# IDENTIFICATION, CHARACTERIZATION, AND DYNAMICS OF TRANSPOSABLE ELEMENTS IN *IRIS* AND *BRASSIC*A

by

EDWARD K KENTNER

(Under the Direction of Susan R Wessler)

## ABSTRACT

Transposable elements (TEs) are ubiquitous components of plant genomes, yet their dynamics and the genetic variability due to their activities within populations has not been adequately investigated. This dissertation contains two studies designed to investigate the impact of transposable elements on plant genome evolution. In the first study, TY3/*gypsy*-like retrotransposons were characterized from *Iris fulva* and *I. brevicaulis* and used to develop molecular markers which can be used to study the insertion site polymorphism of the elements. The copy number of these IRRE elements (for IRis RetroElement), is ~$1 \times 10^5$, accounting for ~6-10% of the ~10,000 Mb haploid Louisiana Iris genome. IRRE elements are transcriptionally active in *Iris brevicaulis* and *I. fulva* and their $F_1$ and backcross hybrids, but evidence for increased transcription in hybrid plants was not found. The Iris genome contains many subfamilies of these elements which can be used to generate molecular markers in diverse *Iris* species.

The second study investigates the insertion polymorphism of the CACTA-like DNA transposon *Boc*-1 in *Brassica nigra*. The insertion polymorphism of *Boc*-1 elements was assayed by transposon display in four populations from southern California, and compared to the

polymorphism of presumably neutral control markers. The neutrality of the TE insertions was tested using an approach based on the analysis of fixation indices, and using existing TE population genetic theory that has been applied to elements from *Drosophila*. Based on comparisons among fixation indices, most *Boc*-1 insertions appear to be neutral, but there was evidence of forces preventing individual insertions from reaching high population frequencies. Some explanations for these apparently conflicting results based on the biology of the elements are discussed.

INDEX WORDS: Transposable element, Retrotransposon, CACTA-like DNA transposon, *Iris*, *Brassica*, IRRE, *Boc*-1, Selection, Polymorphism.

# IDENTIFICATION, CHARACTERIZATION, AND DYNAMICS OF TRANSPOSABLE ELEMENTS IN *IRIS* AND *BRASSIC*A

By

EDWARD K KENTNER

B.S., Humboldt State University, 1995

M.A., Humboldt State University, 1997

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2004

# IDENTIFICATION, CHARACTERIZATION, AND DYNAMICS OF TRANSPOSABLE ELEMENTS IN *IRIS* AND *BRASSIC*A

By

EDWARD K KENTNER

Major Professor: Susan R Wessler

Committee: Wyatt W. Anderson
R. Kelly Dawe
James L. Hamrick
Daniel E.L. Promislow

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2004

ACKNOWLEDGEMENTS

In addition to the people mentioned in the acknowledgements sections of the individual chapters I would like thank Julie Nelson, whose name was inadvertently omitted from the published version of Chapter 2, for her help with the flow cytometry. Of course, I also thank my family and Jenny Moore for their support while I was pursuing this degree.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

# CHAPTER 1

# INTRODUCTION AND LITERATURE REVIEW

Transposable elements (TEs) are mobile DNA sequences that have been found in almost every organism that has been examined for their presence. Across eukaryotic organisms there is a broad range in the amount of these repetitive sequences present in genomes, but it is not uncommon for them to account for the majority of chromosomal DNA (Kidwell 2002). From the standpoint of the gross composition of genomes, mobile DNA is clearly important, but its evolutionary significance remains the subject of debate (*e.g.* Bowen and Jordan 2002; Doolittle and Sapienza 1980; Kidwell and Lisch 2000; McDonald 1998; Orgel and Crick 1980). Unlike typical cellular genes, TEs experience selection on two levels. Because TEs proliferate within the host genome independently of host reproduction, these sequences undergo selection at the DNA level for replication ability. However, if a TE causes an alteration of gene function that affects host fitness it will experience selection at the level of the host organism. The evolutionary dynamics of TEs are therefore likely to be the result of evolution occurring at both of these levels, and much of the debate regarding the importance of TEs is related to varying views on the relative importance of each.

Although TEs were first discovered in plants (McClintock 1948) and there have been numerous studies of plant elements, few studies have looked at the effects of TEs on plant fitness and the consequential forces exerted on the TEs by this interaction. The broad goal of the research described in this dissertation was to contribute to the understanding of this aspect of TE

biology. The project described in Chapter 2 was initiated to investigate the possibility that TEs could be activated by interspecific hybridization, while Chapter 3 seeks to infer the fitness effects of TEs by looking at the frequencies of individual insertions in natural populations.

The two major classes of eukaryotic mobile elements are defined by differences in their mechanism of transposition (Feschotte *et al.* 2002). Class I elements do not excise during transposition but produce a new DNA copy from an RNA transcript using the element-encoded enzyme, reverse transcriptase. Class I elements therefore increase in copy number during the process of transposition and can become very abundant. Class II, or DNA elements, transpose using a "cut-and-paste" mechanism in which the element excises from one chromosomal location and moves to another. Some Class II elements, particularly miniature inverted-repeat transposable elements (MITEs), can also attain very high copy numbers through mechanisms that are not completely resolved (Feschotte *et al.* 2002). Within both of these classes are several distinct types of TEs. However, the research described in this dissertation is limited to one type of Class I element, LTR retrotransposons, and one type of Class II element, CACTA-like DNA transposons.

**LTR Retrotransposons:** LTR retrotransposons are Class I elements named for the long terminal repeat that typically flank the protein encoding sequence of the transposon. The coding sequence contains several genes that are synthesized as a polyprotein then cleaved into functional proteins that perform the functions necessary for transposition (reviewed in Kumar and Bennetzen 1999). There are two major subtypes of LTR retrotransposons distinguished by the order of genes within the polyprotein. Formally, these are called the Metaviridae and Pseudoviridae (Hull 1999; Pringle 1999), but are most frequently referred to as Ty3/*gypsy*-like and Ty1/*copia*-like LTR retrotransposons, respectively, after the names of yeast and *Drosophila*

elements typical of each group. Both types are prevalent in plant genomes (Kumar and Bennetzen 1999), are present in all species that have been examined (Brandes *et al.* 1997; Flavell *et al.* 1992), and often compose the majority of chromosomal DNA in plants with large genomes (Bennetzen 2002).

Plant LTR retrotransposons have several features that make them interesting subjects for evolutionary studies. Class I elements do not excise once inserted into a genomic location. This property, and the low probability of independent insertions into the same target site makes markers based on these elements phylogeneticaly informative because each insertion has a known ancestral state (*i.e*. not present; Cook and Tristem 1997; Tatout *et al.* 1999). Under normal growth conditions, plant LTR retrotransposons are typically silenced, but become transcriptionally active when the host experiences stress (Grandbastien 1998; Wessler 1996). A variety of stresses such as tissue culture (Hirochika 1993), wounding, pathogen attack, and chemical treatments (Beguiristain *et al.* 2001; Mhiri *et al.* 1999; Mhiri *et al.* 1997) have been shown to activate plant LTR retrotransposons. Because plants do not set aside a germ line early in development, it is possible for plants experiencing stress conditions to incur heritable mutations due to the movement of LTR retrotransposons. Lastly, LTR sequences contain promoter and enhancer elements, and can influence the transcription of adjacent genes (Kashkush *et al.* 2002). Therefore, the movement of these elements may not only cause disruption of gene function by inserting directly into a gene, but also may impact the expression of adjacent genes, raising the possibility that insertions may occasionally cause beneficial mutations (Kidwell and Lisch 2001).

Barbara McClintock (1984) proposed that "wide crosses" could exert genomic stress, called "genome shock", leading to the activation of TEs. The first support for this hypothesis

was the dramatic increase in the number of retrotransposons observed in the sterile offspring of an intergeneric cross between two wallaby species (Waugh O'Neil *et al.* 1998). The increase in element transposition was correlated with a genome wide loss of DNA methylation in the hybrid. The finding sparked investigations seeking evidence of the same phenomenon in several Eutherian mammal species, but no evidence of TE activation was found (Robinson *et al.* 2000; Roemer *et al.* 1999). In *Drosophila*, transposition of the retrotransposon *Osvaldo* has been shown to increase in interspecific hybrids (Labrador *et al.* 1999). Although vascular plant species are well known for their ability to produce hybrids (Stebbins 1959), few investigations of TE activation in plant hybrids have been published (Kentner *et al.* 2003; Liu and Wendel 2000), and conclusive evidence for the phenomenon has not been found.

**CACTA-like DNA transposons:** The CACTA superfamily of DNA transposons (also called the *En/Spm* superfamily) is named for the sequence found at the ends of the terminal inverted repeats (TIRs) flanking the internal region of these TEs. To date, members of this superfamily have only been found in plants, where several families have been identified after elements have inserted into genes, causing visible phenotypes (reviewed in Kunze and Weil 2002). Some details of the transposition mechanism of CACTA elements have been worked out for the maize elements *En/Spm* and are quite complex. Autonomous elements code for two proteins that are alternatively spliced from the same RNA precursor. One of these, TNPA, acts as both an activator and repressor of element activity, depending on its concentration. Nonautonomous elements either lack or have mutated coding regions, but can move in response to proteins produced by an autonomous element elsewhere in the genome as long as critical *cis*-acting sequences in the TIRs remain intact (Kunze and Weil 2002). An interesting feature of these TEs is that some nonautonomous elements contain signals allowing them to be spliced out of RNAs

into which they have been transcribed (*e.g.* Menssen *et al.* 1990; Raboy *et al.* 1989), thereby minimizing their deleterious effects on the host. However, if an active autonomous element is present in the genome, transcripts of host genes containing the non-autonomous elements are completely suppressed. There have been no published population-genetic studies of CACTA elements to date.

**TE population genetics:** Although there are many examples of TE sequences becoming co-opted to perform functions for the host organism (reviewed in Kidwell and Lisch 2001), it is clear that the majority of TEs persist in host genomes as genomic parasites (Doolittle and Sapienza 1980; Orgel and Crick 1980) that spread among individuals via the sexual reproduction of their hosts. An early population genetic model predicted that actively transposing TEs will spread rapidly through a diploid randomly mating population even if they reduce host fitness by as much as 50% (Hickey 1982). Among rotifers, only the asexual Class Bdelloidea are free of retrotransposons, supporting the prediction that sexual reproduction is required for the persistence of TEs (Arkhipova and Meselson 2000). As there has never been a documented case of the horizontal transmission of TEs in plants (which unlike animals have a cell wall), sexual reproduction may be the only mechanism by which TEs spread in plant populations (Wright and Finnegan 2001). Mating systems are highly variable in plants, and are predicted to influence the polymorphism and rate of removal of TEs in populations (Charlesworth and Charlesworth 1995; Wright and Schoen 1999). The details of TE behavior at the cellular level can therefore only partially explain the biology of these genomic parasites and much stands to be learned from investigations into their population genetics.

Almost all of the empirical research into the population genetics of TEs published to date has involved species of the genus *Drosophila* (reviewed in Charlesworth and Langley 1989). In

*D. melanogaster* most families of TEs have relatively low copy numbers (in the hundreds or less), yet half of all spontaneous mutations are caused by the movement of TEs (Finnegan 1992). As most of these mutations are deleterious, the expectation is that individual insertions should not rise to high population frequencies as long as populations are of sufficient size. Theory has been developed to infer the strength of selection against TE insertions using the frequency distribution of individual insertions of a family of elements (*e.g.* Charlesworth and Charlesworth 1983; Langley *et al.* 1983). For all of the TE families that have been studied in *D. melanogaster*, individual insertions rarely reach high population frequencies (Bartolome *et al.* 2002; Charlesworth and Langley 1989), but this has not held true for TEs in other *Drosophila* species (*e.g.* Hey 1989; Lepetit *et al.* 2002).

Similar detailed population genetic data is generally not available for most types of organisms, but in mammals, including humans, many TE insertions appear to be ancient, are fixed within species (Batzer *et al.* 1996; Boissinot *et al.* 2000; Eickbush and Furano 2002), and are therefore not likely to be strongly deleterious. In plants observations are available for only one family of Class II elements (Wright *et al.* 2001). Clearly, more data is necessary before generalizations regarding the average fitness effects of insertions can be made.

**Dissertation outline:** The research goal uniting the chapters of this dissertation was to gain a better understanding of the dynamics of TEs in natural populations of plants. The research described in Chapter 2 was initiated to investigate the possibility that TEs could become activated by interspecific hybridization in the Louisiana Irises. The majority of the work in this chapter is devoted to the characterization of a group of TY/*gypsy* LTR retrotransposons (named IRRE for IRis REtrotransposon) that occur in high copy numbers in the genomes of both *I. brevicaulis* and *I. fulva*. Two families of related Ty3/*gypsy*-like LTR retrotransposons were

characterized using PCR and genomic library screens. These elements account for 6-10% of the ~9,650 mb Iris genome and are transcriptionally active in *Iris brevicaulis*, *I. fulva*, and in their $F_1$ and backcross hybrids. However, no evidence for increased transcription in hybrid plants was detected. Transposon display or S-SAP (Van den Broeck *et al.* 1998; Waugh *et al.* 1997) markers specific for the insertions of two subfamilies of IRRE elements were developed. These markers are based on a modified amplified fragment length polymorphism (AFLP Vos *et al.* 1995) protocol that uses a ligated adaptor primer and a primer specific for the terminal sequence of an element family to amplify fragments consisting of the element ends and adjacent genomic DNA. Transposon display markers anchored in two subfamilies of IRRE elements were shown to be highly polymorphic in wild-collected individuals of *I. brevicaulis* and *I. fulva*.

