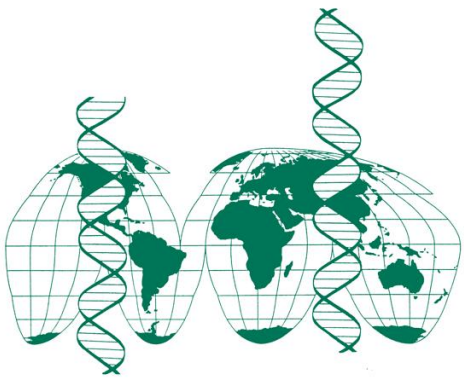
ISSN 0016-6707, Volume 138, Number 6

CODEN GENEA3 Vol. 138 No. 6 June 2010 ISSN 0016 6707

Genetica

An International Journal
of Genetics and Evolution



Springer

Genetica (2010) 138:567–578
DOI 10.1007/s10709-010-9449-9

Genomic screening in dioecious "yerba mate" tree (*Ilex paraguariensis* A. St. Hill., Aquifoliaceae) through representational difference analysis

Alexandra Marina Gottlieb · Lidia Poggio

Received: 30 March 2009 / Accepted: 25 September 2009 / Published online: 10 March 2010

Abstract The "yerba mate" tree, *Ilex paraguariensis,* is a functionally dioecious crop species with economic relevance in several South American countries. We report a genomic screening accomplished through representational difference analysis (RDA) in male and female *I. paraguariensis* trees. The aim of the present paper was to investigate the occurrence of sex-related genomic differences in order to develop an early gender detection molecular method that could help reducing energy inputs during the "yerba mate" processing and that could be suitable for breeding programs. An intra-experiment redundancy was detected via SSCP analysis and sequence characterization. Taking together both reciprocal RDA assays, fragments isolated can be discriminated into three main categories. The first category of fragments shows spurious affinities with available deposited sequences and could be considered as specific to *I. paraguariensis*. The second category comprises sequences identified as organellar or ribosomal plant DNA. Sequences grouped in the third category involve clones akin to conserved domains of retrotransposons (RNaseH, integrases and/or chromodomains) from at least two distinct lineages of Ty3/Gypsy retrotransposons and one from Ty1/Copia retroelements, which in addition are associated to sex determination regions of the Solanaceae, Caricaceae and Salicaceae. A contig sequence was assembled that codes for an integrase core domain and a chromodomain. The phylogenetic analysis of the so-called IPRE (for *I. paraguariensis* retroelement) integrase domain indicates that it belongs to the Del lineage of the Chromoviridae. This is the first report of mobile elements isolated and detected from the "yerba mate" tree. Although RDA derived fragments, so far tested, have been retrieved from both sexes with similar sequences, association to sex related regions cannot be completely discarded. Implications of present results are further discussed.

Keywords *Ilex paraguariensis* · Gender detection · RDA · Transposable elements · Chromoviridae

Introduction

The "yerba mate" tree, *Ilex paraguariensis* A. St. Hill. (Aquifoliaceae Bartl.) is an economically important crop widely employed in several South American countries, like Argentina, Brazil, Paraguay and Uruguay, to prepare a stimulating beverage called "mate". The species shows a basal phylogenetic position, relative to other *Ilex* species, that contradicts its current classification (Gottlieb et al. 2005). The "yerba mate" tree is functionally dioecious, exhibiting diclinous flowers in which one of the sexes is sterile or abortive (Giberti 2001). Both, the mechanisms

Electronic supplementary material The online version of this article (doi:10.1007/s10709-010-9449-9) contains supplementary material, which is available to authorized users.

A. M. Gottlieb (✉) · L. Poggio
Laboratorio de Citogenética y Evolución (LaCyE), Departamento de Ecología, Genética y Evolución, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Intendente Güiraldes y Costanera Norte s/n, 4to. Piso, Pabellón II, Ciudad Universitaria, C1428EHA Ciudad Autónoma de Buenos Aires, Argentina
e-mail: gottlieb@ege.fcen.uba.ar

A. M. Gottlieb · L. Poggio
Consejo Nacional de Investigaciones Científicas y Técnicas, Avenida Rivadavia 1917, C1033AAJ Ciudad Autónoma de Buenos Aires, Argentina

that triggers development of one set of organs and arrestment of the other set, and the sex determination system operating in *I. paraguariensis* are currently unknown. It has been suggested that hormonal and/or environmental factors might be involved in sex determination in this species (Winge 1997); however, this hypothesis has not been empirically tested. Even though cytogenetic studies in "yerba mate" reported regular male meiosis ($2n = 40$) in which no heteromorphic bivalents were detected (Barral et al. 1995; Greizerstein et al. 2004), *Ilex* species remain largely unexplored on classical and molecular cytogenetics. Given that there is a wealth of basic data lacking for *I. paraguariensis* and that several systems of sex determination have been described in plants (Vyskot and Hobza 2004; Ming et al. 2007; Jamilena et al. 2008, among others), almost any sex determination system and genetic mechanism are plausible to take place in "yerba mate". For instance, it could be postulated the occurrence of homomorphic sex chromosomes, at any evolutionary stage, or a multilocus autosomal system could be invoke.

We report a screening of male and female *I. paraguariensis* genomes through the representational difference analysis (RDA) technique. By applying this technique, we expect to gather sex-related sequence data. The RDA methodology was originally developed to identify polymorphisms in human neoplasia (Lisitsyn et al. 1993), but it has also been successfully applied in many plants. For instance, it allowed the identification of male-specific fragments in *Silene latifolia* (Donnison et al. 1996). Zoldos et al. (2001) reported the use of this technique to investigate genomic differentiation between two species of *Quercus*, and Panaud et al. (2002) to study genomic differences between rice (*Oryza sativa*) and foxtail millet (*Setaria italica*) and to clone several rice-specific transposable elements. Likewise, Nekrutenko and Baker (2003) applied RDA in cotton (*Gossypium hirsutum*) to develop markers specific to the A and D subgenomes and to elucidate the genome composition of allopolyploid species. More recently, Sabot et al. (2004) isolated AT-rich repeated sequences and transposable elements from the allohexaploid *Triticum aestivum*.

The aim of the present paper was to investigate the occurrence of sex-related genomic differences in male and female *I. paraguariensis* as an initial step to develop an early gender detection molecular method. The sex of these plants cannot be determined until plants flower, which could represent three to 10 years post seed germination. Therefore, the development of such a method will permit discrimination of fruit producing plantlets, which in turn, will benefit crop producers by reducing energy costs during the "yerba mate" processing. In addition, an early sex detection system will be helpful for breeding programs.

Materials and methods

Plant material

The leaf samples of *I. paraguariensis* were provided by Banco de Germoplasma de Yerba Mate y Té, at the Estación Experimental INTA Cerro Azul (EEINTA CA; Misiones, Argentina) and were preserved in silica-gel. Most "yerba mate" accessions available originate from seeds pooled according to geographical provenance. In order to work with a manageable amount of genetic variation, we selected argentine accessions No 50 (R. Prov. 210, Cnia. Alberdi, Depto. San Ignacio, Misiones) and No 51 (R. N. 14, Mun. Campo Viera, Depto. Oberá, Misiones), which involve registered half-sibling male and female plants. We employed in total, three male and four female plants. Fifteen additional samples, detailed below, were used for further studies. The gender of the trees was determined in the field by Ing. Agr. Dr. GC Giberti.

