Acta Societatis Botanicorum Poloniae

Authors' contributions
MA performed the experiments, wrote the paper, and prepared manuscript; AK designed and coordinated experiments, performed the experiments, analyzed the data

ORIGINAL RESEARCH PAPER

Tracking of intercalary DNA sequences integrated into tandem repeat arrays in rye *Secale vavilovii*

Magdalena Achrem[1,2]*, Anna Kalinka[1,2]

[1] Department of Cell Biology, Institute for Research on Biodiversity, Faculty of Biology, University of Szczecin, Wąska 13, 71-415 Szczecin, Poland
[2] Molecular Biology and Biotechnology Center, Faculty of Biology, University of Szczecin, Wąska 13, 71-415 Szczecin, Poland

* Corresponding author. Email: biolmagda@poczta.onet.pl

Abstract

The structure of repetitive sequences of the JNK block present in the pericentromeric region of the 2RL chromosome was studied in *Secale vavilovii*. Amplification of sequences present between the JNK sequences led to the identification of seven abnormal DNA fragments. Two of these fragments showed high similarity to the glutamate 5-kinase gene and putative alcohol dehydrogenase gene of trypanosomatid from the genus *Leishmania*, whose presence can be explained by horizontal gene transfer (HGT). Other fragments were similar to mitochondrial gene for ribosomal protein S4 in plants and to the glycoprotein (G) gene of the IHNV virus. Presumably, they are pseudogenes inserted into the JNK heterochromatin region. Within this region, also fragments similar to the rye repetitive sequence and chromosome 3B in wheat were found. There is no known mechanism that would explain how foreign sequences were inserted into the block region of tandem repetitive sequences of the JNK family.

Keywords

intercalary sequences; *Secale*; repetitive sequences; JNK family; heterochromatin; horizontal gene transfer

Introduction

Understanding the diversity, complexity, and evolution of eukaryotic genomes is impossible without an accurate knowledge of their repetitive sequences [1–3]. This especially applies to plants, in which the proportion of repetitive sequences in the genome varies significantly, and may even reach 97% of the nuclear DNA [4–6]. This translates to significant differences in the size of the genomes of these organisms [7,8]. It is believed that a large number of these sequences ensures the integrity of the genome in higher eukaryotes [9], inter alia, by affecting the expression of genes, organization of chromatin, processes such as recombination, DNA replication, or cell division [10]. Studying the distribution, origin, and function of repetitive sequences helps to understand their potential and role in the organization and evolution of eukaryotic genomes [11]. The results of recent studies have provided a lot of new information about the origin and functioning of repetitive sequences in the genome [12–16]. Repetitive DNA is highly polymorphic, which is caused by amplification and rapid evolution. This leads to changes in genomic structure and contributes to enlargement of the genome, thus affecting its complexity [17,18].

Repetitive sequences can occur in tandem or dispersed arrangement. Family of repetitive sequences can be commonly present in a given genus or be specific to a species or a chromosome. Repetitions may be present only in certain regions in the chromosome, as in the case of telomeres, or scattered throughout the genome [19]. Repetitive

sequences may occur in a dozen or a few thousand repetitions even within individual genus, e.g., family of JNK repetitive sequences in *Secale vavilovii* on chromosome 2R is repeated 4000 times, while it is dispersed in other rye species and repeated about 20 times [20,21]. The JNK sequence was named after rye (*Secale cereale* L.) landrace 'JNK', originally collected by Dr. H. Kishikawa in Japan [20].

The JNK sequence is a heterochromatin block identified in 2R chromosomes of a local Japanese cultivar, *Secale cereale* L. [20] and inbred lines of *Secale vavilovii* Grossh. [21] using Giemsa reagent. These are highly methylated sequences, rich in GC and composed of tandem repeats of 1200 bp in size [20]. Blocks of the JNK family are flanked by R173 repeats characteristic for rye. Downstream, R173 is adjacent to JNK in an antiparallel orientation, while upstream in a parallel orientation [22]. The JNK sequence shows high similarity to the fragment of 5S rDNA and the *Angela* retrotransposon. Probably it is a set of defective genes, which are deposited in one region on chromosome 2RL [23].

A study showed that the blocks of repetitive sequences do not have a uniform structure, and foreign sequences are often inserted between them [22]. An example can be the *Ty3-gypsy CRR* (centromeric retrotransposons of rice) rice retrotransposons and *CRM* (centromeric retrotransposons of maize) in corn, which preferentially transpose into the centromeric satellite DNA [24].

The exact mechanism that would explain how foreign sequences are tandemly inserted between repetitive sequences is unknown. Therefore, extending knowledge on the heterochromatin is invaluable particularly in the light of the present research which shows that these sequences are rich in regulatory RNA coding sequences, playing an important role in the epigenetic regulation of gene expression [25,26]. This fact indicates that the study of unknown sequences occurring in repetitive sequence regions is also necessary to understand the mechanisms of the genome control and phenotypic expression [27].

The current study aimed at understanding the structure of a block of repetitive JNK sequences. Revealing the structure of repetitive sequence blocks in detail may help in understanding the mechanism of accumulating them in one place in the genome, which is associated with decreased vigor and fertility of inbred *S. vavilovii* lines with the JNK block. It can be said that in order to know and understand the organization and functioning of the genome, its repetitive sequences must also be well understood.

Material and methods

Plant material

We performed studies with two inbred lines of *Secale vavilovii* Grossh. (109 and 116) obtained by self-pollination of plants, whose anthers and caryopses were mosaic colored, and in 2R chromosomes additional heterochromatin band was present [23,28,29]. Plants were cultivated in a plant house of the West Pomeranian University of Technology in Szczecin. Plants were cultivated in containers (volume: 10 L) filled with loamy sand with the addition of Azofoska fertilizer in the dose 2 g/L. Caryopses were sown to the containers in January in a greenhouse in which the temperatures were as follows: 20°C – day, 16°C – night. Young plants were exposed to vernalization (1–4°C) in a greenhouse. In March, plants were put outside the greenhouse and in a first decade of May were transferred again to the greenhouse. To ensure self-pollination of plants, appropriate cellophane insulators were applied. The drip system was used during the vegetation. Fungicides and insecticides were used symptomatically. Plants were fertilized twice, at the stage of tillering and booting (NPK fertilizer 20-20-20). Caryopses from the plants were collected and coleoptiles were obtained under laboratory conditions.

DNA isolation

Rye genomic DNA was extracted from etiolated coleoptiles (1 g per sample) as described by Kalinka and Achrem [22]. The purity and quantity of the extracted DNA was evaluated using the NanoDrop 2000c spectrophotometer (Thermo Scientific, Lithuania).

Amplification of sequences located among JNK tandem repeats

The sequences of primers used in the PCR reactions are given in Tab. 1. Various combinations of these primers were tested (Tab. 2). Each primer was designed for inverted PCR. Thus, by using forward and reverse primer designed for JNK sequence we were able to amplify each sequence interposed between two JNK repeats. Primers were designed [22] with the FastPCR software v. 6.2.89 [30,31], using an inverted PCR option, for a JNK sequence (GenBank: AB008922.1).

Tab. 1 JNK-specific primers.

Name	Sequence (5′→3′)	JNK region	Amplification direction
revGW1	TCCGCCCAGATCGAGGCTCGCCTGAATGAA	488–517	Upstream
revGW2	CGACAAGTGGCAGTGCGAATCTTGGCGCA	161–189	Upstream
revGW3	TCAGAGAGTGGCGTTTTGGCACACGGATGC	18–47	Upstream
forGW1	TTCATTCAGGCGAGCCTCGATCTGGGCGGA	488–517	Downstream
forGW2	TTCCTTGGTTGCCACTTGCCAGGATGGGAG	611–640	Downstream
forGW3	GTCGGCATCCACAGCTCCGGCTTCAGAGAT	816–845	Downstream

Tab. 2 Combinations of forGW and revGW primers used to amplify the sequences between the JNK sequences.

