crassostrea gigas* Thunberg, 1793 was the first sequenced and assembled bivalve genome. Because of high individual polymorphism and abundant repetitive sequences, estimated to build 36% of the genome, fosmid pooling, next-generation sequencing, and hierarchical assembly were combined in this work[23]. *C. gigas* genome is replete with transposase and reverse-transcriptase gene fragments and their transcripts, indicating importance of transposition processes in shaping the genome[23,24]. In a recent analysis of the black-shelled Pacific oyster (a variety of the Pacific oyster *C. gigas*), long and short reads sequencing was combined and even a higher content (48%) of repetitive sequences was revealed than in the previous study[25]. It was concluded that both strains are highly variable and divergent in content of their repetitive sequences.

C. gigas has a low level of heterochromatin observed as a weak centromeric C-band on one chromosome pair, and a telomeric C-band on the other[26]. This observation is in agreement with the low abundance of classical satDNAs, as predicted by Zhang et al.[23]. The most abundant satDNA in *C. gigas* is Cg170, with ~ 166 bp long monomers, occupying 1–4% of the genome, and detected in centromeric regions of only some chromosome pairs[27,28]. A similar satDNA was identified independently in seven oyster species belonging to the genera *Ostrea* and *Crassostrea*, and named HindIII satDNA denoting the restriction site by which it was detected[29]. Although described independently, both satDNAs can be considered as subfamilies of one divergent family that can be unified under the name Cg170/HindIII[30]. A preliminary study on a limited sample of Cg170 satDNA arrays extracted from the genome assembly indicated their association with members of the *Helitron superfamily*[31]. Furthermore, Cg170/HindIII monomers are similar to the central repeats of the miniature inverted-repeat element (MITE) *Pearl*, which is widespread in bivalves[32], and according to its structural characteristics was later re-categorized as a non-autonomous *Helentron*[21].

A detailed view of the genomic inventory of satDNAs started to accumulate in different animal and plant species by combining advanced sequencing methods and specialized bioinformatics tools (for example[3,33–37]). However, information about content, distribution, and composition of short arrays of TRs located in euchromatic genome compartments and related or resembling satDNAs (therefore named satDNA-like sequences), and about genome environment in which they reside, remains limited and shown on a few species, mostly *Drosophila* and beetles[34,38–42].

In this work, the genome assembly of *C. gigas*[23] was searched for all TRs that resemble satDNAs according to criteria of monomer length. This strategy revealed short arrays of TRs dispersed throughout the genome. We focused on the 13 similarity-grouped clusters of most abundant arrays (49% of all detected), studied their adjacent genomic sequences, and found that they altogether can be characterized as non-autonomous elements of the *Helentron* superfamily (*HINE*). According to our best knowledge, by starting from a general inventory of satDNA-like TRs in a putative euchromatic (assembled) genome fraction, we show for the first time that divergent satDNA-like arrays of one species are all linked with *HINE* TEs as their carriers. In total, we identified a family of 11 elements, determined by flanking sequences associated with arrays assigned only to one or, exceptionally, two clusters of satDNA-like TRs.

Results

Detection and grouping of tandem repeats in *C. gigas*. The strategy used to detect and characterize satDNA-like TRs and their flanking sequences in the *C. gigas* genome is shown in Fig. 1.

Screening of the sequenced genome (oyster.v9.fa) and filtering out TRs composed of at least two monomers of the length between 100 and 500 bp revealed 14,591 arrays. Comparison of the number of arrays with the number of monomers in the array shows that the most abundant are those with 2 or 3 monomers (8,224 out of 14,591, or 56.36%; Fig. 2a). The number of detected arrays drops dramatically with increasing number of monomers, and only 2,282 (15.64%) arrays hold ≥ 5 monomers. In total, the set of 14,591 arrays is composed of 51,024 monomers; among them, those with length between 160 and 180 bp are dominant, constituting 51.41% of all monomers (Fig. 2b).

The selected arrays were grouped using 70% sequence similarity as the threshold value (Fig. 1). Grouping resulted in 393 clusters composed of 9,902 arrays, while the remaining 4,689 arrays were too divergent to allow clustering under these conditions. For further analysis we selected the 13 largest clusters, CL1-13, encompassing 7,151 arrays and representing the majority (72%) of clustered arrays (Supplementary Table S1). Altogether, arrays of TRs in clusters CL1-13 constitute ~ 4.8 Mb or 0.85% of the sequenced *C. gigas* genome (Table 1).

From clusters CL1-13 we extracted a total of 32,582 monomers. Monomer consensus sequences were derived by multiple sequence alignment of all repeats in each cluster (with exceptions for the most abundant CL1 and CL2, where > 70% repeats were used; Table 1). Alignment of the monomer consensus sequences is presented in Fig. 3a.

The monomer length in the studied clusters is predominantly 140–180 bp (Table 1), in agreement with the preferred length detected in the analysis of the initial set of 51,024 monomers. Exceptions are the monomers in clusters CL6 and CL12, with an average length of 146 and 138 bp, respectively. The average monomer copy number in arrays of clusters CL1-13 is up to 5, slightly higher than that shown for the whole set. In particular, arrays with ≥ 5 monomers are more abundant in clusters CL1, CL2, CL3, CL4, CL6, and CL10 (Table 1). The longest array in the whole set, with 29.3 monomers, was found in the cluster CL2.

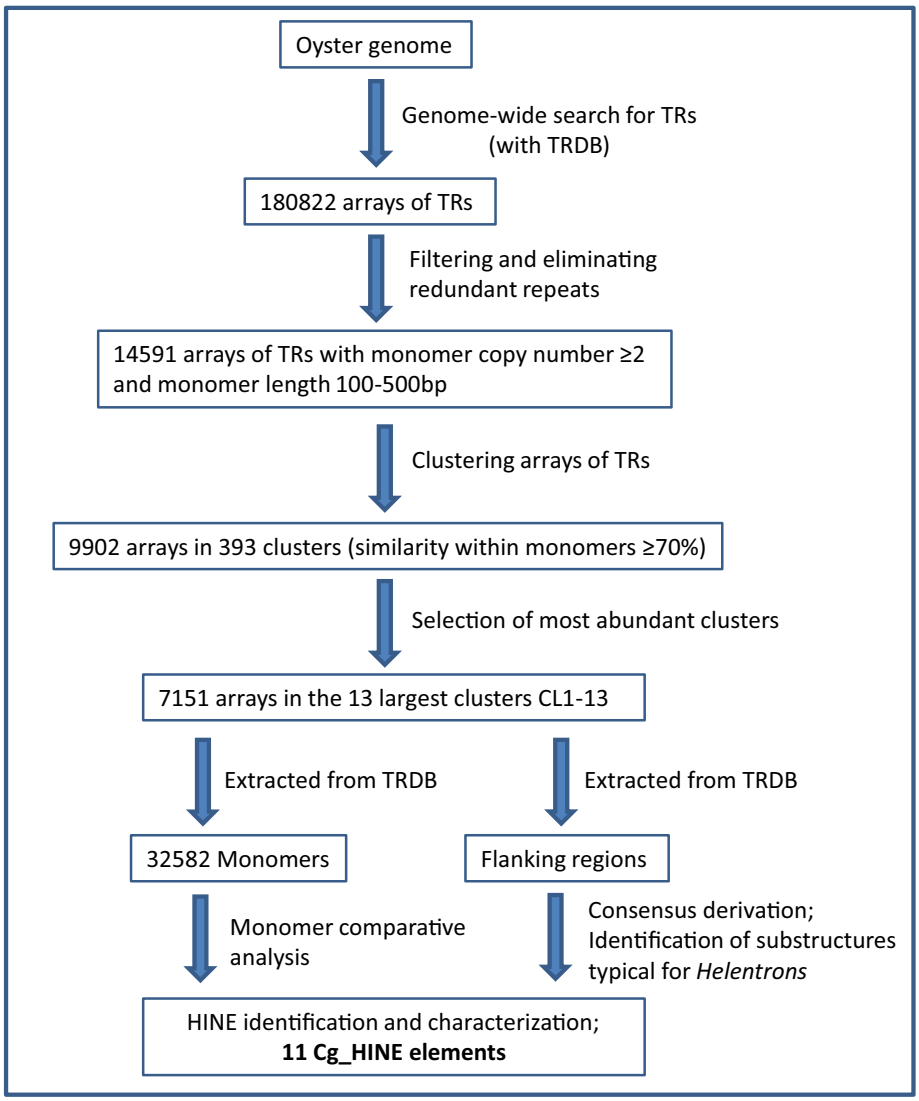

Figure 1. Workflow of the genome-wide identification of tandem repeats and their flanking sequences in the assembled genome of the Pacific oyster *Crassostrea gigas*.