DNA extraction

The DNeasy Plant kit (QIAGEN) was used for DNA extraction following manufacturer's instructions. Quality control and quantification was carried out by agarose gel (0.8% w/v) electrophoresis and by comparison with a DNA molecular-size standard (Lambda EcoRI/HindIII, Promega Corp.). Gels were stained with ethidium bromide and photographed under UV light.

Representational difference analysis

Two pools of genomic DNA were independently prepared. The female pool comprised 100 ng of genomic DNA from each female accession (50-4; 50-6; 51-2 and 51-5); whereas the male pool consisted of *ca*. 133 ng from each male sample (50-7; 50-10 and 51-6). Considering that we ignore which gender is the heterogametic one, we performed two reciprocal assays. In the first assay, the female pool was employed as tester DNA and was used in limiting amount, whereas the male pool was used as driver DNA to block sequences shared by both sexes. Hereafter this experiment will be referred to as RDA-1. In the second assay, the male pool was used as the tester and the female pool as the driver, hereafter RDA-2. Briefly, the RDA procedure involves two major steps; the first reduces the genome complexity by producing amplicons via restriction digestion and adapter ligation followed by PCR amplification. The second step enriches the samples in differential regions via subtractive hybridization followed by amplification of a target sequence population (Allen et al. 2003). This technique relies on the occurrence of either missing restriction sites or genome rearrangements, such as

deletions, insertions, duplications and translocations (Chen et al. 1998). We followed the RDA protocol of Panaud et al. (2002) and performed three rounds of subtractive hybridization and enrichment, employing the oligonucleotide adaptor sets R, N and J, described in that work. All PCR amplification runs were conducted in an Eppendorf Mastercycler (Perkin-Elmer Corp.). The first round of subtractive hybridization and enrichment for both RDA assays, involved the mix of the N-ligated tester with R-ligated driver DNA, in a tester/driver ratio of 1:100. Prior to the second round the N-adaptors were removed from the tester and replaced by the J-adaptor set. In this round, a 1:100 and 1:1,000 tester/driver ratios were assayed for the RDA-1 experiment; whereas a 1:100 and 1:900 ratios were used for the RDA-2. In the third round, the J-adaptors were removed from the tester and replaced by the N-adaptor set. In this cycle, a 1:500 and 1:1,000 ratios were used for the RDA-1; and for RDA-2 we tested 1:100, 1:1,000 and 1:10,000 ratios. Provisional results were checked at each step by electrophoresis in 2% (w/v) agarose gels, as previously described.

Cloning and miniprep

The final PCR products of each RDA experiment were cloned using the pGEM-T Easy Vector Cloning Kit (Promega) following manufacturer's instructions. The population of RDA-1 amplicons derived from the 1:500 tester/driver ratio was employed for cloning; whereas in the RDA-2 the 1:1,000 product was cloned. In both cases, these populations were obtained from the third round of subtractive hybridization and enrichment. Purification of the plasmid vector plus the insert was achieved using the QIAprep Spin Miniprep kit (QIAGEN). Clones were named correlatively within each assay.

Single strand conformational polymorphism

Assessment of the sequence variation of the inserts was accomplished via single strand conformational polymorphism (SSCP) analyses (Orita et al. 1989). For this aim, we followed the protocol employed in Lia et al. (2007). The rationale is that fragments with the same nucleotide sequence exhibit identical SSCP banding pattern. PCR amplicons were obtained from each purified clone by using universal primers T7 and SP6. Air-dried gels were scanned and banding patterns were visually analyzed. Thus, clones were clustered according to their banding patterns.

Sequence characterization

From each SSCP group at least one clone was selected for characterization. Nucleotide sequences were obtained on an Applied Biosystems automated 3130 XL DNA sequencer (Perkin Elmer). Chromatograms were proofread and sequences manually edited with the BioEdit Sequence Alignment Editor program (Hall 1999). Boundaries of the plasmid vector and of the oligonucleotide adaptors were determined and eliminated from each sequence, plus all sequences were compared to each other, and the GC content was calculated. Blast searches were performed for each clean sequence using the BlastN, TBlastX and BlastX algorithms (http://www.ncbi.nlm.nih.gov/Blast). From these comparisons, we recorded the number of significant hits, the alignments maximum score, the percentage of identity between query and retrieved sequences and its description or annotation. An additional BlastX search was conducted against the cores database of the Gypsy DataBase (GyDB; Llorens et al. 2008). Sequence matches were accepted with *E*- value thresholds (probability of random match) $\leq 10^{-4}$. Occurrence of open reading frames was investigated with the application ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf). The RepeatMasker program (version open-3.1.9; Smit, Hubley & Green, at http://repeatmasker.org) was used for screening for interspersed repeats and low complexity DNA sequences, assuming queries to be *Arabidopsis thaliana* (RepBase update 20080801). Sequences obtained were deposited in Genbank (Tables 1 and 2).

PCR evaluation of RDA fragments

In order to evaluate via PCR the occurrence in both sexes of selected RDA fragments, we designed primers for those fragments using the PRIMER3 program (release 1.1.0, Rozen and Skaletsky 2000; Table 1, see Supplementary materials). As DNA templates we used 15 additional samples that were previously obtained (Gottlieb et al. 2005) and correspond to EEINTA CA accessions No 1 ♂; 1 ♀; 34 ♀; 34 ♂; 37 ♂; 1-74 ♀ (Guaraní, Misiones, Argentina); 100 ♀ (Paraná, Brazil); 138 ♀ (Rio Grande do Sul, Brazil); 28 ♀; 28 ♂; 27♀(Santa Catarina, Brazil); and SI-16 ♂; SI-19 ♀; SI-49 ♂; Y383 ♀ from the Establecimiento Las Marías (Misiones, Argentina). PCRs were done using 1 µl of DNA, 2.5 µl buffer 10X (Invitrogen), 0.2 mM of each dNTP (Invitrogen), 1.5 mM $MgCl_2$ (Invitrogen), 60 ng of each right and left primer, 0.5 U of Taq polymerase (Invitrogen) in a final volume of 25 µl. PCRs were carried out for 1 min at 94°C, followed by 25 cycles of 45 s at 94°C, 45 s at 52°C and 1 min at 72°C, and 5 min at 72°C. PCR products were visualized by gel electrophoresis (1.5% w/v) as described above.