Forward primer	Reverse primer	PCR product lenght (bp)
forGW1	revGW2	893
forGW1	revGW3	751
forGW2	revGW1	1098
forGW2	revGW2	770
forGW2	revGW3	628
forGW3	revGW1	893
forGW3	revGW2	565
forGW3	revGW3	423

The expected length of the PCR product was included, assuming that two JNK sequences are adjacent to each other and are located as direct repeats (counted for 1192-bp-long JNK sequence, GenBank accession No. AB008922.1).

PCR reactions

Polymerase chain reactions were performed in 50-μL reaction mixture containing: 100 ng DNA, 1× Advantage 2 PCR Buffer [with $Mg(OAc)_2$ 35 mM], 0.2 mM dNTP (Fermentas, Lithuania), primer forward and reverse (each 0.2 μM), and 1× Advantage 2 Polymerase Mix (Clontech Laboratories, USA). The PCR temperature profile was as follows: seven cycles at 94°C for 20 s, 72°C for 3 min, and 32 cycles at 94°C for 25 s, 67°C for 3 min, ending at 67°C for 7 min. PCR was performed using Peltier Thermal Cycler (Bio-Rad, USA). Positive and negative controls were included in each experiment.

Gel electrophoresis. Amplification products were separated in 1.5% agarose gel in 1× TBE buffer at 80 V for approximately 3 h. Midori Green DNA Stain (Nippon Genetics,

Germany) was added to the gel to final concentration suggested by the manufacturer. The gel images were captured with the Gel Doc XR system (Bio-Rad). The size of the products was determined by comparison with MASSRULER DNA Ladder Mix (Thermo Scientific). Bands were scored and analyzed with Quantity One software (Bio-Rad).

Sequencing. Distinct bands of different lengths than the expected (contiguous sequences JNK; Tab. 2) were excised from the gel and purified using PrepEase Gel Extraction Kit (Affymetrix, USA) according to the manufacturer's instructions. Sequencing was performed at Genomed S.A. (Poland).

The sequences were analyzed using BLASTN 2.2.26+ [29] software.

Results

The PCR reaction, in addition to the products of expected size, i.e., the adjacent JNK sequences, amplified sequences of different size (Fig. 10). Seven atypical fragments were identified located between the sequences of the JNK family.

One of the fragments with a length of 551 bp (accession No. KY853403) showed very high identity (93%) to the DNA sequence of chromosome 26, of *Leishmania braziliensis* (Fig. 1, Tab. 3) and *Leishmania panamensis*. Within this sequence, a region in between 157–551 nt was 96% similar to the mRNA sequence of the putative glutamate 5-kinase *Leishmania braziliensis* gene (Fig. 2, Tab. 3). The entire mRNA sequence of this gene in *Leishmania* has 795 nt and the identified fragment overlapped with the initial 394 nt (71% overlap) of this sequence. Identity between the two sequences is very high (96%).

A sequence of 827 bp (accession No. KY853404) located between the JNK showed the greatest similarity to the putative alcohol dehydrogenase mRNA of *Leishmania mexicana* (Fig. 3, Tab. 3). The gene encoding this protein is located on chromosome 29, while in other *Leishmania* species (Tab. 3) on chromosome 30. The entire mRNA sequence of this gene in *Leishmania* has a length of 1200 nt, whereas the overlap of the detected 827-bp fragment between 12–862 nt of the dehydrogenase sequence reaches 100% in most species (Tab. 3).

Within the JNK sequences, a sequence was located with a very high similarity to mRNA of the mitochondrial gene encoding ribosomal protein S4 in many plant species, in particular of the family Brassicaceae. The highest identity was found for *Arabidopsis thaliana* gene (96%). The mRNA of this gene has a size of 1089 nt (Tab. 3), and the identified fragment in *Secale* had a length of 376 bp (accession No. KY853405) (Fig. 4), thus this was not the full sequence. The gene encoding this protein has a rather conserved structure, as the analyzed fragment also showed identity to the homologous sequences in other plant species, including *Brassica napus* (94% identity), *Medicago truncatula* (92% identity), *Fragaria iinumae* (91% identity), and *Beta vulgaris* (88% identity).

The analysis also showed the presence of a very unusual fragment (423 bp – accession No. KY853406), having a certain similarity to the glycoprotein gene (G) of infectious hematopoietic necrosis virus (IHNV) of salmonids. This identity was in the range of 84–87% in relation to the sequences present in the GenBank under different access numbers. The highest identity (87%) concerned the glycoprotein gene of the Cro/05 virus strain (Fig. 5). The entire coding sequence of this gene had a length of 1527 nt, the identity applied to
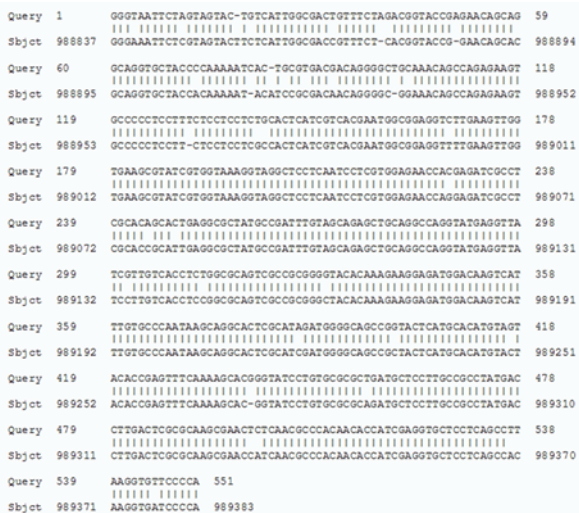

Fig. 1 Comparative alignment of the identified sequence (query) to the sequence of the 26th chromosome of *Leishmania braziliensis* (GenBank accession No. FR799001.1).

Fig. 2 Comparative alignment of the identified sequence (query) to the *Leishmania braziliensis* putative glutamate 5-kinase mRNA (GenBank accession No. XM_001562446.1).

Tab. 3 Sequences showing significant sequence homology to the identified sequence.

Sequence	Description	Range in the sequence	Query coverage	E values	Identities	Gaps	Accession No.
551 bp KY853403	*Leishmania braziliensis* MHOM/BR/75/M2904 – chromosome 26; length: 992961	Range 1: 988837–989383	100%	0.0	516/553 (93%)	8/553 (1%)	FR799001.1
		Range 2: 979844–980387			515/553 (93%)	11/553 (1%)	
	Leishmania panamensis MHOM/PA/94/PSC-1 – chromosome 26; length: 1024527	Range: 1020392–1020934	99%	0.0	509/550 (93%)	10/550 (1%)	CP009395.1
	Leishmania braziliensis MHOM/BR/75/M2904 – putative glutamate 5-kinase mRNA partial mRNA; length: 795	Range: 1–394	71%	7e−178	378/395 (96%)	1/395 (0%)	XM_001562446.1
	Leishmania braziliensis MHOM/BR/75/M2904 – putative glutamate 5-kinase mRNA partial mRNA; length: 795	Range: 1–394	71%	7e−178	378/395 (96%)	1/395 (0%)	XM_001562443.1
	Leishmania panamensis glutamate 5-kinase, putative partial mRNA; length: 795	Range: 1–394	71%	3e−176	377/395 (95%)	1/395 (0%)	XM_010701785.1
827 bp KY853404	*Leishmania mexicana* MHOM/GT/2001/U1103 – putative alkohol dehydrogenase partial mRNA; length: 1200	Range: 12–862	100%	0.0	790/852 (93%)	26/852 (3%)	XM_003877304.1
	Leishmania donovani alkohol dehydrogenase, putative (LDBPK_302100) mRNA, complete cds; length: 1200	Range: 12–862	100%	0.0	762/852 (89%)	26/852 (3%)	XM_003862874.1
	Leishmania infantum JPCM5 putative alkohol dehydrogenase partial mRNA; length: 1200	Range: 12–862	100%	0.0	761/852 (89%)	26/852 (3%)	XM_001467020.1
	Leishmania major strain Friedlin putative alkohol dehydrogenase (LMJF_30_2090); mRNA, complete cds; length: 1200	Range: 12–862	100%	0.0	759/853 (89%)	28/853 (3%)	XM_001684762.1
423 bp KY853405	Infectious hematopoietic necrosis virus strain Cro/05 glycoprotein (G) gene, complete cds	Range: 655–1071	100%	1e−125	366/423 (87%)	6/423 (1%)	EU219616.1