In the process of characterizing the IRRE elements, it was discovered that related elements are present in each of 11 *Iris* species tested, but absent from related genera in the Iridaceae. The research described in Appendix A examines the diversity of IRRE retrotransposons present in the California Iris species complex. The project was initiated to investigate the level of TE polymorphism in natural populations of 10 species of California *Iris* in order to evaluate the utility of IRRE insertions as molecular markers for population and phylogenetic studies. To this end a large number of IRRE element fragments were cloned and sequenced from *I. douglaiana*, *I. bracteata* and *I. inominata*. Many distinct subfamilies were found, including a distinct group of elements apparently not present in the Louisiana Iris which have been designated IRRE3. However, numerous attempts to develop transposon display markers using the sequence information were not successful for unknown reasons.

The difficulties associated with working with the large and uncharacterized Iris genome lead to the decision to change systems in order to address questions concerning the population

genetics of plant TEs. The availability of the TIGR *Brassica oleracea* genomic survey sequence led to the discovery and characterization of several families of recently active TEs in this species (Zhang, 2003). The objective of the research described in Chapter 3 was to investigate the polymorphism of TEs in natural *Brassica* populations in order to make inferences regarding the selective consequences of transposition. Transposon display primers designed to amplify three recently-active families of elements from *B. oleracea* were used to score polymorphisms in *B. nigra* populations from southern California. Although all of the primers produced reproducible polymorphisms in *B. nigra*, only the primers for the CACTA transposon *Boc*-1 produced markers that were anchored in TEs. The majority of polymorphisms generated by the other primers are likely to be neutral and were compared to the insertion polymorphism of *Boc*-1. There was evidence of weak forces removing the *Boc*-1 transposons from the populations, but it was not possible to identify the nature of the selective force(s), or to rule out element excision using insertion frequency data alone. However, based on the comparison of fixation indexes across loci, the majority of *Boc*-1 insertions are likely to be selectively neutral, suggesting that most insertions evolve through a process of random drift, and that excision may be more important than selection in preventing individual insertions from drifting to high population frequencies.

## LITERATURE CITED

Arkhipova, I., and M. Meselson, 2000 Transposable elements in sexual and ancient asexual taxa. Proceedings of the National Academy of Sciences of the United States of America **97:** 14473-14477.

Bartolome, C., X. Maside and B. Charlesworth, 2002 On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster. Molecular Biology and Evolution **19:** 926-937.

Batzer, M. A., S. S. Arcot, J. W. Phinney, M. AlegriaHartman, D. H. Kass *et al.*, 1996 Genetic variation of recent Alu insertions in human populations. Journal of Molecular Evolution **42:** 22-29.

Beguiristain, T., M. A. Grandbastien, P. Puigdomenech and J. M. Casacuberta, 2001 ThreeTnt1 subfamilies show different stress-associated patterns of expression in tobacco. Consequences for retrotransposon control and evolution in plants. Plant Physiology **127:** 212-221.

Bennetzen, J. L., 2002 Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica **115:** 29-36.

Boissinot, S., P. Chevret and A. V. Furano, 2000 L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Molecular Biology and Evolution **17:** 915-928.

Bowen, N. J., and I. K. Jordan, 2002 Transposable elements and the evolution of eukaryotic complexity. Current Issues in Molecular Biology **4:** 65-76.

Brandes, A., J. S. Heslop-Harrison, A. Kamm, S. Kubis, R. L. Doudrick *et al.*, 1997 Comparative analysis of the chromosomal and genomic organization of *Ty1-copia*-like retrotransposons in pteridophytes, gymnosperms and angiosperms. Plant Molecular Biology **33:** 11-21.

Charlesworth, B., and D. Charlesworth, 1983 The Population-Dynamics of Transposable Elements. Genetical Research **42:** 1-27.

Charlesworth, B., and C. H. Langley, 1989 The population genetics of *Drosophila* transposable elements. Annual Review of Genetics **23:** 251-287.

Charlesworth, D., and B. Charlesworth, 1995 Transposable elements in inbreeding and outbreeding populations. Genetics **140:** 415-417.

Cook, J. M., and M. Tristem, 1997 'SINEs of the times' - transposable elements as clade markers for their hosts. Trends in Ecology and Evolution **12:** 295-297.

Doolittle, W. F., and C. Sapienza, 1980 Selfish genes, the phenotype paradigm and genome evolution. Nature **284:** 601-603.

Eickbush, T. H., and A. V. Furano, 2002 Fruit flies and humans respond differently to retrotransposons. Current Opinion in Genetics & Development **12:** 669-674.

Feschotte, C., J. Jiang and S. R. Wessler, 2002 Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics **3:** 329-341.

Finnegan, D. J., 1992 Transposable elements. Current Opinion in Genetics & Development **2:** 861-867.

Flavell, A., E. Dunbar, R. Anderson, S. Pearce, R. Hartley *et al.*, 1992 *Ty1-copia* group retrotransposons are ubiquitous and heterogeneous in higher plants. Nucleic Acids Research **20:** 3639-3644.

Grandbastien, M.-A., 1998 Activation of plant retrotransposons under stress conditions. Trends in Plant Science **3:** 181-187.

Hey, J., 1989 The transposable portion of the genome of *Drosophila algonquin* is very different from that in *D. melanogaster*. Molecular Biology and Evolution **6:** 66-79.

Hickey, D. A., 1982 Selfish DNA: A sexually-transmitted nuclear parasite. Genetics **101:** 519-531.

Hirochika, H., 1993 Activation of tobacco retrotransposons during tissue culture. The EMBO Journal **12:** 2521-2528.

Hull, R., 1999 Classification of reverse transcribing elements: a discussion document. Archives of Virology **144:** 209-214.

Kashkush, K., M. Feldman and A. A. Levy, 2002 Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics **160:** 1651-1659.

Kentner, E. K., M. L. Arnold and S. R. Wessler, 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.

Kidwell, M. G., 2002 Transposable elements and the evolution of genome size in eukaryotes. Genetica **115:** 49-63.

Kidwell, M. G., and D. R. Lisch, 2000 Transposable elements and host genome evolution. Trends in Ecology & Evolution **15:** 95-99.

Kidwell, M. G., and D. R. Lisch, 2001 Perspective: Transposable elements, parasitic DNA, and genome evolution. Evolution **55:** 1-24.

Kumar, A., and J. L. Bennetzen, 1999 Plant retrotransposons. Annual Review of Genetics **33:** 479-532.

Kunze, R., and C. Weil, 2002 The *h*AT and CACTA superfamilies of plant transposons, pp. 565-610 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. American Society for Microbiology Press, Washington, D.C.

Labrador, M., M. Farre, F. Utzet and A. Fontdevila, 1999 Interspecific hybridization increases transposition rates of Osvaldo. Molecular Biology and Evolution **16:** 931-937.

Langley, C. H., J. F. Y. Brookfield and N. Kaplan, 1983 Transposable Elements in Mendelian Populations .1. A Theory. Genetics **104:** 457-471.

Lepetit, D., A. Brehm, P. Fouillet and C. Biemont, 2002 Insertion polymorphism of retrotransposable elements in populations of the insular, endemic species Drosophila madeirensis. Molecular Ecology **11:** 347-354.

Liu, B., and J. F. Wendel, 2000 Retrotransposon activation followed by rapid repression in introgressed rice plants. Genome **43:** 874-880.

McClintock, B., 1948 Mutable loci in maize. Carnegie Institution of Washington Year Book **47:** 155-169.

McClintock, B., 1984 The significance of responses of the genome to challenge. Science **226:** 792-801.

McDonald, J. F., 1998 Transposable elements, gene silencing and macroevolution. Tree **13:** 95-95.

Menssen, A., S. Hohmann, W. Martin, P. S. Schnable, P. A. Peterson *et al.*, 1990 The En/Spm Transposable Element of Zea-Mays Contains Splice Sites at the Termini Generating a Novel Intron from a Dspm Element in the A2 Gene. Embo Journal **9:** 3051-3057.

Mhiri, C., P. J. G. M. De Wit and M. A. Grandbastein, 1999 Activation of the promoter of the Tnt1 retrotransposon in tomato after inoculation with the fungal pathogen *Cldosporium fulvum*. Molecular Plant-Microbe Interactions **12:** 592-603.

Mhiri, C., J. B. Morel, S. Vernhettes, J. M. Casacuberta, H. Lucas *et al.*, 1997 The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. Plant Molecular Biology **33:** 257-266.

Orgel, L. E., and F. H. C. Crick, 1980 Selfish DNA: the ultimate parasite. Nature **284:** 604-607.

Pringle, C. R., 1999 Virus taxonomy - 1999 - The Universal System of Virus Taxonomy, updated to include the new proposals ratified by the International Committee on Taxonomy of Viruses during 1998. Archives of Virology **144:** 421-429.

Raboy, V., H. Y. Kim, J. W. Schiefelbein and O. E. Nelson, 1989 Deletions in a Dspm Insert in a Maize Bronze-1 Allele Alter Rna Processing and Gene-Expression. Genetics **122:** 695-703.

Robinson, T. J., O. Wittekindt, J. J. Pasantes, W. S. Modi, W. Schempp *et al.*, 2000 Stable methylation patterns in interspecific antelope hybrids and the characterization and localization of a satellite fraction in the Alcelaphini and Hippotragini. Chromosome Research **8:** 635-643.

Roemer, I., F. Grutzner, H. Winkig, T. Haff, A. Orth *et al.*, 1999 Global methylation in eutherian hybrids. Nature **40:** 131-132.

Stebbins, L. G., 1959 The role of hybridization in evolution. Proceedings of the American Philosophical Society **403:** 231-251.

Tatout, C., S. Warwick, A. Lenoir and J. M. Deragon, 1999 SINE insertions as clade markers for wild crucifer species. Molecular Biology and Evolution **16:** 1614-1621.

Van den Broeck, D., T. Maes, M. Saur, J. Zetof, P. De Keukeleire *et al.*, 1998 *Transposon Display* identifies individual transposable elements in high copy number lines. The Plant Journal **13:** 121-129.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee *et al.*, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research **23:** 4407-4414.

Waugh O'Neil, R. J., M. J. O'Neil and J. A. Marshall Graves, 1998 Undermethylation associated with retroelement activation and chromosome remodeling in a interspecific mammalian hybrid. Nature **393:** 68-72.

Waugh, R., K. McLean, A. J. Flavell, S. R. Pearce, A. Kumar *et al.*, 1997 Genetic distribution of BARE-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Molecular and General Genetics **253:** 687-694.

Wessler, S. R., 1996 Plant retrotransposons: Turned on by stress. Current Biology **6:** 959-961.

Wright, S., and D. Finnegan, 2001 Genome evolution: Sex and the transposable element. Current Biology **11:** R296-R299.

Wright, S. I., Q. H. Le, D. J. Schoen and T. E. Bureau, 2001 Population dynamics of an *Ac*-like transposable element in self- and cross-pollinating *Arabidopsis*. Genetics **158:** 1279-1288.

Wright, S. I., and D. J. Schoen, 1999 Transposon dynamics and the breeding system. Genetica **107:** 139-148.

# CHAPTER 2

# CHARACTERIZATION OF HIGH COPY NUMBER RETROTRANSPOSONS FROM THE LARGE GENOMES OF THE LOUISIANA IRIS SPECIES AND THEIR USE AS MOLECULAR MARKERS[1]

[1]Kentner, E. K., Arnold, M. L. and S. R. Wessler. 2003. Genetics. 164: 685–697.
Reprinted here with permission of publisher.

ABSTRACT

The Louisiana Iris species *Iris brevicaulis* and *I. fulva* are morphologically and karyotypically distinct yet frequently hybridize in nature. A group of high copy number TY3/*gypsy*-like retrotransposons was characterized from these species and used to develop molecular markers which take advantage of the abundance and distribution of these elements in the large Iris genome. The copy number of these IRRE elements (for IRis RetroElement), is ~1 $\times 10^5$, accounting for ~6-10% of the ~10,000 Mb haploid Louisiana Iris genome. IRRE elements are transcriptionally active in *Iris brevicaulis* and *I. fulva* and their $F_1$ and backcross hybrids. The LTRs of the elements are more variable than the coding domains and can be used to define several distinct IRRE subfamilies. Transposon display or S-SAP markers specific to two of these subfamilies have been developed and are highly polymorphic among wild-collected individuals of each species. As IRRE elements are present in each of 11 Iris species tested, the marker system has the potential to provide valuable comparative data on the dynamics of retrotransposition in large plant genomes.