Comparisons of tandem repeats from clusters CL1-13. Nucleotide sequence diversity among monomers within a cluster ranges from 8.5% in CL9 to 21% in CL4, and differences among variants are mostly due to nucleotide substitutions. Sequence comparisons revealed that only monomers from clusters CL1, CL2, CL10, and CL13 share relevant similarity along the whole length (Group 1 in Fig. 3a). Particularly, CL1 and CL2 monomers are 89.8% identical in their consensus sequences, while they share ~ 70% identity with those from clusters CL10 and CL13 (Supplementary Table S3). Phylogenetic analysis confirmed, even in the case of the highly similar CL1 and CL2, grouping of monomers into four clearly distinctive clusters and homogeneity of arrays (Supplementary Fig. S1). Similarity is also shared between consensus sequences of monomers from clusters CL1, CL2, CL10, and CL13 and consensus sequences of Cg170[27] and HindIII satDNAs[29] (Supplementary Table S3). In this regard, monomers from these clusters can be considered as novel subfamilies of the Cg170/HindIII satDNA family. In addition, CL1, CL2, CL10, and CL13 share two relatively conserved sequence segments with monomers from the cluster CL4 (boxed segments in Group 1, Fig. 3a), although the rest of their sequence is dissimilar (see also Supplementary Table S3).

Our survey also revealed that the CL4 monomer consensus sequence is 97.6% similar to that of SAT-2_CGi, the *C. gigas* DNA sequence annotated in Repbase as a satDNA. However, according to our knowledge, this satDNA was not further characterized. A more detailed insight into the SAT-2_CGi Repbase entry (668 bp) revealed 4 tandemly repeated monomers, about 168 bp long, sharing 95.4% of mutual similarity[43]. Monomer consensus sequences from other clusters did not reveal similarities with any known satDNA.

Tandem repeats from clusters CL1-13 are parts of *Cg_HINE* mobile elements. To characterize the genomic environment in which the detected satDNA-like arrays reside, we analyzed their flanking regions.

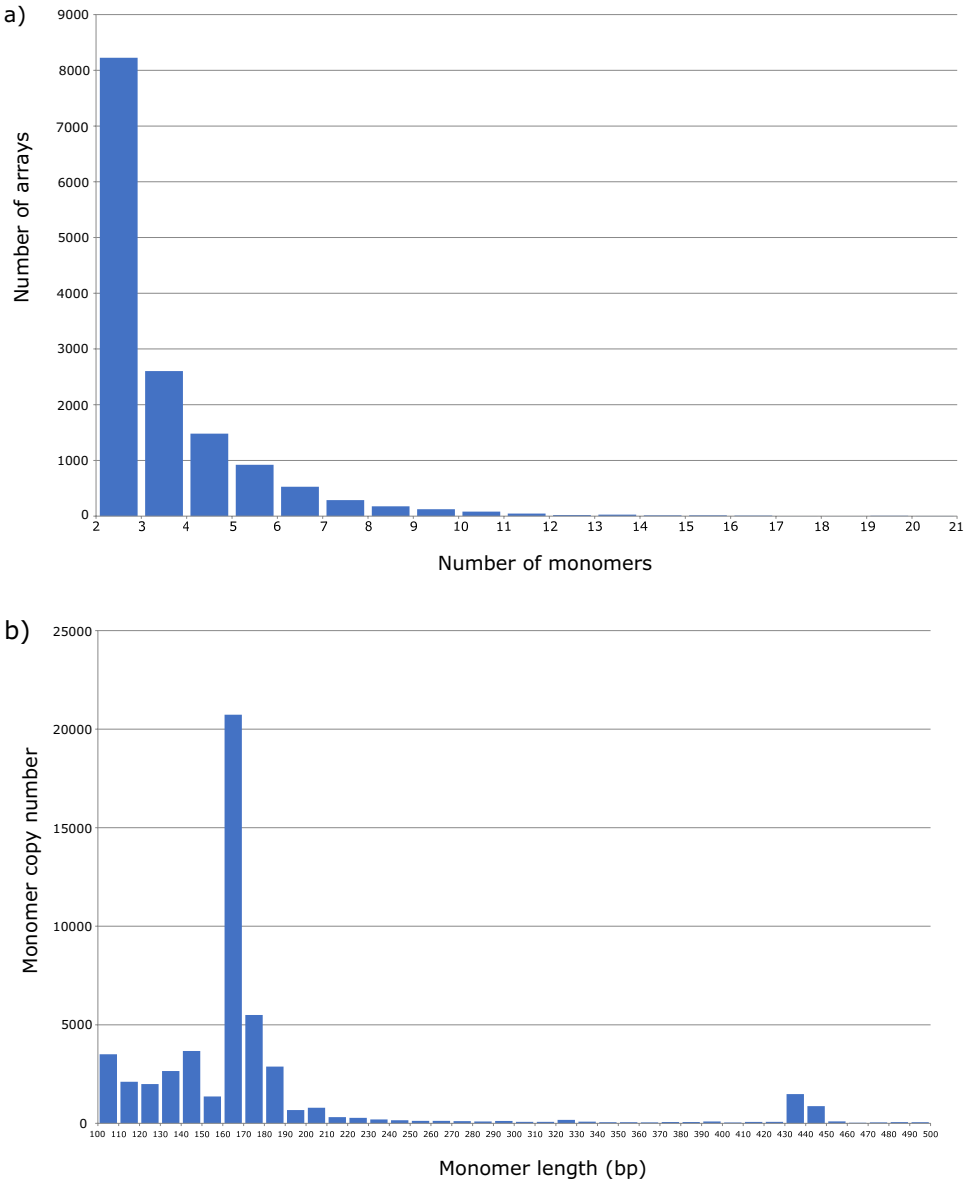

Figure 2. Correlation between number of monomers, monomer length, and number of arrays. Number of arrays plotted as a function of number of monomers (**a**), monomer copy number plotted as a function of monomer length (**b**).

Alignments of sequences flanking satDNA-like arrays revealed 50 to 100 bp long stretches of similarities that enabled derivation of left and right consensus sequences (LF and RF, respectively) for each cluster (Fig. 3b, c). These sequences were used as queries in a search through Repbase (Supplementary Table S4), which indicated high similarities with non-autonomous mobile elements assigned as members of the *Helitron* superfamily from *C. gigas*[43].