Tab. 3 Continued

Sequence	Description	Range in the sequence	Query coverage	E values	Identities	Gaps	Accession No.
376 bp KY853406	*Arabidopsis thaliana* ribosomal protein S4 mRNA, complete cds; mitochondrial; length: 1089	Range: 169–545	100%	1e−169	361/377 (96%)	1/377 (0%)	EU999008.1
	Brassica nigra mito-chondrial DNA, com-plete sequence; length: 232145	Range: 147039–147415	100%	1e−159	355/377 (94%)	1/377 (0%)	AP012989.1
	Brassica napus ribo-somal protein S4 rps4 gene; complete cds; length: 2269	Range: 913–1289	100%	1e−159	355/377 (94%)	1/377 (0%)	AF055707.1
	Oryza sativa Indica Group mitochondrion, complete genome; length: 637692	Range 1: 35161–35499	89%	1e−120	303/339 (89%)	1/339 (0%)	JF281153.1
		Range 2: 452590–452928			303/339 (89%)	1/339 (0%)	
770 bp KY853407	*Secale cereale* clone 16 tandem repeat region; length: 984	Range: 104–884	100%	0.0	722/781 (92%)	11/781 (1%)	AY522382.1
	Triticum aestivum chromosome 3B, genomic scaffold, cul-tivar 'Chinese Spring'; length: 774434471	Range 1: 176123151–176123833	98%	0.0	589/692 (85%)	16/692 (2%)	HG670306.1
		Range 2: 328034791–328035521			617/738 (84%)	14/738 (1%)	
		Range 3: 401473601–401474352			632/760 (83%)	17/760 (2%)	
		Range 4: 329796729–329797425			598/707 (85%)	22/707 (3%)	
	Triticum aestivum cDNA, clone: WT009_G09, cultivar 'Chinese Spring'; length: 1357	Range: 633–1130	64%	3e−128	423/505 (84%)	13/505 (2%)	AK334095.1
	Secale cereale Re-volver-1 pseudogene for transposase; length 4000	Range: 3642–4000	47%	6e−165	348/362 (96%)	3/362 (0%)	AB124639.1
772 bp KY853408	*Triticum aestivum* chromosome 3B, ge-nomic scaffold, cultivar 'Chinese Spring'	Range: 555138951–555139737	100%	3e−173	642/794 (81%)	29/794 (3%)	HG670306.1
	Triticum turgidum subsp. *durum* clone BAC 219E24 genomic sequence	Range: 151583–152272	89%	4e−152	561/695 (81%)	11/695 (1%)	EF081027.1

Tab. 3 Continued

Sequence	Description	Range in the sequence	Query coverage	*E* values	Identities	Gaps	Accession No.
	Triticum aestivum chromosome 3B specific BAC library, BAC clone TaaCsp3BFhA_0111F24	Range: 1425–1523	12%	2e−25	89/99 (90%)	0/99 (0%)	HF541873.1
808 bp KY853409	*Secale cereale* clone HK1-59 repeat region	Range: 782–536	97%	0.0	686/787 (87%)	35/787 (4%)	KY327931.1
	Triticum aestivum chromosome 3B, genomic scaffold, cultivar 'Chinese Spring'	Range: 484582015–484582807	100%	0.0	684/816 (84%)	31/816 (3%)	HG670306.1
	Triticum aestivum chromosome 3B-specific BAC library, contig ctg1030b	Range: 549454–550226	99%	0.0	659/793 (83%)	23/793 (2%)	FN564435.1

Fig. 3 Comparative alignment of the identified sequence (query) to the *Leishmania mexicana* putative alcohol dehydrogenase mRNA (GenBank accession No. XM_003877304.1).

Fig. 4 Comparative alignment of the identified sequence (query) to the *Arabidopsis thaliana* ribosomal protein S4 mRNA (GenBank accession No. EU999008.1).

Fig. 5 Comparative alignment of the identified sequence (query) to the infectious hematopoietic necrosis virus strain Cro/05 glycoprotein (G) gene (GenBank accession No. EU219616.1).

```
Query  181  GACTGTATGACTTGAGTGTTGGGTCATGTGACCCCTACCTGTATGGATATGTTATGTATG  240
            ||||||||||||  ||||||  ||||||||| ||||||||||||| ||||||||||||||
Sbjct  284  GACTGTATGACTTGTGTGTTGGGTCATGGGACCCCTACCTGTACGGATATGTTATGTATG  343

Query  241  GCTCTTGGAGGCCTCTTAAATAAAGTACTTTGAGTCATAGAGTTCTGTTGTGATGCAATGT  300
            ||||||||||||||  |||||||||||||||||| ||||||||||||||||||||||||||
Sbjct  344  GCTCTTGGAGGCCTTTTAAATAAAGTACTTTGAGTCATAGAGTTCTGTTGTGATGCAATGT  403

Query  301  TGTATTTGCACATATTGAGCATA-TTGCGTGTATGTTTGAAATGCTTGGTATGTGTGGGA  359
            |||||||||||||||||||||||| ||| | || |||||||||||||||||| ||||||||
Sbjct  404  TGTATTTGCACATATTGAGCATATTTGTGTGAATGATTGAAATGCTTGGTATGTGTGGGA  463

Query  360  TCCGACACCTAGTTGTTTTCCTTTGGTAGCCTACCTTACAGGGAAATGCCTCCTAGTGAT  419
            |||||||||||||||||||||||||||||||||  |||||||||||||||||||||||| |
Sbjct  464  TCCGACACCTAGTTGTTTTCCTTTGGTAGCCTATCTTACAGGGAAATGCCTCCTAGTGTC  523

Query  420  TCCAATGAGCCTTGGTAGCTTGCTACTGCTCGGAATACGTAGGCTGGCCGGCATGTGTCC  479
            || |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||
Sbjct  524  TCTAATGAGCCTTGGTAGCTTGCTACTGCTCGGAACACGTAGGCTGGCCGGCATGTGTCC  583

Query  480  TTCTTCGTTCCTGTGTCTGTCCCTTAGGGGAAATGTCACGCGTTGTTCCTTTAAG-----  534
            |||||||||||||||| |||||||| |||||||||||||||||||| |||||||||
Sbjct  584  TTCTTCGTTCCTGTGTCTGTCCCTTACGGGAAATGTCACGCGTGTTCCTTTACGGGAAA  643

Query  535  T---CCAT-GTAGCTTGCTACGACTTGGGTTCATACGTTGATGACCGACACGTTCGTTGC  590
            |    || |  ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  644  TGAGCCTTGGTAGCTTGCTACGACTTGGGATCATACGTTGATGACCGACACATTCGTTGC  703

Query  591  TCGGTCATGGATGCCTGTCTCGTGTAATCTAGTGCCACTTTTGGTTTATAACTAGCCATGT  650
            | |||||||||||||||||||||| |||| |||||||||||||||||||||||||||||||
Sbjct  704  TGGGTCATGGATGCCTGTCTCGTGTAATTTAGTGCCACTTTTGGTTTATAACTAGCCATGT  763

Query  651  CGGCCCGGGTTSTCTGTCATATGGAWGCYAGCGACACTGTCATTTKRGRGYGCCAAAATG  710
            |||||||||||| |||| ||||||||| |||||||||||||||| ||| | ||||||||||
Sbjct  764  CGGCCCGGGTTCTCTATCATATGGATGCTAGCGACACTGTCATATACGTGTGGCAAAATG  823

Query  711  CCCAAACGSGCATRGGCCAGGTAAGGGAGCACCAGAGCGGTCACCGTGAA-GAGGCCGCA  769
            | ||||| ||||| |||||||||||| |||||||  ||||||||||||||| ||||||||||
Sbjct  824  CGCAAACGGTCCCGGGCCAGGTAAGGTGGCACCCGTGGGGTTACCGTGCGTGAGGCCGCA  883

Query  770  A  770
            |
Sbjct  884  A  884
```