# INTRODUCTION

The majority of chromosomal DNA in plants with large genomes is repetitive and is likely composed of various classes of mobile elements (Flavell *et al.* 1974; Joseph *et al.* 1990; Kidwell 2002). Although many classes of elements contribute significantly to overall genome size (*e.g.* Leeton and Smyth 1993), recent results from the grasses suggest that LTR retrotransposons comprise the largest fraction of genomic DNA (Bennetzen 2002; San Miguel *et al.* 1996; SanMiguel and Bennetzen 1998; SanMiguel *et al.* 1998). In grasses with relatively large genomes such as maize and barley, over 60% of the genome is composed of LTR retrotransposons (Meyers *et al.* 2001; SanMiguel and Bennetzen 1998; Vicient *et al.* 1999), while in the smaller rice genome the proportion is 30-35% (N. Jiang. unpublished data).

LTR retrotransposons are class I mobile elements related to infectious retroviruses (Malik *et al.* 2000). There are two major types of LTR retrotransposons, Ty1/*copia*-like (Pseudoviridae) and Ty3/*gypsy*-like (Metaviridae) (Hull 1999; Pringle 1999) which are categorized by the order of genes within the *pol* polyprotein (Kumar and Bennetzen 1999). Members of both types are ubiquitous in plant genomes (Flavell *et al.* 1992; Levin 2002). Unlike class II, or DNA elements, which excise from a chromosomal location and insert elsewhere in the genome, class I elements transpose through an RNA intermediate so that a single genomic copy can potentially be the source of numerous new insertions (Kumar and Bennetzen 1999).

Plant retrotransposons have been shown to be activated by several forms of stress to the host plant including wounding, tissue culture, pathogen attack, and chemical treatment (Feschotte *et al.* 2002; Grandbastien 1998). Wide crosses may also be a source of genomic stress leading to the activation of elements (McClintock 1984). Homoploid interspecific hybridization has been shown to activate LTR retrotransposons in wallabies (Waugh O'Neil *et al.* 1998) and

Drosophila (Labrador *et al.* 1999). However element activation was not detected in other homoploid interspecific crosses (*e.g.* Robinson *et al.* 2000; Roemer *et al.* 1999).

The Louisiana Iris species complex has a long history as a model system for studying the evolutionary implications of natural hybridization (*e.g.* Anderson 1949; Arnold 1997; Arnold 2000; Riley 1938). The complex consists of four species, *I. brevicaulis*, *I. fulva*, *I. hexagona*, and the rare hybrid species *I. nelsonii* (Arnold 1993; Randolph 1966). Hybrids involving the first three taxa are common in southeastern Louisiana, especially in areas of recent habitat disturbance (Randolph *et al.* 1967). The work described here is focused on *I. brevicaulis* and *I. fulva,* which are morphologically and karyotypically distinct (Randolph *et al.* 1961), but can produce vigorous hybrids with high fitness (Burke *et al.* 1998).

The goal of this study was to characterize LTR retrotransposons from the large Iris genome in order to take advantage of the abundance and distribution of these elements for the development molecular markers useful for hybridization and speciation research. Two families of related Ty3/*gypsy*-like LTR retrotransposons were characterized using PCR and genomic library screens. These IRRE elements (for IRis RetroElement) account for 6-10% of the ~9,650 mb Iris genome and are transcriptionally active in *Iris brevicaulis*, *I. fulva*, and in their $F_1$ and backcross hybrids. IRRE elements were detected in each of 11 Iris species tested, but not in several related genera. Transposon display or S-SAP primers specific to two subfamilies of IRRE elements were used to generated large numbers markers in *I. brevicaulis* and *I. fulva*, and the technique can be adapted for use in other *Iris* species as well.

## MATERIALS AND METHODS

**Materials:** All material from Louisiana Iris species (*I. brevicaulis*, *I. fulva*, *I hexagona*, *I. nelsonii*) was obtained from wild-collected plants maintained at the University of Georgia Plant Biology Department greenhouses. Other species were collected from natural populations in Georgia (*I. cristata*, *I. verna*, and *Sisyrinchium* sp.) and California (*I. bracteata*, *I. crysophylla*, *I. douglasiana*, *I. missouriensis*, and *I. longipetala*) or were obtained from plants cultivated in the University of Georgia Plant Biology Department greenhouses (*Acidanthera bicolor* and *Neomarica longifolia*). Seed of the genome size standard *Allium cepa* cv. Ailsa Craig was provided by Michael Bennett (Royal Botanical Garden Kew).

**Nucleic acid extraction:** DNA was extracted using the CTAB procedure of Doyle and Doyle (1987) as modified by Soltis *et al*. (1991) followed by treatment with Rnase A. Total RNA was prepared from leaf or root tissue using the RNeasy plant RNA extraction kit (Qiagen, Valencia, CA) and poly A+ RNA was purified from ~600 μg of leaf RNA using an Oligotex mRNA purification kit (Qiagen, Valencia, CA). First strand cDNA was obtained using the Superscript cDNA synthesis kit (Invitrogen, Carlsbad, CA).

**Cloning procedures:** Repetitive elements from *Iris fulva* and *I. brevicaulis* were isolated by constructing small insert (~200-900bp) genomic libraries for each species in the plasmid vector pBlueScript II (Strategene, La Jolla CA.) following partial digestion of genomic DNA with *Sau*3A-I. Libraries were probed with sheared $\alpha$ $^{32}$P labeled total genomic DNA (random primers labeling kit, Invitrogen, Carlsbad, CA) from either *I. fulva* or *I. brevicaulis*. Plasmid clones showing homology to retrotransposons in database searches were used to probe phage libraries constructed by cloning ~5-10 kb *Sau*3A-1 genomic fragments into the lambda ZAP

express phage vector (Stratagene, La Jolla CA). PCR products were cloned using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA).

**Polymerase Chain Reaction:** Retrotransposon fragments containing the 3' end of the integrase domain and the 5' end of the 3' LTR were amplified using the primer pair LTRSCREENF (CACAYTTGTTYGACTCGTRAGG)/LTRSCREENR (TYRTGCAAGATGTACTTGCC). PCR amplifications were performed on 50-200ng of genomic DNA in 30 μl reaction volumes containing 1.5 units of Amplitaq DNA polymerase (Perkin Elmer/Applied Biosystems, Foster City, CA ), 0.2 mM each dNTP, 1.5 mM $MgCl_2$, and the buffer supplied with the enzyme. Cycling conditions were 94° for 3 min, followed by 32 cycles of 94° for 45 sec, 52° for 45 sec, 72° for 1 min, and ending with 72° for 6 min. RT-PCR was performed using the same primers and cycling conditions except that 1μl of first strand cDNA or 1μl of the DNase-treated template RNA for cDNA synthesis (negative controls) was used as a template.

**Transposon display:** Total genomic DNA (~500 ng) was digested overnight at 37° with an excess (50 units) of *EcoR*I. Standard *EcoR*I AFLP adapters (Vos *et al.* 1995) were ligated overnight at 25° using 5 units of T4 DNA ligase and the buffer supplied by the manufacturer (Invitrogen, Carlsbad CA). Nested element-specific primers were used in the preamplification and selective amplification reactions with the selective primers closer to the element ends than the preamplification primers.

Preamplification reactions contained 10 pmol of primers homologous to the adapters plus two selective bases (Vos *et al.* 1995) and 10 pmol of primer homologous to the LTR-end of either IRRE1-A1 (CCAAACCAAACCAAGCCACACTAAACC) or IRRE1-C (ACAGGAACACRTTCCAATTACGT). Reactions were performed in 30 μl containing 3 μl of 2:1 diluted restriction/ligation reaction, 1.5 U Ampli*Taq* DNA polymerase (Perkin

Elmer/Applied Biosystems, Foster City, CA ), 0.2 mM each dNTP, and 2.5 mM $MgCl_2$, and the buffer supplied with the enzyme. The cycling conditions were 72° for 2 min, 94° for 3 min, followed by 30 cycles of 94° for 30 sec, 56° for 30 sec (IRRE1-A1) or 51° for 30 sec (IRRE1-C), 72° for 1 min, and a final elongation of 72° for 3 min.

Selective amplifications were performed in 10 μl containing 1 μl of the 10:1 diluted preamplification reaction, 5 pmol of adapter primer plus four selective bases, 3 pmol $^{33}$P-labeled IRRE1-A1 primer (CGTATAAAATACGTACACAAGAG) or IRRE1-C primer (TCCAATTACGTATAAAATACG), 1.5 U Ampli*Taq* DNA polymerase (Perkin Elmer/Applied Biosystems, Foster City, CA ), 0.2 mM each dNTP, and 2.5 mM $MgCl_2$, and the buffer supplied with the enzyme. The cycling conditions were 94° for 3 min, followed by 30 cycles of 94° for 30 sec, 56° for 50 sec (IRRE1-A1) or 51° for 30 sec (IRRE1-C), 72° for 1 min, and a final elongation of 72° for 3 min. The amplification products were run on polyacrylamide sequencing gels and visualized by autoradiography.

**DNA sequencing and analysis:** DNA clones from the plasmid library screens were sequenced by the Molecular Genetics Instrumentation Facility at the University of Georgia. Lambda clones and cloned PCR products were sequenced using the Big Dye terminator sequencing kit (Perkin Elmer/Applied Biosystems, Foster City, CA) on an ABI 377 automated DNA sequencer (Perkin Elmer/Applied Biosystems, Foster City, CA ). A primer walking strategy was employed to sequence the lambda clones and universal sequencing primers were used to sequence the cloned PCR products. DNA and amino acid sequences were aligned with the ClustalW Service at the European Bioinformatics Institute (http://www2.ebi.ac.uk/clustalw) using the default parameters, and GeneDoc (www.psc.edu/biomed/genedoc) was used to manually edit and box-shade the alignments. Neighbor-joining trees were constructed using MEGA 2.1

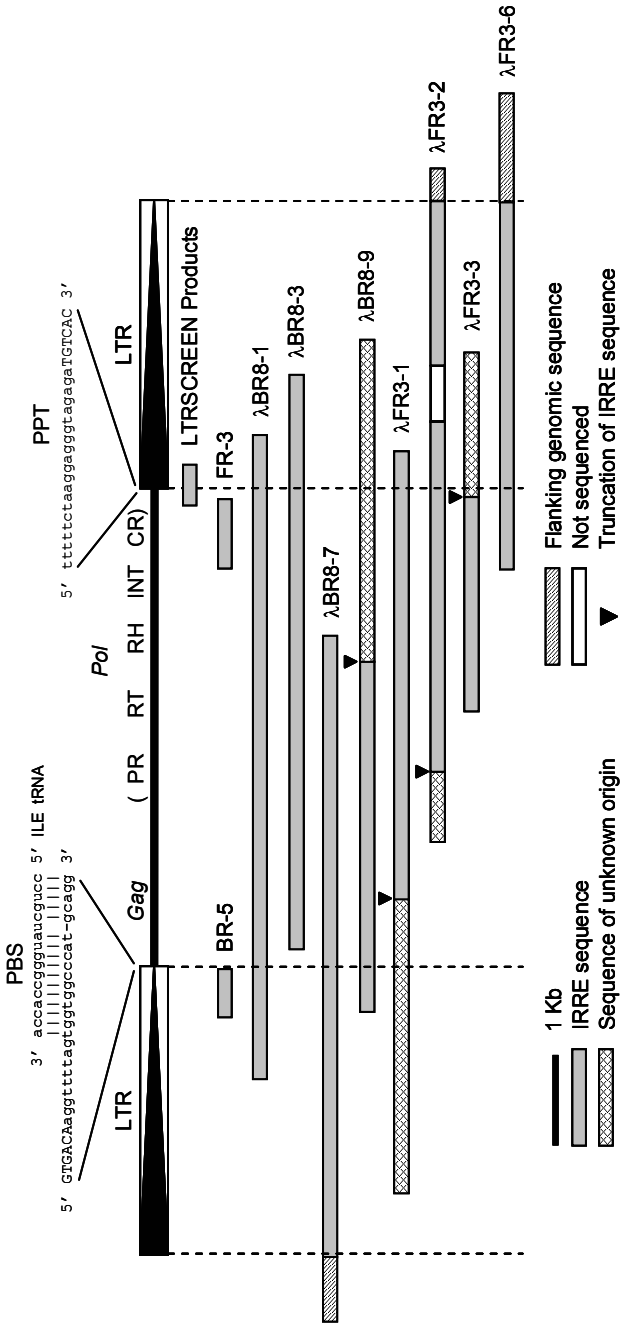
(http://www.megasoftware.net/), and the sliding window analysis was carried out using DnaSp (Rozas and Rozas 1999).

**Flow cytometry**: Nuclear DNA content was measured by flow cytometry according to Galbraith *et al* (1997). Following the recommendations of Johnston *et al* (1999), nuclei were prepared simultaneously with those of the plant genome size standard *Allium cepa* cv. Ailsa Craig, and stained with propidium iodide. Peak fluorescence was measured with a FACSCalibur flow cytometer (BD Biosciences San Jose, CA) using an excitation wavelength of 488nm. Cellquest v.3.2.1 (BD Biosciences San Jose, CA) was used to analyze the peaks and Flowjo 3.5.4 (Tree Star, Inc. San Carlos, CA) was used to generate histograms. Nuclear DNA content was calculated from the peak means according to the formula in Galbraith *et al* (1997) using 2C = 33.55pg as the value for the genome size standard (Johnston *et al.* 1999). All analyses were preformed in the CTEGD flow cytometry facility at the University of Georgia.

