Theor Appl Genet (2010) 120:491–508
DOI 10.1007/s00122-009-1170-7

ORIGINAL PAPER

Analysis of transposons and repeat composition of the sunflower (*Helianthus annuus* L.) genome

Andrea Cavallini · Lucia Natali · Andrea Zuccolo · Tommaso Giordani · Irena Jurman · Veronica Ferrillo · Nicola Vitacolonna · Vania Sarri · Federica Cattonaro · Marilena Ceccarelli · Pier Giorgio Cionini · Michele Morgante

Received: 4 April 2009 / Accepted: 27 September 2009 / Published online: 14 October 2009

Abstract A sample-sequencing strategy combined with slot–blot hybridization and FISH was used to study the composition of the repetitive component of the sunflower genome. One thousand six hundred thirty-eight sequences for a total of 954,517 bp were analyzed. The fraction of sequences that can be classified as repetitive using computational and hybridization approaches amounts to 62% in total. Almost two thirds remain as yet uncharacterized in nature. Of those characterized, most belong to the *gypsy* superfamily of LTR-retrotransposons. Unlike in other species, where single families can account for large fractions of the genome, it appears that no transposon family has been amplified to very high levels in sunflower. All other known classes of transposable elements were also found. One family of unknown nature (contig 61) was the most repeated in the sunflower genome. The evolution of the repetitive component in the *Helianthus* genus and in other Asteraceae was studied by comparative analysis of the hybridization of total genomic DNAs from these species to the sunflower small-insert library and compared to gene-based phylogeny. Very little similarity is observed between *Helianthus* species and two related Asteraceae species outside of the genus. Most repetitive elements are similar in annual and perennial *Helianthus* species indicating that sequence amplification largely predates such divergence. *Gypsy*-like elements are more represented in the annuals than in the perennials, while *copia*-like elements are similarly represented, attesting a different amplification history of the two superfamilies of LTR-retrotransposons in the *Helianthus* genus.

Sequences from *Helianthus annuus* randomly sheared genomic DNA library and sequences of genes used for phylogenetic analyses are available at the URL: https://services.appliedgenomics.org/sequences-export/26-Helianthus/.

Communicated by A. Bervillé.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-009-1170-7) contains supplementary material, which is available to authorized users.

A. Cavallini · L. Natali · T. Giordani · V. Ferrillo
Genetics Section, Department of Crop Plant Biology, University of Pisa, Pisa, Italy

A. Zuccolo · M. Morgante (✉)
Department of Crop and Environmental Sciences, University of Udine, Viale delle Scienze 208, 33100 Udine, Italy
e-mail: michele.morgante@uniud.it

I. Jurman · F. Cattonaro · M. Morgante
Istituto di Genomica Applicata, Parco Scientifico e Tecnologico Luigi Danieli, Udine, Italy

N. Vitacolonna
Department of Mathematics and Informatics, University of Udine, Udine, Italy

V. Sarri · M. Ceccarelli · P. G. Cionini
Department of Cellular and Environmental Biology, University of Perugia, Perugia, Italy

Introduction

Improved knowledge of genome composition, especially of its repetitive component, generates information that is of importance in both theoretical and applied research, such as for improving strategies for genetic and physical mapping of genomes and for the discovery and development of molecular markers. Moreover, knowledge of genome composition is a prerequisite for the annotation steps in sequencing projects both of ESTs (Expressed Sequence Tags) and of genomic regions.

Differences in genome size of plant species result mainly from differences in the amount of repetitive DNA. The majority of repetitive DNA sequences are transposons of different classes and subclasses (Wicker et al. 2007). Besides the amount of repetitive DNA, also the composition of the repetitive component of the genome may vary between species. For example, abundant non-LTR elements (LINEs and SINEs) have been identified in the human genome (International Human Genome Sequencing Consortium 2001), while their occurrence in plant genomes appears to be reduced compared to the abundance of LTR-retrotransposons (The Arabidopsis Genome Initiative 2000; Meyers et al. 2001; Vitte and Bennetzen 2006). Among plant species, differences were found in the composition of the repetitive component in the genomes that have been sequenced: Arabidopsis, rice, poplar, and grapevine (The Arabidopsis Genome Initiative 2000; Goff et al. 2002; Tuskan et al. 2006; The French-Italian Public Consortium for grape genome characterization 2007), that are relatively small. Some data are also available for medium-large genome sized species such as maize, wheat, and cotton (Meyers et al. 2001; Hawkins et al. 2006; Vitte and Bennetzen 2006; Charles et al. 2008), showing differences in the evolution of these genomes.

Many questions remain about the distribution of repetitive sequences and the overall genome organization in plants with medium-large genomes. Grasses are by far the group of plants where most information has been collected for species with large genomes, such as maize, barley, wheat. Dicotyledons other than *Arabidopsis* and *Gossypium* species have in general been given little attention, despite their great economic importance, and very little information is available on genome composition and organization in the Compositae family, which is very large and includes very important crop species, for example, the sunflower.

The sunflower (*Helianthus annuus* L.) belongs to the genus *Helianthus*, whose origin ranges between 4.75 and 22.7 million years, as estimated from RFLP analysis of cpDNA (Schilling 1997). In their classification, based on morphological and crossability analyses, Schilling and Heiser (1981) included 67 species, annual or perennial, 50 native to North America and 17 to South America. Within the genus, the extant lineages arose between 1.7 and 8.2 million years ago (Schilling 1997), i.e., this genus originated relatively recently.

Phylogenetic analyses suggested a close relationship between *Helianthus* and both *Viguiera* and *Tithonia* (Soltis and Soltis 2000; Schilling 2001). Based on the geographic distributions of its closest relatives, the genus *Helianthus* likely originated in Mexico, with subsequent migration through North America (Schilling et al. 1998). The origin of cultivated sunflower was found in the eastern regions of North America (Harter et al. 2004), although some controversial recent evidences have shown an earlier presence of domesticated sunflower in Mexico, suggesting a second domestication event in this area (Lentz et al. 2008). Several species of *Helianthus* are known to be of hybrid origin (Rieseberg 1995).

Schilling and Heiser (1981) excluded South American species from *Helianthus* genus and subdivided it into four sections, section *Annui* (comprising only diploid and annual species), section *Agrestes* (comprising only the diploid and annual *H. agrestis*), section *Ciliares* (with two series, *Ciliares* and *Pumili*, comprising six perennial species), and section *Atrorubentes* (with five series, *Angustifolii*, *Atrorubentes*, *Divaricati*, *Gigantei*, and *Microcephali*, comprising 30 perennial and one annual species, *H. porteri*).

Molecular studies have been made to clarify the relationships among *Helianthus* species. RFLP analyses of chloroplast DNA evidenced four sections, although nodes between sections were not well supported: one including the annual *H. agrestis*, another including the annual *H. porteri*, a third (sect. *Helianthus*) with all other annuals, and a fourth including all perennials (Schilling 1997). In a subsequent work, Schilling et al. (1998), using rDNA ITS sequences, found little differentiation among most *Helianthus* species. Sossey-Alaoui et al. (1998), by RAPD analysis, obtained clear-cut separations of three main sections, *Helianthus* (annuals), *Atrorubentes* and *Ciliares* (both perennials), and separate positions of *H. agrestis* and *H. porteri*. Recently, analysis of rDNA ETS sequence confirmed the occurrence of a monophyletic annual section *Helianthus*, a two-lineage polyphyletic section *Ciliares*, the monotypic section *Agrestis*, and a perennial and polyphyletic section *Divaricati* (= *Atrorubentes*) (Timme et al. 2007).

Many approaches can be used to characterize genome structure. Fluorescent in situ hybridization (FISH) has been used to investigate the distribution of repetitive elements in plant genomes (Pearce et al. 1996; Heslop-Harrison et al. 1997; Zhang et al. 2005). Genomic sample sequencing is a more direct approach and a rapid mean of assessing genome organization (Brenner et al. 1993; Elgar et al. 1999; Meyers et al. 2001; Hawkins et al. 2006). Sequence analysis of libraries of randomly generated small genomic DNA clones can provide an unbiased sample that allows the determination of genome composition and genome properties.

A detailed structural analysis of the sunflower genome is still to be performed. Santini et al. (2002) described the occurrence of one *gypsy* and one *copia* LTR-retrotransposon subfamilies in the genome of sunflower and other *Helianthus* species. However, the frequency and distribution of many specific repetitive sequences has not been determined in this species. To this end, we used a sample-sequencing strategy combined with slot–blot hybridization

and FISH to study the composition of the repetitive component of the sunflower genome. Moreover, the evolution of the repetitive component in the *Helianthus* genus and in other Asteraceae was studied by comparative analysis of the hybridization of total genomic DNAs from these species to the sunflower small-insert library.

Materials and methods

Plant material and DNA isolation

The sunflower inbred line HCM was used for analysis of genome composition. This line was developed at the Dept. of Crop Plant Biology of Pisa University after 20 self-pollination cycles, starting from an open-pollinated cultivar, and it is a highly homozygous line. Seeds were washed in tap water and germinated on moist paper in Petri dishes and plants were grown in the open air. Young leaves were collected and DNA purification was carried on according to Doyle and Doyle (1989) with some modifications.

Leaf portions were homogenized in liquid nitrogen in a mortar and lysed in hexadecyl trimethyl ammonium bromide (CTAB) isolation buffer, 1 ml/g tissue. Samples were incubated at 60°C for 30 min with occasional gentle swirling, and DNA was then extracted once with chloroform: isoamyl alcohol (24:1, v/v).

After centrifugation (6,200 × *g*) for 15 min at 4°C, nucleic acids were precipitated from the aqueous phase by adding two-third volume of cold isopropanol and then centrifuged at 6,200 *g* for 10 min at 4°C, washed in 70% (v/v) cold ethanol by centrifuging at 6,200 *g* for 5 min at 4°C, and resuspended in TE buffer.

Genomic DNA was purified by ethidium bromide-CsCl density gradients. Solid CsCl and ethidium bromide were added to the nucleic acid solution up to final concentrations of 0.8 g/ml and 150 μg/ml, respectively. The solution was centrifuged at 137,500 *g* for 60 h at 15°C in a Beckman L5-65 ultracentrifuge using a 50-Ti rotor and the DNA band, visualized under long-wave UV illumination, and collected. Ethidium bromide was then removed by gentle inversion of the solution with n-buthanol, and CsCl was eliminated by dialysis against water at 4°C. Finally, DNA was precipitated from the aqueous phase by adding 0.1 volume of NaCl saturated solution and two volumes of cold ethanol, centrifuged and washed as above, and resuspended in water.