RDA fragments linkage relationship

To elucidate the relative order of selected clones, a series of PCR runs were undertaken on male and female

Table 1 Sequence characterization and statistics from Blast searches performed on sequences derived from assay RDA-1

RDA clones[a]; Genbank accession number	Insert length (in bp)	Insert GC content[b] (in %)	BlastN results[c]	TBlastX results[d]	BlastX results[d]	GyDB Blastx results[e]
1B9 GS598722	265	**46.62**	0; **0.055**; Human; AL162294.8; 90%; 12%	0; **1.1**; *Sus scrofa*; NM001110429.1; 55%; 63%; (33%)	NS	NS
1A10 GS598719	246	51.33	58; 2×10^{-110}–6×10^{-8}; *Nicotiana tabacum* NADH mitochondrial; BA000042.1; 96%; 99%	52; 10^{-42}–10^{-4}; *N. tabacum* NADH mitochondrial; BA000042.1; 98%; 98%; (74%)	NT	NT
1B2 GS598720	327	**43.87**	5; 5×10^{-68}–2×10^{-9}; *Cucurbita pepo* atp9 mitochondrial gene; AY305266.1; 80%; 81%	26; 3×10^{-19}–10^{-18}; *C. pepo* atp9 mitochondrial gene; AY305266.1; 89%; 89%; (27%)	NT	NT
1B6 GS598723	327	**48.77**	100; 5×10^{-16}–10^{-12}; *Plantago major* rDNA 5S; AY684361.1; 100%; 15%	13; 2×10^{-6}–10^{-4}; *Perilla frutescens* rDNA 5S; EF673040.1; 53%; 67%; (39%)	NT	NT
1B10 (1A5; 1A9; 1A11) GS598725	274	50.92	100; 2×10^{-129}–6×10^{-123}; *Ilex opaca* rDNA 18S; AF206938.1; 97%; 100%	>100;3×10^{-50}–10^{-48}; *I. opaca* rDNA 18S; AF206938.1;95%; 95%; (100%)	NT	NT
1A2 (1A4; 1A13; 1A14; 1A18) GS598724	237	**49.15**	100; 7×10^{-20}–9×10^{-6}; *Vitis vinifera* genomic DNA; AM471443; 76%; 64%	100; 6×10^{-17}–4×10^{-15}; *Solanum lycopersicum* genomic DNA; CU915747.3; 60%; 71%; (80%)	100; 6×10^{-14}–3×10^{-11}; hypothetical protein *V. vinifera*; CAN69016.1; 70%; 90%; (92%)	INT_Retrosat-2; 10^{-14}; *Oryza sativa*; AF111709; 52%; 70%
1B13 GS598721	273	**42.65**	>100; 7×10^{-8}–10^{-4}; *Lotus japonicus* genomic DNA; AP009686.1; 91%; 16%	>100; 10^{-23}–3×10^{-19}; *Medicago truncatula* genomic DNA; AC225524.4; 54%; 75%; (92%)	100; 4×10^{-18}–4×10^{-14}; Integrase from putative polyprotein *Solanum habrochaites*; ACB45081.1; 52%; 70%; (92%)	[e1]
1B14 (1A6; 1A7; 1A16; 1B7) (see Contig)	293	54.11	27; 6×10^{-15}–10^{-4}; *Populus trichocarpa* genomic DNA; AC214276.1; 76%; 41%	>100; 4×10^{-25}–10^{-16}; *Iris brevicaulis* IRRE retrotransposon; AY245368.1; 58%; 74%; (99%)	100; 6×10^{-25}–2×10^{-16}; hypothetical protein *V. vinifera*; CAN69016.1; 59%; 74%; (96%)	INT_Peabody; 7×10^{-11}; *Pisum sativum*; AF083074; 66%; 88%
1B15 (1A20) (see Contig)	345	51.3	>100; 10^{-25}–2×10^{-16}; *Petunia integrifolia* self-incompatibility (S2-RNase and SLF2) locus; AY136628; 70%; 73%	>100; 7×10^{-52}–6×10^{-46}; *Carica papaya* Y chromosome; AM778093.1; 68%; 83%; (99%)	100; 5×10^{-42}–6×10^{-36}; Integrase from hypothetical protein *V. vinifera*; CAN69016.1; 66%; 82%; (99%)	INT_Retrosat-2; 7×10^{-40}; *Oryza sativa*; AF111709; 57%; 77%

[a] Between brackets we indicate redundant clones

[b] In bold GC contents <50%

[c] Number of significant hits; *E*-value range for retrieved significant hits or for the first 100 significant hits (when correspond), in bold we indicate the non significant *E*-value retrieved from the best scored hit; accession name of the best scored hit; accession number; percentage of identity between query sequence and best hit; percentage of query coverage between query sequence and best hit

[d] Number of significant hits; *E*-value range for retrieved significant hits or for the first 100 significant (when correspond), in bold we indicate the non significant *E*-value retrieved from the best scored hit; accession name of the best scored hit; accession number; percentage of identity between query and retrieved sequence; percentage of positive shared residues query and retrieved sequences; calculated query coverage, between query sequence and best hit, is indicated between brackets (= (total number of matched residues * 3 * 100)/insert length)

[e] Results shown correspond to the best scored hit: name of the Ty3/Gypsy retrotransposon domain retrieved (INT = integrase, CHR = chromodomain, RNAseH = ribonuclease H); E-value; host name; accession number; percentage of identity between query and retrieved sequences; percentage of positive shared residues

[e1] Showing clone 1B13 affinity to a Copia element, the GyDB result could be misleading (INT_MuLV; 10^{-5}; Murine Leukemia virus; AF221065; 32%; 70%), and thus were excluded from the Table. NT, not tested. NS, no significant similarity was found

Table 2 Sequence characterization and statistics from Blast searches performed on sequences derived from assay RDA-2

RDA clones[a]; Genbank accession number	Insert length (in bp)	Insert GC content[b] (in %)	BlastN results[c]	TBlastX results[d]	BlastX results[d]	GyDB Blastx results[e]
2A1 (2B13) GS598738	226	57.33	0; **0.57**; *Canis familiaris*; AC147784.3; 92%; 12%	0; **0.18**; *Danio rerio*; CR749162; 45%; 59% (49%)	0; **0.57**; *Bifidobacterium gallicum*; ZP03446131.1; 44%; 61% (48%)	NS
2A6 (2A12; 2A14; 2A16; 2B8; 2B10) GS598728	205	56.37	0; **0.041**; *Pan troglodytes*; AC190209.3; 93%; 14%	0; **0.79**; *Rhodobacter sphaeroides*; CP000143.1; 43%; 73% (44%)	0; **8.1**; *Strongyloncentrotus purpuratus*; XP001191724.1; 48%; 57% (51%)	NS
2A8 GS598727	210	**42.38**	0; **0.11**; *Juncus effusus*; AY493568; 100%; 14%	0; **1.8**; *V. vinifera* genomic DNA; AM465566.2; 66%; 83%; (26%)	NS	NS
2B5 GS598731	178	50.85	0; **0.43**; *R. sphaeroides*; CP000661.1; 84%; 21%	1; 10^{-4}; *V. vinifera* genomic DNA; AM483312.1; 38%; 59% (83%)	0; **0.029**; *V. vinifera* hypothetical protein; CAN67018.1; 35%; 57% (94%)	NS
2B14 GS598733	148	**44.22**	0; **4.2**; *Citrobacter koserii*; CP000822.1; 85%; 22%	0; **0.011**; *Rattus norvegicus*; AC118839.5; 53%; 61% (53%)	0; **6.3**; Vibrionales bacterium; ZP01816618.1; 38%; 67% (63%)	NS
2B17 (2A4; 2A7; 2A17; 2B15) GS598730	225	55.72	0; **0.14**; *Mus musculus*; AC117724.14; 88%; 17%	0; **0.14**; *R. sphaeroides*; CP000143.1; 33%; 52% (68%)	0; **4.8**; *Burkholderia ambifaria*; YP_001811337.1; 36%; 52% (91%)	NS
2B21 GS598735	135	**45.14**	0; **0.33**; *Juncus effusus*; AY493568.1; 100%; 15%	0; **3.2**; *Papio anubis*; AC098812.3; 50%; 60% (62%)	NS	NS
2A15 GS598739	241	**48.24**	2; 10^{-35}–10^{-10}; *Oryza japonica* genomic DNA; AK110296.1; 76%; 91%	1; 2×10^{-9}; *O. japonica* genomic DNA; AK110296.1; 66%; 77% (45%)	NS	NS
2A5 (2A13; 2B2; 2B7; 2B16) GS598729	225	56.30	>100; 4×10^{-105}–5×10^{-104}; *Alnus glutinosa* chloroplastic rDNA 16S; U46209.1; 98%; 100%	>100; 2×10^{-43}; *Betula pendula* var. *carelica* chloroplastic rDNA 16S; GQ284849.1; 98%; 98% (99%)	NT	NT
2A10 GS598734	190	**44.97**	>100; 6×10^{-89}–3×10^{-87}; *Ilex repanda* nuclear rDNA 26S; AY727932.1; 98%; 100%	>100; 1×10^{-35}–3×10^{-34}; *I. repanda* nuclear rDNA 26 s; AY727932.1; 96%; 96% (99%)	NT	NT
2B18 GS598737	221	**39.0**	100; 9×10^{-107}–2×10^{-97}; *Ilex pseudobuxus* rbcl gene; X98736.1; 99%; 100%	>100; 2×10^{-43}; *I. pseudobuxus* rbcl gene; X98736.1; 98%; 98% (99%)	NT	NT
2B4 GS598732	207	52.91	0; **1.8**; *Danio rerio*; CR388413.20; 96%; 12%	>100; 2×10^{-15}–2×10^{-8}; *V. vinifera* genomic DNA; AM467085.2; 54%; 68% (93%)	35; 2×10^{-12}–10^{-4}; *V. vinifera* hypothetical protein chromodomain; CAN68184.1; 54%; 68% (93%)	CHR_Del1; 3×10^{-6}; *Lilium henryi;* X13886; 48%; 75%