**Copy number determination:** Two methods, dot blot hybridizations and a genomic library screen were used to determine the copy number of the retrotransposon internal domains and LTRs. Two probes labeled with $\alpha$ $^{32}$P by random priming (Invitrogen, Carlsbad, CA) were used. The internal domain probe was FR-3, a 761bp plasmid clone containing the end of the integrase core domain and the downstream sequence (chromodomain) ending approximately 100bp before the start of the 3' LTR. The LTR probe was BR-5, a 542bp plasmid clone containing an LTR fragment ending 10bp before the 3' end of the LTR (Figure 2.1).

Serial dilutions of FR-3 or BR-5 and genomic DNA from *I. brevicaulis*, *I. fulva*, *I. hexagona*, and *I. nelsonii* were spotted onto GeneScreen hybridization membranes (NEN, Boston, MA) using a dot blot apparatus (BRL, Gaithersburg, MD). Two replicate dots containing 1ng, 10ng, 25ng, 50 ng, and 100ng of genomic DNA were made for each species (for

Figure 2.1. Reconstructed IRRE element based on a series of overlapping λ clones. Long-terminal-repeats (LTR) are depicted as 3.2 kb in length based on the length of the alignment between the LTRs of FR3-6 (2786 bp) and BR8-7 (3,003 bp). The actual range in size of these repeats among IRRE subfamilies is unknown. The sequence of the putative primer binding site (PBS) and its homology to the isoleucine tRNA of *Lupinus leuteus* (X06459) is presented at the top of the figure, as are the sequence of the polypurine tract (PPT) and LTR ends (in capital letters). The relative positions of the plasmid copy number probes (BR-5 and FR3) and the LTRSCREEN PCR products used to define IRRE subfamilies are also depicted. *Pol*, polyprotein; PR, protease; RT, reverse transcriptase; RH, RNAse H; INT, integrase; CR, chromodomain.


PBS
3′ accaccggguaucgucc 5′ ILE tRNA
5′ GTGACAaggtttttagtggtggcccat-gcagg 3′
LTR
Gag
( PR
RT
RH
INT
CR)
Pol
PPT
5′ tttttctaaggagggtagagaTGTCAC 3′
LTR
LTRSCREEN Products
FR-3
λBR8-1
λBR8-3
λBR8-9
λFR3-2
λFR3-6
BR-5
λBR8-7
λFR3-1
λFR3-3
Flanking genomic sequence
Not sequenced
Truncation of IRRE sequence
1 Kb
IRRE sequence
Sequence of unknown origin

a total of ten dots per species). Internal domain and LTR spots were also replicated twice and contained 0.01ng, 0.05ng, 0.125ng, 0.25ng, and 0.5ng of either FR-3 or BR-5. The total amount of DNA in each spot was adjusted to 100ng with salmon sperm DNA, and the DNA was bound to the membrane using ultraviolet light. DNAs were quantified by fluorimitry (Hoefer Scientific, San Francisco, CA), adjusted to the same concentration, then checked on agarose gels stained with ethidium bromide before making the final dilutions. Two identical blots were probed with either FR-3 or BR-5 before a final wash of 0.1× SSC and 0.1% SDS at 65° for 15 min. Hybridization signals from each dot were quantified with a STORM phosphoimager (Molecular Dynamics, Piscataway, NJ ) and the average number of counts per copy in the FR-3 and BR-5 dots was used to calculate the total number of copies present in each genomic dot. The genome size measurements obtained by flow cytometry for each species were then used to calculate the number of genomes per dot, and the number of copies of each probe per genome was determined by dividing the number of copies per genomic dot by the number of genomes per dot. As regressions of DNA quantity vs. hybridization signal were nearly perfectly linear ($R^2$ >0.99, data not shown) for each series of dots, the copy numbers reported are the average copy number calculated from all dots of a given species.

Copy number estimates were obtained for FR-3 and BR-5 by screening the *I. brevicaulis* primary lambda phage library (average insert size of ≈6900bp) and counting the number of positive plaques. Replicate filters were made so that the fraction of the library screened was identical for both probes. A total of 3192 plaques containing ~22 Mb were screened and copy number calculated by dividing the number of positive plaques by the proportion of the genome screened (~0.11%).

**Gel blot analysis**: DNA gel blot analysis was performed using GeneScreen hybridization transfer membranes (NEN, Boston, MA) following the manufacturers "salt transfer protocol" for transferring DNA to the membrane, and the "aqueous hybridization buffer for DNA" protocol for prehybridization and hybridization. Following overnight hybridization at 65° the membranes were washed twice with 2× SSC and 1% SDS at 60° for 15 min before a final 15 min wash at 25° with 0.1× SSC.

RNA gel blot analysis was performed as described by Seeley (1992) using approximately 5µg of polyA+ RNA isolated from *I. brevicaulis* leaf tissue. The blot was subjected to a final wash in 5mM Tris-HCl pH 8.0 and 0.1% SDS at 65° for 15 min.

## RESULTS

**Isolation and characterization of Iris LTR retrotransposons**: The cloning strategy for isolating Iris retrotransposons was based on the expectation that the highest copy number repeats should be LTR retrotransposons. High copy repetitive sequences were isolated from small-insert *I. brevicaulis* (IB) and *I. fulva* (IF) genomic libraries by probing with sheared total DNA from the genome used to construct the library. Sixteen IB clones and 12 IF clones were recovered and seven randomly chosen clones from each species were confirmed to be repetitive by DNA gel blot hybridization (data not shown) before all 28 clones were fully sequenced. BLASTX searches revealed that ten of the 28 clones share sequence similarity with the coding regions of LTR retrotransposons in the public databases.

To obtain the LTR sequence information necessary for the development of the primers for transposon display (see below), lambda phage libraries were constructed and probed. Two clones, FR3 (*fulva* repeat 3) and BR8 (*brevicaulis* repeat 8), were chosen to probe *I. fulva* and *I. brevicaulis* lambda phage libraries, respectively, based on their high level of amino acid

similarity to Ty3/*Gypsy*-like elements in the databases. FR3 is a 761bp integrase/chromodomain fragment and BR8 is a 426bp RNaseH fragment. Both probes hybridized strongly with >5% of the plaques screened. Six *I. fulva* clones hybridizing to the FR3 probe (λFR3s) and eight *I. brevicaulis* clones hybridizing to the BR8 probe (λBR8s) were chosen for DNA sequencing. Three of the λFR3 clones and four of the λBR8 clones were fully sequenced and the rest of the clones were partially sequenced from each end. The sequencing of a clone was abandoned when it became clear that it did not contain fragments useful for defining the LTR ends (our primary objective), or in a few cases, when regions that were difficult to sequence were encountered.

The sequence of the larger fragments contained in the lambda clones revealed that both of the probes were fragments of elements belonging to closely related Ty3/*Gypsy*-like retrotransposons. The elements were named IRRE for IRis RetroElement, using the naming scheme that has been applied to rice ("RIRE", Nakajima *et al.* 1996), barley ("BARE", Manninen and Schulman 1993), oat ("OARE", Kimura *et al.* 2001), and other LTR retrotransposons. The elements contain genes arranged in the order typical of Ty3/*Gypsy*-like elements (Kumar and Bennetzen 1999) and contain a putative chromatin binding domain (Malik and Eickbush 1999) downstream of the integrase gene (Figure 2.2). Consistent with this observation, IRRE elements group with other plant chromodomain-containing LTR retrotransposons in a neighbor-joining tree based on an amino acid alignment of the reverse transcriptase (RT) domain (Figure 2.3). The phylogenetic analysis also revealed two well supported groups of Iris elements (>90% of bootstrap replications) that were named IRRE1 and IRRE2 (Figure 2.3). Following the recommendation of Bowen and McDonald (1999) these two groups of elements are referred to as distinct "families" because they display greater than 10% divergence in the amino acid sequence of their RT domains. The two families are also clearly

Figure 2.2. Amino acid alignments of protein domains from IRRE and other chromodomain-containing plant retrotransposons. The IRRE1 and IRRE2 sequences presented are consensus sequences. There was insufficient data to reconstruct the IRRE2 consensus for the chromodomain, and the sequence presented is for IRRE1. The alignment of each protein domain consists of the residues present in the "core domain" defined by Pfam database (http://pfam.wustl.edu/). Other LTR retrotransposons: *del1*, *Lilium henryi* (X13886); *dea1*, *Ananas comosas* (Y12432); *RIRE3*, *Oryza sativa* (AB014738); *Legolas*, *Arabidopsis thaliana* (AC007730.1).

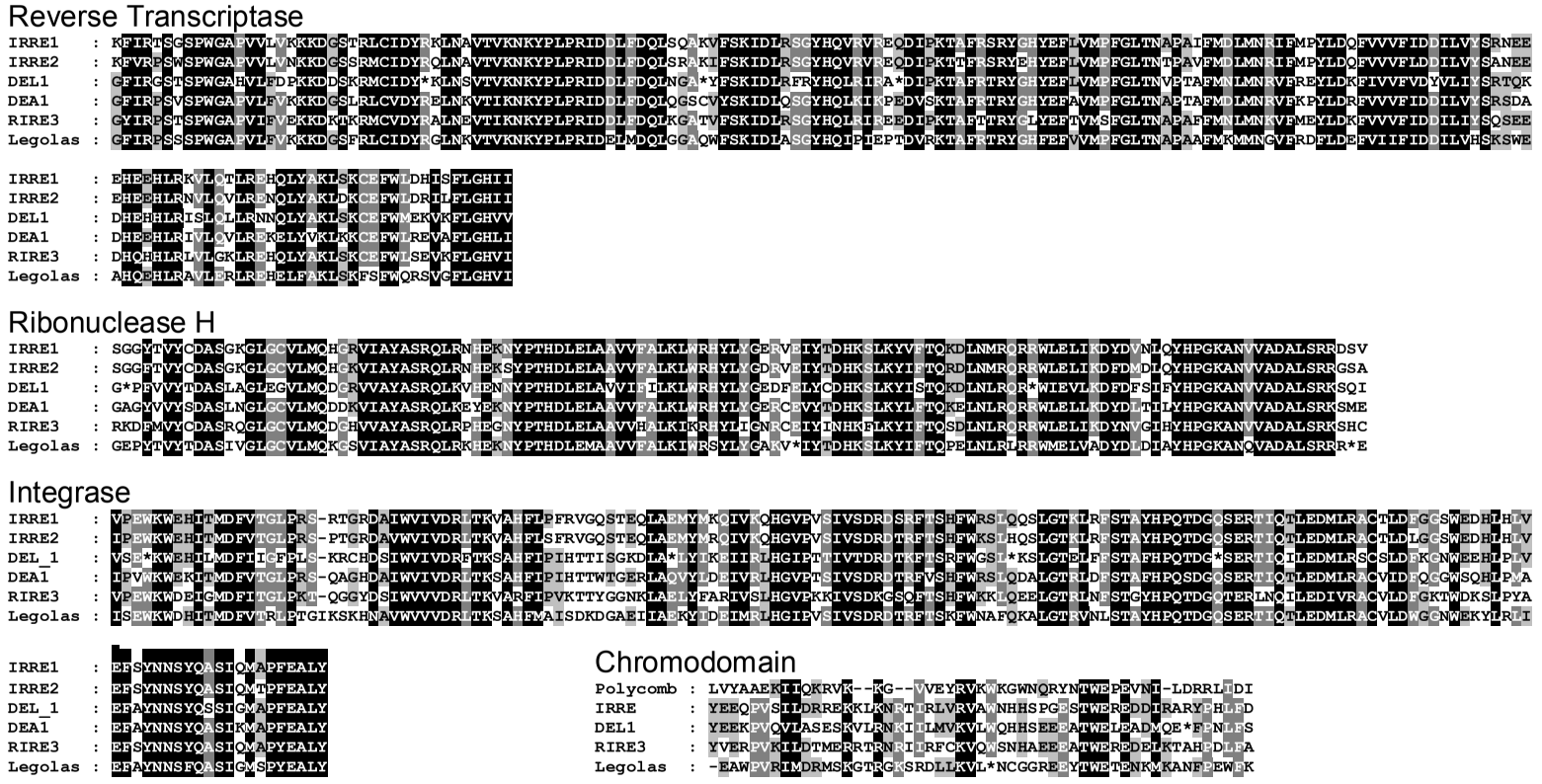
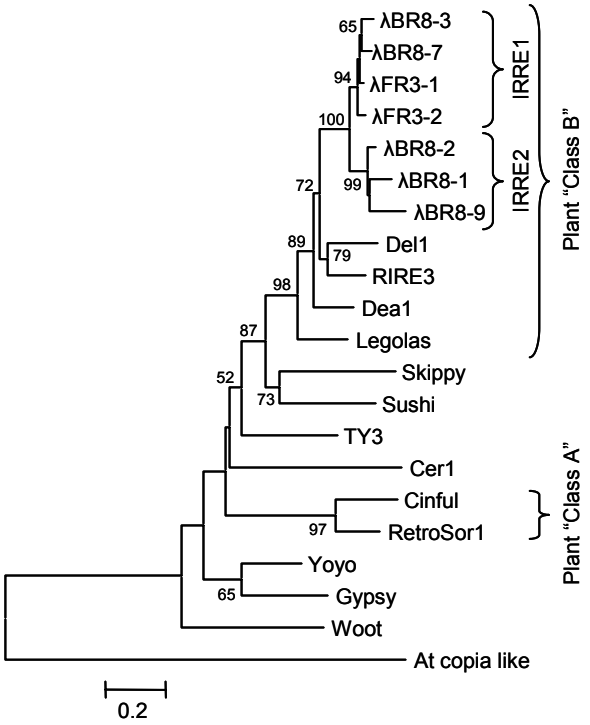