Fig. 6 Comparative alignment of the identified sequence (query) to the *Secale cereale* repetitive sequence (GenBank accession No. AY522382.1).

```
Query  1     GAGGAAAACCCTGCCYCYCTGCTGATGGGTTMTACATAGACGTTGACGTCGACGAGTAGCTT  60
             ||||| |||||||| |||||||||||||||||| ||||||||||| ||||| |||||||||||
Sbjct  3642  GAGGAGAACCCTGCCCCTGCTGATGGGTTCTACATAGACGGTGACGGCGACGAGTAGCTT  3701

Query  61    GACACCCAGACGGAGATCTGGAGCTTGTAAAGGGCCTTGTAGATAGTCAGGCTACACCAA  120
             |||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3702  GACA-CCAGACGGAGATCTGGAGCTTGTGAAGGGCCTTGTAGATAGTCAGGCTACACCAA  3760

Query  121   GCCTGTTTTATTTTCTAAGTTGTCTGTACTCAGACATGTAGCTTCCGCTTGTGCTTGTAT  180
             |||||||||||||||||||||||||||||||||||||||||||| |||||||||||||||
Sbjct  3761  GCCTGTTTTATTTTCTAAGTTGTCTGTACTCAGACATGTAGCTTTCGCTTGTGCTTGTAT  3820

Query  181   GACTGTATGACTTGAGTGTTGGGTCATGTGACCCCTACCTGTATGGATATGTTATGTATG  240
             |||||||||||||||| ||| ||||||||||||||||| ||||||||||||||||||||||
Sbjct  3821  GACTGTATGACTTGAGTGTT-GGTCATGTGACCCCAACCTGTATGGATATGTTATGTATG  3879

Query  241   GCTCTTGGAGGCCTCTTAAATAAAGTACTTTGAGTCATAGAGTTCTGTTGTGATGCAATGT  300
             |||||||||||||||||||||||||||||||| |||||||||||| |||||||||||||||
Sbjct  3880  GCTCTTGGAGGCCTCTTAAATAAAGTACTTTGAATCATAGAGTTCTGTTGTGATGCAATGT  3939

Query  301   TGTATTTGCACATATTGAGCATATTGCGTGTATGTTTGAAATGCTTGGTATGTGTGGGAT  360
             ||||||||||||||||||||||||||| |||||| |||||||||||||||||| |||| ||
Sbjct  3940  TGTATTTGCACATATTGAGCATATTGCGTGTATGATTGAAATGCTTGGTATGTCTGG-AT  3998

Query  361   CC  362
             ||
Sbjct  3999  CC  4000
```

Fig. 7 Comparative alignment of the identified sequence (query) to the *Secale cereale* Revolver-1 pseudogene for transposase (GenBank accession No. AB124639.1).

366 of 423 nt (87%, *E* value = 1e−125) at the number of gaps amounting to 6 (1%).

Among the identified sequences, there also was a large fragment of a previously identified repetitive rye sequence (Fig. 6). The identified fragment was slightly shorter (770 bp – accession No. KY853407) than the sequence in the GenBank (984 bp), but the identity was high (92%; Tab. 3). A similar sequence is located in the wheat genome in chromosome 3B (Tab. 3). Interestingly, cDNA sequence data also exist from this wheat region (Tab. 3). Part of this sequence was also identified in rye as a transposase pseudogene of the Revolver-1 element (Fig. 7, Tab. 3). The analyzed fragment is similar to the terminal portion of this pseudogene.

Two of the identified fragments (772 bp – accession No. KY853408 and 808 bp – accession No. KY853409, respectively) showed similarity to different DNA regions of the wheat chromosome 3B (Fig. 8, Fig. 9). In both cases, many fragments with different overlap were found, whereas the maximum identity was 81–84% at 100% overlap of the sequence. The sequence similarity was significant (*E* values <1e−50). Based on recent studies it can be concluded, that 808-bp-long fragment is a part of repetitive motif (1759-bp-long) found also in *Secale*. The identity of 808-bp sequence to the newly described motif (KY327931.1) in *Secale cereale* amounted 87% (*E* value = 0).

Discussion

A thorough analysis of the JNK sequence block in the *S. vavilovii* genome showed that it is flanked by R173-2 or R173-3 sequences characteristic of rye. In addition, it was found that the JNK sequences did not form a homogeneous arrangement [22]. The research has shown that between the tandemly arranged sequences, there are other sequences, showing often very high similarity to gene sequences characteristic of other organisms. Amplification of the sequences lying between the JNK sequences showed that two atypical fragments present between the sequences of the JNK family are similar to the genes (alcohol dehydrogenase and glutamate 5-kinase) occurring in trypanosomes of the genus *Leishmania*. It may seem very strange, but it should be kept in mind that the family Trypanosomatidae comprises protozoan parasites that infect not only animals, but also different plants [32,33]. Different insects serve as vectors for plant Trypanosomatidae [34–37]. In 2014, Porcel et al. [38] described the genomic sequence of two plant trypanosomes. As shown by comparison of the genomes of the two parasites with other trypanosomes, they shared a common simplified organization of the genome. The genomes of Trypanosomatidae are organized into large policistronic gene clusters [39–41]. What is characteristic, almost all protein-coding genes lack introns. Introns were found only in poli(A) polymerase and ATP-dependent RNA helicase (DEAD/H) genes in *T. brucei*, *T. cruzi*, and *Leishmania* spp. However, in *Phytomonas* even these genes do not have introns [38]. This fact supports the results obtained in this study, since the identified sequences were similar to both DNA and mRNA of alcohol dehydrogenase and glutamate 5-kinase. It can be hypothesized that there are/ were some trypanosomes of the genus *Phytomonas*, which are/were parasites of rye and horizontal gene transfer (HGT) has occurred of some sequences from the genome of these protozoa into the plant genome. *Phytomonas* parasitize on plants of the family Poaceae [42], and although according to the current data, there are no trypanosomes

Fig. 8 Comparative alignment of the 772-bp-long sequence (query) to the *Triticum aestivum* chromosome 3B sequence (GenBank accession No. HG670306.1).

Fig. 9 Comparative alignment of the 808-bp-long sequence (query) to the *Triticum aestivum* chromosome 3B sequence (GenBank accession No. HG670306.1).

Fig. 10 Agarose (1.5%) gel electrophoresis of PCR-amplified fragments located among JNK tandem repetitive sequences. M – MASSRULER DNA Ladder Mix #SM0403 (Thermo Scientific); 1–5 – representative PCR product sets obtained using different combinations of primers: forGW2/revGW2 (1), forGW3/revGW2 (2), forGW3/revGW3 (3), forGW1/revGW2 (4), forGW2/revGW2 (5); the most prominent bands (marked with the arrows) represent the DNA fragments, in which two JNK sequences adjoin directly; longer fragments were excised from the gel, purified, and sequenced.

detected infecting rye, they have been discovered in maize [42,44], in which they reproduce in kernels. Because trypanosome genes (with some exceptions) do not contain introns and the sequences that we found in JNK sequence block are partial, it is difficult to say whether the copy of DNA or cDNA has been inserted.