Table 2 continued

RDA clones[a]; Genbank accession number	Insert length (in bp)	Insert GC content[b] (in %)	BlastN results[c]	TBlastX results[d]	BlastX results[d]	GyDB Blastx results[e]
2B6 (2B11) GS598726	223	**46.85**	17; 2×10^{-6}–10^{-4}; *Daucus carota* genomic DNA; FJ148001.1; 72%; 60%	>100; 2×10^{-40}–2×10^{-37}; *D. carota* genomic DNA; FJ147957.1; 64%; 79% (71%)	100; 8×10^{-11}–10^{-9}; *V. vinifera* hypothetical protein; CAN72485.1; 52%; 83% (71%)	RNAseH_Tat4-1; 4×10^{-15}; *A. thaliana*; AB005247; 60%; 80%
2B9 (see Contig)	214	55	11; 6×10^{-7}–3×10^{-4}; *Ananas comosus* Ty3/gypsy LTR retrotransposon Dea1; Y12432; 69%; 68%	>100; 3×10^{-17}–5×10^{-15}; *V. vinifera* genomic DNA; AM475021.2; 64%; 88% (94%)	>100; 5×10^{-14}–3×10^{-9}; hypothetical protein *V. vinifera*; CAN69016.1; 64%; 88% (94%)	INT_Retrosat-2; 2×10^{-11}; *Oryza sativa*; AF111709; 40%; 60%
2B20 GS598736	235	50	7; 4×10^{-9}–10^{-4}; *Solanum tuberosum* resistance gene cluster; AF265664.1; 74%; 45%	>100; 10^{-21}–10^{-15}; *V. vinifera* genomic DNA; AM480525.2; 57%; 74% (100%)	100; 3×10^{-17}–7×10^{-7}; hypothetical protein *Vitis vinifera*; CAN68184.1; 56%; 73% (100%)	INT_Peabody; 8×10^{-8}; *Pisum sativum*; AF083074; 57%;74%
Contig[f] GU129908	1,320	**48.94**	>100; 6×10^{-100}–10^{-45}; *P. trichocarpa* genomic DNA; AC182684.2; 67%; 74%	>100; 6×10^{-177}–2×10^{-156}; *V. vinifera* genomic DNA; AM440321.1; 68%; 83% (61%)	>100; 3×10^{-157}–2×10^{-131}; *V. vinifera* hypothetical protein integrase domain; CAN69016.1; 64%; 78% (92%)	INT_Peabody; 10^{-115}; *P. sativum*; AF083074; 57%;77%

[a] Between brackets we indicate redundant clones

[b] In bold GC contents <50%

[c] Number of significant hits; *E*-value range for retrieved significant hits or for the first 100 significant hits (when correspond), in bold we indicate the non significant *E*-value retrieved from the best scored hit; accession name of the best scored hit; accession number; percentage of identity between query sequence and best hit; percentage of query coverage between query sequence and best hit

[d] Number of significant hits; *E*-value range for retrieved significant hits or for the first 100 significant (when correspond), in bold we indicate the non significant *E*-value retrieved from the best scored hit; accession name of the best scored hit; accession number; percentage of identity between query and retrieved sequence; percentage of positive shared residues query and retrieved sequences; calculated query coverage, between query sequence and best hit, is indicated between brackets (= (total number of matched residues * 3 * 100)/insert length)

[e] Results shown correspond to the best scored hit: name of the Ty3/Gypsy retrotransposon domain retrieved (INT = integrase, CHR = chromodomain, RNAseH = ribonuclease H); *E*-value; host name; accession number; percentage of identity between query and retrieved sequences; percentage of positive shared residues

[f] Results shown correspond to the contig assembled sequence (see text). NT, not tested. NS, no significant similarity was found

templates using designed primers (Table 1, see Supplementary materials). PCR amplification conditions were those indicated in the previous section; left and right primers were combined in the following general way: clone *i* left (L)-primer + clone *ii* right (R)-primer; clone *ii* L-primer + clone *i* R-primer; clone *i* L-primer + clone *ii* L-primer; and clone *i* R-primer + clone *ii* R-primer. Left/right PCR products were gel purified using QIAquick Gel Extraction kit (QIAGEN) and automatically sequenced as mentioned above. A contig assembly was obtained by using the BioEdit program. Blast searches were as previously stated.

Phylogenetic analysis

In order to establish the phylogenetic relationship of the contig-assembled sequence we performed a phylogenetic analysis including 11 Ty3/Gypsy retrotransposon protein sequences from a wide array of hosts (monocot, dicot, vertebrate and fungi) and several lineages according to Marin and Llorens (2000) and GyDB (http://gydb.uv.es/phylogeny.php?tree=int). Retroelement′s name, host and Genbank accession number, are as follows: CaRep (*Cicer arietinum,* CAC44142); CRM (*Zea mays,* AAC33526); Dea1 (*Ananas comosus,* CAA70342.1); Del (*Lilium henryi,* X13886); Diaspora (*Glycine max,* AAO23078); Galadriel (*Lycopersicon esculentum,* AAD13304.1); Peabody (*Pisum sativum,* AF083074); Retrosat-2 (*Oryza sativa* subsp. *indica,* AAD27547); Sushi-ichi (*Takifugu rubripes,* AAC33526); Ty3 (*Saccharomyces cerevisiae,* Q99315.3), and an unnamed mobile element from *Vitis vinifera* (CAN69016.1). Retroelements Athila (*Arabidopsis thaliana,* AAM15254), Cinful (*Zea mays,* AAD11615) and RetroSor-1 (*Sorghum bicolor,* AAD19359) were included as outgroups. These retroelements belong to the Athila/Tat lineage and are distantly related to the ingroup (Marín and Lloréns 2000).