Reverse Transcriptase
Ribonuclease H
Integrase
Chromodomain
IRRE1
IRRE2
DEL1
DEA1
RIRE3
Legolas
Polycomb
IRRE
DEL1
RIRE3
Legolas

Figure 2.3. Phylogenetic relationships of Louisiana Iris retrotransposons to other Ty3/*gypsy*-like retrotransposons. The tree is based on an amino acid alignment of the reverse transcriptase domain using the neighbor joining algorithm and was rooted using a *copia*-like element from *Arabidopsis thaliana* (AAG51258). Branches receiving less than 50% support in 1000 bootstrap replications are not shown. Plant "Class A" and "Class B" groups after Marin and Llorens (2000) correspond to previously describe clades of Ty3/*gypsy*-like elements. The scale bar depicts Poisson corrected distances. Elements and the organisms from which they were isolated are: *del1*, *Lilium henryi* (X13886); *dea1*, *Ananas comosas* (Y12432); *RIRE3*, *Oryza sativa* (AB014738); *Legolas*, *A. thaliana* (AC007730.1); *Skippy*, *Fusarium oxysporum* (AAA88791); *Sushi*, *Fugu rubripes* (AAC335260); *Cer1*, *Caenorhabditis elegans* (AAA50456); *TY3*, *Saccharomyces cerevisiae* (S69842); *Cinful*, *Zea mays* (T14595); *RetroSor1*, *Sorghum bicolor* (AAD19359); *Yoyo*, *Ceratitis capitata* (AAC28743); *Gypsy, Drosophila melanogaster* (GNFFG1); *Woot, Tribolium castaneum* (AAC47271).


λBR8-3
λBR8-7
λFR3-1
λFR3-2
λBR8-2
λBR8-1
λBR8-9
Del1
RIRE3
Dea1
Legolas
Skippy
Sushi
TY3
Cer1
Cinful
RetroSor1
Yoyo
Gypsy
Woot
At copia like
65
94
100
72
99
89
79
98
87
52
73
97
65
0.2
IRRE1
IRRE2
Plant "Class B"
Plant "Class A"

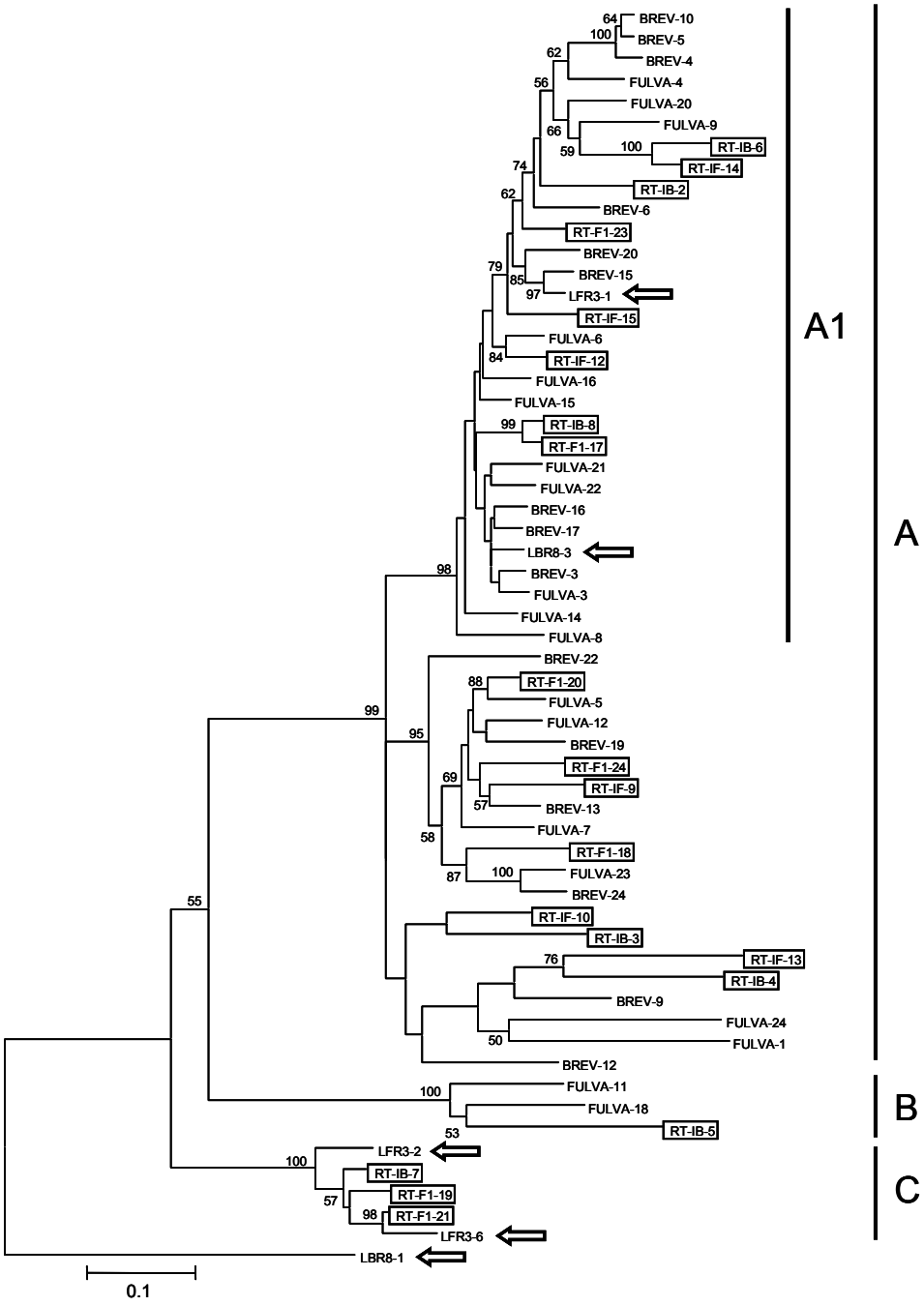
differentiated by amino acid substitutions in the additional protein core domains presented in Figure 2.2.

While retrotransposon proteins are well conserved and easily recognizable, LTR sequences are highly variable in length and in primary sequence and generally cannot be identified for uncharacterized elements using database searches. Instead, LTRs must be defined as direct repeats flanking the coding region of an element. Attempts to define the IRRE LTRs using this strategy were complicated by the length of the LTRs relative to the average insert size of the libraries from which the clones were derived (lambda phage library average insert sizes: *I. fulva* ≈6200bp, *I. brevicaulis* ≈6900 bp), so a complete IRRE sequence was reconstructed from a series of overlapping lambda clones representing paralogous copies of the element (Figure 2.1). Variable, but identifiable direct repeats of ~2.8–3.0kb flanking the coding region of several clones were identified as likely LTRs. The LTRs end in the typical 5' TG preceded by a polypurine tract (PPT) and in a 3' CA followed by a primer binding site (PBS). The putative PBS is most similar to the cytoplasmic isoleucine tRNA from *Lupinus luteus* (Figure 2.1; (Barciszewska *et al.* 1988)), which is unusual, as the PBS of most (but not all) plant retrotransposons is derived from a methionine tRNA (Kumar and Bennetzen 1999). LTRs typically end in short inverted repeats, and the putative IRRE LTRs end in the 6bp inverted repeat 5'TGTCAC/GTGACA3'. For additional confirmation that the LTR ends had been properly defined, the sequences flanking the putative LTR ends were compared among all of the clones containing these sequences (13 clones, both plasmid and lambda). In all cases, the sequence similarity between the clones either dropped off abruptly at the end of the LTRs (representing the flanking genomic DNA) or continued into the coding regions (either the *gag* or the integrase) of the element, as expected (data not shown).

**The Iris genome contains diverse subfamilies of IRRE elements:** Alignment of the LTR sequences from the lambda clones clearly indicated that the IRRE1 and IRRE2 families can be divided into subfamilies of elements sharing diagnostic nucleotide residues at many positions. In order to further define these subfamilies and to derive the LTR-end consensus sequences necessary for the design of transposon display primers, the PCR primer pair LTRSCREENF/LTRSCREENR was used to amplify IRRE fragments consisting of the non-coding region after the stop codon of the *pol* domain and the first ~280bp of the 3' LTR (Figure 2.1). These primers are degenerate and were designed to amplify as many IRRE variants as possible given the available sequence information. A total of 34 of these PCR products were cloned from genomes of *I. brevicaulis* and *I. fulva* and sequenced, revealing remarkable LTR diversity among IRRE elements. The relationships among IRRE subfamilies as defined by these LTR sequences is presented in the neighbor joining tree of Figure 2.4. While the adjacent internal domain is relatively conserved among all of the sequenced PCR products, the LTR sequences contain numerous insertion/deletion polymorphisms of ~3-30bp that are most often shared by several sequences. The overall size of the region corresponding to these PCR products among 59 genomic and cDNA sequences (see below) varies from 382 to 498 bp.

Retrotransposons are transcribed from promoter elements typically located in the 5' end of the LTR. To identify potential IRRE promoter sequences, each PCR product was analyzed with eukaryotic promoter prediction software (http://www.fruitfly.org/seq_tools/promoter.html). A region of the LTR was consistently identified as a likely (score ≥ 0.90) TATA box and transcriptional start site. To confirm this result, a sliding window analysis of nucleotide diversity across the alignment of genomic PCR products was performed to search for conserved, and therefore possible functional domains within the IRRE LTR sequences (Figure 2.5). Two highly

Figure 2.4. Neighbor joining tree of IRRE LTR ends and adjacent internal region. The tree is based on the alignment of RT-PCR and genomic PCR products from the primer pair LTRSCREENF/LTRCREENR and of lambda clones containing the same region, which includes ~200bp of the internal domain downstream of the *pol* stop codon and the first ~280 bp of the 3' LTR. RT-PCR products are boxed and arrows indicate the lambda clones. Numbers on the branches represent the percent bootstrap support calculated from 1000 replicates, with values <50% not shown. The tree is rooted with the IRRE2 element λBR8-1. The scale bar depicts distances based on the Kimura two-parameter substitution model. Bars and letters (A,A1,B,C) to the right of the tree indicate IRRE1 subfamily designations.


BREV-10
BREV-5
BREV-4
FULVA-4
FULVA-20
FULVA-9
RT-IB-6
RT-IF-14
RT-IB-2
BREV-6
RT-F1-23
BREV-20
BREV-15
LFR3-1
RT-IF-15
FULVA-6
RT-IF-12
FULVA-16
FULVA-15
RT-IB-8
RT-F1-17
FULVA-21
FULVA-22
BREV-16
BREV-17
LBR8-3
BREV-3
FULVA-3
FULVA-14
FULVA-8
BREV-22
RT-F1-20
FULVA-5
FULVA-12
BREV-19
RT-F1-24
RT-IF-9
BREV-13
FULVA-7
RT-F1-18
FULVA-23
BREV-24
RT-IF-10
RT-IB-3
RT-IF-13
RT-IB-4
BREV-9
FULVA-24
FULVA-1
BREV-12
FULVA-11
FULVA-18
RT-IB-5
LFR3-2
RT-IB-7
RT-F1-19
RT-F1-21
LFR3-6
LBR8-1
0.1
A1
A
B
C

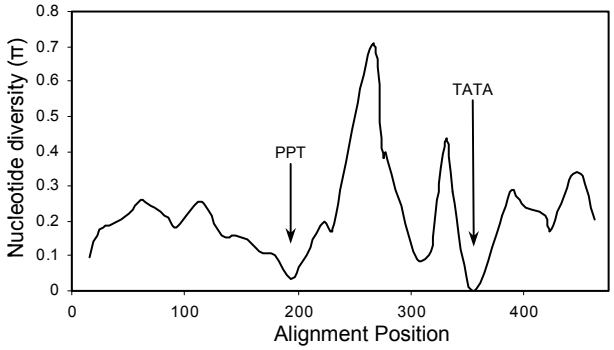



Figure 2.5. Nucleotide diversity across an alignment of 34 LTR sequences and adjacent non-coding internal domain. Sliding window analysis of genomic LTRSCREEN products show regions of low diversity corresponding to the PPT/LTR start (PPT), and to the predicted promoter sequence (TATA). A 25 bp window length and step size of 12 bp was used with the pairwise gap deletion option of DNAsp selected.

conserved regions were detected, one of them corresponding to the putative PPT/LTR-end, and the other located ~150 bp downstream in the alignment. This second conserved region corresponds to the putative promoter sequences independently identified by the promoter prediction software.