Construction of a *Helianthus* genomic library

Five micrograms of *Helianthus* genomic DNA was sheared according to the nebulization protocol of Wilson and Mardis (1997). DNA was precipitated and resuspended in 40 μl of TE. Fragment ends were repaired in a 50-μl reaction mix using Pfu DNA Polymerase (Stratagene) and 1 × Pfu buffer (Stratagene), following the protocol suggested by the manufacturer. End-repaired nebulized DNA was electrophoresed on a 1% agarose gel. Fragments were selected in the size range of 0.8–1.5 Kb and DNA was purified from the gel and ligated into pPCR-Script Amp SK(+) (Stratagene) according to the manufacturer's instructions in a 10-μl mix. One microliter of the ligation mix was electroporated into *E. coli* ElectroMAX DH10B Cells (mcrA, mcrB, mcrC, mrr; Invitrogen), using the BioRad GenePulser® II electroporator, in a 0.1 cm cuvette at the conditions of 2.0 kV, 200 Ω, 25 μF. The average insert size after cloning was about 2 Kb and insert from 1,248 clones were selected for sequencing from both directions. Two hundred of the clones turned out to contain no insert after sequence analysis.

Isolation of gene sequences

DNA sequence-based phylogeny was constructed comparing four gene sequences isolated by PCR from genomic DNAs from the species listed in Table 1. The four selected genes encode a dehydrin (HaDhn1a, Giordani et al. 2003, EMBL Accession AJ002741), a drought-responsive-element binding protein (DREB2, Diaz-Martin et al. 2005, EMBL Accession AY508007), an early light-induced protein (ELIP, Ouvrard et al. 1996, EMBL Accession X92646), a non-specific lipid transfer protein (LTP, Ouvrard et al. 1996, and EMBL Accession X92648). These four genes were chosen because designed primers produced a unique fragment to be sequenced when used on genomic DNA of the sunflower HCM inbred line. PCR amplifications of the four genes were performed using these oligonucleotides pairs: 5′-ATATGGCAAACTACCGAGGAGA TAAG-3′ (sense) and 5′- GTGAAACCACATACAAAAC AAAA1-3′ (antisense) for HaDhn1a; 5′-CGAAGAAG GGTTGTATGAAAG-3′ (sense) and 5′-AAACCAAGAC CCAACTCCTC-3′ (antisense) for DREB2; 5′- CAACCGA CGCTTCCAAAAC -3′ (sense) and 5′-AGCACTCTTTGT TTCTATGATTCTT-3′ (antisense) for ELIP; and 5′-TG GCAAAGATGGCAATGATG-3′ (sense) and 5′-ATCAA AGACACATACACATCCATA-3′ (antisense) for LTP. Sequences were amplified using 100 ng of genomic DNA as a template; thermocycling was performed at 94°C for 30 s, 55°C for 30 s and 72°C for 60 s, for 30 cycles, using FIREPol thermostable DNA polymerase (Biodyne). Since wild *Helianthus* species are presumably heterozygous, the amplified fragments were cloned into a pGEM-T Easy plasmid vector (Promega) before sequencing.

Table 1 Classification of Asteraceae species whose genomic DNA was used as probes

Species	*Helianthus* species classification		Genome size[a]	Source[b]
	Series	Section		
Helianthus annuus	*Annui*	*Helianthus*	3.30	DBPA
H. argophyllus	*Annui*	*Helianthus*	4.43	NCRPIS
H. debilis	*Annui*	*Helianthus*	3.30	NCRPIS
H. petiolaris	*Annui*	*Helianthus*	3.40	NCRPIS
H. ciliaris	*Ciliares*	*Ciliares*	na	NCRPIS
H. pumilus	*Pumili*	*Ciliares*	na	NCRPIS
H. atrorubens	*Atrorubentes*	*Divaricati*	na	NCRPIS
H. giganteus	*Corona-solis*	*Divaricati*	4.83	NCRPIS
H. simulans	*Angustifolii*	*Divaricati*	na	NCRPIS
H. tuberosus	*Corona-solis*	*Divaricati*	12.55	DBPA
Viguiera multiflora			na	NCRPIS
Tithonia rotundifolia			na	NCRPIS

na not available

[a] In pg (= 1C) The value for *H. annuus* HCM line was taken from Cavallini et al. (1986), the other values from the Plant DNA *C*-value Database, Kew, UK (http://www.kew.org/cval/)

[b] *DBPA* Dipartimento di Biologia delle Piante Agrarie, Pisa, Italy, *NCRPIS* North Central Regional Plant Introduction Station, Ames, FL, USA

DNA sequencing

DNA for sequencing was prepared from selected transformants using the Montage Plasmid Miniprep Kit (Millipore). DNA sequencing was performed using an Applied Biosystems 3730 DNA Analyzer, the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) and standard M13 forward and M13 reverse primers. In other experiments, gene fragments were amplified by PCR and directly sequenced using the same instrument and the specific primers as reported above.

BLAST and clustering analyses of sequences

All sequences were subjected to BLASTN and BLASTX analysis against the non-redundant nucleotide and protein GenBank databases (June 1st release), respectively. All sequences were also subject to BLASTX analysis against a curated database (M. Morgante, unpublished) of 397 plant transposable elements encoded proteins that were classified according to the criteria provided by Wicker et al. (2007). A BLAST *E* value of 10^{-5} was deemed significant to classify sunflower sequences. All sequences and their Blast results can be accessed at https://www.appliedgenomics.org/sunflowers/. All sequences were compared to each other to detect additional repetitive sequences that did not show homology to known repeated sequences but did overlap to each other by using the CAP3 sequence assembler (Huang and Madan 1999) under relaxed stringency parameters.

Southern blot hybridisation

Five μg of genomic DNA was digested with the restriction enzymes *Bst*NI, *Eco*RII, *Msp*I, *Hpa*II, *Sma*I, *Xma*I, *Bgl*II, *Dra*I, *Eco*RI, *Hind*III, *Sau*3A, *Pst*I (Amersham, Roche), electrophoresed in 0.8% agarose gel, and blotted on positively charged nylon membranes (Roche) according to Sambrook et al. (1989).

Probes to be used for Southern analysis were digoxigenin-labeled by PCR using 1 × PCR buffer, 0.5 μl *Taq* DNA polymerase (Promega), dNTP labeling mix (final concentrations 200 μM dATP, 200 μM dCTP, 200 μM dGTP, 190 μM dTTP, 10 μM digoxigenin-11-dUTP, alkalin labile; Roche), 2.5 mM $MgCl_2$, 0.8 μM each forward and reverse primers designed on cloned sequence, 2 ng plasmid DNA derived from selected clones as template (total volume 50 μl). Samples were heated at 94°C for 4 min, then 30 PCR cycles were performed: 94°C for 30″, 58°C for 30″, 72°C for 1 min. Final extension was performed at 72°C for 7 min. PCR products were purified with Wizard SV gel and PCR clean-up system (Promega).

Hybridizations were performed using 15 ng/ml probe at 65°C for 12 h in deionized water, 5 × SSC, 2% blocking reagent (Roche), 0.02% SDS, and 0.2% SLS. Filters were washed twice in 2 × SSC, 0.1% SDS for 15 min at room temperature, once in 1 × SSC 0.1% SDS for 30 min at 68°C and once in 0.3 × SSC, 0.1% SDS for 30 min at 68°C. The temperature of the final wash was calculated in order to ensure hybridization of DNA sequences sharing at least 90% similarity with the probe. Detection was

performed using the DIG-Nucleic Acid Detection Kit (Roche) according to the manufacturer's instructions.

Dot blot hybridization and calculation of sequence copy number

Dot blots were prepared by applying dilution series of DNA to nylon filters (Hybond N^+; Amersham) using a Bio-Dot apparatus (Biorad). Using a *C*-value estimation of 3.30 pg DNA (Cavallini et al. 1986), *H. annuus* genomic DNA was spotted in a dilution series from 50 to 25,600 genomes. Similarly, PCR products of ~ 1 kb, derived from clones selected according to their sequence, were solubilized in sterile distilled water with NaOH to a final concentration of 0.4 M in a dilution series representing 7.35×10^6 to 3.76×10^9 copies and spotted on the same membranes.

PCR products were digoxigenin-labeled by PCR using specific primers as described above and used as DNA probes at a concentration of 15 ng/ml of hybridization solution. Hybridization and post-hybridization washes, and detection of digoxigenin in DNA-DNA hybrids were performed as above. Estimation of the copy number of the sequence probed in the genomic DNA was carried out as already described (Santini et al. 2002). The amounts of loaded genomic DNAs were checked as described by Santini et al. (2002).

Cytological preparations and in situ hybridization

Sunflower seeds were germinated in damp vermiculite, and seminal roots were collected, treated with a 0.04% aqueous solution of colchicine (Sigma) for 4 h at room temperature and fixed in ethanol-acetic acid 3:1 (v/v). Fixed root apices were then treated with a solution of pectinase (20%; Sigma) and cellulase (4%; Calbiochem) in citrate buffer pH 4.6 for 2 h at 37°C, and squashed in a drop of 60% acetic acid. After removing the coverslip by the solid CO_2 method, the preparations were air dried.

In situ hybridization was performed according to Schwarzacher et al. (1989) with minor modifications. The DNA of nuclei and chromosomes was denatured in a thermal cycler for 7 min at 70°C and the preparations were incubated overnight at 37°C with 2 ng/µl of heat-denatured DNA probes which were labeled with digoxigenin-11-dUTP by PCR or by nick-translation. The digoxigenin at the hybridization sites was detected by using sheep anti-digoxigenin-fluorescein. The preparations were then counterstained with a 2% solution of DAPI (4′,6-diamidino-2-phenylindole) in McIlvaine buffer pH 7.0 and mounted in antifade solution (AFI; Citifluor). Metaphase chromosomes were studied in images captured by a CCD camera using a Leica Q500MC image analyzer.

Hybridization to gridded small-insert library

Forty microliter of plasmid DNA that had been isolated for sequencing from each of the clones of the sunflower small insert library was first linearised by overnight digestion with *Eco*RI (4 units) in a total volume of 50 µl. DNA was then denatured for 10 min at 91°C and gridded at moderate density (4 × 4) in duplicate using a Beckman Biomek 2000 replicator tool onto Nylon membranes that had been presoaked in denaturation buffer. Filters were then denatured for 3 min in 1.5 M NaCl, 0.5 M NaOH, neutralized for 15 min in 1.5 M NaCl, 0.5 TrisHCl pH8, and rinsed in 5 × SSC. Filters were then exposed to UV light for 2.5 min.