Sequence alignment was generated using ClustalX 2.0.9 program (Larkin et al. 2007) with PAM matrices (Dayhoff 1979) and remaining parameters as default. Phylogenetic distance analyses and bootstrapping were performed using programs implemented in the PHYLIP 3.66 package (Felsenstein 2005). We applied a Neighbor-Joining method (Saitou and Nei 1987) employing the NEIGHBOR program, the PROTDIST program was used to generate the distance matrix selecting the PAM model of aminoacid substitution. To assess bootstrap support, 500 replicates were produced with SEQBOOT and were processed as described. Information was condensed through a majority rule consensus tree using the CONSENSE program. Trees were visualized with TreeView version 1.6.6 (Page 2001).

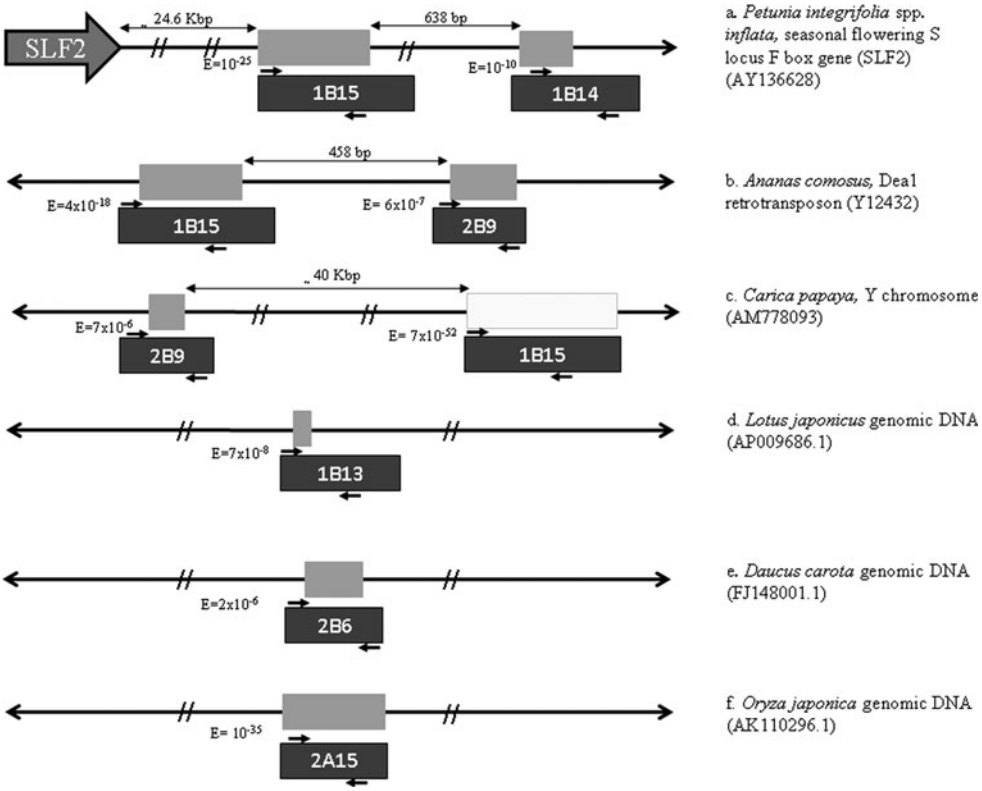


Fig. 1 Schematic representation showing the relationship between *I. paraguariensis* selected RDA fragments and sequences retrieved from Genbank. Full length RDA clones are shown as *dark grey boxes*; *grey boxes* represent the clone coverage over retrieved sequence (*straight lines*); the *light gray box* represents a protein level comparison. Numbers above *double headed arrows* are calculated distances, in base pairs (bp) or kilobase pairs (Kbp). *Small black arrows* represent primer pairs of Table 1 (see Supplementary materials)

Results

Seventy-five recombinant clones derived from the RDA assays performed on *I. paraguariensis* were characterized. Nucleotide variation of the inserts was evaluated by SSCP analysis and therefore the clones were grouped according to their banding pattern (not shown). For assay RDA-1, 14 SSCP groups were formed, and at least one representative clone from each SSCP group was sequenced and characterized (in total 21). Sequence characterization and comparison through bioinformatics tools allowed the distinction of nine sequence types (Table 1). Of these, one type (clone 1B9) shows either spurious affinities or no significant similarity with any deposited sequences in Genbank; four types (represented by clones 1A10, 1B2, 1B6 and 1B10) were identified as plant mitochondrial or ribosomal DNA. Remaining four types (clones 1A2, 1B13, 1B14 and 1B15) comprise sequences that, at the protein level, are related to hypothetical polyproteins and retrotransposons-like sequences. Within this latter group of sequences, clones 1B15 and 1B14 show association to plant sex-related regions. For instance, the best-scored nucleotide alignment for clone 1B15 indicates 70% similarity with a genomic region linked to the self-incompatibility locus of *Petunia integrifolia* ssp. *inflata* (Table 1, Fig. 1a). Subsequent significant hits comprise plant accessions lacking annotation, Ty3/Gypsy retrotransposons from dicots and monocots (Fig. 1b) and a 72% identity with an X chromosome region of *Carica papaya,* among others. At the protein level, clone 1B15 exhibits 83% positive matches with a fragment of the Y chromosome of *Carica papaya* (Fig. 1c). BlastX and RepeatMasker searches also indicate similarity to Ty3/Gypsy elements, which the GyDB search matched to the integrase domain of Retrosat-2 (Table 1). The clone 1B14 exhibits 76% identity with a *Populus trichocarpa* genomic clone located in poplar's gender determination locus (Tuskan GA, personal communication); it also shows 74% similarity to *P. integrifolia* ssp. *inflata* (Fig. 1a). At the protein level, clone 1B14 exhibits 74% of positive matches with a monocot Ty3/Gypsy retrotransposon. The GyDB search indicates affinity with the integrase domain of Peabody (Table 1); the RepeatMasker search failed to detect repetitive regions. Clones 1A2 and 1B13 have affinity to plant genomic clones and at the protein level to retrotransposon polyproteins. Clone 1A2 exhibits similarity with a region of the integrase domain of Retrosat-2, not overlapping with that shown by clone 1B15. As to clone 1B13, the Repeat Masker search resulted in that 33% of the query sequence has 69% similarity with the integrase domain of a Ty1/Copia retrotransposon from *Arabidopsis* (ATCOPIA23; accession AB019224).

Likewise, clones derived from assay RDA-2 were clustered in 24 groups according to their SSCP banding pattern. Then, at least one representative clone of each SSCP group was sequenced and characterized (in total 30) using bioinformatics tools. In this case, sequence characterization and comparison allowed us to distinguish 15 sequence types (Table 2). Of these, seven types (represented by clones 2A1, 2A6, 2A8, 2B5, 2B14, 2B17 and 2B21) show non-significant matches with deposited sequences in Genbank. In contrast, one type (clone 2A15) has significant hits with plant sequences lacking annotation, whereas three other types (clones 2A5, 2A10 and 2B18) were identified as plant ribosomal or plastid DNA. Remaining four types (clones 2B4, 2B6, 2B9 and 2B20) were associated to fragments of plant retrotransposons. In particular, clone 2B9 shows 69% nucleotide similarity with a monocot Ty3/Gypsy retrotransposon (Fig. 1b), and at a subsequent hit a 74% identity with a segment of the Y chromosome of *C. papaya* was also evidenced (Fig. 1c). At the protein level, the clone 2B9 appeared most similar to an integrase domain of the Retrosat-2 element. For clone 2B20, the best match rendered 74% nucleotide similarity with a *Solanum tuberosum* gene cluster associated to pathogens resistance, and at the protein level, clone 2B20 matched putative polyproteins and Ty3/Gypsy retroelements, in particular to the integrase domain of Peabody. In addition, BlastX results allowed the detection of an overlap, not otherwise revealed, between clones 2B20 and 2B4 (84% similarity, covering 63% of clone 2B4 length). The latter and clone 2B6 were recognized related to monocot retrotransposon conserved domains solely through BlastX and GyDB searches. Particularly, clone 2B4 has similarity to the chromodomain of Del1, whereas clone 2B6 shows affinity to the RNaseH of Tat4-1.