**Louisiana Iris genome size:** Because an estimation of total genome size is required to calculate the copy number of IRRE elements, flow cytometry was used to measure the C-values of each of the four hybridizing Louisiana Iris species (Figure 2.6). The values measured for each species (*I. brevicaulis*, 2C=19.75 pg; *I. fulva*, 2C=19.57 pg; *I. hexagona*, 2C=19.59 pg; *I. nelsonii*, 2C=20.04 pg) are comparable to the available data for other *Iris* species (Iris median =19.05 pg Bennett *et al.* 1998). All of the Louisiana Iris species appear to have similarly sized genomes, but valid comparisons at the species level are not possible because only a single individual of each was measured. The size of these genomes is large relative to other angiosperms, as *I. fulva* has a larger genome than ~81% of the approximately 3400 species that have been measured (Bennett *et al.* 1998).

**Copy number estimation of IRRE elements:** Two methods were used to determine the genomic copy number of IRRE elements, dot blot hybridizations and a genomic library screen. Estimates were obtained for each of the four hybridizing Louisiana Iris species using dot blots, and independently estimated for *I. brevicaulis* by screening the primary phage library used to isolate the IRRE clones (the *I. fulva* library was amplified and was therefore inappropriate for copy number determination). The results obtained for both methods are presented in Table 2.1 and indicate that between $6.5 \times 10^4$ and $1 \times 10^5$ copies of IRRE elements are present per haploid

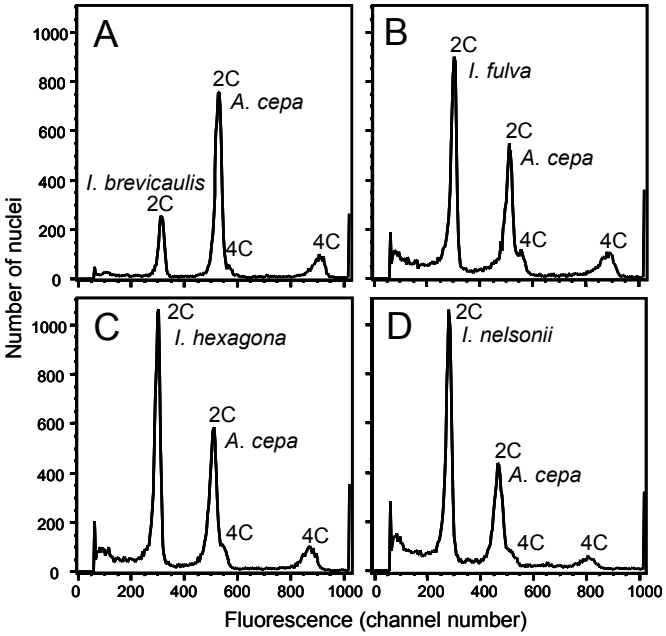



Figure 2.6. Flow histograms of nuclear DNA content in the Louisiana Iris with *Allium cepa* cv Ailsa Craig nuclei used as the genome size standard. Coefficients of variation for Iris and Allium 2C peaks respectively are (A) 3.28%, 2.25%; (B) 3.49%, 2.58%; (C) 3.40%, 2.68%; (D) 3.91%, 2.97%.

Table 2.1. IRRE copy number.

| Iris species | INT | LTR | LTR/INT |
|---|---|---|---|
| *I. brevicaulis*[a] | 204,048 | 374,530 | 1.84 |
| *I. brevicaulis*[b] | 128,355 ±9,777 | 180,098 ±27,937 | 1.40 |
| *I. fulva*[b] | 188,459 ±14,652 | 253,384 ±43,900 | 1.34 |
| *I. hexagona*[b] | 163,843 ±22,451 | 198,612 ±46,202 | 1.21 |
| *I. nelsonii*[b] | 154,752 ±28,582 | 209,642 ±45,701 | 1.35 |

[a] Genomic library screen

[b] Dot blot. The mean ± the standard deviation of ten replicate dots is reported.

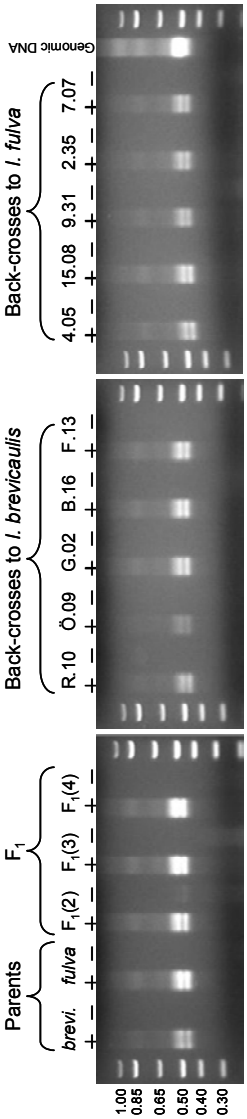
genome. Assuming an average element size of 11 kb, IRRE sequences are estimated to account for approximately 6-10% of the Louisiana Iris genome.

**Estimation of the number of solo LTRs:** Recombination between the LTRs of a retrotransposon can result in the loss of internal sequences, leaving behind a solo LTR. The ratio of intact elements to solo LTRs for the barley retrotransposon *BARE-1* has been shown to be highly variable among barley species, with the excess of LTR sequences reported to be seven to 42-fold more than the expected two to one ratio (Vicient *et al.* 1999). In an effort to determine the ratio of intact IRRE elements to solo LTRs in the Louisiana Iris, the copy number of both an internal and an LTR probe were calculated (Table1). However, in all cases, fewer LTR sequences than the expected minimum ratio of two to one were detected based on the calculated number of internal regions (Table 2.1). This result is most likely due to the rapid divergence of the noncoding LTR sequences relative to the more highly conserved element coding domains as the LTR sequences of the lambda clones differ by as much as 30% in the probe region and the hybridization wash conditions were stringent. If the LTRs evolve faster than the internal domains then the LTR probe would hybridize to fewer IRRE subfamilies than the internal probe, resulting in an underestimate of the number of LTRs. To test for this possibility, comparisons of nucleotide similarity for the region homologous to the LTR probe (BR5) and for the element protein core domains were made among all pairs of lambda clones containing the appropriate sequences. In these comparisons, the nucleotide sequence of the region homologous to the LTR probe is significantly more divergent among element copies than the coding regions (randomization test, $P<0.001$), suggesting that region of the IRRE LTR corresponding to the probe evolves at a faster rate than the element coding regions.

**IRRE elements are trancriptionally active:** In order to test for the possible transcriptional activity of IRRE elements an *I. fulva* × *I. brevicaulis* interspecific mapping population was assayed using RT-PCR. The parents (*i.e.* "pure" *I. brevicaulis*, and *I. fulva*), several $F_1$ plants and five back crosses to each parent were assayed for IRRE transcripts using the LTRSCREEENF/LTRSCREENR primer pair. Transcripts were present in all of the genotypes tested (Figure 2.7). Contamination by genomic DNA was ruled by negative controls which used the DNase-treated RNA as the amplification template. For *I. brevicaulis* transcripts were also detected on northern blots (data not shown), but only when a relatively large amount of polyA+ RNA (~5μg) was used, suggesting that IRRE transcripts are not particularly abundant. To verify that the amplified bands represent IRRE fragments, 21 cloned PCR products were sequenced from *I. brevicaulis*, *I. fulva*, and from an $F_1$ hybrid between them. The two bands evident in all of the RT-PCR reactions represent different subfamilies of elements containing insertion/deletion polymorphisms with the larger band representing at least two sequence variants that result in similar overall fragment length.

**IRRE retrotransposons are useful molecular markers:** One of the primary reasons for characterizing LTR retrotransposons from Louisiana Iris species was to develop transposon display or S-SAP markers (Van den Broeck *et al.* 1998; Waugh *et al.* 1997). This marker technology is attractive for use in the Louisiana Iris system because it takes advantage of the abundant repeats that are characteristic of plants with large genomes. To this end, PCR primers were developed based on the consensus sequence of the ends of the LTRs for two IRRE1 subfamilies, IRRE1-A1 and IRRE1-C (Figure 2.4). The subfamily designated IRRE1-A1 contains both genomic and RT-PCR products, whereas the IRRE1-C subfamily contains only

Figure 2.7. RT-PCR amplification of Iris retrotransposon sequences. Lanes marked "+" contained cDNA as the PCR template while "–" lanes contained RNA untreated with reverse transcriptase as a control for contamination by genomic DNA. Lane designations in the back-cross panels refer to individual genotypes from the mapping population. Products include the end of the putative chromatin binding domain and the first ~280 bp of the LTR. Double bands are the results of insertion-deletion polymorphisms among LTR sequences.


Parents
brevi. fulva
+ – + –
F1
F1(2) F1(3) F1(4)
+ – + – + –
Back-crosses to I. brevicaulis
R.10 Ö.09 G.02 B.16 F.13
+ – + – + – + – +
Back-crosses to I. fulva
4.05 15.08 9.31 2.35 7.07
+ – + – + – + – +
Genomic DNA
1.00
0.85
0.65
0.50
0.40
0.30

RT-PCR products and two of the lambda clones. The LTR sequence of the two subfamilies is divergent in the region suitable for transposon display primer sites, enabling the design of subfamily specific primers.

In order to test the level of polymorphism of the IRRE retrotransposon-based markers, ten wild-collected individuals each of *I. fulva* and *I. brevicaulis* were screened using primers specific for the IRRE1-A1 and IRRE1-C subfamilies (Figure 2.8). Both sets of primers amplified numerous bands from each species, and a high proportion of these bands are polymorphic among the individuals tested (Table 2.2). Several of the monomorphic bands appear to be species specific markers (Table 2.2) which are particularly useful for studying natural hybridization. However, transposon display generates dominant markers, and high frequency insertions may not be distinguishable from fixed insertions when the number of individuals sampled is small. Assuming that the monomorphic bands in the sample are fixed, the proportion of polymorphic loci is significantly different between the two species for both elements (exact test: IRRE1-A1, $P < 0.0001$; IRRE1-C, $P = 0.023$). No significant difference in the level of polymorphism between the two element subfamilies was detected within either species (exact test: *I. brevicaulis*, $P = 0.813$; *I fulva*, $P = 0.085$), suggesting that the timing and/or level of retrotranspositional activity is not dramatically different for the two subfamilies. The majority of bands generated for both subfamilies are likely to represent individual loci as they segregate in normal Mendelian ratios in a separate set of linkage mapping experiments using these markers (A. Bouck, E. Kentner, R. Peeler, M. Arnold, S. Wessler unpublished data).

**IRRE retrotransposons are present in many Iris species:** To investigate the taxonomic distribution of IRRE LTR retrotransposons, we assayed their presence in 11 Iris species and in

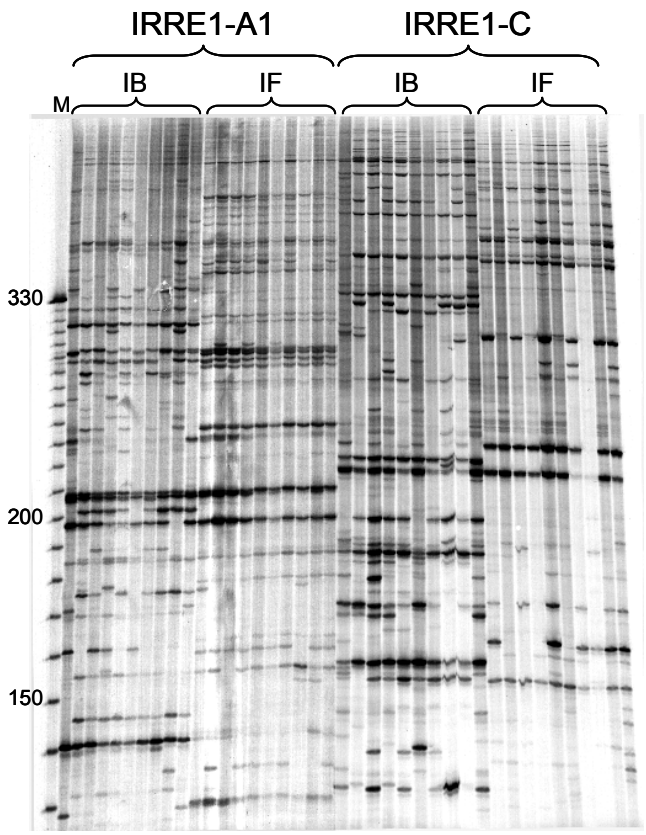
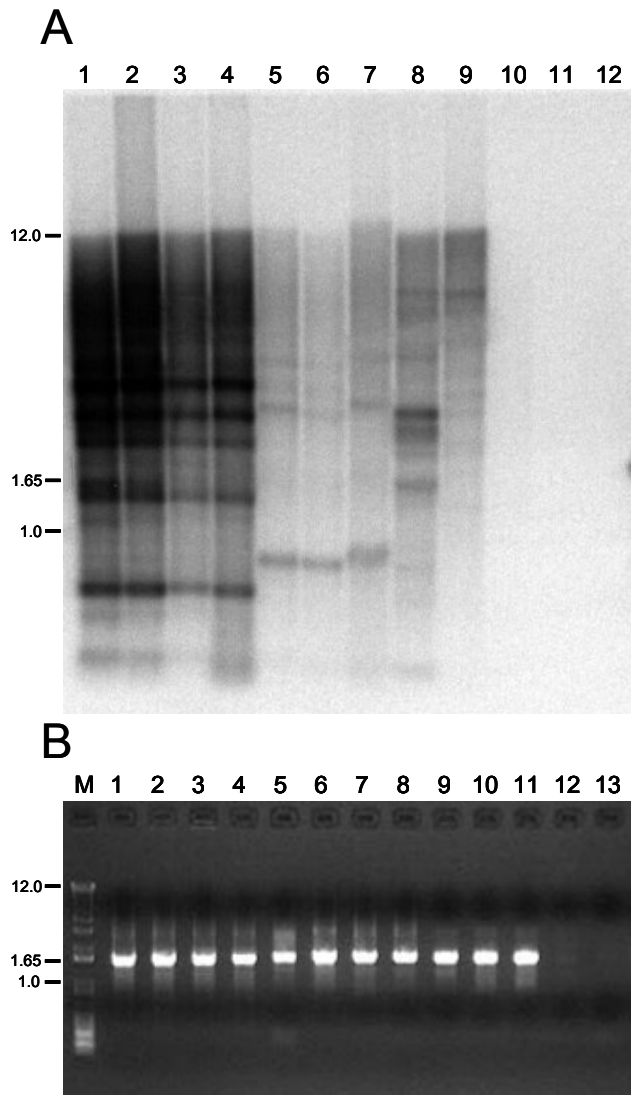



Figure 2.8. Transposon display using primers specific to the two element subfamilies IRRE1-A1 and IRRE1-C. Both sets of reactions used the four selective bases CTAT. Ten wild-collected individuals of each species were used. IB, *I. brevicaulis*. IF, *I. fulva*. Size markers to the left are in number of base pairs.