A total of 1,380 clones were arrayed on the membranes that were probed using total labeled genomic DNA from different species (Table 1). Total genomic DNA from each species was isolated from young leaves as described above and digoxigenin-labeled by the random primed DNA labeling technique using a DIG DNA Labeling Kit (Roche) according to the manufacturer's recommendations.

Hybridization and detection were performed as already described. Labeled lambda DNA was also used as control probe. The relative hybridization intensity for each spot in macroarrays was analyzed by eye and quantified in arbitrary units in the range 0–3, where 0 is for not labeled, 1 for slightly labeled, 2 for labeled, and 3 for heavily labeled. Pairwise comparisons between species were made for each clone by plotting the intensity of hybridization estimated as above. The correlation coefficient R was calculated and used as a parameter to indicate the level of relationship between two species and matrices were built. Cluster analysis was conducted on R estimates using the UPGMA (unweighted pair-group method arithmetic average) procedure of the NTSYS-pc program, version 2.02 (Rohlf 1998). The resulting clusters were expressed as dendrograms.

The similarity between species was also estimated according to the following formula:

$$1 - \Sigma\left|\left(X_i - X_j\right)\right|/N$$

where X_i is the hybridization intensity of each clone to DNA of species i, X_j is the hybridization intensity of the same clone to DNA of species j, and N is the total number of clones. Also these coefficients of similarity were used to build matrices and dendrograms as described above.

Construction of a gene-based phylogenetic tree

Gene sequences were aligned using CLUSTALW (Thompson et al. 1994). The multiple sequence alignments were locally adjusted manually. Alignments are available from the authors upon request. Relationships among different species

were investigated using the neighbor-joining method (NJ), employing the PHYLIP program package version 3.572 (Felsenstein 1989). Using the SEQBOOT program, 100 versions of the original alignment were generated; then, trees was generated using the DNADIST and the NEIGHBOR programs. The CONSENSE program was used to obtain the consensus dendrogram.

Difference between hybridization-based and gene sequence-based dendrograms were tested using the Mantel matrix correspondence test (Mantel 1967).

Results

Sample sequencing of genomic DNA and repeat composition of the sunflower genome

A library of small-insert genomic DNA was constructed using nebulized sunflower total genomic DNA from the inbred line HCM. After vector masking and trimming of low-quality regions and assembling of overlapping forward and reverse reads from the same genomic clone (79 cases), 1,638 sequences for a total of 954,517 bp, corresponding to 0.03% of the sunflower genome, were available for analysis. We calculated an average GC content of 39% for the entire small insert library. All sequences were subject to BLASTN and BLASTX analysis against the non-redundant nucleotide and protein GenBank databases, respectively. Moreover, all sequences were compared to each other to detect additional repetitive sequences that did not show homology to known repeated sequences but did overlap to each other. The results of these different analytical approaches allowed us to estimate the abundance and composition of the repetitive fraction of the sunflower genome (Table 2). Approximately 48% of the library sequences were classified as repetitive. Twenty-one percent showed similarity to known transposable elements, mostly not from sunflower and belonging to class I retrotransposons. Within class I elements, LTR-retrotransposons dominated with a large prevalence of *gypsy*-type elements over *copia*-type ones. Non-LTR elements such as LINEs were rare and accounted only for 0.4% of all sequences. Class II elements, i.e., DNA transposons, were also rare and corresponded to only 0.7% of the library sequences. Tandemly arranged repeats were found in 1.5% of all sequences, most of which corresponded to rDNA repeats. Six additional sequences containing tandemly arranged repeats were identified using dot plot analysis (Table 2). Sequences containing direct repeats were also annotated using dot plot analysis and amounted to 11 instances (data not shown). A large fraction of the sequences (419, corresponding to 25.6%) were recognized as repetitive by virtue of their similarity to at least another uncharacterized sunflower sequence, either within the nebulized library (344 sequences that belong to clusters obtained using the CAP3 assembler) and/or among the sunflower repeated sequences available within GenBank (75 sequences), but did not show similarity to previously described transposable elements and were therefore classified as unknown repeats. These are presumably repetitive elements that await further classification. Among the sequences of the nebulized library, 3.9% showed significant similarity to previously described genes and 47% of all sequences remained uncharacterized and could not be classified into any of the previously described classes. A significant fraction of these unclassified sequences (14.0% of all sequences) did appear repetitive when they were experimentally hybridized to sunflower total genomic DNA (see below), raising the total fraction of sequences that can be classified as repetitive using a combination of computational and laboratory approaches to 62% in total (Fig. 1).

Table 2 Computational analysis of small insert genomic library sequences

Classification	No. of occurrences	Genomic sequence fraction (%)
Repeated sequences	782	47.74
Transposable elements	338	20.63
Class I (retrotransposons)	326	19.90
Order LINE	7	0.43
Order LTR		
Superfamily *Copia*	58	3.54
Superfamily *Gypsy*	255	15.57
Unknown superfamily	6	0.37
Class II (DNA transposons)	12	0.73
Superfamily Tc1/Mariner	8	0.49
Superfamily Mutator	1	0.06
Superfamily hAT	2	0.12
Unknown superfamily	1	0.06
Tandem repeats	25	1.53
Tandem repeats	6	0.37
rDNA	19	1.16
Unknown repeats	419	25.58
Similarity to Genes	64	3.91
Chloroplast DNA	21	1.28
Mitochondrial DNA	2	0.12
Unknown	769	46.95
Total	1,638	100.00

The classification is based on results of BLASTN and BLASTX analysis against the non-redundant nucleotide and protein GenBank databases, respectively; on dot plot analysis of single sequences; on clustering of all sequences using the CAP3 assembler

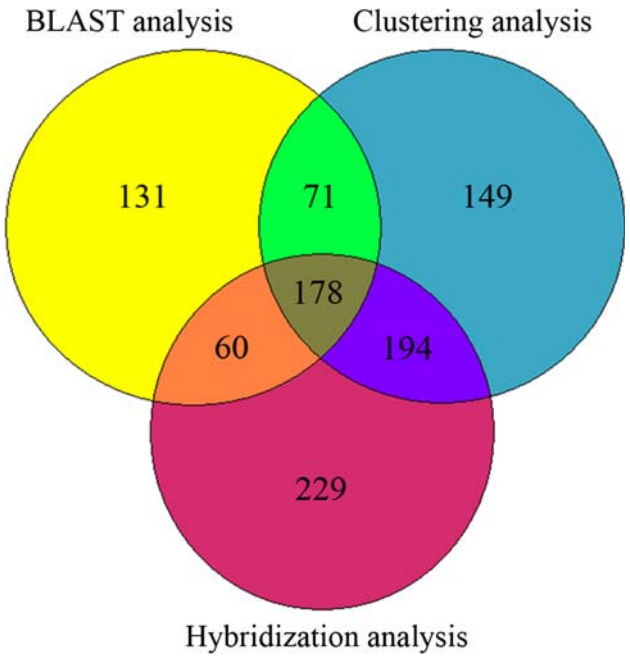

Fig. 1 Venn diagram showing the number of sunflower genomic sequences classified as being repetitive according to different experimental evidences. Blast analysis refers to sequences that showed a BLAST *E* value of 10^{-5} or less when subject to BLASTN and BLASTX analysis against the non-redundant nucleotide and protein GenBank databases, respectively. Cluster analysis refers to sequences that did overlap to each other when compared by using the CAP3 sequence assembler under relaxed stringency parameters. Hybridization analysis refers to sequences that showed a positive hybridization signal when hybridized to *H. annuus* total genomic DNA (See Materials and methods for details)

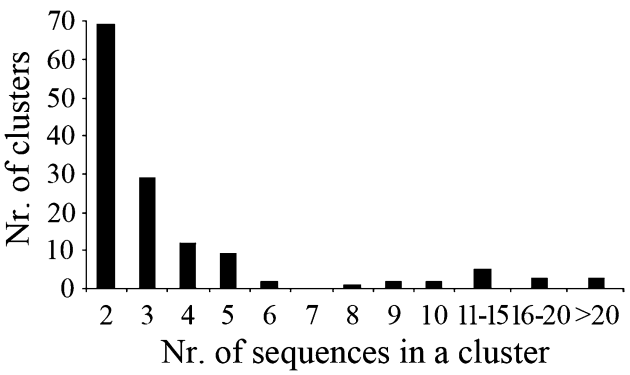

Fig. 2 Size distribution of clusters of genomic sequences obtained using the CAP3 assembler. The histogram depicts the number of sequence clusters (*Y* axis) containing a specified number of genomic sequences (*X* axis)

The clustering of sequences from the nebulized library using the CAP3 assembler allowed us to estimate the abundance of specific repetitive DNA families and also to reconstruct putative consensus sequences from the contiguous sequences (contigs) that were produced for each cluster. A total of 592 nuclear non-genic sequences fell into contigs composed of two or more sequences (Fig. 2) with a significant number of contigs (13) including ten or more (up to 45) sequences, for which the contig length ranged from 1.8 to 5.1 kb. Out of these 13 largest contigs, seven were classified as unknown repeats (including the largest of all clusters comprising 45 sequences, Contig 61) and six were classified as belonging to the *gypsy* superfamily of LTR-retrotransposons.

We explicitely searched for MITEs in our set of genomic sequences by using the FindMITE software and found eight putative examples that were then subject to further experimental analysis.

Abundance and genomic organization of selected repeated DNA sequences

Slot–blot hybridizations were performed to analyze the redundancy of the inserts of many clones from the small-insert library. Such clones were selected based on the occurrence of direct repeats into their sequence (eight clones), on their sequence similarity to DNA transposons (five clones) or to LINEs (seven clones), and on their structural similarity to MITEs (eight clones). Moreover, based on clustering analysis, other clones were selected, one belonging to the most numerous contig (contig 61), which was classified as "unknown" after BLAST analysis, and other clones belonging to contigs showing high similarity to LTR retrotransposons, both *gypsy* (nine clones) and *copia* (six clones). The copy number per haploid genome (1C) for each analyzed sequence is reported as supplementary material. Some of the sequences were also used as probes in Southern blot hybridization experiments to gain additional information on the organization and dispersion pattern in the genome of the family to which they belong.

The sequence belonging to Contig 61, a putative repeated family of unknown nature, was the most repeated in the sunflower genome, with a redundancy of 27,000 copies per haploid genome. Assuming the contig length (4,849 bp) as the minimal length of this repeat, this sequence family should represent 4.11% of the sunflower genome. Southern analysis confirmed high redundancy, with heavy labeling and many bands (Fig. 3a). The use of isoschizomeres with different sensitivity to cytosine methylation in the target site revealed that this repeated element is highly methylated.