Considering all 24 sequence types, no clear association was detected between the sequence type and GC content (Tables 1 and 2).

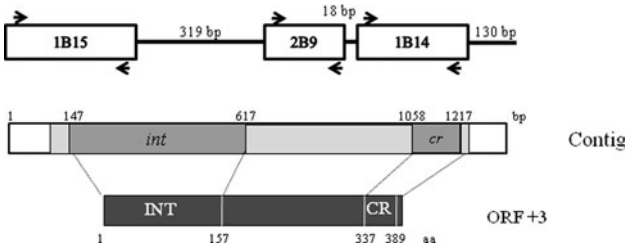


Fig. 2 Schematic representation of the *Ilex paraguariensis* retrotransposon sequence. *White boxes* represent RDA clones; distances are indicated in base pairs (bp). *Bold arrows* stands for designed left/right primers from Table 1 (see Supplementary material). The position of predicted integrase core domain (*INT*) and chromodomain (*CR*) have been delimited by comparison with pfam00665 database and chromodomain alignments of Marin and Llorens (2000) and Kentner et al. (2003)

PCR evaluation of RDA fragments

Based on results of the previous section, we evaluated through PCR amplification the occurrence of seven RDA selected fragments on additional male and female *I. paraguariensis* samples, using specially designed primers (Table 1, supplementary material). Fragment selection included five putative retroelements (represented by inserts of clones 1B13, 1B14, 1B15, 2B6 and 2B9) and two fragments akin to unidentified plant DNA (clones 2A8 and 2A15). On the one hand, PCR amplification of fragments 1B14, 1B15 and 2B9, gave a single distinct band of the expected size in all templates, irrespective of the gender of the plants. On the other hand, PCR amplification of fragments 1B13, 2A8, 2A15 and 2B6 rendered variable results; some templates gave single bands, whereas others produced multiple bands, irrespective of the gender of the samples (not shown). In all cases, sequence comparison shows no significant differences between the sexes.

RDA fragments linkage relationship

Linkage relationships, suggested in Fig. 1 for RDA fragments 1B14, 1B15 and 2B9, were tested through PCR amplifications. Twelve PCR mixes were assayed using combinations of the primers for clones 1B14, 1B15 and 2B9. Only those PCRs that employed left and right primers from different RDA clones (i.e., clone i L-primer + clone ii R-primer) gave discrete bands. Primer combinations left/left and right/right yielded multiple bands with different intensities (not shown) and were not considered for analysis. Nineteen partially overlapping fragments were sequenced and used for assembling a contig of 1320 base pairs (bp). Deduced relative linkage of clones 1B14, 1B15 and 2B9 is shown in Fig. 2. General results from the Blast searches performed on the assembled sequence indicate a 67% nucleotide identity with *P. trichocarpa* genomic clones (Table 2). Translation of the contig sequence identified an uninterrupted open reading frame (ORF) encoding 395 amino acids, which contains the integrase motif $DX_{60}DX_{35}E$ described by Thomson et al. (1998) and a chromodomain (Fig. 2). All other reading frames had stop codons and coded for 30–50 amino acids. Database searches showed that the ORF was most similar (78% positive matches) to an unnamed *Vitis vinifera* putative polyproteins and to the integrase domain of the element peabody (Table 2).

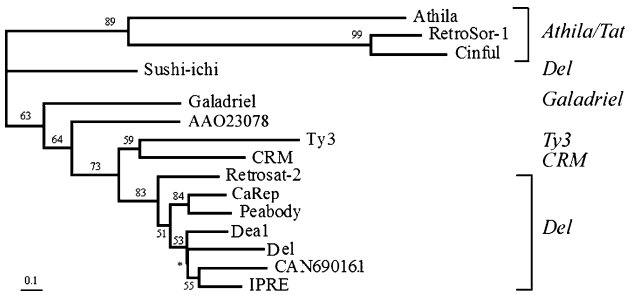


Fig. 3 Neighbor-joining tree of integrases from several Ty3/Gypsy retrotransposons lineages. Numbers stands for bootstrap values, branches receiving <50% support in 500 replicates are not shown; the *asterisk* indicates a branch that collapses in the majority rule consensus bootstrap tree. Integrase sequence names, hosts and accession numbers are detailed in "Materials and Methods" section. IPRE stands for *Ilex paraguariensis* retroelement. The retrotransposon lineages are indicated in *italics*. The *scale bar* depicts branch lengths based on PAM substitution model

Phylogenetic analysis

To determine the evolutionary placement of the retrotransposon-like sequence isolated from *I. paraguariensis* (IPRE, for *Ilex paraguariensis* retroelement), we aligned the region corresponding to the integrase domain of IPRE with other 14 Ty3/Gypsy integrase protein sequences, spanning 157 amino acid residues, and performed a phylogenetic analysis. The phylogenetic tree obtained is depicted in Fig. 3. It shows a strongly supported relationship of the IPRE integrase to elements from the Chromoviridae, particularly from plant Del lineage. Though poorly supported, IPRE shares a most recent common ancestor with the *V. vinifera* integrase. Integrase sequences of retroelements Athila, RetroSor-1 and Cinful, included as outgroups, belong to the Athila/Tat lineage that lack chromodomains. The sequence AAO23078 correspond to an integrase named Diaspora from *Glycine max*; it is described in Llorens et al. (2008) and in the Gypsy DataBase belonging to the Athila/Tat lineage. Since this sequence was found to include a chromodomain, it was excluded from the potential outgroup and was transferred to the ingroup, where it shows no clear relationship to any representative lineage.

Discussion

In this paper, we performed a screening of *Ilex paraguariensis* male and female genomes through the RDA technique and attempted sequence characterization using bioinformatic tools, with the aim of gathering sex-related sequence data. We found that several genomic regions were repeatedly recovered generating an intra-experiment redundancy, which was detected primarily via SSCP analysis and further by sequence characterization. The redundancy encountered has twofold implications. On the one hand, it indicates that the SSCP analysis overestimated the actual sequence variation of cloned RDA fragments, and on the other hand, it suggests that isolated regions involve repetitive DNA.

Taking together the results from both RDA assays, the sequences obtained could be discriminated into three main categories. The first category involves eight fragments that show spurious affinities with sequences deposited in public databases. For this category, it is observed that the region matched is less than 25% of the fragment length, and that the minimum *E*-value retrieved is 110 times higher than the a priori established threshold. Therefore, these eight sequences could be regarded as unidentified genomic regions specific to *I. paraguariensis*, following the criteria applied by Nagarajan et al. (2008) in papaya. The second category comprises sequences identified as organellar or ribosomal plant DNA. These latter products could be due to a combination of factors such as: (1) lack of sufficient driver competition during hybridization; (2) inefficient annealing in the complex mixture after the subtraction step; and/or (3) trapping of a fraction of the tester molecules in heteroduplexes (Kuvbachieva and Goffinet 2002). Certainly, additional RDA experiments using other restriction enzymes and more stringent tester/driver ratios may ensure a more comprehensive coverage of the "yerba mate" genome, but will not resolve the deficiency for reference sequences or of completely annotated sequences. Their utility as markers for gender discrimination remains to be accomplished.