The identification of fragments similar to the G glycoprotein gene of infectious hematopoietic necrosis virus (IHNV) of salmonids seem equally unusual as finding sequences similar to *Leishmania* genes. However, a similar hypothesis can also be postulated in this case. IHNV is a member of the genus *Novirhabdovirus* and belongs to the family Rhabdoviridae [45,46]. The family Rhabdoviridae currently includes six genera: *Vesiculovirus*, *Lyssavirus*, and *Ephemerovirus*, which were isolated from various animal species [47], *Novirhabdovirus* infecting fish, and *Cytorhabdovirus* and *Nucleorhabdovirus* infecting arthropods and plants. Genomes of all rhabdoviruses are a single-stranded RNA (−) which encodes five structural proteins designated as N, P, M, G, and L, which are: nucleoprotein (N), phosphoprotein (P), matrix protein (M), glycoprotein (G), and polymerase (L). [48]. Division of each genus into species is supported by comparisons of nucleotide and amino acid sequences of one or more genes [49,50], however, the sequence of the entire genome is available only for a few species. This might also explain why the sequence identified in rye is similar to the gene found in the virus occurring in fish. The similarity of rhabdoviruses infecting plants and fish should be established. Bourhy et al. [51] conducted a phylogenetic analysis of 38 rhabdoviruses of all genera based on the sequence of the polymerase gene (L). As a result of this study, it was found that the viruses classified to the genera *Novirhabdovirus*, *Cytorhabdovirus*, and *Nucleorhabdoviruses* were closely related and formed one clade in the dendrogram, evidently differing from the other three genera that formed the second clade. This fact suggests that the fragment identified in rye genome, showing a strong similarity to the glycoprotein gene (G) of IHNV, would probably also be very similar to that gene of plant viruses, but such data are not available in the GenBank.

A sequence has been also identified within the JNK sequence block showing very high sequence similarity to sequence of the mitochondrial ribosomal protein S4 of

Arabidopsis thaliana involved in translation. In this case, it is obvious that in plants the protein coding sequences may be similar.

Some mitochondrial genes exist in multiple copies in plant mitochondrial genomes due to the rearrangement of the genome or a gene duplication. An example might be *Petunia* gene *atp9* [52] or the *cox2* gene in *Petunia* and sugar beet [53,54]. The latter has been transferred from the mitochondrial genome into the nuclear genome in plants [55]. It is believed that after the transfer into the nuclear genome, these genes are inactivated, which then often leads to the loss of their sequence. There is some kind of the transition state, in which both copies, located in the nucleus and the mitochondrion, coexist and are fully functional. After some time, redundant copy is inactivated or eliminated from the genome. Remnants of ribosomal proteins pseudogenes (or their total absence) in mtDNA correlated with recently transferred functional copies to the nucleus [56]. The example of gene undergoing the transition state is ribosomal protein L5 gene in wheat [57]. The copy transferred the nuclear genome can take over the function, then the mtDNA copy degenerates. Alternatively, if nuclear copy loses its function during transition state, a copy located in mtDNA retains function. A good example of these both possibilities is gene encoding S19 protein. In most of the cereals, this gene is present only in nuclear DNA, but in rice it is located in mitochondrial genome [58]. Here, we report for the first time about the presence of ribosomal protein S4 pseudogene in the nuclear DNA. The sequence shows similarity both to DNA and mRNA because *rps4* do not contain introns [56]. As indicated above, it could be the remnant of the functional copy transferred from the mtDNA. It is quite surprising, as it is believed, that the rps4 protein is encoded only the gene located exclusively in mitochondrial genome [56]. There are no literature data about presence of this gene's sequence in nuclear genome. Successful transfers between mitochondrial and nuclear genomes are a rare phenomenon, the examples are *rps14* and *rps19* [59]. The most common cause of ineffective transfer is displacement of incomplete sequence. This may perfectly explain the presence of defective sequences in the nuclear DNA [60].

Sequences similar to mitochondrial sequences located in the nuclear genome are called NUMTs (nuclear mitochondrial pseudogenes) and have been localized in the genomes of eukaryotes [61–63]. Some of these sequences are highly conserved, other contain deletions or duplications. They vary in size and are randomly distributed in different chromosomes, and may be arranged in tandems. In many plant genomes, scientists identified large NUMT insertions [64], but also short sequences, which are sometimes organized in clusters [65,66]. Although occasionally NUMT genes become functional in the nucleus [67], most insertions of organelle DNA in the nucleus is mutated and partially deleted [66]. Studies in many plant species have shown that NUMTs are located preferentially in the centromeres in species with a small genome (*Arabidopsis* and rice) [68] or on B chromosomes in rye [69], while in species with a larger genome they are more fragmented and located along the entire chromosomes, as in additional heterochromatin in the long arm of chromosome 2R in *S. vavilovii*. Perhaps the sequence showing high similarity to the mRNA sequence of the mitochondrial ribosomal protein S4 in *S. vavilovii*, similar to the genes of plant species such as *Arabidopsis thaliana*, *Brassica nigra*, *B. napus*, *Medicago truncatula*, *Fragaria iinumae*, or *Beta vulgaris*, is a NUMT gene that has mutated and was inserted into the heterochromatin region. Although the NUMT inserts have been considered little functional [70,71], a study in humans has shown that by mutagenicity they may contribute to the damage of the functional genome integrity [72].

The wide variety of sequences types which are inserted among tandem repetitive JNK sequences is astonishing. In such cases, a possibility of contamination should be always considered. The are some reports showing the sources of contamination. For example, there is high risk of contamination if the DNA is isolated from roots. It is very difficult to remove the rhizospere from roots. The rhizosphere contains many bacteria, protozoa, and nematodes [73]. Being aware of the possibility of error, we isolated DNA from the coleoptiles grown under sterile conditions, and good laboratory practice was followed throughout the procedure as well. However, it could be still not enough to avoid errors caused by contamination. Also, when the sequences of nuclear DNA are studied by PCR reaction, there is a risk to amplify the sequences from cpDNA or mtDNA. The primers were designed to repetitive elements, thus, the possibility of unspecific annealing was high. We reduced it significantly by designing long primers

with high melting temperature and, as a control, nested PCR was performed. Moreover, we analyzed only these PCR products, which were flanked by JNK sequence fragments. Doing this, we were confident that we found sequences inserted among two JNK copies in the genome of *S. vavilovii*. In the case of such analyses, it is particularly important to avoid errors [74] and make false assumptions about horizontal gene flow [75].