Among tandem repeats, the two assayed sequences were estimated to be present in 2,600 and 7,800 copies per genome, respectively. Southern blot hybridizations using the latter of these two sequences as probe confirmed its redundancy and showed the expected ladder pattern after digestion with cytosine methylation sensitive enzymes (Fig. 3b). The copy number of sequences containing direct

repeats ranged from less than 50 to 3,000. The hybridization pattern of the most redundant of these sequences indicated interspersion in the genome and a high degree of methylation (Fig. 3c).

Sequences putatively belonging to DNA transposons occurred in 210 to 8,000 copies per haploid genome. However, it is to be noted that the similarity of HAG003K19 (the only sequence belonging to this class which showed high redundancy) to previously identified DNA transposons is low and limited to a poorly described *Medicago truncatula* hypothetical transposon. It is therefore possible that this sequence has been misclassified as a DNA transposon. The hybridization pattern of this clone reveals a high degree of cytosine methylation (Fig. 3d).

Sequences classified as LINEs and MITEs resulted poorly represented in the sunflower genome. Only two sequences showing homology to LINE elements and one putatively classified as MITE showed some degree of repetitiveness (120, 135, and 700 copies per haploid genome, respectively), that was confirmed by Southern hybridization analysis (Fig. 3e, h).

LTR-retrotransposons were the most represented group of transposable elements in the sunflower genome. We estimated the copy number of sequences belonging to six *gypsy* and three *copia* retrotransposon putative families, corresponding to the most numerous clusters we classified as LTR-retrotransposons. When using a coding portion of the retrotransposon as a probe, the copy number per haploid genome of sequences belonging to the *gypsy* families ranged from 5,000 to 23,900, and that of sequences belonging to the *copia* families from 1,000 to 5,100. These results confirm those of computational sequence analysis that showed *copia* retrotransposons to be generally much less repeated than *gypsy* ones in sunflower. Southern analysis using these sequences as probes showed patterns typical of retrotransposon elements, with a higher number of bands for *gypsy* than for *copia* ones (Fig. 3f, g). Digestion of genomic DNA with methylation sensitive enzymes indicated that also LTR-retrotransposon sequences are highly methylated.

When we were able to identify the putative LTR region within a contig corresponding to selected retrotransposon, we used also this region as a probe. Since two LTRs occur in a retroelement, the copy number of LTRs should be twice compared to that of the corresponding coding portion. Putative LTRs from *gypsy* elements were equally or less numerous than the corresponding retrotransposon coding sequence, probably due to nucleotide mutations, deletions, and insertion of nested retrotransposons (SanMiguel et al. 1996). For one of the three *copia* families (Contig 131), the redundancy of the LTR region was 6.5-fold higher than that of the coding portion. This family might contain a high proportion of solo-LTRs that have been described in many plant species (Vicient et al. 1999) as the result of inter-LTR homologous recombination events.

In situ hybridization reveals the chromosomal organization of repetitive sequences

The distribution of different repetitive sequences in the chromosome complement of *H. annuus* was studied by fluorescence in situ hybridization. No clear signal was observed after hybridization with many probes, probably due to their limited length and the low redundancy level of related sequences in the genome, possibly coupled with a disperse distribution in the chromosomes. Appreciable hybridization signals were obtained with some probes.

Scattered labeling over all chromosomes, indicating wide dispersal of DNA sequences, was observed after many of the successful hybridizations, irrespective of the putative classification of the sequences used as probes (data not shown). This pattern was particularly clear when hybridizing sequences belonged to Contig 61 (Fig. 4a), and this result confirms that this sequence family is the most repeated family in the sunflower genome.

When the clones containing tandem repeats were hybridized, one (HAG004N15) revealed discrete locations on all chromosomes (Fig. 4b) as previously reported (Ceccarelli et al. 2007). Though the small size of sunflower chromosomes prevents in some cases a precise definition of their structure, the other tandem repeat containing clone (HAG002P01) revealed an interspersed distribution with possible enhanced hybridization at the centromeric regions of many chromosomes (Fig. 4h). A similar labeling pattern was also observed after hybridization with HAG004G03 (Fig. 4c), a sequence containing direct repeats, showing that interspersed sequences recognized by both probes may have preferential paracentromeric localization.

Hybridization signals were very faint when using clone HAG003M16, which was putatively classified as MITE, as a probe. However, the picture was clear enough to indicate that the sequences recognized by this probe are scattered along the length of all chromosomes, without any obvious preferential location at given chromosome regions (Fig. 4e). Since the element is present in 700 copies, it is also possible that this sequence is flanked in the clone by other (non-MITE) repetitive DNA.

A disperse distribution of DNA sequences recognized by probes which were putatively classified as parts of mobile elements was observed. No preferential localization at any chromosome region was visible after hybridization with HAG003K19 (Fig. 4d), whose nucleotide

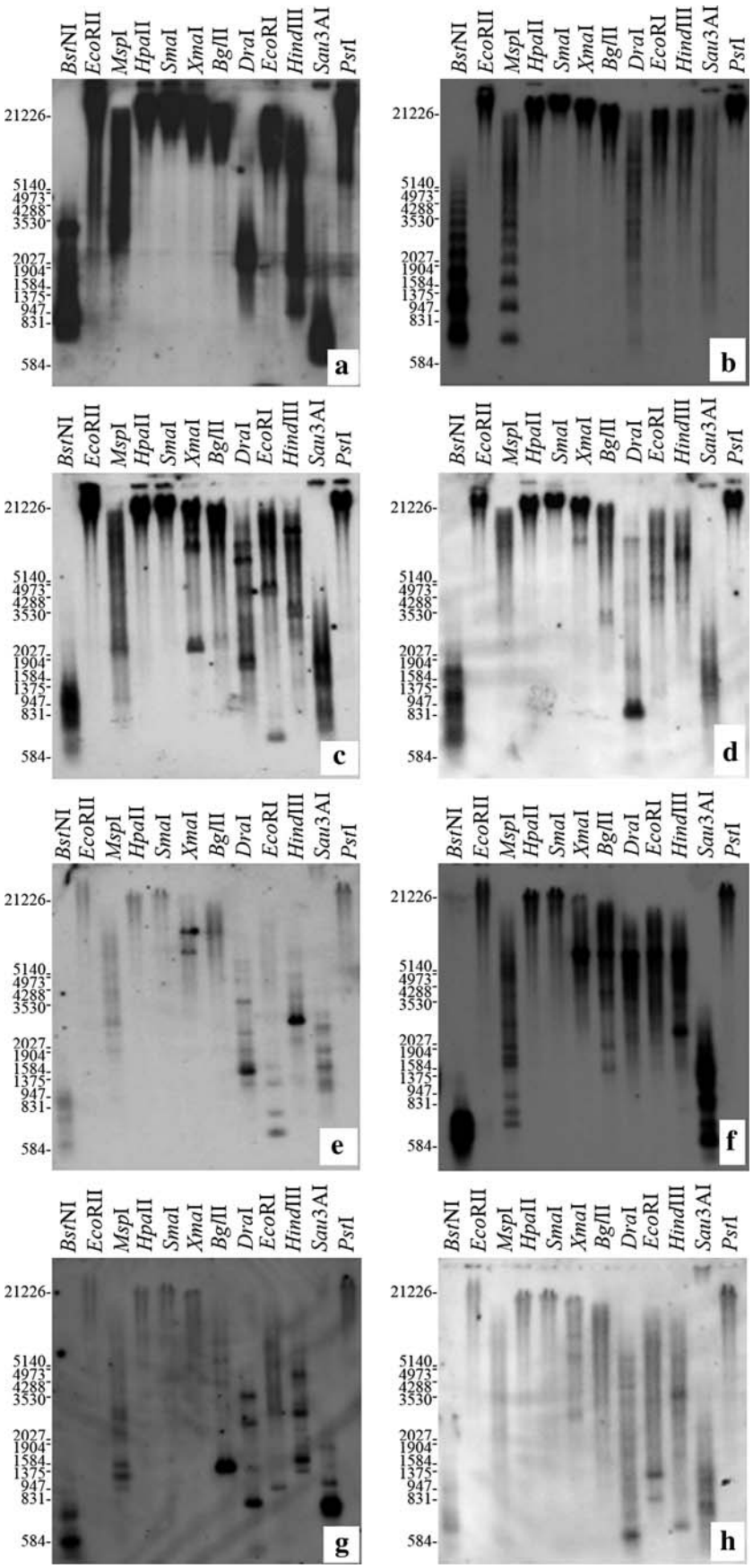

Fig. 3 Southern blots of *H. annuus* genomic DNA digested with different restriction enzymes and hybridized to digoxigenin labeled DNA probes (**a** HAG002K07 clone, corresponding to a portion of Contig 61; **b** HAG004N15 clone, containing tandem repeats; **c** HAG004G03 clone, containing direct repeats; **d** HAG003K19 clone, containing part of a putative DNA transposon; **e** HAG003M16 clone, containing a putative MITE; **f** HAG003L09 clone, corresponding to a portion of contig 104, a *gypsy* retrotransposon; **g** HAG004L03 clone, corresponding to a portion of contig 82, a *copia* retrotransposon; **h** HAG004L21 clone, containing part of a putative LINE). Molecular weight marker sizes are reported on the left (in Kbp)