Table 2.2. Polymorphism detected by IRRE transposon display

| Element subfamily | Species | Total Number of bands | Percentage of polymorphic loci | Number of species specific bands |
|---|---|---|---|---|
| IRRE1-A1 | *I. brevicaulis* | 63 | 84.1 | 2 |
| | *I. fulva* | 53 | 45.3 | 3 |
| IRRE1-C | *I. brevicaulis* | 61 | 82 | 6 |
| | *I. fulva* | 56 | 62.5 | 8 |

Figure 2.9. Survey of IRRE-like retrotransposons in the genus Iris and in other Iridaceae. (A) Genomic DNA gel blot hybridization of the integrase/chromodomain probe FR3 to the genomic DNAs of 12 species of Iridaceae. All lanes contain approximately 2 μg of genomic DNA digested with *Eco*RI: *I. brevicaulis* (lane 1), *I. fulva* (lane 2), *I. hexagona* (lane 3), *I. nelsonii* (lane 4), *I. bracteata* (lane 5), *I. douglassiana* (lane 6), *I. inominata* (lane 7), *I. cristata* (lane 8), *I. verna* (lane 9), *Acidantthera bicolor* (lane 10), *Neomarica longifolia* (lane 11), and Sisyrinchium sp. (lane 12). (B) PCR amplification IRRE fragments (integrase core domain plus ~280bp of the 3' LTR) from Iris species. *I. brevicaulis* (lane 1), *I. fulva* (lane 2), *I. nelsonii* (lane 3), *I. hexagona* (lane 4), *I. verna* (lane 5), *I. cristata* (lane 6), *I. missouriensis* (lane 7), *I. longipetala* (lane 8), *I. douglassiana* (lane 9), *I. bracteata* (lane 10), *I. inominata* (lane 11), *Sisyrinchium* sp. (lane 12), *Neomarica longifolia* (lane 13).


A
1 2 3 4 5 6 7 8 9 10 11 12
12.0
1.65
1.0
B
M 1 2 3 4 5 6 7 8 9 10 11 12 13
12.0
1.65
1.0

three other genera of Iridaceae by PCR and/or southern hybridizations (Figure 2.9). The results for both techniques were consistent in all cases. IRRE elements are present in all members of the genus *Iris* examined, although the hybridization signal on Southern blots is much stronger in the Louisiana Iris than it is in other members of the genus (Figure 2.9). This result could be due to the sequence divergence of IRRE elements in the genomes of more distantly related Iris, or to lower IRRE copy number in these genomes, or both. For the California Irises (Figure 2.9A, lanes 5-7) preliminary results obtained by sequencing IRRE PCR products suggest that the lower hybridization signal may be due to sequence divergence (E. Kentner unpublished data).

## DISCUSSION

The IRRE elements are typical Ty3/*Gypsy*-like LTR retrotransposons that occur in high copy number in the genomes of each of the four species of hybridizing Louisiana Iris. LTR retrotransposons are major components of plant genomes, and the phylogenetic relationships among a diverse set of these elements or element fragments from many plant species have been determined (*e.g.* Marin and Llorens 2000). In order to determine the evolutionary placement of the IRRE elements, we aligned IRRE RT sequences with those of representatives from each major clade of the existing phylogenies and generated a neighbor joining tree. The group of elements to which the IRRE elements belong was originally identified by Wright and Voytas (1998 "Plant Branch 1"), with additional elements assigned to it by Marin and Lorens (2000 "Plant Class B"). The members of this group are characterized by having a putative chromatin binding domain downstream of the integrase gene. The clade seems to be ancient and ubiquitous in plants since it includes elements from monocots, dicots and gymnosperms. As IRRE elements are most closely related to elements from other monocots (Figure 2.3), it appears that the primary mode of transmission for these elements has been vertical.

Several subfamilies of IRRE elements can be distinguished based on the sequence variation in their LTR-ends. This variation is similar to the variation documented among the Tnt1 subfamilies of tobacco (Casacuberta *et al.* 1997; Casacuberta *et al.* 1995; Vernhettes *et al.* 1998) and among copies of the Retrolyc1 retrotransposon of Lycopersicon (Araujo *et al.* 2001), for which the promoter region is variable among subfamilies while the adjacent internal region is more highly conserved. For Tnt1, this promoter variation is correlated with the expression of specific subfamilies in response to different stress-associated signaling molecules (Beguiristain *et al.* 2001), suggesting that adaptive promoter variants have arisen through the error prone process of retrotransposition (Beguiristain *et al.* 2001; Casacuberta *et al.* 1997; Preston 1996). Although this study has identified putative promoter elements within the IRRE LTRs, it is currently unknown whether the variation present in these sequences has an influence on the replication cycle of the retrotransposons. From a practical standpoint, it is fortunate that the most highly variable LTR region among IRRE subfamilies (Figure 2.5) corresponds to the optimal region for transposon display primers. This has facilitated the development of subfamily-specific primers that will make comparative studies of polymorphism among element subfamilies possible.

Although the 3' ends of the IRRE LTRs are less variable than the 5' ends containing the putative promoter elements, the sequence of the 3' end of the LTR corresponding to the copy number probe is more variable among IRRE copies than the sequence of the internal probe. Given the level of LTR variation among IRRE subfamilies and the size of the Iris genome, the accurate quantification of the ratio of intact elements to solo LTRs may require an alternative strategy such as the construction and screening of BAC libraries, which would be very difficult considering the size of the Iris genome. As discussed by Meyers (2001), there are limitations to

measuring copy number with hybridization-based techniques because both sequence divergence among repetitive elements and copy number can affect the hybridization signal. Also, the accuracy of the dot blot technique for determining copy number is dependent on the quantification of the DNA in the dots unless a probe for a single-copy gene is used as an internal control. The copy number estimates from the library screening do not depend on DNA quantification and may be more accurate. However, both methods indicate that at least 800 Mb of the haploid Iris genome is composed of IRRE elements.

At approximately 0.75-1.0 $\times 10^5$ copies, IRRE elements are abundant in the Louisiana Iris genome, but well within the range that has been observed for LTR retrotransposons in other plant genomes. For example, the Ty3/*gypsy*-like element Huck accounts for ~10% of the 2.5 $\times 10^9$ bp maize genome with a copy number exceeding $1 \times 10^5$ (Meyers *et al.* 2001). An element closely related to IRRE, *del*1, is present in $1.3 \times 10^4$ copies in *Lilium henryi*, but the copy number is variable among Lilium species and is not correlated with the relationships between species (Joseph *et al.* 1990). IRRE copy number also appears to be variable among Iris species (Figure 2.9), although sequence divergence may also contribute to the hybridization pattern. The copy number of *BARE-1* in wild barley is correlated with microclimatic conditions and varies more than three-fold among individuals within a single canyon (Kalendar *et al.* 2000). It will be interesting to compare the BARE-1 results to the situation in Iris by investigating the insertional polymorphism of IRRE elements in natural Iris populations.

To date, only a handful of plant LTR retrotransposons have been shown to be transcriptionally active, with activation most often associated with biotic or abiotic stresses (Feschotte *et al.* 2002; Grandbastien 1998). The exceptions seem to be the *BARE-1* element from barley and the related *OARE-1* from oat for which low levels of transcription are detectable

under normal growing conditions (Kimura *et al.* 2001; Suoniemi *et al.* 1996), although *OARE-1* transcription is also up regulated by stress (Kimura *et al.* 2001). Like these elements, the high copy IRRE elements are expressed under normal growing conditions (Figure 2.7). However, retrotransposition can be controlled post-transcriptionally, and transcribed elements do not necessarily produce new insertions (Curcio and Garfinkel 1999). The maize genome appears to have reached its present size through recent bursts of retrotransposon activity (SanMiguel and Bennetzen 1998). An interesting, but unresolved question is whether the Iris genome has reached its present size through such bursts of retrotransposition or through continuous element activity as suggested by the transcription data.

McClintock (1984) predicted that interspecific hybridization may be a form of genomic stress that could lead to the mobilization of transposable elements. Indeed, interspecific hybridization in wallabies is associated with genome-wide loss of DNA methylation and a massive amplification of retrotransposons within a single generation (Waugh O'Neil *et al.* 1998). A less dramatic, but significant, increase in retrotransposition has also been documented following interspecific hybridization in Drosophila (Labrador *et al.* 1999). To investigate the possibility that hybridization between Louisiana Iris species could lead to the transcriptional activation of IRRE retrotransposons, a backcross interspecific mapping population was assayed for element-encoded transcripts by RT-PCR. However, transcripts were present in all of the hybrid and pure species individuals tested, and there was no evidence that previously quiescent IRRE elements were activated following hybridization. While the RT-PCR results seem to rule out the kind of retrotransposition burst observed in wallaby hybrids, no conclusion can be reached regarding more subtle changes in the level of transcription in the various hybrids with the current data. If retrotansposition is occurring in the hybrids, new insertions may be very

difficult to detect given the number of existing IRRE elements in the Iris genome. To date, no genetic evidence for new insertions has been observed in a large sample of backcross hybrids in a mapping population that has been genotyped extensively with IRRE transposon display markers (A. Bouck, E. Kentner, R. Peeler, M. Arnold, S. Wessler unpublished data).

Transposon display markers were developed for two subfamilies of IRRE elements and insertional polymorphism was assayed in wild collected individuals of *I. brevicaulis* and *I. fulva*. For markers derived from both subfamilies of elements, the proportion of polymorphic loci is higher for *I. brevicaulis* than for *I. fulva*. The allozyme data of Arnold *et al* (1990) show the same trend in species level polymorphism with *I. brevicaulis* containing a higher proportion of polymorphic loci (54%) than *I. fulva* (45%). The standing level of polymorphism detected by any marker system can be influenced by many aspects of a species' population biology (*e.g.* Charlesworth and Wright 2001; Hamrick and Godt 1996), but this fact has often been ignored in the literature in favor of arguments equating the insertional polymorphism of transposons with recent element activity. That the timing of insertion events cannot necessarily be inferred from the existence of polymorphism has been clearly demonstrated in maize, where sequencing data have shown that polymorphisms generated by a burst of retrotransposition estimated to have occurred 2-3 million years ago are still segregating in modern North American maize lines (Fu and Dooner 2002). Currently, very little data exist on the population genetics of plant retrotransposons. If, in contrast to the situation in Drosophila where most euchromatic retrotransposon insertions are likely to be deleterious (Bartolome *et al.* 2002; Carr *et al.* 2002; Charlesworth and Langley 1989), the average IRRE insertion is neutral with respect to plant fitness, then the polymorphism of IRRE insertions is likely to be influenced by the population biology of the Iris species, as suggested by the allozyme data. Although no data currently exist

pertaining to the fitness effects of IRRE insertions, it is difficult to imagine these elements attaining such a high copy number if new insertions are most often deleterious.