To explore the features of sequences that flank satDNA-like arrays in more detail, we excluded the array part, and, separately for each element, constructed LF-RF chimeric segments. The LF consensus sequence regularly ends with a microsatellite-like segment, separated from the first repeat in the array of TRs by a 10 to 500 bp long segment that is highly variable in DNA sequence and length. RF consensus sequence was identified as a 60 bp long segment following the 3′ end of the last repeat in an array, sharing > 80% similarity among elements within each cluster (Fig. 3c). The consensus sequences of these identified elements have typical HINE substructures[20]: 5′ subTIR, IR (complementary to the subTIR), and a microsatellite in LF; and 3′ subTIR and a palindrome in RF (Fig. 3b, Supplementary Table S2). The presence of these substructures was confirmed in the majority of array flanking regions (Supplementary Table S2). Accordingly, we named the identified *C. gigas* elements *Cg_HINE*. In total, we were able to distinguish 11 different *Cg_HINE* elements that carry arrays of TRs from clusters CL1-13. Exceptionally, sequences flanking arrays in clusters CL10 and CL13, and in clusters CL11 and CL12, were

Cluster	Number of arrays	Arrays with ≥5 monomers (% of total arrays)	Total length of arrays (kb)	Proportion in the genome assembly (%)**	Total number of monomers	Average number of monomers / array	Average GC (%)	Consensus monomer length (bp)
CL1	1,397 (900*)	507 (36.29%)	1,067,806	0.19	6,432 (4,881*)	4.6	39.22	167
CL2	1,031 (950*)	453 (43.94%)	855,355	0.15	5,183 (5,985*)	5.05	36.85	166
CL3	703	234 (33%)	531,465	0.1	3,209	4.56	31.16	167
CL4	684	167 (24.4%)	460,902	0.08	2,754	4.03	36.51	170
CL5	643	62 (9.64%)	359,046	0.06	2005	3.12	31.6	181
CL6	598	123 (20.5%)	346,142	0.06	2,384	3.99	39.57	147
CL7	564	42 (7.44%)	345,312	0.06	1957	3.47	41.24	178
CL8	488	47 (9.63%)	263,021	0.05	1635	3.35	32.96	162
CL9	307	25 (8.14%)	182,211	0.03	1,059	3.45	33.2	173
CL10	227	64 (28.19%)	153,409	0.03	927	4.08	34.66	167
CL11	213	0	82,801	0.01	509	2.39	30.91	162
CL12	156	1 (0.64%)	47,771	0.01	356	2.28	44.92	138
CL13	140	4 (2.86%)	62,802	0.01	380	2.72	32.29	168
Total CL1-13	7,151	1729	4,758,043	0.85	32,582	3.62	35.77	165

Table 1. Features of arrays of TRs grouped by similarity in 13 clusters in the genome of the Pacific oyster *Crassostrea gigas*. *Due to technical constrains of the Tandem Repeat Database, it was not possible to extract CL1 and CL2 monomers which would belong to all detected arrays, but from the majority of arrays. For the monomer consensus derivation, the maximal number of monomers which could be obtained for those two clusters was used: 4,881 CL1 monomers (belonging to 900 arrays of CL1) and 5,985 CL2 monomers (belonging to 950 arrays of CL2). **The calculation for the proportion of analyzed arrays (bp) in the genome assembly was done using the genome assembly size of 559 Mb[23].

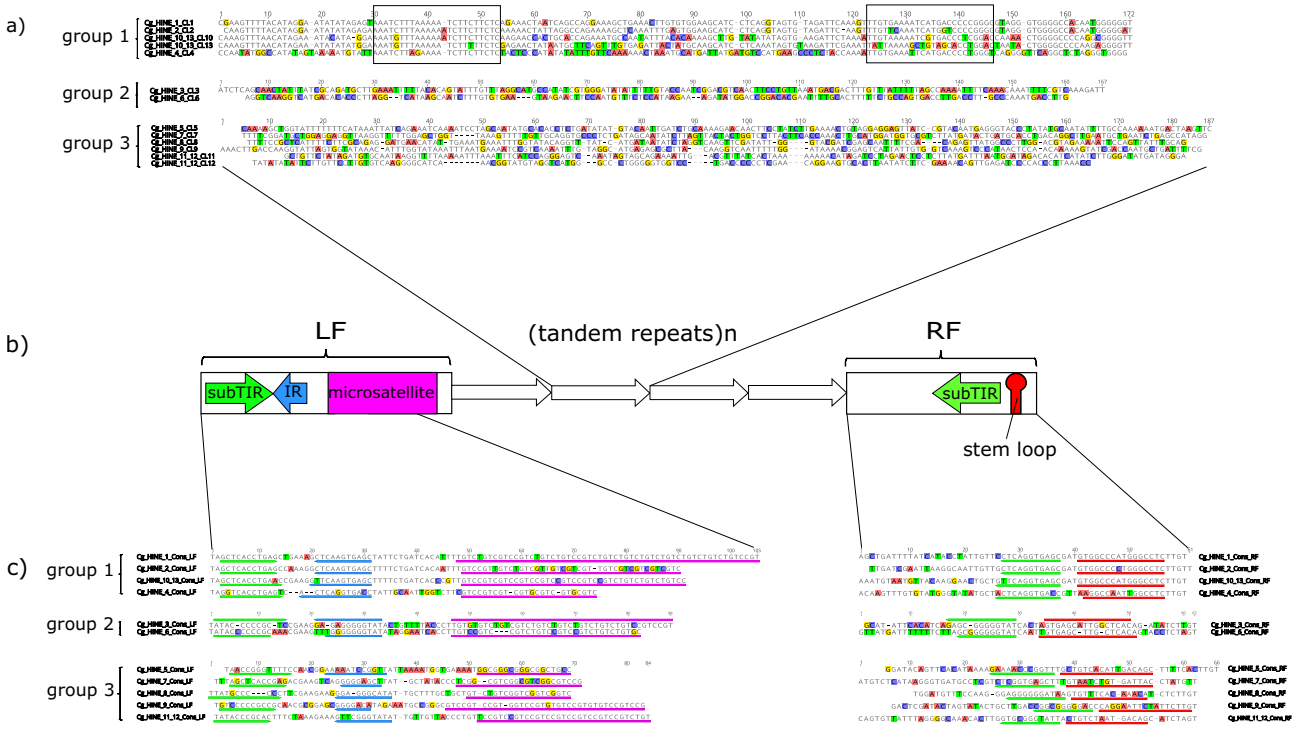

Figure 3. Structural characteristics and sequence comparisons of *Cg_HINE* elements. Consensus sequences of monomers belonging to each cluster are shown in (**a**). A general scheme of all depicted *CG_HINE* elements is presented in the central part of this figure (**b**). Consensus sequences of left and right flanking sequences (LF and RF, respectively) are presented in (**c**). Furthermore, elements are grouped according to sequence similarity. Group 1 form *Cg_HINE*s that share similarity in all element parts. In group 2 there are elements similar in flanking segments but not in monomers building TRs. Group 3 form elements divergent in their nucleotide sequences. Sequence segments corresponding to the structural elements in LF and RF are underlined with the same color as used in schematic presentation in (**b**). In group 1 monomer consensus sequences (**a**) boxed are sequence segments with reduced variability compared to monomers from the cluster CL4 (see text for explanation). In all alignments, differences present in less than half of nucleotides at each position are colored.