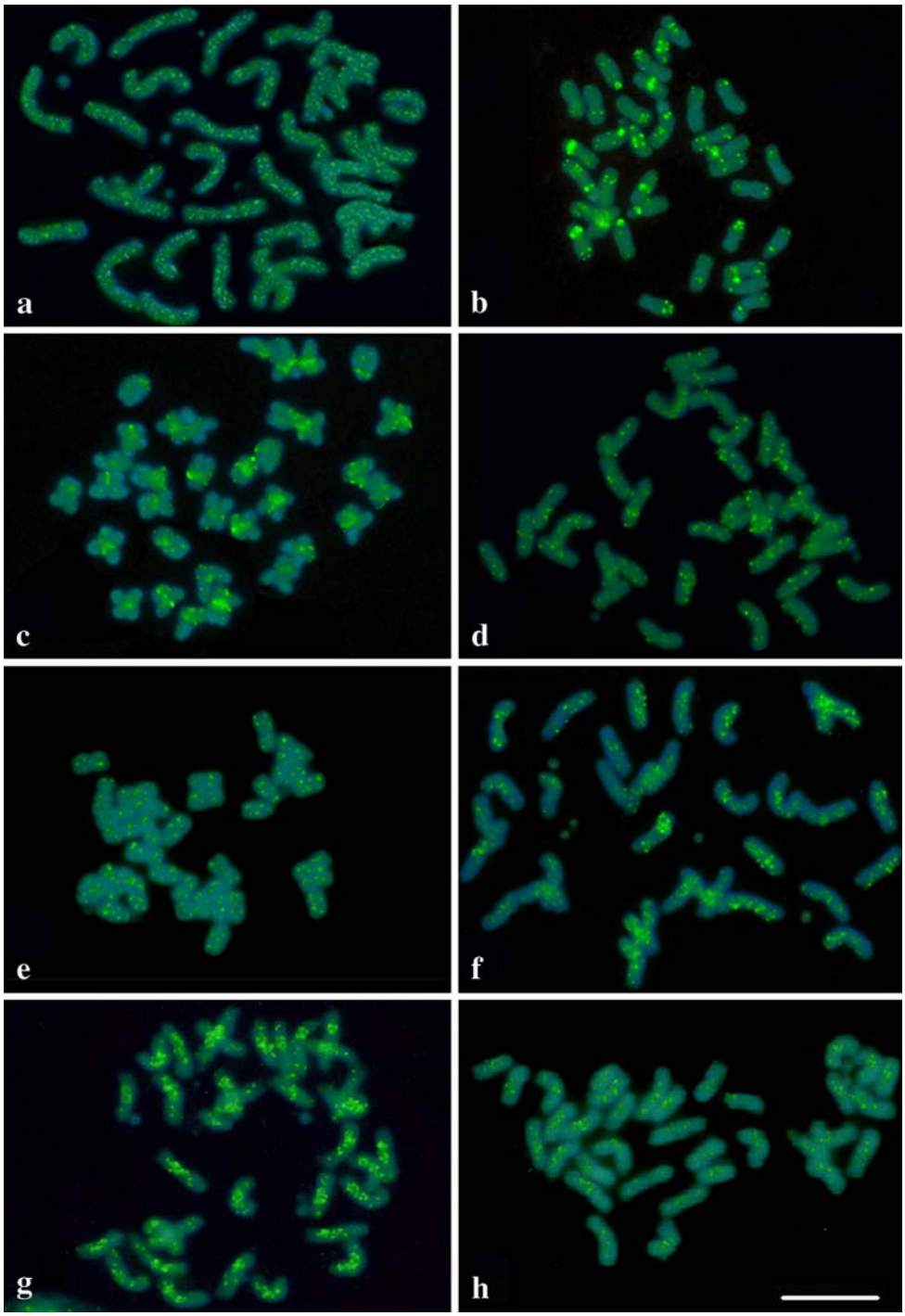

Fig. 4 Electronic overlapping of the images of metaphase plates of *H. annuus* after staining with 4′,6-diamidino-2-phenylindole (*light blue*) and hybridization (*yellow; fluorescein*) with: **a**: HAG002K07 clone, corresponding to a portion of contig 61; **b**: HAG004N15 clone, containing tandem repeats; **c**: HAG004G03 clone, containing direct repeats; **d**: HAG003K19 clone, containing part of a putative DNA transposon; **e**: HAG003M16 clone, containing a putative MITE; **f**: HAG004K09 clone, corresponding to a portion of Contig 3, a *copia* retrotransposon; **g**: HAG002A17 clone, corresponding to a portion of Contig 79, a *gypsy* retrotransposon; **h**: HAG002P01 clone, containing tandem repeats. *Bar* represents 10 μm

sequence showed a low degree of similarity to putative DNA transposons, and with a sequence (HAG004K09) belonging to Contig 3, which showed similarity to *copia* LTR-retrotrasposons (Fig. 4f). Possible preferential hybridization to centromeric chromosomal regions was observed in addition to a dispersed pattern when a sequence (HAG002A17) belonging to Contig 79, which showed similarity to *gypsy* retrotransposons, was hybridized (Fig. 4g).

Analysis of the repetitive DNA component in the genus *Helianthus* and in related species

DNAs from clones from the small-insert library were arrayed on Nylon filters and probed using labeled genomic DNA from ten different *Helianthus* species (annuals and perennials) and two related Asteraceae species (Table 1). Labeled lambda DNA was also used as a control probe: no hybridization was observed.

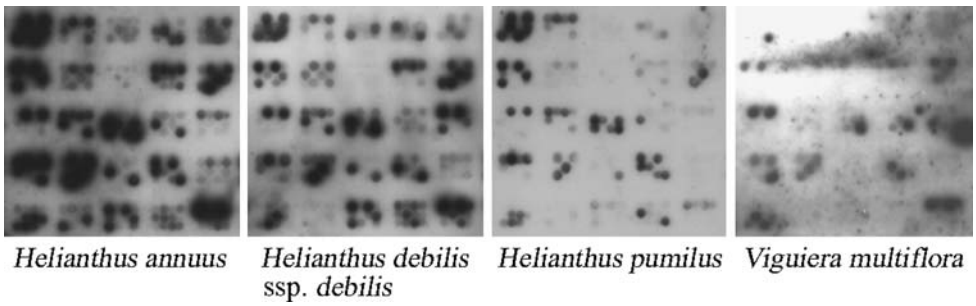

Fig. 5 Examples of hybridization patterns of labeled genomic DNA from different species to 150 clones of the sunflower small-insert library, spotted in a non-regular duplicate arrangement

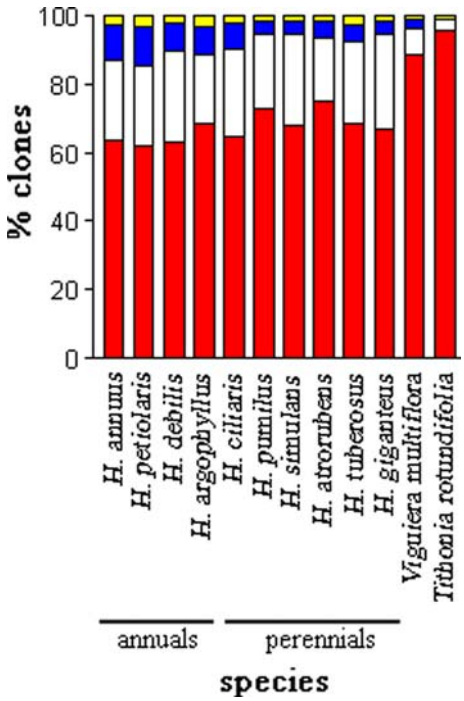

Fig. 6 Percentage of clones showing different intensity of hybridization signal after hybridization with labeled genomic DNAs of ten *Helianthus* species and two Asteraceae species not belonging to *Helianthus* genus. Hybridization signal intensity in arbitrary units: 0, lack of signal (*red*); 1, low-intensity signal (*white*); 2, medium-intensity signal (*blue*); and 3, strong-intensity signal (*yellow*)

The relative hybridization intensity for each spot in the nylon arrays was visually analyzed and quantified in the range 0–3 (see Materials and methods). It is to be considered that lack of a strong hybridization signal might not accurately reflect a low copy number for DNA families because of sequence divergence among family members. After probing the library with the DNA of the genotype from which the library was made (sunflower inbred line HCM), 63.4% clones showed no hybridization signal, 23.4% clones showed low-intensity signals, and 10.5% medium-intensity signals; only a small proportion of clones (36/1344, 2.7%) hybridized strongly, i.e., contained highly redundant sequences. Twelve of the 36 most intense spots were produced by sequences of Contig 61, 8 by sequences of *gypsy* retrotransposons; 5 by rDNA fragments. The remaining 11 intense signals of hybridization were produced by unknown sequences.

Hybridization with genomic DNA of different *Helianthus* species produced similar hybridization patterns (see Fig. 5 as an example). The distribution of hybridization intensities for each clone in the species analyzed is reported in Fig. 6. It can be observed that this distribution is similar in all *Helianthus* species tested, even if the number of clones showing no hybridization signal slightly increases in the perennials, especially in *H. pumilus* and *H. atrorubens*.

A few hybridization signals were obtained when using labeled genomic DNA of *Viguiera multiflora* (11.61% of clones) or of *Tithonia rotundifolia* (4.46%). The majority of clones producing medium or strong signals when using DNA of *Helianthus* species were those belonging to Contig 61. The patterns of clones belonging to Contig 61 are very similar to that observed in sunflower; however, stronger hybridization signals are found when using genomic DNA of perennial than annual species (not shown). This result should indicate that amplification of this sequence occurred in the progenitor of *Helianthus* genus, and, after splitting between annuals and perennials, amplification has gone on in the perennial ancestral and/or loss of sequences has occurred in the annual ancestral. Some of these clones showed strong hybridization signals also to *Viguiera multiflora* genomic DNA. Hence, Contig 61 is apparently highly repeated also in this species, suggesting that the initial amplification of this repeat predates the origin of the *Helianthus* genus.

Other analyses were performed on clones containing sequences similar to LTR-retrotransposons. It was observed that the majority (54.4%) of *gypsy* retrotransposon sequences in the library hybridized to all *Helianthus* species, both annuals and perennials (Table 3), and 16.2% hybridized also to the DNA of *Tithonia* or *Viguiera*. Some (15.0%) *gypsy* retrotransposon sequences hybridized only to annual *Helianthus* species. It is also to be noted that 10.0% of *gypsy* sequences hybridized only to perennials, showing that such sunflower retrotransposon sequences are more abundant in perennials. The same trend is observed for *copia* retrotransposon sequences (Table 3). These results indicate that the majority of LTR-retrotransposon families

Table 3 Results of hybridization of sunflower small-insert library clones containing retrotransposon sequences to total genomic DNA of annual and perennial *Helianthus* species and of other Asteraceae

Groups of species whose genomic DNA hybridizes to *H. annuus* retrotransposon sequences	*Gypsy*		Copia	
	nr.	%	nr.	%
Helianthus annuals only	24	15.0	5	11.9
Both *Helianthus* annuals and perennials	87	54.4	20	47.6
Helianthus annuals and perennials, and other Asteraceae	26	16.2	9	21.4
Helianthus perennials only	16	10.0	5	11.9
Other Asteraceae only	2	1.3	1	2.4
Other*	5	3.1	2	4.8
Total	160		42	

Clones showing lack of signal when hybridized to any genomic DNA were excluded

* Other combinations of groups of species

Table 4 Comparisons of hybridization signal intensities of sunflower small-insert library clones containing retrotransposon sequences when hybridized to total genomic DNA of *H. annuus*, other *Helianthus* annual species, and perennial *Helianthus* species

Hybridization comparison	*Gypsy*		*Copia*	
	nr.	%	nr.	%
H. annuus $\gg$ other annuals	–	–	–	–
H. annuus $>$ other annuals	33	20.4	9	21.4
H. annuus $=$ other annuals	106	65.4	26	61.9
H. annuus $<$ other annuals	22	13.6	7	16.7
H. annuus $\ll$ other annuals	1	0.6	–	–
Total	162		42	
Annuals $\gg$ perennials	15	9.2	–	–
Annuals $>$ perennials	63	38.9	7	16.7
Annuals $=$ perennials	79	48.8	33	78.5
Annuals $<$ perennials	5	3.1	2	4.8
Annuals $\ll$ perennials	–	–	–	–
Total	162		42	

Comparisons were made between *H. annuus* and other annuals (mean of three species), up, and between annuals (including *H. annuus*, mean of four species) and perennials (mean of six species), bottom. Hybridization signal intensity as described in the text

$=$, signal intensity differing less than 0.5 units; $>$ or $<$, signal intensity differing more than 0.5 and less than 1.5 units; $\gg$ or $\ll$, signal intensity differing more than 1.5 units

amplified in the *Helianthus* ancestor before the splitting between annuals and perennials occurred.