It is easier to explain the presence of rye or wheat sequences within the repeated JNK blocks. Research has shown that between the tandemly arranged sequences, there are other sequences, showing often very high similarity to gene sequences characteristic of other organisms. This is confirmed by numerous studies, and it is believed that mobile genetic elements are most frequently observed within repetitive sequences. An example is heterochromatin knobs in the corn genome composed of 180-bp tandem repeats that are separated by retrotransposon sequences [76] or location of the *Athila* retrotransposon at 180-bp repeats in pericentromeric regions of all *Arabidopsis* chromosomes [77]. The presence of inserts between the repetitive sequences has been also found in rye chromosomes. Fragments were tested between the pSc200 and pSc250 monomers [78], and the obtained fragments were in the length range from 35 to 250 bp and showed similarity (80–90%) to various repetitive elements described earlier, such as barley *BAGY-1* [79] or *cereba* retrotransposons [80], or sequences associated with telomeres in wheat [78,81]. A fragment similar to the terminal portion of the transposase pseudogene Revolver-1 element has been identified among the sequence occurring within the JNK blocks. The *Revolver* sequence is present in high copy number in some species. Twenty thousand copies of this sequence were found in *Dasypyrum villosum* and *Secale* sp. as well as in diploid *Triticum monococcum* species or tetraploid wheat species [27]. However, in the hexaploid wheat, the number of copies of *Revolver* element is very low [82]. Another study has shown that the *Revolver* family of mobile elements in rye is highly dispersed and clones of this element are chromosomally-specific. It has been found that those elements possess transcriptional activity in rye. There is a vast variation in the copy number, length, and level of activity among these elements in different cereal genomes, clearly indicating that they are in constant evolution [27]. Many mobile genetic elements are present in high copy number in the eukaryotic genomes, but the vast majority of them is inactive, and only a small part of them retains the ability to transpose [17,83]. The *Revolver* element is characterized by a very high variation (especially with respect to the length of the sequence), number, and activity between the different cereal genomes, which indicates its constant evolution [27]. During the evolution, the majority of inserted elements become truncated and non-functional. Many plants show the strong tendency to enlarge the size of the genome. Most often it is accomplished by amplification, rearrangements, differentiation, and insertions of repetitive sequences. The perfect example is large subtelomeric heterochromatin blocks in *Secale*, which consist of rye-specific repetitive sequences [84]. During the evolution of tandemly repeated sequences, insertion of different types of sequences may occur. The evolution of heterochromatin blocks is coupled with an enrichment in foreign sequences which were inserted by transposition or RNA-dependent DNA transfer [85,86]. Viral sequences are thought to integrate rarely to plant genome and it can be achieved either by the viral integrase or by illegitimate recombination [87]. The newly inserted sequence starts to evolve together with the tandem repetition and may become a part of a new unit, which will be then amplified [88]. Foreign sequences can spread throughout the genome by unequal recombination or gene conversion [85,89]. Insertion of coding sequence into tandem repetitive sequences usually does not favor its expression, these regions are usually epigenetically silenced. This may protect genome from potential instability and genetic redundancy. In this work, we found different types of sequences inserted among tandem repetitive motifs. It is not clear whether this is the original site of their insertion or they appeared as a result of, e.g., irregular recombination. Due to the fact that unequal crossing-over is assumed to be the primary evolutionary force acting on satellite sequences [9], there is a high possibility that JNK sequence block is the secondary location of atypical insertions. Along with this, a new question arises, especially in the case of pseudogenes, whether they were initially functional or if there exist other copies in the rye genome. According to the literature data most of the transfers are nonfunctional and evolve neutrally [25–27]. They can be characterized as dead-on-arrival horizontal gene transfer (doaHGT). Further studies are needed to answer these questions.

Conclusions

The study helped to better understand the structure of repetitive sequences of the JNK block. The work demonstrated a remarkable complexity of this heterochromatin region, even though it did not allow to understand the mechanism which would explain how foreign sequences were inserted within the blocks of repetitive sequences of the JNK family. It can be hypothesized that they are fragments of different genes or pseudogenes, woven into heterochromatin sequences that silence them, as it is known that the active pseudogenes can regulate the activity of genes by ncRNA. In addition, it can be assumed that the R173 sequence, which flanks the JNK block sequences, acts as a barrier that does not allow the spread of the heterochromatin state to adjacent regions [22].

References

1. Shapiro JA, von Sternberg R. Why repetitive DNA is essential to genome function. Biol Rev. 2005;80:227–250. https://doi.org/10.1017/S1464793104006657

2. Bennetzen JL. Patterns in grass genome evolution. Curr Opin Plant Biol. 2007;10:176–181. https://doi.org/10.1016/j.pbi.2007.01.010

3. Morgante M, de Paoli E, Radovic S. Transposable elements and the plant pan-genomes. Curr Opin Plant Biol. 2007;10:149–155. https://doi.org/10.1016/j.pbi.2007.02.001

4. Flavell RB, Bennett MD, Smith JB, Smith DB. Genome size and proportion of repeated nucleotide-sequence DNA in plants. Biochem Genet. 1974;12:257–269. https://doi.org/10.1007/BF00485947

5. Murray MG, Peters DL, Thompson WF. Ancient repeated sequences in the pea and mung bean genomes and implications for genome evolution. J Mol Evol. 1981;17:31–42. https://doi.org/10.1007/BF01792422

6. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Behan A, et al. Nested retrotransposons in the intergenic regions of the maize genome. Science. 1996;273:765–769. https://doi.org/10.1126/science.274.5288.765

7. Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res. 2006;16:1262–1269. https://doi.org/10.1101/gr.5290206

8. Neumann P, Koblizkova A, Navratilova A, Macas J. Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. Genetics. 2006;173:1047–1056. https://doi.org/10.1534/genetics.106.056259

9. Mehrotra S, Goyal V. Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. Genomics Proteomics Bioinformatics. 2014;12:164–171. https://doi.org/10.1016/j.gpb.2014.07.003

10. Li W, Zhang P, Fellers JP, Friebe B, Gill BS. Sequence composition, organization, and evolution of the core Triticeae genome. Plant J. 2004;40:500–511. https://doi.org/10.1111/j.1365-313X.2004.02228.x

11. Plohl M. Those mysterious sequences of satellite DNAs. Period Biol. 2010;112:403–410.

12. Jurka J, Kapitonov VV, Kohany O, Jurka MV. Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet. 2007;8:241–259. https://doi.org/10.1146/annurev.genom.8.080706.092416

13. Jiang J, Birchler JA, Parrott WA, Dawe RK. A molecular view of plant centromeres. Trends Plant Sci. 2003;8:570–575. https://doi.org/10.1016/j.tplants.2003.10.011

14. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 2013;14:R10. https://doi.org/10.1186/gb-2013-14-1-r10

15. Dechyeva D, Schmidt T. Molecular organization of terminal repetitive DNA in *Beta* species. Chromosome Res. 2006;14:881–897. https://doi.org/10.1007/s10577-006-1096-8

16. Burgess-Beusse B, Farrell C, Gaszner M, Litt M, Mutskov V, Recillas-Targa F, et al. The

insulation of genes from external enhancers and silencing chromatin. Proc Natl Acad Sci USA. 2002;99:16433–16437. https://doi.org/10.1073/pnas.162342499

17. Kumar A, Bennetzen JL. Plant retrotransposons. Annu Rev Genet. 1999;33:479–532. https://doi.org/10.1146/annurev.genet.33.1.479

18. Heslop-Harrison JS. Comparative genome organization in plants: from sequence and markers to chromatin and chromosomes. Plant Cell. 2000;12:617–635. https://doi.org/10.1105/tpc.12.5.617

19. Rao SR, Trivedi S, Emmanuel D, Merita K, Hynniewta M. DNA repetitive sequences-types, distribution and function: a review. J Cell Mol Biol. 2010;8:1–11. https://doi.org/10.1016/j.gpb.2014.07.003

20. Nagaki K, Tsujimoto H, Saskuma T. A novel repetitive sequence, termed the JNK repeat family, located on an extra heterochromatic region of chromosome 2R of Japanese rye. Chromosome Res. 1999;6:95–101. https://doi.org/10.1023/A:1009226612818

21. Rogalska SM, Achrem M, Słomińska-Walkowiak R, Filip E, Skuza L, Pawłowska J, et al. Polymorphism of heterochromatin bands on chromosomes of rye *Secale vavilovii* Grossh. lines. Acta Biol Crac Ser Bot. 2002;44:111–117.