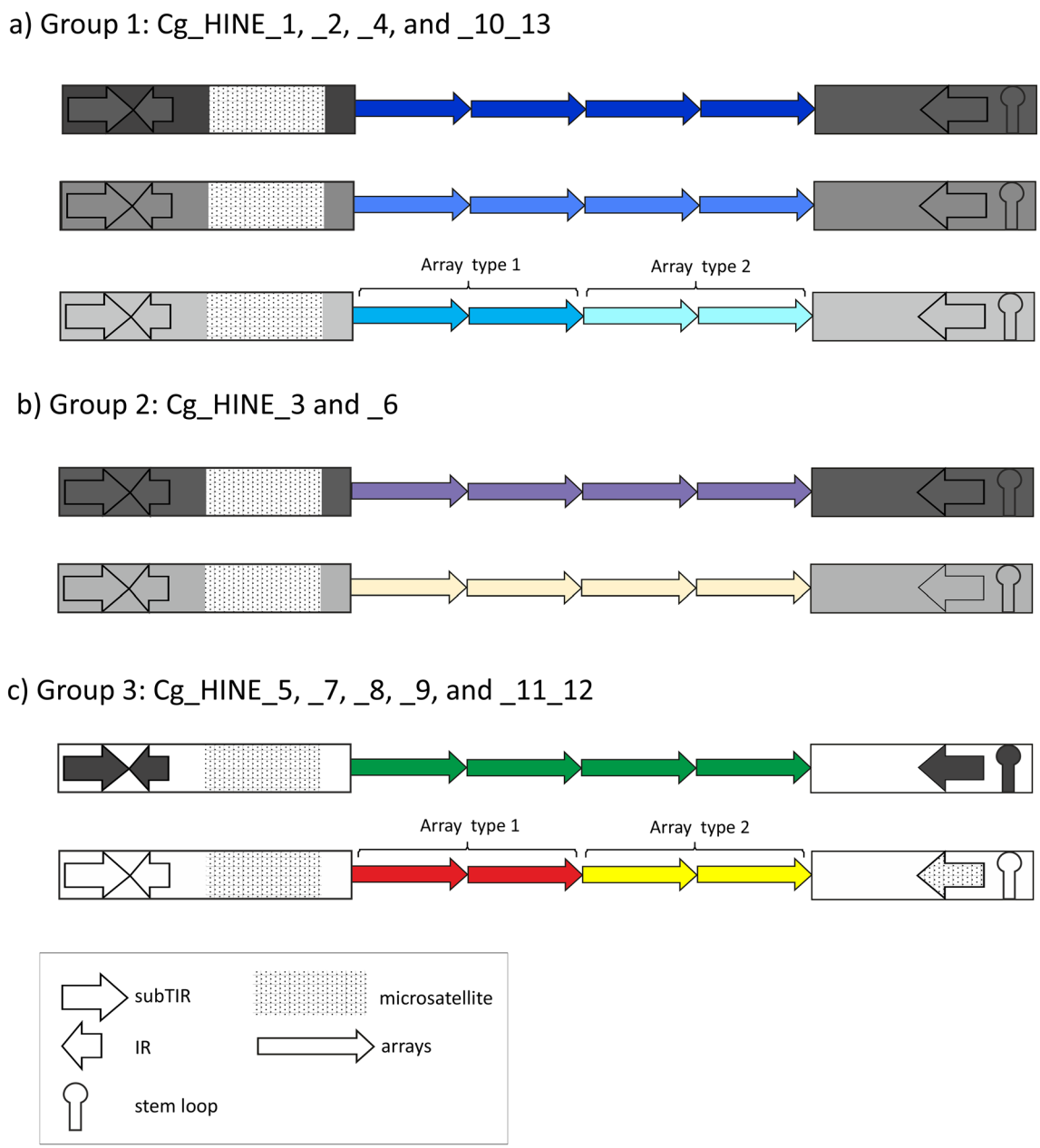

Figure 4. Schematic presentation of groups of *Cg_HINE* elements. Related elements of group 1 are shown in (**a**), related in flanking sequences but with divergent TRs of group 2 are in (**b**), and divergent in all segments (group 3) are in (**c**). Shown are also elements with two arrays of TRs, detected in group 1 and 3. Grey tones indicate sequence similarity in flanking segments. Tones of a color in monomers indicate similarity, while different colors indicate unrelated monomer sequences.

analyzed together owing to the fact that annotations revealed a predominant organization of these pairs of arrays in a consecutive order (Fig. 4).

Detailed analyses showed that perfect subTIRs are present in the majority of *Cg_HINE_9* and *Cg_HINE_10_13* elements, and 1 mismatch on different positions dominates in the rest. In all other elements subTIRs with 1 mismatch are predominant, while perfect subTIR sequences were found in < 50% of the examined flanking sequence pairs. The subTIR sequence, normally 11–12 bp long, is in LFs accompanied by 7–12 bp long complementary IR sequence, separated by a short segment of 1–12 bp. In turn, the subTIR in RF is followed by a palindrome (Fig. 3b, c, Supplementary Table S2). All of the examined 3′ palindromes have the potential of forming stem and loop structures (not shown).

Microsatellite sequences are mostly built of 4-nucleotide motifs as repeat units, although those with 3- and up to 8-nucleotide repeats were found (Supplementary Table S2). The LF sequence of several elements (*Cg_HINE_1, _3, _6, _8, _10_13, and _11_12*) share similar microsatellite sequences (GTCY, GTCC and GTCK). In all examined elements, the microsatellite array is short and highly variable in length and nucleotide sequence.

Many ambiguous nucleotide positions, insertions and deletions in the nucleotide sequence indicate high level of mutations eroding this segment (Fig. 3c).

Grouping of detected *Cg_HINE* elements. We further grouped *Cg_HINE* elements according to sequence similarities among their flanking segments and/or among their constitutive TRs, taking also into consideration the organizational patterns of TRs within elements (Fig. 4).

The group 1 (*Cg_HINE_1, _2, _4, and _10_13* in Fig. 3), characterized by similarity that extends throughout all element modules, is presented schematically in Fig. 4a. Related subfamilies of Cg170 satDNA are comprised of central repeats of elements in this group (Fig. 3a, Supplementary Table S3). Differences among TRs are accompanied by element-specific diagnostic changes in LF and RF consensus sequences, including in the sub-TIR, IR, stem and loops, and microsatellites (Fig. 3c, Supplementary Table S2). The highest similarity is between *Cg_HINE_1 and _2* which share the most similar monomers in their TRs (Supplementary Table S3). The most divergent in this group is *Cg_HINE_4*, sharing with others sequence similarity in the substructures of flanking segments (Fig. 3c), while its monomers are divergent, except in the two motifs shared with the rest (Fig. 3a). This group also includes *Cg_HINE_10_13*, which carries two consecutive arrays of related monomers, CL10 and CL13, instead of only one (Fig. 4a, Supplementary Table S3). The two arrays continue directly one after the other in the majority of *Cg_HINE_10_13* elements, although in a small fraction (~5%) they are separated by up to about 1.2 kb of anonymous DNA sequences.

Group 2 includes two elements, *Cg_HINE_3* and *_6*. In their flanking regions they differ in a way comparable to that of elements of the first group, although their TRs are completely unrelated (Fig. 3, 4b).

Elements *Cg_HINE_5, _7, _8*, and *_9* make up group 3 (Fig. 3, 4c). Modules of elements in this group are unrelated among themselves and with any other element in the studied sample. In each of these elements, satDNA-like TRs are derived from a single cluster of unique sequences, except in *Cg_HINE_11_12* which incorporates two arrays of TRs in a manner as described for *Cg_HINE_10_13* but with unrelated monomers (Fig. 4c).

We have also addressed the question of orientation of TRs with regard to their flanking sequences in the *Cg_HINE* element. Alignment of all elements shows the same orientation of flanking regions and their corresponding TRs, thus emphasizing the regularity of the proposed organizational pattern. This feature was observed in all studied elements without exception.

Insertion sites and genomic distribution of *Cg_HINE* elements. To characterize insertion sites of *Cg_HINE*s, we constructed "empty site" chimeric fragments (i.e. without the *HINE* element) coupled at the element ends determined previously. For this, 50 bp long stretches upstream and downstream of an element were taken and used in a BLAST search throughout the genome assembly. Combining this analysis and inspecting the LF and RF consensus sequences, an insertion preference for T-rich regions was observed for all *Cg_HINE* elements. In general, TT dinucleotides are suggested to be the preferential insertion site (Supplementary Fig. S2). Furthermore, we did not find any indication of TSD at the insertion site. These two features are consistent with both *Helentrons* and *Helitrons*[21,44].

To find *C. gigas* genomic sequences uninterrupted with *Cg_HINE* elements, we used "*Cg_HINE*-empty" chimeric constructs as a query. This search revealed the existence of uninterrupted segments of high similarity (>90%) in the genome. For some chimeric constructs, the results disclosed many highly similar or identical hits, indicating insertions of *Cg_HINE*s into repetitive regions. Some of them could be identified as fragments of other TEs (mostly DNA transposons, but also some non-LTR retrotransposons; data not shown).