Hybridization signal intensity of retrotransposon sequences was also compared between *H. annuus* and the other annual species, and between *Helianthus* annuals and perennials (Table 4). The majority of *gypsy* retrotransposons are equally redundant in *H. annuus* and the other annual species. This percentage is reduced when comparing annuals to perennials. A slightly different pattern is found for *copia* retrotransposons (Table 4): these retroelements are for the vast majority equally redundant between *H. annuus* and other annuals (61.9%), and also between annuals and perennials (78.5%).

Hybridization signal intensity on each clone depends on both sequence conservation and copy number of that sequence in a given species. Anyway, comparisons between hybridization patterns using genomic DNAs of different species can help to establish relationships between cultivated sunflower and other *Helianthus* species or other Asteraceae.

In our experiments, pairwise comparisons between species were made for each clone by plotting the intensity of hybridization estimated as above. The correlation coefficient R was calculated and used as a parameter to indicate the level of relationship between two species. Cluster analysis was conducted on *R* estimates and the resulting clusters were expressed as a dendrogram (Fig. 7). When coefficients of similarity were estimated in a different way (see Materials and methods), the dendrograms turned out to be identical in topology (data not shown).

The dendrograms obtained from the analysis of hybridization (i.e., based on the sequence redundancy) are reported in Fig. 7. In Fig. 8, a dendrogram based on DNA sequences of four single-copy genes (encoding a dehydrin, a drought-responsive-element binding protein, an early light-induced protein, and a non-specific lipid transfer protein) summing up to a total of 2,744 bp, obtained by NJ method, is reported for comparison. The dendrogram obtained from the analysis of all clones (Fig. 7a) confirms the subdivision of the genus between annual and perennial species. *H. annuus* is placed in an intermediate position between annuals and perennials. Because of the autonomous nature of transposable elements, coevolution between retrotransposons and the host genome is often not complete (Capy et al. 2000). Macroarray hybridization data were therefore also analyzed excluding clones (229/1344) containing retrotransposon and rDNA sequences (15/1344),

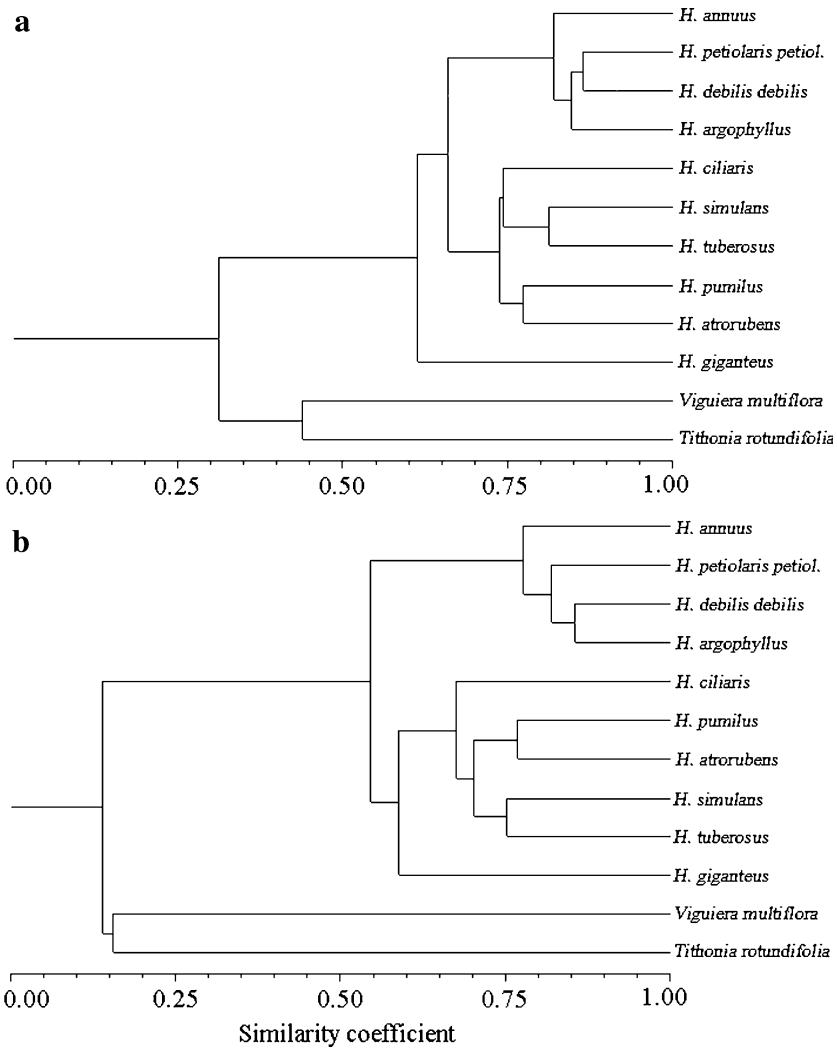

Fig. 7 Dendrograms obtained by cluster analysis of correlation coefficients between signal intensities after hybridisation of genomic DNAs of different species to all clones of the library (**a**) and to the clones containing retrotransposon sequences only (**b**)

which are usually redundant and much conserved. The resulting dendrogram was very similar to that obtained using the complete set of clones (data not shown). Despite possible reservations on the use of retrotransposon sequences to infer relationships among species, hybridization data of retrotransposon sequences-containing clones only produced a dendrogram showing a single topological difference that places *Helianthus giganteus* closer to the other perennials (Fig. 7b), as observed also from sequence-based phylogeny (Fig. 8).

The genetic similarity matrices based on hybridization filters (i.e., evidencing the repeat component of genomes) and on gene sequences were plotted between pairs of species (Fig. 9) and compared using the Mantel test. Significant coefficients were established both using hybridization on all clones or on retrotransposon-containing clones only ($r = 0.8923$ and $r = 0.9109$ for $P = 0.0008$, respectively), suggesting that DNA amplification of the repetitive elements largely predates the species divergence in this genus. Though such a good concordance between gene sequence- and hybridization-based phylogenies, some differences can be observed. This indicates that either precise relationship is still to be established among *Helianthus* species or that the specific mode of evolution of repeated sequences does not match exactly the phylogenetic relationships between species.

Discussion

Sample sequencing of a small-insert genomic library from sunflower provided a set of sequences that were used to analyze the composition of the sunflower genome in terms of types and abundance of repetitive elements.

In accordance with results from other plant species studied so far, the largest recognizable component of the repetitive fraction of the sunflower genome appears to be represented by Class I transposable elements, i.e., retrotransposons and particularly LTR retrotransposons. LINE elements appear to be equally rare as in maize

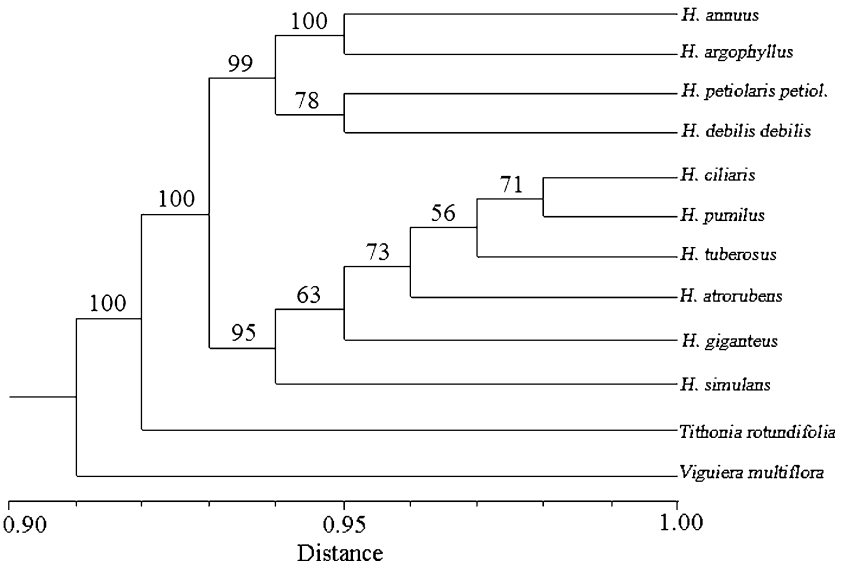

Fig. 8 Majority rule bootstrap consensus tree, based on nucleotide sequences corresponding to four single-copy gene sequences subjected to NJ analysis. Percent bootstrap values are listed at each node

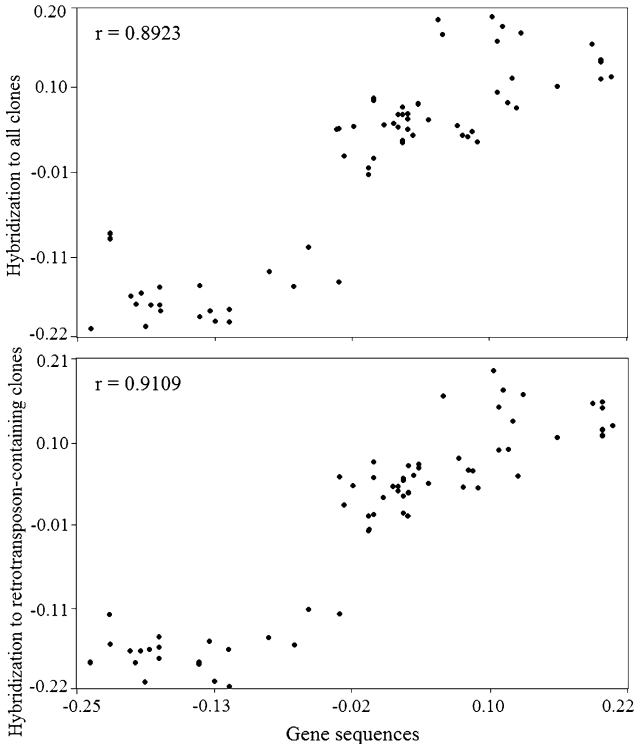

Fig. 9 Two-dimensional representations of genetic distances estimated by gene sequence data and hybridization of genomic DNAs of different species to all small insert library clones (above) or only to retrotransposon-containing clones (below). The distance matrix correlation according to Mantel test is reported