22. Kalinka A, Achrem M. Analysis of the flanking sequences of the heterochromatic JNK region in *Secale vavilovii* Grossh. chromosomes. Biol Plant. 2015;59:637–644. https://doi.org/10.1007/s10535-015-0531-0

23. Achrem M, Rogalska SM, Kalinka A. Possible ancient origin of heterochromatic JNK sequences in chromosomes 2R of *Secale vavilovii* Grossh. J Appl Genet. 2010;51(1):1. https://doi.org/10.1007/BF03195704

24. Nagaki K, Cheng Z, Ouyang S, Talbert PB, Kim M, Jones KM, et al. Sequencing of a rice centromere uncovers active genes. Nat Genet. 2004;36:138–145. https://doi.org/10.1038/ng1289

25. Sunker R, Girke T, Zhu JK. Identification and characterization of endogenous small interfering RNAs from rice. Nucleic Acids Res. 2005;33:4443–4454. https://doi.org/10.1093/nar/gki758

26. Siomi H, Siomi MC. Interactions between transposable elements and Argonautes have (probably) been shaping the *Drosophila* genome throughout evolution. Curr Opin Genet Dev. 2008;18:181–187. https://doi.org/10.1016/j.gde.2008.01.002

27. Tomita M, Okutani A, Beiles A, Nevo E. Genomic, RNA, and ecological divergences of the *Revolver* transposon-like multi-gene family in Triticeae. BMC Evol Biol. 2011;11:269. https://doi.org/10.1186/1471-2148-11-269

28. Rogalska SM, Cybulska-Augustyniak J, Mroczyk W, Cierniewska A, Obuchowski W. Investigations of contents of proteins and pentosans in kernels of rye *Secale vavilovii* Grossh. Bulletin IHAR. 1993;187:93–98.

29. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7:203–214. https://doi.org/10.1089/10665270050081478

30. Kalendar R, Lee D, Schulman AH. FastPCR software for PCR, in silico PCR, and oligonucleotide assembly and analysis. DNA cloning and assembly methods. Methods Mol Biol. 2014;1116:271–302. https://doi.org/10.1007/978-1-62703-764-8_18

31. Kalendar R, Lee D, Schulman AH. Java web tools for PCR, in silico PCR, and oligonucleotide assembly and analysis. Genomics. 2011;98(2):137–144. https://doi.org/10.1016/j.ygeno.2011.04.009

32. Camargo EP. *Phytomonas* and other trypanosomatid parasites of plants and fruit. Adv Parasitol. 1999;42:29–112. https://doi.org/10.1016/S0065-308X(08)60148-7

33. Daniels JP, Gull K, Wickstead B. Cell biology of the trypanosome genome. Microbiol Mol Biol Rev. 2010;74:552–569. https://doi.org/10.1128/MMBR.00024-10

34. Dollet M. Plant diseases caused by flagellate protozoa (*Phytomonas*). Annu Rev Phytopathol. 1984;22:115–125. https://doi.org/10.1146/annurev.py.22.090184.000555

35. Dollet, M. Phloem-restricted trypanosomatids form a clearly characterized monophyletic group among trypanosomatids isolated from plants. Int J Parasitol. 2001;31:459–467. https://doi.org/10.1016/S0020-7519(01)00157-6

36. Camargo EP, Wallace G. Vectors of plant parasites of the genus *Phytomonas* (Protozoa, Zoomastigophorea, Kinetoplastida). In: Harris KF, editor. Advances in disease vector research. New York, NY: Springer; 1994. p. 333–359. https://doi.org/10.1007/978-1-4612-2590-4_12

37. Luise C, Dollet M, Mariau D. Research into hartrot coconut, a disease caused by *Phytomonas* (Trypanosomatidae), and into its vector *Lincus* sp. (Pentatomidae) in Guiana. Oléagineux. 1986;41:437–449.

38. Porcel BM, Denoeud F, Opperdoes F, Noel B, Madoui MA, Hammarton TC, et al. The streamlined genome of *Phytomonas* spp. relative to human pathogenic kinetoplastids reveals a parasite tailored for plants. PLoS Genet. 2014;10:e1004007. https://doi.org/10.1371/journal.pgen.1004007

39. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome *Trypanosoma brucei*. Science. 2005;309:416–422. https://doi.org/10.1126/science.1112642

40. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, *Leishmania major*. Science. 2005;309:436–442. https://doi.org/10.1126/science.1112680

41. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, et al. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science. 2005;309:409–415. https://doi.org/10.1126/science.1112631

42. Jaskowska E, Butler C, Preston G, Kelly S. *Phytomonas*: trypanosomatids adapted to plant environments. PLoS Pathog. 2015;11(1):e1004484. https://doi.org/10.1371/journal.ppat.1004484

43. Jankevicius SI, Almeida ML, Jankevicius JV, Cavazzana M, Attias M, de Souza W. Axenic cultivation of trypanosomatids found in corn (*Zea mays*) and in phytophagous hemipterans (*Leptoglossus zonatus* Coreidae) and their experimental transmission. J Euk Microbiol. 1993;40:576–581. https://doi.org/10.1111/j.1550-7408.1993.tb06110.x

44. Teixeira MMG, Takata CSA, Conchon I, Campaner M, Camargo EP. Ribosomal and kDNA markers distinguish two sub-groups of *Herpetomonas* among old species and new trypanosomatids isolated from flies. J Parasitol. 1997;83:5845. https://doi.org/10.2307/3284317

45. Bernard J, Bremond M. Molecular biology of fish viruses: a review. Vet Res. 1995;26:341–351.

46. Walker PJ, Benmansour A, Calisher CH, Dietzgen R, Fang RX, Jackson AO, et al. Family Rhabdoviridae. In: van Regenmortel MHV, Fauquet CM, Bishop DHL, editors. Virus taxonomy: classification and nomenclature of viruses. Seventh report of the International Committee on Taxonomy of Viruses. Berlin: Springer; 2000. p. 563–583.

47. Tordo N, Benmansour A, Calisher C, Dietzgen RG, Fang RX, Jackson AO, et al. Rhabdoviridae. In: Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA, editors. Virus taxonomy. Eigth report of the International Committee on Taxonomy of Viruses. London: Elsevier; 2004. p. 623–644.

48. Dale JL, Peters D. Protein composition of the virions of five plant rhabdoviruses. Intervirology. 1981;16:86–94. https://doi.org/10.1159/000149252

49. Badrane H, Tordo N. Host switching in *Lyssavirus* history from the Chiroptera to the Carnivora orders. J Virol. 2001;75:8096–8104. https://doi.org/10.1128/JVI.75.17.8096-8104.2001

50. Wang Y, Cowley JA, Walker PJ. Adelaide River virus nucleoprotein gene: analysis of phylogenetic relationships of ephemeroviruses and other rhabdoviruses. J Gen Virol. 1995;76:995–999. https://doi.org/10.1099/0022-1317-76-4-995

51. Bourhy H, Cowley JA, Larrous F, Holmes EC, Walker PJ. Phylogenetic relationships among rhabdoviruses inferred using the L polymerase gene. J Gen Virol. 2005;86:2849–2858. https://doi.org/10.1099/vir.0.81128-0

52. Rothenberg M, Hanson MR. Different transcript abundance of two divergent ATP synthase subunit 9 genes in the mitochondrial genome of *Petunia hybrida*. Mol Gen Genet. 1987;209:21–27. https://doi.org/10.1007/BF00329831

53. Pruitt KD, Hanson MR. Cytochrome oxidase subunit II sequences in *Petunia* mitochondria: two intron-containing genes and an intron-less pseudogene associated with cytoplasmic male sterility. Curr Genet. 1989;16:281–291. https://doi.org/10.1007/BF00422115

54. Senda M, Harada T, Mikami T, Sugiura M, Kinoshita T. Genomic organization and sequence analysis of the cytochrome oxidase subunit II gene from normal and male-sterile mitochondria in sugar beet. Curr Genet. 1991;19:175–181. https://doi.org/10.1007/BF00336484