In addition, preliminary BLAST survey through GenBank using *Cg_HINE* elements indicated similarities in non-coding regions of some *C. gigas* genes. Regions of similarity correspond to entire elements or to their deletion derivatives (Supplementary Table S5). As an illustration, the *bmpr1* gene contains a whole *Cg_HINE* element with a short central array comprised of about two monomers similar to the CL2 consensus (80%). In the close vicinity of the *Gigasin-2* gene, a *Cg_HINE* element with 1.5 monomers similar to the CL6 consensus (73%) was found. The *bindin* gene incorporates three truncated *Cg_HINE* elements, one containing 8.7 monomers, averaging 76% similarity to the CL4 consensus sequence. The other two are ~4.5 kb distant and 99.7% identical one to the other. They contain arrays of 5.8 monomers, with 76% of average similarity to the CL1 consensus sequence.

Assessing the organizational patterns of *Cg_HINE* elements revealed their integration into assembled genomic sequences in both orientations. Because of the general association of examined TRs with *Cg_HINE* elements, element distribution was approximated by mapping sequences identical to arrays in CL1-13 onto *C. gigas* pseudochromosomes[45]. A dense interspersed pattern has been shown for each studied sequence, and no preference to any assembled chromosome or to any particular chromosomal segment could be detected (Supplementary Fig. S3).

Discussion

In the present work we revealed sequences repeated in tandem in the genome assembly of the Pacific oyster *Crassostrea gigas*[23]. Because of our interest in understanding patterns and drivers of genome-wide dispersal of sequences that might be related to satDNAs, we limited our analysis to repeats between 100 and 500 bp in size, as most commonly found in satDNAs[46–48], including bivalves' satDNAs[30]. The search for TRs in genome assemblies introduces another limitation. Namely, long arrays of satDNAs, built of highly similar sequences repeated in tandem and characteristic for heterochromatic regions and centromeres[2] are generally misrepresented or overlooked in genome outputs due to difficulties in discerning their sequential order and length[49]. However, such assemblies offer a reliable platform when a genome-wide "euchromatic" distribution of short arrays of TRs and sequences associated with them are specifically targeted[38,39,41,50,51].

The 13 most abundant clusters of short arrays of satDNA-like TRs (CL1-13, Table 1) detected in our survey were assigned to the 0.85% of the assembled *C. gigas* genome, and form 18% of sequences repeated in tandem as anticipated by Zhang et al.[23]. Furthermore, clusters CL1-13 comprise 49% of all arrays we have detected as 100–500 bp long tandemly repeated monomers. The remaining arrays are either present in a small number per cluster, or are too different to be clustered at all, representing putative singletons. The 13 clusters analyzed in this study can therefore be considered as a representative sample in illustrating the genome-wide organizational patterns of satDNA-like TRs in the *C. gigas* genome assembly.

Monomer lengths in clusters CL1-13 exist in a narrow range, on average 140–180 bp. The arrays are mostly comprised of up to 5 monomers, and those with ≥ 5 make up only 9% of the studied sample, the longest array containing only 29 monomers. The model proposed by Scalvenzi and Pollet[52] on *Xenopus* frogs predicts a predominance of short arrays of satDNA-like TRs in putative euchromatic genomic segments, as obtained in our analysis of the *C. gigas* genome assembly. According to this model, the limited array length is favored because of the inverse correlation between number of TRs and mobility of TEs that may be involved in their dispersal. This observation, however, does not exclude that some of the repeats can also be builders of long arrays of classical satDNAs, associated with heterochromatic fractions, not included in the genome assembly.

Analysis of flanking segments revealed regular association of short satDNA-like arrays in all 13 clusters with sequences that have structural signatures of *HINEs*, non-autonomous TEs of the *Helentron* superfamily. Accordingly, they also lack TSD and have oligo-T segments as the preferential insertion site[20,21,44]. The best-studied *HINE* elements are nevertheless *DINE*-1 and its derivatives, which are widespread in *Drosophila*[20,53,54]. Their centrally-located TRs can also be found in the form of classical satDNAs, hypothesizing the general role of TR-carrying *Helentrons* in satDNA expansion[39].

The Cg170/HindIII satDNA family is the most abundant in *C. gigas*, comprising 1–4% of the genome, and located in the centromeric regions of some chromosomes[27–29]. It is therefore not surprising that monomer variants of this satDNA family appear in some of the *Cg_HINE* elements. The average number of repeats in the elements carrying Cg170/HindIII monomers is slightly higher than the number of repeats in other *Cg_HINE* elements, but with a clearly increased number of arrays that contain ≥ 5 monomers (Table 1). In this regard, we can speculate that non-Cg170/HindIII monomers enriched in arrays containing ≥ 5 monomers in *Cg_HINE_3* and *Cg_HINE_6* may represent repeats in expansion or copies of undetected classical satDNA candidate sequences in this species.

It must be noted that some elements classified as *HINE* were already detected in bivalves. The *Pearl* family was described in the cupped oyster *Crassostrea virginica* and the blood ark *Anadara trapezia* and it had been anticipated that related elements could also be present in *C. gigas*[32]. *Pearl* elements carry short arrays of up to six ~ 160 bp long central repeats, some of them being similar to monomers of classical satDNAs found in other bivalve species, including Cg170 of *C. gigas*. Of comparable architecture are also *DTC84* of the clam *Donax trunculus*[55] and the element *MgE* in the Mediterranean mussel *Mytilus galloprovincialis*[56].

Analysis of the relationships among 11 *Cg_HINE* elements in *C. gigas* can help to understand drivers of satDNA-like TR dispersal and evolution. In the studied sample, two characteristics turned out to be common to all of them. First, the orientation of an array with regard to the flanking sequences is always the same in every element, without exception, indicating a single event in TR formation. Second, association of satDNA-like sequences from a particular cluster with a specific pair of flanking sequences is consistent (but not vice-versa, see below). In addition, three groups of intraspecific relationships defined according to similarities among element modules (LF-array-RF) can be discerned.

Group 1 is formed by elements similar in flanking segments and in associated satDNA-like repeats (Fig. 4a). Accumulated mutations should allow subsequent spread of variants if they still retain the structural requirements needed for replication[16]. Concurrent accumulation of changes along the whole element length suggests a persistence of association between flanking modules and satDNA-like central repeats (in this case related to the Cg170/HindIII) emerging from the formation of the ancestral copy. It can be further hypothesized that changes accumulated in the array of satDNA-like TRs in the course of element evolution may be a source of the subfamilies of Cg170/HindIII satDNA. Comparably, *DINE*-1 elements in *Drosophila willistoni* diverged into three subtypes, with changes both in subTIRs and TRs[53]. Concurrent accumulation of differences along whole element lengths has also been observed in interspecies comparisons of *Pearl* elements *CvE* of *C. virginica* and *MgE* of the Mediterranean mussel *Mytilus galloprovincialis*[56].

In group 2, related flanking segments incorporate unrelated satDNA-like TRs (Fig. 4b), indicating independent incorporation events into flanking sequences of common origin. At the interspecific level, TRs of different origin associated with related flanking segments were observed in *Drosophila DINE*-1 elements[53].

Group 3 is formed by *Cg_HINE* elements unrelated in nucleotide sequences of all modules (Fig. 4c). Elements in this subset could therefore be considered as *HINE* families that arose independently in the genome. In addition, four *Pearl* elements detected in *C. virginica*[32] are also of independent origin. It can be concluded that the existence of unrelated *Cg_HINE* elements indicate multiple, and probably not rare events of independent element acquisition.

A special case is represented by *C. gigas* elements that incorporate two arrays of TRs instead of only one, originating either from related or from unrelated clusters (Fig. 4a and c). Multiple central arrays can be formed by recombination of elements that share flanking modules but not the central repeats. An alternative hypothesis is that a junction fragment containing segments of two divergent satDNA arrays became a source of double arrays integrated into a single *Cg_HINE* element. Abrupt junctions between two repetitive sequences that may be candidates for such a scenario were observed in bivalves[31], as well as in other species (for example[57–59]).