(Meyers et al. 2001) but rarer than in the much smaller genome of grape (The French-Italian Public Consortium for grape genome characterization 2007), or in the similarly sized genomes of *Gossypium* species (Hawkins et al. 2006), and of legumes (Hill et al. 2005; Macas et al. 2007). Within LTR-retrotransposons, the *gypsy* superfamily appears to be more than fourfold more abundant than the *copia* one. Plant species show a varying relative abundance of *gypsy* versus *copia* elements, ranging from a 3:1 ratio in rice (The International Rice Genome Sequencing Project 2005) to a 1:2 ratio in grapevine (The French-Italian Public Consortium for grape genome characterization 2007). Maize shows an approximately equal abundance of the two classes (Meyers et al. 2001), similar to other cereal species with large genomes such as wheat and barley (Paux et al. 2006; Vicient et al. 2005). Species of the *Gossypium* genus show a variable proportion of *gypsy* versus *copia* elements with *gypsy* elements prevailing in species with larger genome sizes. Class II elements (DNA transposons) amounted to a minor fraction of the sunflower genome, similar to what was observed in maize, in legumes, and in *Gossypium,* but in contrast to observations made in rice and *Brassica* where they make up 12 and 6% of the genome, respectively (Jiang et al. 2004; Jiang and Wessler 2001). The genic fraction was estimated to represent about 4% of the sunflower genome, a proportion that is similar to that estimated for maize (5%, Meyers et al. 2001), where, however, the identification of putative genes in the genomic sequences was aided by the availability of a large set of EST sequences.

Both the clustering of sequences using the CAP3 assembler and the copy number estimation using slot blot hybridizations revealed that the only families of repetitive sequences reaching high copy number in the sunflower genome are either LTR-retrotransposons of the *gypsy* superfamily or unidentified sequences. The most abundant family appears to be that corresponding to Contig 61 that extends over 4.8 kb and reaches a number of 27,000 copies per haploid genome. Sequence analysis of Contig 61 revealed no structural feature that could help in the

classification of this family of sequences that therefore awaits further characterization. No *copia* retrotransposon family reaches a copy number of 10,000: this is in marked contrast to what is observed in large cereal genomes where most of the families that are predicted to have more than 40,000 copies per haploid genome are of the *copia* superfamily (Vitte and Bennetzen 2006). The prevalence of *gypsy* elements among the high abundance families of repetitive sequences makes the sunflower genome more similar to that of *Gossypium* species with large genome sizes, where the expansion has been specifically attributed to the amplification of *gypsy* LTR-retrotransposons (Hawkins et al. 2006). A marked difference in copy number distribution of repetitive families is observable between sunflower on one hand and cereals with large genomes on the other, due to the lack of families that have reached very large copy number in sunflower. Families present in more than 50,000 copies have been found in both dicots and monocots (Meyers et al. 2001; Vicient et al. 2005; Neumann et al. 2006; Vitte and Bennetzen 2006).

A good concordance in the abundance estimates was observed when comparing the results of computational analysis with those of slot blot and Southern blot hybridization experiments. Southern blot hybridization experiments revealed a general hypermethylation of all the different families of repetitive sequences analyzed, in accordance with a large body of data showing hypermethylation of repeats in plants (Rabinowicz et al. 2005).

FISH analysis of selected repeated sequences allowed us to gain further insight into the genome organization of the sunflower. Preferential localization around the centromeres is suggested for the analyzed *gypsy*-like LTR-retrotransposon, as already observed in *Helianthus* species for another *gypsy* element by Santini et al. (2002). It can be observed that *gypsy* and *copia* elements have often complementary patterns of preferential localization, with the former located mostly in the centromeric regions, or around the centromeres, as for example in the genus *Beta* (Gindullis et al. 2001) and in cereals (Presting et al. 1998; Li et al. 2004; Liu et al. 2008), and the latter being rare or absent around the centromeres (Heslop-Harrison et al. 1997; Pich and Schubert 1998). Santini et al. (2002) also reported that a *copia*-like element was rare or absent at the centromeric chromosome regions and abundant at the chromosome ends. The present FISH analysis, using another *copia*-element, showed on the contrary a disperse distribution along all chromosomes, suggesting that different subfamilies of *copia* retrotransposons may have a partly different chromosomal organization.

Satellite DNA represents only 0.37% of the library, and, presumably, of the *Helianthus annuus* genome. The two tandem-repeated sequences analyzed in this work show a copy number lower than 10,000 copies per haploid genome, i.e., relatively low for such sequences. It is possible that some of the sequences containing direct repeats are blocks of tandem repeats in which some repeat units underwent sequence degeneration. No subtelomeric repeats were found, probably because of the relatively small dataset. Two sequences containing direct or tandem repeats showed a disperse distribution with possible preferential centromeric localization. Such a disperse distribution was unexpected for such sequences. Probably, these sequences, whose copy numbers were estimated to be 3,000 and 2,600 per haploid genome, respectively, form a number of short tandem arrays scattered along all chromosomes. Actually, no hybridization signal was obtained after labeling these sequence probes by nick-translation, when the incorporation of labeled nucleotides per sequence is much lower than using polymerase chain reaction. Such a structure is also confirmed by the absence of apparent DAPI bands in sunflower metaphase chromosomes that we observed during FISH preparation. The organization of the repetitive units of satellite DNA of sunflower appears to be unusual in plants. A similar organization was observed in *Zamia paucijuga* (Cafasso et al. 2003).

Hybridization of the small-insert library clones to labeled total genomic DNA from *Helianthus annuus* provided yet another mean to estimate abundance in the genome: the intensity of the hybridization signal should be proportional to the copy number of the specific sequence. Only a small proportion of the library hybridized strongly, indicating high copy sequences that corresponded to Contig 61, *gypsy* retrotranposons, rDNA and unidentified sequences. Similar observations were made in maize in the same type of experiment (Meyers et al. 2001). Forty percent of the sequences belonged to clones that gave a detectable hybridization signal, indicating their repetitive nature. Approximately a third of these had not been previously detected as repetitive by either BLAST-based analysis or by clustering analysis.

Using a combination of computational and laboratory approaches (BlastN and BlastX homology searches, clustering of sequences using the CAP3 assembler, hybridization to total genomic DNA), we estimated that at least 62% of the genome is made of repetitive sequences of which almost two-thirds remain as yet uncharacterized in nature. In a previous study, reassociation kinetics of sunflower DNA evidenced that repetitive DNA should account for around 60% of the genome; $C_o t$ values indicated that the majority of repeated sequences should be considered as medium repeated, with complexity ranging from 1,000 to 10,000 (Cavallini et al. 1986). This estimate is lower than those previously obtained for other plant species with large genome sizes such as maize (77% repetitive DNA, using a very similar approach; Meyers et al. 2001) and wheat (70% from the analysis of BAC clone sequences; Wicker et al.

2001). The identification of transposable elements, the largest component of the repetitive fraction in plant species, was however more difficult in sunflower than in maize due to the paucity of sequences of previously described and annotated elements. While a fraction of the coding portions of the elements may have been recognized through the BlastX homology searches, any of the non-coding portions (e.g., the long terminal repeat regions of LTR-retrotransposons) may have been much more difficult to detect using BlastN homology searches against elements from other species due to the high rate of sequence evolution of transposable elements (Ma and Bennetzen 2004). Internal clustering of genomic sequences and/or hybridization to genomic DNA appear to be more effective strategies to identify repetitive sequences than BLAST homology searches in a species where transposable elements have not previously been well characterized, even though they do not still allow to classify the majority of sequences according to their structural or functional characteristics.

Hybridization analysis was also useful to elucidate the evolution of the repetitive component of *Helianthus annuus* and its relationships to that of other species from the Asteraceae family. The similarly sized annual species that we analyzed all show a very similar composition in repeat sequences, leading to the conclusion that the majority of the repetitive elements may have been in proportions similar to those seen in sunflower at the time of divergence of these species and that therefore the genome amplification also largely predates such divergence. *H. annuus* diverged from *H. petiolaris*, one of the annual species we analyzed, between 750,000 and one million years ago (Rieseberg et al. 1991): this provides a lower bound for the repetitive sequence amplification events. A more accurate dating of amplification events of the LTR-retrotransposon component will require a comparison of the two LTR sequences from single elements that could be obtained from the sequencing of large genomic regions (SanMiguel et al. 1996).

Perennial species show some differences in hybridization signals to the gridded small insert library that result in a lower number of strongly hybridizing sequences that could either be attributed to differential amplification of these sequences in the annual species or more simply to sequence divergence that prevents cross hybridization. An additional difference between annual and perennial species is visible in the LTR-retrotransposon class, where *gypsy* like elements are more represented in the annuals than in the perennials: such a difference is not visible for *copia*-like elements, attesting a different amplification history of the two superfamilies of LTR-retrotransposons in the *Helianthus* genus. That retrotransposon superfamilies are subjected to different amplification histories during the evolution of a species has been recently reported in wheat, in which *copia* and *gypsy* superfamilies are differently represented in the A and B genomes (Charles et al. 2008). In sunflower, the difference observed between annuals and perennials confirms that the amplification of *gypsy* retrotransposons occurred before the splitting between annual and perennial species, but in part continued also after such splitting. Only minor amplification should have taken place after the separation among annual species, as shown by the high proportion of retrotransposon sequences that are equally redundant between *H. annuus* and other annuals. On the contrary, most of *copia* retrotransposons presumably amplified before the divergence within the *Helianthus* genus. Alternatively, the amplification of *gypsy* elements might have preceded that of *copia* ones, so that higher sequence divergence occurs in *gypsy* than in *copia* retroelements.

Different genome sizes have been reported for some of the *Helianthus* species analyzed in this work (see Table 1). Among species at the same ploidy level, *H. annuus*, *H. debilis,* and *H. petiolaris* show similar genome size, while *H. argophyllus* and *H. giganteus* have larger (around 30%) genomes. Some retrotransposon containing clones of the *H. annuus* small insert library were shown to be especially redundant in *H. giganteus*; therefore, they possibly contributed to the expansion of the genome of this species, as already reported for *Oryza australiensis*, in which the amplification of retrotransposons has recently determined a large increase in genome size, in comparison with other species of rice (Piegu et al. 2006).