55. Nugent JM, Palmer JD. RNA-mediated transfer of the *coxII* gene from the mitochondrion to the nucleus during flowering plant evolution. Cell. 1991;66:473–481. https://doi.org/10.1016/0092-8674(81)90011-8

56. Bonen L, Calixte S. Comparative analysis of bacterial-origin genes for plant mitochondrial ribosomal proteins. Mol Biol Evol. 2006;23:701–712. https://doi.org/10.1093/molbev/msj080

57. Sandoval P, Leon G, Gomez I, Carmona R, Figueroa P, Holuigue L, et al. Transfer of *RPS14* and *RPL5* from the mitochondrion to the nucleus in grasses. Gene. 2004;324:139–147. https://doi.org/10.1016/j.gene.2003.09.027

58. Fallahi M, Crosthwait J, Calixte S, Bonen L. Fate of mitochondrially located S19 ribosomal protein genes after transfer of a functional copy to the nucleus in cereals. Mol Genet Genomics. 2005;273:76–83. https://doi.org/10.1007/s00438-004-1102-9

59. Sanchez H, Fester T, Kloska S, Schroder W, Schuster W. Transfer of *rps19* to the nucleus involves the gain of an RNP-binding motif which may functionally replace RPS13 in *Arabidopsis* mitochondria. EMBO J. 1996;15:2138–2149.

60. Manfredi G, Fu J, Ojaimi J, Sadlock JE, Kwong JQ, Guy J, et al. Rescue of a deficiency in ATP synthesis by transfer of *MTATP6*, a mitochondrial DNA-encoded gene, to the nucleus. Nat Genet. 2002;30:394–399. https://doi.org/10.1038/ng851

61. Ayliffe MA, Scott NS, Timmis JN. Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. Mol Biol Evol. 1998;15:738–745. https://doi.org/10.1093/oxfordjournals.molbev.a025977

62. Tourmen Y, Baris O, Dessen P, Jacques C, Malthiery Y, Reynier P. Structure and chromosomal distribution of human mitochondrial pseudogenes. Genomics. 2002;80(1):71–77. https://doi.org/10.1006/geno.2002.6798

63. Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: mitochondrial DNA copies (NUMTS) in sequenced nuclear genomes. PLoS Genet. 2010;6:e1000834. https://doi.org/10.1371/journal.pgen.1000834

64. Roark LM, Hui AY, Donnelly L, Birchler JA, Newton KJ. Recent and frequent insertions of chloroplast DNA into maize nuclear chromosomes. Cytogenet Genome Res. 2010;129:17–23. https://doi.org/10.1159/000312724

65. Richly E, Leister D. NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. Mol Biol Evol. 2004;21:1972–1980. https://doi.org/10.1093/molbev/msh210

66. Noutsos C, Richly E, Leister D. Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. Genome Res. 2005;15:616–628. https://doi.org/10.1101/gr.3788705

67. Stegemann S, Bock R. Experimental reconstruction of functional gene transfer from tobacco plastid genome to the nucleus. Plant Cell. 2006;18:2869–2878. https://doi.org/10.1105/tpc.106.046466

68. Michalovova M, Vyskot B, Kejnovsky E. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. Heredity. 2013;111:314–320. https://doi.org/10.1038/hdy.2013.51

69. Martis MM, Klemme S, Banaei-Moghaddam AM, Blattner FK, Macas J, Schmutzer T, et al. Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. Proc Natl Acad Sci USA. 2012;109:13343–13346. https://doi.org/10.1073/pnas.1204237109

70. Perna NT, Kocher TD. Mitochondrial DNA: molecular fossils in the nucleus. Curr Biol. 1996;6:128–129. https://doi.org/10.1016/S0960-9822(02)00441-4

71. Hazkani-Covo E, Sorek R, Graur D. Evolutionary dynamics of large *numts* in the human genome: rarity of independent insertions and abundance of post-insertion duplications. J Mol Evol. 2003;56:169–174. https://doi.org/10.1007/s00239-002-2390-5

72. Ricchetti M, Tekaia F, Dujon B. Continued colonization of the human genome by mitochondrial DNA. PLoS Biol. 2004;2:1313–1324. https://doi.org/10.1371/journal.pbio.0020273

73. Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P, Fernández-Pozo N. Why assembling plant genome sequences is so challenging. Biology. 2012;1(2):439–459. https://doi.org/10.3390/biology1020439

74. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome

sequencing projects. PeerJ. 2014;2:e675. https://doi.org/10.7717/peerj.675

75. Willerslev E, Mourier T, Hansen AJ, Christensen B, Barnes I, Salzberg SL. Contamination in the draft of the human genome masquerades as lateral gene transfer. DNA Seq. 2002;13(2):75–76. https://doi.org/10.1080/10425170290023392

76. Ananiev EV, Phillips RL, Rines HW. Complex structure of knob DNA on maize chromosome 9: retrotransposon invasion into heterochromatin. Genetics. 1998;149:2025–2037.

77. Pelissier T, Tutois S, Tourmente S, Deragon JM, Picard G. DNA regions flanking the major *Arabidopsis thaliana* satellite are principally enriched in *Athila* retroelement sequences. Genetica. 1996;97:141–151. https://doi.org/10.1007/BF00054621

78. Alkhimova OG, Mazurok NA, Potapova TA, Zakian SM, Heslop-Harrison JS, Vershinin, AV. Diverse patterns of the tandem repeats organization in rye chromosomes. Chromosoma. 2004;113:42–52. https://doi.org/10.1007/s00412-004-0294-4

79. Panstruga R, Buschges R, Piffanelli P, Schulze-Lefert P. A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. Nucleic Acids Res. 1998;26:1056–1062. https://doi.org/10.1093/nar/26.4.1056

80. Presting GG, Malysheva L, Fuchs J, Schubert I. A *TY3*/*GYPSY* retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. Plant J. 1998;16:721–728. https://doi.org/10.1046/j.1365-313x.1998.00341.x

81. Mao L, Devos KM, Zhu I, Gale M. Cloning and genetic mapping of wheat telomere-associated sequences. Mol Genet Genomics. 1997;254:584–591. https://doi.org/10.1007/s004380050455

82. Tomita M, Shinohara K, Morimoto M. Revolver is a new class of transposon-like gene composing the Triticeae genome. DNA Res. 2008;15:49–62. https://doi.org/10.1093/dnares/dsm029

83. Grandbastien MA. Activation of plant retrotransposons under stress conditions. Trends Plant Sci. 1998;3:181–187. https://doi.org/10.1016/S1360-1385(98)01232-1

84. Flavell R. Sequence amplification, deletion and rearrangement: major sources of variation during species divergence. In: Dover GA, Flavell R, editors, Genome evolution. London: Academic Press; 1982. p. 301–323.

85. Dover GA. Molecular drive: a cohesive mode of species evolution. Nature. 1982;299:111–117. https://doi.org/10.1038/299111a0

86. Dover GA. Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. Trends Genet. 1986;2:159–165. https://doi.org/10.1016/0168-9525(86)90211-8

87. Jakowitsch J, Mette MF, van der Winden J, Matzke MA, Matzke AJM. Integrated pararetroviral sequences define a uniqueclass of dispersed repetitive DNA in plants. Proc Natl Acad Sci USA. 1999;96(23):13241–13246. https://doi.org/10.1073/pnas.96.23.13241

88. John B, Miklos G. The eukaryote genome in development and evolution. London: Allen and Unwinl; 1988.

89. Dover GA, Tautz D. Conservation and divergence in multigene families: alternatives to selection and drift. Philos Trans R Soc Lond B Biol Sci. 1986;312:275–289. https://doi.org/10.1098/rstb.1986.0007