The genesis of TRs in TR-carrying elements can be explained in the light of two scenarios discussed by[52]. According to the first, precursor satDNA sequences are captured ("filled") and further propagated by an element, while according to the second, TRs are formed from the element's intrinsic sequences. Analysis of the

acquisition of sequence segments by the insect *Helitrons* favor the filler DNA model, proposing that internal segments are integrated into an element by multiple insertions[17]. Such events might also explain the formation of double arrays as observed in the two *Cg_HINE* elements. In addition, divergent central repeats carried by related flanking modules can be a consequence of insertion of potential monomer segment(s) and concurrent excision of the previously existing sequence. This process can be based on motifs in satDNA sequences recognized by transposase-related proteins[60], as explained for monomer replacements observed in a root-knot nematode satDNA[58]. Similar cut-and-replace events were also anticipated in our previous analysis of Cg170 satDNA junction regions[31].

Autonomous *Helentrons* can be assumed to be putative partners of *Cg_HINE* elements. Nearly-perfect identity marked autonomous *Helentrons* as partners of three *DINE*-1 elements in *Drosophila*[20]. In this regard, a *Helentron*-type Rep motif 2, a signature of autonomous *Helentrons*, has been detected in *C. gigas* genome data[21]. We found 10 *C. gigas* autonomous elements that harbor the *Helentron*-type Rep motif 2 in Repbase but could not relate any of them with the *Cg_HINE*s (not shown), so the nature of their relationship, if any, remains unresolved. In addition, some *Cg_HINE*s were found integrated into repetitive regions that may represent other putative TEs or may be the result of segmental duplications. Frequent integration into other TEs as new drivers of spread is a feature expected for TR-carrying *Helentrons*[18].

Analysis of the genomic dispersal of arrays in clusters CL1-13 revealed their apparently uniform distribution on all *C. gigas* pseudochromosomes, which comprise about 50% of the genome[45]. Our preliminary search indicated insertions of *Cg_HINE*s within non-coding regions of some genes. Functions of these genes are related to early embryogenesis (*bmpr1* gene)[61], fertilization (*bindin* gene)[62] and oyster defence system (*Ecsit* and *Gigasin-2* genes)[63]. It can be expected that also fragmented *Cg_HINE* elements, isolated arrays of TRs or monomer fragments could be found dispersed throughout the genome, affecting genes and/or their regulatory regions. *Helitrons* and *Helentrons* in general, whether they carry internal TRs or not, have a strong influence on gene expression, not only by frequent gene capturing but also by inserting themselves close to the gene[17,53,64]. Therefore, the abundance of satDNA-like TRs as parts of *Cg_HINE* elements suggests they have a high impact on *C. gigas* genome evolution and function.

Conclusion

We searched the genome assembly of the Pacific oyster *C. gigas* for TRs that resemble satDNAs in their monomer length. In the euchromatic genome fraction the detected satDNA-like TRs are composed of only short arrays. The most abundant clusters of TRs (49% of all detected) have a monomer length in a narrow range of 140–180 bp, characteristic for classical satDNAs. We found the most abundant satDNA-like arrays of TRs in the *C. gigas* genome assembly integrated as central repeats of non-autonomous *HINE* elements. Each group of satDNA-like arrays is associated with element-specific flanking sequences, making altogether a unique *Cg_HINE* element. The ability to follow the evolution of whole elements indicates stability once a relationship between the satDNA-like TRs and their flanking sequences was established. Sequences related to the most abundant satDNA Cg170 of *C. gigas*[27,28] were also found as short satDNA-like arrays in some of the 13 studied *Cg_HINE* elements, showing close interrelations between these two classes of repetitive sequences, TE and satDNAs. Information obtained in this study promote bivalves as a second model system, after *Drosophila*, in analysis of non-autonomous TR-carrying *Helentrons*, a still poorly understood group of TEs using RCR mechanism in their spread.

Materials and methods

Detection and grouping of tandem repeats. The assembled *C. gigas* genome sequence (oyster.v9.fa) was downloaded from https://gigadb.org/dataset/100030 and uploaded into Tandem Repeats Database (TRDB) available at https://tandem.bu.edu/cgibin/trdb/trdb.exe65. Tandem repeats (TRs) were extracted using default parameters: alignment parameters 2,7,7 (match, mismatch, indels) and 50 as the minimum alignment score. The resulting arrays of TRs were filtered using the following criteria: pattern size ≥ 100 and ≤ 500 bp and repeat copy number ≥ 2 (Supplementary Table 1). Filtered arrays were processed using the redundancy tool with redundancy by period set at 50% overlap to eliminate multiple reporting of repeats (i.e. in cases when one repeat is part of another one). The clustering tool, implemented in TRDB, was used to group arrays of TRs that share at least 70% similarity under the following conditions: cutoff value set to 70, heuristical algorithm, DUST (to filter low complexity regions), and PAM (default values) options included. Clusters were ordered in descending order according to the total number of arrays, so the first cluster, CL1, contained the highest number of arrays (Table 1). For further analysis, arrays belonging to a specific cluster were downloaded from TRDB, and processed in Geneious 9.0.4 (Biomatters, Ltd). Multiple sequence alignments were performed to obtain monomer consensus sequence for each cluster. *C. gigas* pseudochromosomes[45] were annotated using local databases holding all arrays from clusters CL1-13. Only 100% identical arrays were annotated on the 10 pseudochromosomes.

Defining DNA sequences flanking tandem repeats. In order to explore the genomic environment of recognized TRs, sequences surrounding arrays were extracted from the corresponding scaffolds using TRDB. Up to 4,000 bp was extracted from each array side, in dependence to the array position in the scaffold and the scaffold length. Arrays of TRs positioned at the very end of a scaffold were excluded from further analysis, as well as those containing stretches of "Ns" in the flanking sequence. The consensus left and right flanking sequences (LF and RF, respectively) have been determined in a series of multiple alignments, performed separately for arrays in each cluster. Using Map to Reference Tool implemented in Geneious 9.0.4 (Biomatters, Ltd), consensus LF and RF sequences were further used to anchor alignment of all LF-array-RF sequence segments in the corresponding cluster. In the following step, sequence segments matching consensus sequences were extracted and used in additional alignments, in order to obtain refined LF and RF consensus sequences for arrays of TRs in each clus-

ter. In this way, number of flanking sequences used for the final derivation of LF and RF consensus sequences of each element was > 50%, and similarity according to the consensus was > 70% (Supplementary Table 2).

Detection of substructures in sequences flanking tandem repeats. In order to detect substructures, LF and RF sequences of each array in the cluster were merged into chimeric constructs using a local script. Because of high sequence variability, instead of using consensus sequences, this was done for each LF and RF sequence pair, preserving their original orientation. Chimeric LF-RF sequence sets were imported into Inverted Repeats Database (IRDB; https://tandem.bu.edu/cgi-bin/irdb/irdb.exe). A search for inverted repeat (IR) was performed using default values, with the exception of alignment parameters set to 2,5,7 and the minimum alignment score set to 14. Results were filtered by position of IRs in a way that one of the pairs (left first index) is present in LF and the other (right first index) is in RF. Similarity of IRs was set to 90%. Sets of filtered IRs were downloaded and further analyzed by multiple sequence alignments. IRs that were the most prominent according to sequence similarity and abundance were selected, and, in order to check their appropriate positioning, annotated in chimeric LF-RF constructs using the Motif Search Tool implemented in Geneious 9.0.4 (Biomatters, Ltd), allowing 1 mismatch. This procedure enabled identification of IR and subterminal inverted repeat (subTIR) structures.

For palindrome search, we used programs *einverted* and *palindrome* in the EMBOSS package[66] (https://www.hgmp.mrc.ac.uk/Software/EMBOSS/). Secondary structures formed by palindromes were predicted by the program *DNA fold* implemented in Geneious 9.0.4 (Biomatters, Ltd).