Sequences grouped in the third category involve fragments related to a diverse array of annotated plant accessions, mainly associated to conserved domains (namely, RNaseH, integrases and chromodomains) from Ty3/Gypsy and Ty1/Copia retroelements. Interestingly, these fragments appeared highly allied to sex-related locus or to regions located in sex chromosomes of species that belong to plant families distantly related to the Aquifoliaceae. For instance, we found similarity with a region linked to the self-incompatibility locus of *Petunia integrifolia* ssp. *inflata* (Solanaceae). This is consistent with Wang et al. (2004) who reported that surrounding the centromeric S_2 locus of *Petunia,* there is a higher than expected density (76%) of transposons and other repetitive sequences, in comparison to other self-incompatible Solanaceae. We also found similarity with regions belonging to both sex chromosomes from dioecious *Carica papaya* (Caricaceae), which has a pair of primitive homomorphic sexual chromosomes (Liu et al. 2004). The similarity encountered could be due mainly by the high retroelement content indicated for papaya's sex chromosomes (Yu et al. 2008). Moreover, the sequencing of this fruit plant revealed that more than 52% of its genome comprises mainly long terminal repeat sequences (Ming et al. 2008) and that Ty3/Gypsy retrotransposons are the most abundant type of mobile elements in the papaya genome (27.8%; Nagarajan et al. 2008). Likewise, we found identity with several accessions from the dioecious tree *Populus trichocarpa* (Salicaceae). Poplar was the first woody plant to be completely sequenced (Tuskan et al. 2006) becoming a suitable reference organism for genetic studies. Yin et al. (2008) associated the peritelomeric region of chromosome XIX to a gender determination locus and hypothesized that this chromosome could represent an incipient sex chromosome. Our retrotransposon sequence shows affinity to several poplar's linkage groups, including the linkage group XIX (Tuskan GA, personal communication). The fact that 42% of the assembled genome of this tree consists of repetitive elements (Tuskan et al. 2006) could account for the extensive similarity evidenced.

RDA sequences so far tested have been retrieved from both sexes through PCR amplifications without significant sequence differentiation, and thus they do not discriminate the gender of the plants. However, association of RDA sequences to sex related regions cannot be completely discarded. Cytogenetic studies on *I. paraguariensis* failed to distinguish heteromorphic pairs (Barral et al. 1995; Greizerstein et al. 2004), but the occurrence of homomorphic sex chromosomes is still plausible. Presently, we can discard linkage of tested RDA fragments to the single dose chromosome (Y or W) although a linkage to the double dose chromosome (X or Z) remains to be carefully verified. Alternatively, if some RDA sequences are associated to the autosomes and are differentially distributed over the male and female genomes, then they would be useful as gender indicators. In this respect, it is interesting to note that 62% of the sequences recovered from the assay RDA-1, expected to enrich in female sequences, correspond to retrotransposons (above-mentioned third category). In contrast, the majority of the sequences (57%) recovered from the assay RDA-2, aimed to enrich in male sequences, correspond to the first category (i.e.: unidentified sequences). If this is not due to a bias in the representation or cloning steps, these results could suggest an actual difference between the sexes in their relative abundance.

Presently, we report for the first time the isolation from an *Ilex* genome of at least two distinct lineages of Ty3/Gypsy and one lineage of Ty1/Copia retrotransposon sequences. A similar coexistence of distinct lineages of plant retrotransposons has been found in other plants (Matsunaga et al. 2002; Vitte and Panaud 2005). According to the classification of Llorens et al. (2008), clones 1A2, 2B4 and 2B20 and the linked clones 1B14, 1B15 and 2B9, are representatives of the Del lineage, whereas clone 2B6 represents the Athila/Tat lineage. Ty3/Gypsy elements commonly exist in high copy numbers in plant genomes. Due to their replicative transposition mechanism, they can stimulate genome rearrangements like inversions, duplications or deletions, shaping the host genome (Bennetzen et al. 2005). Numerous studies have focused on plant transposable elements (Feschotte et al. 2002; Kentener et al. 2003; Matsunaga et al. 2002, among others) and over

150 LTR retrotransposon families have been fully characterized in more than 20 species of angiosperms (Vitte and Panaud 2005). The phylogenetic analysis performed here of the integrase domain, shows that our *I. paraguariensis* retroelement (named IPRE) integrase shares common ancestry with representatives of the Chromoviridae of the Del lineage. Retroelement integrases are zinc finger endonucleases that play a key role in the integration into the host genome (Llorns et al. 2008). Some Ty3/Gypsy elements described in genomes of plants, fungi, and vertebrates incorporate an additional domain, known as chromodomain, which is similar to certain chromatin-interacting proteins (Gorinsĕk et al. 2005; Malik and Eickbush 1999). This finding inspired the term Chromoviridae to cluster chromodomain-containing integrase elements (Marin and Llorens 2000). Consistent with this, we detected a chromodomain in our IPRE sequence. The phylogenetic analysis presented here also revealed that retroelements CaRep of *Cicer arietinum* and the unnamed element of *Vitis vinifera,* could be ascribed to the Del lineage of Chromoviridae, though a chromodomain was not originally reported for them. Similarly, the pineapple Dea1 element was reported within Chromoviridae by Marin and Llorens (2000), noting that the chromodomain is not found. According to the phylogenetic analyses of other retrotransposon coding regions, the Chromoviridae clade is ancient and ubiquitous in plants, since the clade includes elements from monocots, dicots and gymnosperms (Marin and Llorens 2000; Llorens et al. 2008). The fact that no stop codons have been encountered in the reading frame of the IPRE integrase suggests that it is still potentially able to transpose. However, remaining coding sequences should be analyzed to verify this hypothesis. Retrotransposons of the Ty1/Copia type are also ubiquitous elements, and are principally distinguished from Ty3/Gypsy elements by the order of occurrence of encoded proteins (Rogers and Pauls 2000). Herein we gathered data that suggests the presence of this type of mobile elements within *Ilex* genomes.

We are currently undertaken in situ hybridization experiments in order to determine the pattern of binding of retrotransponson probes over the entire chromosome complement of the "yerba mate" tree, and to determine if this pattern allows gender distinction. In addition, *I. paraguariensis* specific regions will be tested as potential probes for in situ hybridization experiments, but their suitability will depend on their copy number. These studies would help elucidate the mechanism of sex determination and the development of a gender discrimination system.

Acknowledgments The authors are indebted to L. D. Belingheri, S. D. Prat Kricun, and R. M. Mayol from EE INTA Cerro Azul (Misiones, Argentina) who kindly provided specimens from living collections; to G. Giberti who determined the gender of the plants and to E. R. Hasson and to G. A. Tuskan whose invaluable suggestions helped us to improve a previous version of this manuscript and to V. A. Confalonieri for critical reading of the current version of the manuscript. We are also grateful to an anonymous referee for helpful comments. Several grants from the Consejo Nacional de Investigaciones Científicas y Técnicas (PIP 5927), the University of Buenos Aires (EX–317) and the Agencia Nacional de Promoción Científica y Tecnológica (BID 1728 OC-AR PICT 14119 and 14170) are gratefully acknowledged.