The two other *Asteraceae* species that we analyzed, *Viguiera multiflora* and *Tithonia rotundifolia,* share only a small fraction of their repetitive sequence component with *H. annuus*. Phylogenetic analyses have previously suggested a common progenitor for *Helianthus*, *Viguiera,* and *Tithonia* (Sossey-Alaoui et al. 1998; Soltis and Soltis 2000; Schilling 2001). Our results indicate a closer relationship to *Helianthus* of *Viguiera* than *Tithonia*.

Cluster analysis based on hybridization signal intensities, i.e., based on the degree of repetitiveness and on the sequence similarity of the *H. annuus* genomic sequences in the different species tested, produced dendrograms that are in significant concordance with a gene sequence-based phylogeny, though some differences are observed in the position of species within annual and perennial groups of *Helianthus*, probably related to the autonomous nature of the repetitive component of the genome, that probably underwent some changes after *Helianthus* speciation. Contrary to the expectation that the molecular clock-independent evolution and amplification of LTR-retrotransposons may hinder their utilization to infer phylogenetic relationships, the dendrogram obtained using only retrotransposon-containing clones was in an as good an agreement with the supposed relationships among

species within *Helianthus* as the dendrogram obtained after hybridization to all clones of the library.

We have shown that the sunflower genome has a composition resembling that of the similarly sized maize genome in terms of repeat types and ages; however, it appears that, unlike in maize, where single families can make up 10% or more of the genome, no single transposable element family has been amplified to very high levels in sunflower. Species distribution of repetitive elements showed that, within the *Helianthus* genus, annual species share the same repetitive sequence component with *H. annuus* while perennial ones are more differentiated, and that very little similarity is observed outside of the genus.

Acknowledgments This work was supported by PRIN-MIUR, Projects "Caratterizzazione della componente ripetitiva di genomi complessi in specie vegetali: modelli per angiosperme e gimnosperme" and "Variabilità di sequenza ed eterosi in piante coltivate".

References

Brenner S, Elgar G, Sandford R, MacRae A et al (1993) Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. Nature 366:265–268

Cafasso D, Cozzolino S, De Luca P, Chinali G (2003) An unusual satellite DNA from *Zamia paucijuga* (Cycadales) characterised by two different organisations of the repetitive unit in the plant genome. Gene 311:71–79

Capy P, Gasperi G, Biémont C, Bazin C (2000) Stress and transposable elements: co-evolution or useful parasites? Heredity 85:101–106

Cavallini A, Zolfino C, Cionini G, Cremonini R, Natali L et al (1986) Nuclear DNA changes within *Helianthus annuus* L.: cytophotometric, karyological and biochemical analyses. Theor Appl Genet 73:20–26

Ceccarelli M, Sarri V, Natali L, Giordani T, Cavallini A et al (2007) Characterization of the chromosome complement of *Helianthus annuus* by in situ hybridization of a tandemly repeated DNA sequence. Genome 50:429–434

Charles M, Belcram H, Just J, Huneau C, Viollet A et al (2008) Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics 180:1071–1086

Diaz-Martin J, Almoguera C, Prieto-Dapena P, Espinosa JM, Jordano J (2005) Functional interaction between two transcription factors involved in the developmental regulation of a small heat stress protein gene promoter. Plant Physiol 139:1483–1494

Doyle JJ, Doyle JL (1989) Isolation of plant DNA from fresh tissue. Focus 12:13–15

Elgar G, Clark MS, Meek S, Smith S et al (1999) Generation and analysis of 25 Mb of genomic DNA from the pufferfish *Fugu rubripes* by sequence scanning. Genome Res 9:960–971

Felsenstein J (1989) PHYLIP: phylogeny inference package. Cladistics 5:164–166

Gindullis F, Desel C, Galasso I, Schmidt T (2001) The large scale organization of the centromeric region in *Beta* species. Genome Res 11:253–265

Giordani T, Natali L, Cavallini A (2003) Analysis of a dehydrin encoding gene and its phylogenetic utility in *Helianthus*. Theor Appl Genet 107:316–325

Goff SA, Ricke D, Lan TH, Presting G et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Harter AV, Gardner KA, Falush D, Lentz DL, Bye R et al (2004) Origin of extant domesticated sunflowers in eastern North America. Nature 430:201–205

Hawkins JS, Kim HR, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res 16:1252–1261

Heslop-Harrison JS, Brandes A, Taketa S, Schmidt T, Vershinin AV et al (1997) The chromosomal distribution of Ty1-*copia* group retrotransposable elements in higher plants and their implications for genome evolution. Genetica 100:197–204

Hill P, Burford D, Martin DMA, Flavell AJ (2005) Retrotransposon populations of *Vicia* species with varying genome size. Mol Genet Genom 273:371–381

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Jiang N, Wessler SR (2001) Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. Plant Cell 13:2533–2564

Jiang N, Feschotte C, Zhang X, Wessler SR (2004) Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Curr Opin Plant Biol 7:115–119

Lentz DL, Pohl MD, Alvarado JL, Tarighat S, Bye R (2008) Sunflower (*Helianthus annuus* L.) as a pre-Columbian domesticate in Mexico. Proc Natl Acad Sci USA 105:6232–6237

Li W, Zhang P, Fellers JP, Friebe B, Gill BS (2004) Sequence composition, organization, and evolution of the core Triticeae genome. Plant J 40:500–511

Liu Z, Yue W, Li D, Wang RR, Kong X, Lu K, Wang G, Dong Y, Jin W, Zhang X (2008) Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres. Chromosoma 117:445–456

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA 34:12404–12410

Macas J, Neumann P, Navrátilová A (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. BMC Genomics 8:427–442

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11:1660–1676

Neumann P, Koblizkova A, Navratilova A, Macas J (2006) Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. Genetics 173:1047–1056

Ouvrard O, Cellier F, Ferrare K, Tousch D, Lamaze T et al (1996) Identification and expression of water stress- and abscisic acid-regulated genes in a drought-tolerant sunflower genotype. Plant Mol Biol 31:819–829

Paux E, Roger D, Badaeva E, Gay G, Bernard M et al (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J 48:463–474

Pearce SR, Harrison G, Li D, Heslop-Harrison J et al (1996) The Ty1-*copia* group retrotransposons in *Vicia* species: Copy number, sequence heterogeneity and chromosomal localisation. Mol Gen Genet 250:305–315

Pich U, Schubert I (1998) Terminal heterochromatin and alternative telomeric sequences in *Allium cepa*. Chromosome Res 6:315–321

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A et al (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res 16:1262–1269

Presting GG, Malysheva L, Fuchs J, Schubert I (1998) A Ty3/*gypsy* retrotransposon-like sequence localized to the centromeric regions of cereal chromosomes. Plant J 16:721–728

Rabinowicz D, Citek R, Budiman MA, Nunberg A, Bedell JA et al (2005) Differential methylation of genes and repeats in land plants. Genome Res 15:1431–1440

Rieseberg LH (1995) The role of hybridization in evolution: old wine in new skins. Am J Bot 82:944–953

Rieseberg LH, Beckstrom-Sternberg SM, Liston A, Arias DM (1991) Phylogenetic and systematic inferences from chloroplast DNA and isozyme variation in *Helianthus* sect. *Helianthus* (Asteraceae). Syst Bot 16:50–76

Rohlf FJ (1998) NTSYS-pc. Numerical taxonomy and multivariate analysis system (version 2.02 j). Exeter Software, Setauket

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd edn. Cold Spring Harbor Laboratory, Cold Spring Harbor

SanMiguel P, Tikhonov A, Springer PS, Edwards KJ, Lee M et al (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765–768

Santini S, Cavallini A, Natali L, Minelli S et al (2002) Ty1/*copia*- and Ty3/*gypsy*-like DNA sequences in *Helianthus* species. Chromosoma 111:192–200

Schilling EE (1997) Phylogenetic analysis of *Helianthus* (Asteraceae) based on chloroplast DNA restriction-site data. Theor Appl Genet 94:925–933

Schilling EE (2001) Phylogeny of *Helianthus* and related genera. Oleagineaux Corps Gras Lipides 8:22–25

Schilling EE, Heiser CB (1981) Infrageneric classification of *Helianthus* (Compositae). Taxonomy 30:393–403

Schilling EE, Linder CR, Noyes RD, Rieseberg LH (1998) Phylogenetic relationships in *Helianthus* (Asteraceae) based on nuclear ribosomal DNA internal transcribed spacer region sequence data. Syst Bot 23:177–187

Schwarzacher T, Leitch AR, Bennett MD, Heslop-Harrison JS (1989) In situ localization of parental genomes in a wide hybrid. Ann Bot 64:315–324

Soltis ED, Soltis PS (2000) Contributions of plant molecular systematics to studies of molecular evolution. Plant Mol Biol 42:45–75

Sossey-Alaoui K, Serieys H, Tersac M, Lambert P, Schilling EE et al (1998) Evidence for several genomes in *Helianthus*. Theor Appl Genet 97:422–430

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

The French-Italian Public Consortium for grape genome characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

The International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

The International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Thompson JD, Desmond G, Gibson H, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 22:4673–4680

Timme RE, Simpson BB, Linder CR (2007) High-resolution phylogeny for Helianthus (Asteraceae) using the 18 s–26 s ribosomal DNA external transcribed spacer. Am J Bot 94:1837–1852

Tuskan GA, DiFazio S, Jansson S, Bohlmann J et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604

Vicient CM, Kalendar R, Anamthawat-Jónsson K, Suoniemi A, Schulman AH (1999) Structure, functionality, and evolution of the BARE-1 retrotransposon of barley. Genetica 107:53–63

Vicient CM, Kalendar R, Schulman AH (2005) Variability, recombination, and mosaic evolution of the barley BARE-1 retrotransposon. J Mol Evol 61:275–291

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Nat Acad Sci USA 103:17638–17643

Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. Plant J 26:307–316

Wicker T, Sabot F, Hua-Van A, Bennetzen JL et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982

Wilson RK, Mardis ER (1997) Genome analysis: a laboratory manual. vol 1. Analyzing DNA. Cold Spring Harbor Laboratory Press, New York

Zhang D, Yang Q, Bao W, Zhang Y, Han B et al (2005) Molecular cytogenetic characterization of the *Antirrhinum majus* genome. Genetics 169:325–335