Tandem Repeat Finder v4.09.[67] was used for the microsatellite detection and definition (alignment parameters set at 2,3,5). The microsatellite repeat consensus was determined by alignments of all microsatellite sequences within a cluster using MUSCLE (implemented in Geneious).

Similarity search through databases. In order to determine similarities with known repetitive elements or other published sequences, monomer and flanking consensus sequences were queried through Repbase[68] and NCBI GenBank Database. For local blast analysis, a collection of *C. gigas* repetitive elements was made by downloading sequences from Repbase.

Empty site analysis. For each particular *Cg_HINE*, 10 to 20 elements with well-defined sequence segments were selected randomly and used in the empty site analysis. Sequences 50 bp upstream and downstream from the determined element ends were merged into chimeric constructs. These constructs were used as queries in a homology search through the genome assembly (Discontiguous Megablast, max E value = 10). At least 80% identity over 85% of query length was used as criterion for verification of a paralogous site.

Exploring relationships among monomers in clusters. Arrays from clusters with highest mutual sequence similarity (CL1, CL2, CL10 and CL13), were randomly selected to obtain approximately 150 monomers from each of them. To avoid truncated monomers at array beginning and/or end, only those longer than 140 bp, and in the frame with the corresponding consensus sequence were selected. For the ~ 600 monomers, short FASTA headers were derived by renaming monomers in a way to distinguish the cluster and the array, as well as the monomer position in the array. MAFFT v7.017 type of alignment was used for further analyses[69]. The best-substitution model was identified by jModelTest 2.1.4.[70]. The best model was chosen according to the Akaike Information Criterion (TPM1uf + G). For the phylogenetic analysis PhyML 3.0.[71] using 100 bootstrap replicates was run. The obtained maximum likelihood trees were displayed in Fig Tree 1.4.2. Available at https://tree.bio.ed.ac.uk/software/figtree/.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Received: 16 June 2020; Accepted: 11 August 2020
Published online: 15 September 2020

References

1. Biscotti, M. A., Olmo, E. & Heslop-Harrison, J. S. Repetitive DNA in eukaryotic genomes. *Chromosom. Res.* **23**, 415–420 (2015).
2. Plohl, M., Meštrović, N. & Mravinac, B. Satellite DNA evolution. in *Genome Dynamics* (ed. Garrido- Ramos, M. A.) vol. 7 126–152 (Karger AG, Basel, 2012).
3. Garrido-Ramos, M. A. Satellite DNA: an evolving topic. *Genes (Basel).* **8**, (2017).
4. Hartley, G. & O'neill, R. J. Centromere repeats: hidden gems of the genome. *Genes (Basel).* **10**, 233 (2019).
5. Jurka, J., Kapitonov, V. V., Kohany, O. & Jurka, M. V. Repetitive sequences in complex genomes: structure and evolution. *Annu. Rev. Genom. Hum. Genet.* **8**, 241–259 (2007).
6. Kojima, K. K. Structural and sequence diversity of eukaryotic transposable elements. *Genes Genet. Syst.* **94**, 233–252 (2019).
7. Meštrović, N. *et al.* Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosom. Res.* **23**, 583–596 (2015).
8. Macas, J., Koblížková, A., Navrátilová, A. & Neumann, P. Hypervariable 3′ UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. *Gene* **448**, 198–206 (2009).
9. Sharma, A., Wolfgruber, T. K. & Presting, G. G. Tandem repeats derived from centromeric retrotransposons. *BMC Genom.* **14**, 142 (2013).
10. McGurk, M. P. & Barbash, D. A. Double insertion of transposable elements provides a substrate for the evolution of satellite DNA. *Genome Res.* **28**, 714–725 (2018).

11. Vondrak, T. *et al.* Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. *Plant J.* (2019). https://doi.org/10.1111/tpj.14546.
12. Dias, G. B., Svartman, M., Delprat, A., Ruiz, A. & Kuhn, G. C. S. Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of *Drosophila virilis*. *Genome Biol. Evol.* **6**, 1302–1313 (2014).
13. Kapitonov, V. V. & Jurka, J. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet.* **23**, 521–529 (2007).
14. Kapitonov, V. V. & Jurka, J. Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci. USA* **98**, 8714–8719 (2001).
15. Thomas, J., Phillips, C. D., Baker, R. J. & Pritham, E. J. Rolling-circle transposons catalyze genomic innovation in a mammalian lineage. *Genome Biol. Evol.* **6**, 2595–2610 (2014).
16. Grabundzija, I. *et al.* A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nat. Commun.* **7**, 10716 (2016).
17. Han, G. *et al.* Characterization of a novel Helitron family in insect genomes: Insights into classification, evolution and horizontal transfer. *Mob. DNA* **10**, 1–15 (2019).
18. Pritham, E. J. & Feschotte, C. Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1895–1900 (2007).
19. Xiong, W., Dooner, H. K. & Du, C. Rolling-circle amplification of centromeric Helitrons in plant genomes. *Plant J.* **88**, 1038–1045 (2016).
20. Thomas, J., Vadnagara, K. & Pritham, E. J. DINE-1, the highest copy number repeats in *Drosophila melanogaster* are non-autonomous endonuclease-encoding rolling-circle transposable elements (Helentrons). *Mob. DNA* **5**, 18 (2014).
21. Pritham, E. J. & Thomas, J. Helitrons, the eukaryotic rolling-circle transposable elements. *Microbiol. Spectr.* **3**, 1–32 (2015).
22. Fernández Robledo, J. A. *et al.* HHS public access. *Dev. Comp. Immunol.* **92**, 260–282 (2019).
23. Zhang, G. *et al.* The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**, 49–54 (2012).
24. Wang, X. *et al.* Genome-wide and single-base resolution DNA methylomes of the Pacific oyster *Crassostrea gigas* provide insight into the evolution of invertebrate CpG methylation. *BMC Genom.* **15**, 1119 (2014).
25. Wang, X. *et al.* Nanopore sequencing and de novo assembly of a black-shelled Pacific oyster (*Crassostrea gigas*) genome. *Front. Genet.* **10**, 1211 (2019).
26. Bouilly, K., Chaves, R., Leitão, A., Benabdelmouna, A. & Guedes-Pinto, H. Chromosomal organization of simple sequence repeats in the Pacific oyster (*Crassostrea gigas*): (GGAT)4, (GT)7 and (TA)10 chromosome patterns. *J. Genet.* **87**, 119–125 (2008).
27. Clabby, C. *et al.* Cloning, characterization and chromosomal location of a satellite DNA from the Pacific oyster, Crassostrea gigas. *Gene* **168**, 205–209 (1996).
28. Wang, Y., Xu, Z. & Guo, X. A centromeric satellite sequence in the Pacific Oyster (*Crassostrea gigas* Thunberg) identified by fluorescence in situ hybridization. *Mar. Biotechnol.* **3**, 486–492 (2001).
29. López-Flores, I. *et al.* The molecular phylogeny of oysters based on a satellite DNA related to transposons. *Gene* **339**, 181–188 (2004).
30. Šatović, E., Vojvoda Zeljko, T. & Plohl, M. Characteristics and evolution of satellite DNA sequences in bivalve mollusks. *Eur. Zool. J.* **85**, 94–103 (2018).
31. Šatović, E., Vojvoda Zeljko, T., Luchetti, A., Mantovani, B. & Plohl, M. Adjacent sequences disclose potential for intra-genomic dispersal of satellite DNA repeats and suggest a complex network with transposable elements. *BMC Genom.* **17**, 997 (2016).
32. Gaffney, P. M., Pierce, J. C., Mackinley, A. G., Titchen, D. A. & Glenn, W. K. Pearl, a novel family of putative transposable elements in bivalve mollusks. *J. Mol. Evol.* **56**, 308–316 (2003).
33. Macas, J., Neumann, P. & Navrátilová, A. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genom.* **8**, 1–16 (2007).
34. Ruiz-Ruano, F. J., López-León, M. D., Cabrero, J. & Camacho, J. P. M. High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Sci. Rep.* **6**, 28333 (2016).
35. Khost, D. E., Eickbush, D. G. & Larracuente, A. M. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila. *Genome Res.* **27**, 709–721 (2017).
36. Lower, S. S., McGurk, M., Clark, A. G. & Barbash, D. Double insertion of transposable elements provides a substrate for the evolution of satellite DNA. *Curr. Opin. Genet. Dev.* **49**, 70–78 (2018).
37. Silva, B. S. M. L., Heringer, P., Guilherme, B. D., Svartman, M. & Kuhn, G. C. S. De novo identification of satellite DNAs in the sequenced genomes of 2 *Drosophila virilis* and *D. americana* using the RepeatExplorer and 3 TAREAN pipeline. *bioRxiv Prepr. first* **38**, 54–70 (2019).
38. Pavlek, M., Gelfand, Y., Plohl, M. & Meštrović, N. Genome-wide analysis of tandem repeats in *Tribolium castaneum* genome reveals abundant and highly dynamic tandem repeat families with satellite DNA features in euchromatic chromosomal arms. *DNA Res.* **22**, 387–401 (2015).
39. Dias, G. B., Heringer, P., Svartman, M. & Kuhn, G. C. S. Helitrons shaping the genomic architecture of Drosophila: enrichment of DINE-TR1 in α- and β-heterochromatin, satellite DNA emergence, and piRNA expression. *Chromosom. Res.* https://doi.org/10.1007/s10577-015-9480-x (2015).
40. de Lima, L. G., Svartman, M. & Kuhn, G. C. S. Dissecting the satellite DNA landscape in three cactophilic Drosophila sequenced genomes. *G3 Genes Genomes Genet.* **7**, 2831–2843 (2017).
41. Brajković, J. *et al.* Dispersion profiles and gene associations of repetitive DNAs in the euchromatin of the beetle *Tribolium castaneum*. *G3 Genes Genomes Genet.* **8**, 875–886 (2018).
42. Sproul, J. S. *et al.* Dynamic evolution of euchromatic satellites on the X chromosome in *Drosophila melanogaster* and the simulans clade. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msaa078 (2020).
43. Bao, W. and Jurka, J. *DNA transposon from the Pacific oyster genome*. *Repbase Reports* vol. 13 (2013).
44. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
45. Gagnaire, P. A. *et al.* Analysis of genome-wide differentiation between native and introduced populations of the cupped oysters *Crassostrea gigas* and *Crassostrea angulata*. *Genome Biol. Evol.* **10**, 2518–2534 (2018).
46. Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001).
47. Heslop-Harrison, J. S. P. & Schwarzacher, T. Nucleosomes and centromeric DNA packaging. *Proc. Natl. Acad. Sci. USA* **110**, 19974–19975 (2013).
48. Levitsky, V. G., Babenko, V. N. & Vershinin, A. V. The roles of the monomer length and nucleotide context of plant tandem repeats in nucleosome positioning. *J. Biomol. Struct. Dyn.* **32**, 115–126 (2014).
49. Tørresen, O. K. *et al.* Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucl. Acids Res.* **47**, 10994–11006 (2019).
50. Wang, S., Lorenzen, M. D., Beeman, R. W. & Brown, S. J. Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. *Genome Biol.* **9**, R61 (2008).
51. Kuhn, G. C. S., Küttler, H., Moreira-Filho, O. & Heslop-Harrison, J. S. The 1.688 repetitive DNA of Drosophila: concerted evolution at different genomic scales and association with genes. *Mol. Biol. Evol.* **29**, 7–11 (2012).
52. Scalvenzi, T. & Pollet, N. Insights on genome size evolution from a miniature inverted repeat transposon driving a satellite DNA. *Mol. Phylogenet. Evol.* **81**, 1–9 (2014).