References

Allen NL, Penn CW, Hilton AC (2003) Representational difference analysis: critical appraisal and method development for the identification of unique DNA sequences from prokaryotes. J Microbiol Methods 55:73–81

Barral G, Poggio L, Giberti GC (1995) Chromosome numbers and DNA content from *Ilex argentina* (Aquifoliaceae). Bol Soc Arg Bot 30:243–248

Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot 95:127–132

Chen ZJ, Phillips RL, Rines HW (1998) Maize DNA enrichment by representational difference analysis in a maize chromosome addition line of oat. Theor Appl Genet 97:337–344

Dayhoff MO (1979) Atlas of protein sequences and structure. Vol. 5, Supp. 3. National Bioinformatics Research Foundation, Washington, DC

Donnison IS, Siroky J, Vyskot B, Saedler H, Grant SR (1996) Isolation of Y chromosome-specific sequences from *Silene latifolia* and mapping of male sex-determining genes using representational difference analysis. Genetics 144:1893–1901

Felsenstein J (2005) PHYLIP (phylogeny inference package) version 3.6. Distributed by the author, Department of Genome Sciences. University of Washington, Seattle

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3:329–341

Giberti GC (2001) Diferentes aspectos del género *Ilex* (Aquifoliaceae). Corología, arquitectura floral, y posición sistemática. Doctoral Thesis, Universidad de Buenos Aires

Gorinsĕk B, Gubensĕk F, Kordiš D (2005) Phylogenomic analysis of chromoviruses. Cytogenet Genome Res 110:543–552

Gottlieb AM, Giberti GC, Poggio L (2005) Molecular analyses of the genus *Ilex* (Aquifoliaceae) in southern South America, evidence from AFLP and ITS sequence data. Am J Bot 92:352–369

Greizerstein EJ GC, Giberti PoggioL (2004) Cytogenetic studies of southern South American *Ilex* species. Caryologia 57:19–23

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nuc Ac Symp Series 41:95–98

Jamilena M, Mariotti B, Manzano S (2008) Plant sex chromosomes: molecular structure and function. Cytogenet Genome Res 120: 255–264

Kentner EK, Arnold ML, Wessler SR (2003) Characterization of high-copy-number retrotransposons from the large genomes of the Louisiana *Iris* species and their use as molecular markers. Genetics 164:685–697

Kuvbachieva AA, Goffinet AM (2002) A modification of representational difference analysis, with application to the cloning of a candidate in the Reelin signaling pathway. BMC Mol Biol 3:6. This article is available from: http://www.biomedcentral.com/1471-2199/3/6

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948

Lia VV, Bracco M, Gottlieb AM, Poggio L, Confalonieri VA (2007) Complex mutational patterns and size homoplasy at maize microsatellite loci. Theor Appl Genet 115:981–991

Lisitsyn N, Lisitsyn NM, Wigler M (1993) Cloning the differences between two complex genomes. Science 259:946–951

Liu Z, Moore PH, Ma H, Ackerman CM, Ragiba M, Yu Q, Pearl HM, Kim MS, Charlton JW, Stiles JI, Zee FT, Paterson AH, Ming R (2004) A primitive Y chromosome in papaya marks incipient sex chromosome evolution. Nature 427:348–352

Llorens C, Futami R, Bezemer D, Moya A (2008) The gypsy database (GyDB) of mobile genetic elements. Nucleic Acids Res 36:38–46

Malik AH, Eickbush TH (1999) Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. J Virol 73:5186–5190

Marin I, Llorens C (2000) Ty3/Gypsy retrotransposons: description of new arabidopsis thaliana elements and evolutionary perspectives derived from comparative genomic data. Mol Biol Evol 17: 1040–1049

Matsunaga S, Yagisawa F, Yamamoto M, Uchida W, Nakao S, Kawano S (2002) LTR retrotransposons in the dioecious plant *Silene latifolia*. Genome 45:745–751

Ming R, Wang J, Moore PH, Paterson AH (2007) Sex chromosomes in flowering plants. Am J Bot 94:141–150

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452:991–996

Nagarajan N, Navajas-Pérez R, Pop M, Alam M, Ming R, Paterson AH, Salzberg SL (2008) Genome-wide analysis of repetitive elements in papaya. Trop Plant Biol 3–4:191–201

Nekrutenko A, Baker RJ (2003) Subgenome-specific markers in allopolyploid cotton *Gossypium hirsutum*: implications for evolutionary analysis of polyploids. Gene 306:99–103

Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc Natl Acad Sci USA 86:2766–2770

Page RDM (2001) TreeView 1.6.6 (http://taxonomy.zoology.gla.ac.uk/rod/rod.html)

Panaud O, Vitte C, Hivert J, Muzlak S, Talag J, Brar D, Sarr A (2002) Characterization of transposable elements in the genome of rice (*Oryza sativa* L.) using representational difference analysis (RDA). Mol Genet Genom 268:113–123

Rogers SA, Pauls KP (2000) Ty1-copia-like retrotransposons of tomato (*Lycopersicon esculentum* Mill.). Genome 43:887–894

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols: methods in molecular biology. Humana Press, Totowa, pp 365–386

Sabot F, Sourdille P, Bernard M (2004) Detecting specific repeated sequences in large complex genomes by using representative difference analysis and double-probe verification. Plant Mol Biol Rep 22:91a–91j

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Thomson KG, Thomas JE, Dietzgen RG (1998) Retrotransposon-like sequences integrated into the genome of pineapple, *Ananas comosus*. Plant Mol Biol 38:461–465

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

Vitte C, Panaud O (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res 110:91–107

Vyskot B, Hobza R (2004) Gender in plants: sex chromosomes are emerging from the fog. Trends Genet 20:432–438

Wang Y, Tsukamoto T, Yi KW, Wang X, Huang S, McCubbin AG, Kao TH (2004) Chromosome walking in the *Petunia inflata* self-incompatibility (S-) locus and gene identification in an 881-kb contig containing S2-RNase. Plant Mol Biol 54:727–742

Winge H (1997) Conservação genética do erva-mate no Brasil. In: Sturion JA et al. (eds), Anais I Congreso Sul- Americano do Erva Mate. II Reunião Técnica do Cono Sul sobre a cultura do Erva Mate. Curitiba, PR. EMBRAPA, Documentos 33, pp 209–226

Yin T, DiFazio SP, Gunter LE, Zhang X, Sewell MM, Woolbright SA, Allan GJ, Kelleher CT, Douglas CJ, Wang M, Tuskan GA (2008) Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. Gen Res 18:422–430

Yu Q, Hou S, Feltus FA, Jones MR, Murray JE, Veatch O, Lemke C, Saw JH, Moore RC, Thimmapuram J, Liu L, Moore PH, Alam M, Jiang J, Paterson AH, Ming R (2008) Low X/Y divergence in four pairs of papaya sex-linked genes. Plant J 53:124–132

Zoldos V, Siljak-Yakovlev S, Papes D, Sarr A, Panaud O (2001) Representational difference analysis reveals genomic differences between *Q. robur* and *Q. suber*: implication for the study of genome evolution in the genus *Quercus*. Mol Genet Genom 265: 241–243