53. Yang, H.-P. & Barbash, D. A. Abundant and species-specific DINE-1 transposable elements in 12 Drosophila genomes. *Genome Biol.* **9**, R39 (2008).
54. Kuhn, G. C. S. & Heslop-Harrison, J. S. Characterization and genomic organization of PERI, a repetitive DNA in the *Drosophila buzzatii* cluster related to DINE-1 transposable elements and highly abundant in the sex chromosomes. *Cytogenet. Genome Res.* **132**, 79–88 (2011).
55. Šatović, E. & Plohl, M. Tandem repeat-containing MITEs in the clam *Donax trunculus*. *Genome Biol. Evol.* **5**, 2549–2559 (2013).
56. Kourtidis, A., Drosopoulou, E., Pantzartzi, C. N., Chintiroglou, C. C. & Scouras, Z. G. Three new satellite sequences and a mobile element found inside HSP70 introns of the Mediterranean mussel (*Mytilus galloprovincialis*). *Genome* **49**, 1451–1458 (2006).
57. Kuhn, G. C. S., Teo, C. H., Schwarzacher, T. & Heslop-Harrison, J. S. Evolutionary dynamics and sites of illegitimate recombination revealed in the interspersion and sequence junctions of two nonhomologous satellite DNAs in cactophilic Drosophila species. *Heredity (Edinb).* **102**, 453–464 (2009).
58. Meštrović, N. *et al.* Conserved DNA motifs, including the CENP-B Box-like, are possible promoters of satellite DNA array rearrangements in nematodes. *PLoS ONE* **8**, e67328 (2013).
59. Paço, A., Adega, F., Meštrović, N., Plohl, M. & Chaves, R. The puzzling character of repetitive DNA in Phodopus genomes (Cricetidae, Rodentia). *Chromosom. Res.* **23**, 427–440 (2015).
60. Casola, C., Hucks, D. & Feschotte, C. Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. *Mol. Biol. Evol.* **25**, 29–41 (2008).
61. Herpin, A. *et al.* Structural and functional evidence for a singular repertoire of BMP receptor signal transducing proteins in the lophotrochozoan *Crassostrea gigas* suggests a shared ancestral BMP/activin pathway. *FEBS J.* **272**, 3424–3440 (2005).
62. Moy, G. W. & Vacquier, V. D. Bindin genes of the Pacific oyster *Crassostrea gigas*. *Gene* **423**, 215–220 (2008).
63. Zhang, L., Li, L. & Zhang, G. The first identification of molluscan Ecsit in the Pacific oyster, *Crassostrea gigas*, and its expression against bacterial challenge. *Aquac. Res.* **43**, 1071–1080 (2012).
64. Coates, B. S., Hellmich, R. L., Grant, D. M. & Abel, C. A. Mobilizing the genome of lepidoptera through novel sequence gains and end creation by non-autonomous lep1 helitrons. *DNA Res.* **19**, 11–21 (2012).
65. Gelfand, Y., Rodriguez, A. & Benson, G. TRDB - The tandem repeats database. *Nucl. Acids Res.* **35**, 80–87 (2007).
66. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
67. Benson, G. Tandem repeats: a program to analyze DNA sequences. *Nucl. Acids Res.* **27**, 573–580 (1999).
68. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 4–9 (2015).
69. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl. Acids Res.* **30**, 3059–3066 (2002).
70. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772–772 (2012).
71. Guindon, S., Dufayard, J.-F., Lefort, V. & Anisimova, M. New alogrithms and methods to estimate maximum- likelihoods phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

Acknowledgements

We are thankful to Yevgeniy Gelfand for the help with TRDB. We would also like to thank Eva Šatović for critical reading of the manuscript, and Mary Sopta for language editing. This work was performed within the frame of the Croatian Science Foundation project IP-09-2014-3183 to M.P.

Author contributions

T.V.Z., N.M. and M.P. designed the analyses; T.V.Z. performed all analyses; M.Pa. contributed in tandem repeats and phylogenetic analyses; T.V.Z., N.M. and M.P. analyzed the data; all authors participated in the interpretation of the results; T.V.Z. and M.P. prepared the first draft of the manuscript; all authors reviewed and edited the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-71886-y.

Correspondence and requests for materials should be addressed to M.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020