Retrotransposons closely related to the IRRE elements cloned from *I brevicaulis* and *I. fulva* are present in each of 11 Iris species tested. The sample includes a representation of species belonging to the subgenus Limniris (the beardless Iris), and it is likely that all native North American Iris contain these elements. The LTR ends of IRRE elements can be readily amplified from all of these species using the degenerate primers, and these products can be cloned and sequenced using standard techniques. The sequence of the LTR ends can then be used to define additional IRRE subfamilies for transposon display development. As the preliminary sequencing of LTR-ends from several species outside of the series Hexagonae (the Louisiana Iris) have yielded divergent complements of IRRE subfamilies (E. Kentner unpublished data), the application of these markers to other Iris species may require this additional step of subfamily discovery and definition. The markers should be useful for many applications in evolutionary biology and genetics and it will be interesting to compare insertional polymorphism among IRRE subfamilies and among Iris species in order to gain insight into the dynamics of retrotransposition in large plant genomes.

ACKNOWLEDGMENTS

We thank Mike Scanlon, Brad Bernstein and Ryan Peeler for assistance in the laboratory and Ning Jiang, Cedric Feschotte, Scott Cornman, and Amy Bouck for critical reading of the manuscript. EKK was supported by a training grant from the National Science Foundation (DBI 9602223).

LITERATURE CITED

Anderson, E. A., 1949 *Introgressive Hybridization*. John Wiley and Sons, New York, USA.

Araujo, P. G., J. M. Casacuberta, A. P. P. Costa, R. Y. Hashimoto, M. A. Grandbastien *et al.*, 2001 Retrolyc1 subfamilies defined by different U3 LTR regulatory regions in the Lycopersicon genus. Molecular Genetics and Genomics **266:** 35-41.

Arnold, M. L., 1993 *Iris nelsonii* (Iridaceae): Origin and genetic composition of a homoploid hybrid species. American Journal of Botany **80:** 577-583.

Arnold, M. L., 1997 *Natural hybridization and evolution*. Oxford University Press, Oxford, UK.

Arnold, M. L., 2000 Anderson's paradigm: Louisiana Irises and the study of evolutionary phenomena. Molecular Ecology **9:** 1687-1698.

Arnold, M. L., J. L. Hamrick and B. D. Bennett, 1990 Allozyme variation in Louisiana Irises: a test for introgression and hybrid speciation. Heredity **65:** 297-306.

Barciszewska, M., G. Keith, G. Dirheimer and J. Barciszewski, 1988 The Nucleotide-Sequence of Transfer $RNA^{ser}$ and Transfer $RNA^{ile}$ of Yellow Lupin Seeds. Nucleic Acids Research **16:** 8175-8175.

Bartolome, C., X. Maside and B. Charlesworth, 2002 On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. Molecular Biology and Evolution **19:** 926-937.

Beguiristain, T., M. A. Grandbastien, P. Puigdomenech and J. M. Casacuberta, 2001 ThreeTnt1 subfamilies show different stress-associated patterns of expression in tobacco. Consequences for retrotransposon control and evolution in plants. Plant Physiology **127:** 212-221.

Bennett, M. D., A. V. Cox and I. J. Leitch, 1998 Angiosperm DNA C-values database. http://www.rbgkew.org.uk/cval/database1.html

Bennetzen, J. L., 2002 Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica **115:** 29-36.

Bowen, N. J., and J. F. McDonald, 1999 Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. Genome Research **9:** 924-935.

Burke, J. M., T. J. Voss and M. L. Arnold, 1998 Genetic interactions and natural selection in Louisiana Iris hybrids. Evolution **52:** 1304-1310.

Carr, M., J. R. Soloway, T. E. Robinson and J. F. Y. Brookfield, 2002 Mechanisms regulating the copy numbers of six LTR retrotransposons in the genome of *Drosophila melanogaster*. Chromosoma **110:** 511-518.

Casacuberta, J. M., S. Vernhettes, C. Audeon and M. A. Grandbastien, 1997 Quasispecies in retrotransposons: a role for sequence variability in Tnt1 evolution. Genetica **100:** 109-117.

Casacuberta, J. M., S. Vernhettes and M. A. Grandbastien, 1995 Sequence Variability within the Tobacco Retrotransposon Tnt1 Population. Embo Journal **14:** 2670-2678.

Charlesworth, B., and C. H. Langley, 1989 The population genetics of Drosophila transposable elements. Annual Review of Genetics **23:** 251-287.

Charlesworth, D., and S. I. Wright, 2001 Breeding systems and genome evolution. Current Opinion in Genetics & Development **11:** 685-690.

Curcio, M. J., and D. J. Garfinkel, 1999 New lines of host defense: inhibition of Ty1 retrotransposition by Fus3p and NER/TFIIH. Trends in Genetics **15:** 42-45.

Doyle, J. J., and J. L. Doyle, 1987 A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin **19:** 11-15.

Feschotte, C., J. Jiang and S. R. Wessler, 2002 Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics **3:** 329-341.

Flavell, A., E. Dunbar, R. Anderson, S. Pearce, R. Hartley *et al.*, 1992 *Ty1-copia* group retrotransposons are ubiquitous and heterogeneous in higher plants. Nucleic Acids Research **20:** 3639-3644.

Flavell, R. B., M. D. Bennett, J. B. Smith and D. B. Smith, 1974 Genome Size and Proportion of Repeated Nucleotide-Sequence DNA in Plants. Biochemical Genetics **12:** 257-269.

Fu, H. H., and H. K. Dooner, 2002 Intraspecific violation of genetic colinearity and its implications in maize. Proceedings of the National Academy of Sciences of the United States of America **99:** 9573-9578.

Galbraith, D. W., G. M. Lambert, J. Macas and J. Dolezel, 1997 Analysis of nuclear DNA content and ploidy in higher plants. Current Protocols in Cytometry **7.6.1-7.6.22**.

Grandbastien, M.-A., 1998 Activation of plant retrotransposons under stress conditions. **3:** 181-187.

Hamrick, J. L., and M. J. W. Godt, 1996 Effects of life history traits on genetic diversity in plant species. Philosophical Transactions of the Royal Society of London Series B **351:** 1291-1298.

Hull, R., 1999 Classification of reverse transcribing elements: a discussion document. Archives of Virology **144:** 209-214.

Johnston, J. S., M. D. Bennett, A. L. Rayburn, D. W. Galbraith and H. J. Price, 1999 Reference standards for determination of DNA content of plant nuclei. American Journal of Botany **86:** 609-613.

Joseph, J. L., J. W. Sentry and D. R. Smyth, 1990 Interspecies distribution of abundant DNA sequences in Lilium. Journal of Molecular Evolution **30:** 146-154.

Kalendar, R., J. Tanskanen, S. Immonen, E. Nevo and A. H. Schulman, 2000 Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proceedings of the National Academy of Sciences of the USA **97:** 6603-6607.

Kidwell, M. G., 2002 Transposable elements and the evolution of genome size in eukaryotes. Genetica **115:** 49-63.

Kimura, Y., Y. Tosa, S. Shimada, R. Sogo, M. Kusaba *et al.*, 2001 OARE-1, a Ty1-copia retrotransposon in oat activated by abiotic and biotic stresses. Plant and Cell Physiology **42:** 1345-1354.

Kumar, A., and J. L. Bennetzen, 1999 Plant retrotransposons. Annual Review of Genetics **33:** 479-532.

Labrador, M., M. Farre, F. Utzet and A. Fontdevila, 1999 Interspecific hybridization increases transposition rates of Osvaldo. Molecular Biology and Evolution **16:** 931-937.

Leeton, P. R. J., and D. R. Smyth, 1993 An abundant LINE-like element amplified in the genome of *Lilium speciosum*. Molecular and General Genetics **237:** 97-104.

Levin, H. L., 2002 Newly identified retrotransposons of the Ty3/*gypsy* class in fungi, plants, and vertebrates, pp. 684-701 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. American Society for Microbiology Press, Washington D.C.

Malik, H. S., and T. H. Eickbush, 1999 Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. Journal of Virology **73:** 5186-5190.

Malik, H. S., S. Henikoff and T. H. Eickbush, 2000 Poised for contagion: Evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Research **10:** 1307-1318.

Manninen, I., and A. H. Schulman, 1993 Bare-1, a Copia-Like Retroelement in Barley (*Hordeum vulgare* L). Plant Molecular Biology **22:** 829-846.

Marin, I., and C. Llorens, 2000 Ty3/Gypsy retrotransposons: Description of new *Arabidopsis thaliana* elements and evolutionary perspectives derived from comparative genomic data. Molecular Biology and Evolution **17:** 1040-1049.

McClintock, B., 1984 The significance of responses of the genome to challenge. Science **226:** 792-801.

Meyers, B. C., S. V. Tingley and M. Morgante, 2001 Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Research **11:** 1660-1676.

Nakajima, R., K. Noma, H. Ohtsubo and E. Ohtsubo, 1996 Identification and characterization of two tandem repeat sequences (TrsB and TrsC) and a retrotransposon (RIRE1) as genome-general sequences in rice. Genes & Genetic Systems **71:** 373-382.

Preston, B. D., 1996 Error-prone retrotransposition: Rime of the ancient mutators. Proceedings of the National Academy of Sciences of the United States of America **93:** 7427-7431.

Pringle, C. R., 1999 Virus taxonomy - 1999 - The Universal System of Virus Taxonomy, updated to include the new proposals ratified by the International Committee on Taxonomy of Viruses during 1998. Archives of Virology **144:** 421-429.

Randolph, L. F., 1966 *Iris nelsonii*, a new species of Louisiana Iris of hybrid origin. Baileya **14:** 143-169.

Randolph, L. F., J. Mitra and I. S. Nelson, 1961 Cytotaxonomic Studies of Louisiana Irises. Botanical Gazette **123:** 126-131.

Randolph, L. F., I. S. Nelson and R. L. Plaisted, 1967 Negative evidence of introgression affecting the stability of Louisiana Iris species. Memoir - Cornell University Agricultural Experiment Station **398:** 1-56.

Riley, H. P., 1938 A character analysis of colonies of *Iris fulva*, *Iris hexagona* var. *giganticaerulea* and natural hybrids. American Journal of Botany **24:** 649-725.

Robinson, T. J., O. Wittekindt, J. J. Pasantes, W. S. Modi, W. Schempp *et al.*, 2000 Stable methylation patterns in interspecific antelope hybrids and the characterization and localization of a satellite fraction in the Alcelaphini and Hippotragini. Chromosome Research **8:** 635-643.

Roemer, I., F. Grutzner, H. Winkig, T. Haff, A. Orth *et al.*, 1999 Global methylation in eutherian hybrids. Nature **40:** 131-132.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174-175.

San Miguel, P., A. Tikhonov, J. Young-Kwan, N. Motchoulakaia, P. S. Springer *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765-738.

SanMiguel, P., and J. Bennetzen, 1998 Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Annals of Botany **82:** 37-44.

SanMiguel, P., B. S. Gaut, A. Tikhonov, Y. Nakajima and J. L. Bennetzen, 1998 The paleontology of intergene retrotransposons of maize. Nature Genetics **20:** 43-45.

Seeley, K. A., D. H. Byrne and J. T. Colbert, 1992 Red Light-Independent Instability of Oat Phytochrome Messenger- RNA *in vivo*. Plant Cell **4:** 29-38.

Soltis, D. E., P. S. Soltis, T. G. Collier and M. L. Edgerton, 1991 Chloroplast DNA variation within and among genera of the Heuchera group (Saxifragaceae): evidence for chloroplast transfer and paraphyly. American Journal of Botany **78:** 1091-1112.

Suoniemi, A., A. Narvanto and A. Schulman, 1996 The *BARE*-1 retrotransposon is transcribed in barley from an LTR promoter active in transient assays. Plant Molecular Biology **31:** 295-306.

Van den Broeck, D., T. Maes, M. Saur, J. Zetof, P. De Keukeleire *et al.*, 1998 *Transposon Display* identifies individual transposable elements in high copy number lines. The Plant Journal **13:** 121-129.

Vernhettes, S., M.-A. Grandbastien and J. M. Casacuberta, 1998 The evolutionary analysis of the Tnt1 retrotransposon in *Nicotiana* species reveals the high variability of its regulatory sequences. Molecular Biology and Evolution **15:** 827-836.

Vicient, C. M., A. Suoniemi, K. Anamthawat-Jonsson, J. Tanskanen, A. Beharav *et al.*, 1999 Retrotransposon *BARE-1* and its role in genome evolution in the genus *Hordeum*. The Plant Cell **11:** 1769-1784.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee *et al.*, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research **23:** 4407-4414.

Waugh O'Neil, R. J., M. J. O'Neil and J. A. Marshall Graves, 1998 Undermethylation associated with retroelement activation and chromosome remodeling in a interspecific mammalian hybrid. Nature **393:** 68-72.

Waugh, R., K. McLean, A. J. Flavell, S. R. Pearce, A. Kumar *et al.*, 1997 Genetic distribution of BARE-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Molecular and General Genetics **253:** 687-694.

Wright, D. A., and D. F. Voytas, 1998 Potential retroviruses in plants: Tat1 is related to a group of *Arabidopsis thaliana* Ty3/gypsy retrotransposons that encode envelope-like proteins. Genetics **149:** 703-715.

# CHAPTER 3

# POPULATION DYNAMICS OF THE CACTA-LIKE DNA TRANSPOSON *BOC*-1 IN *BRASSICA NIGRA*[2]

[2]Kentner, E. K., Zhang, X. and S. R. Wessler. To be submitted to *Heredity*.

ABSTRACT

The availability of genomic sequence data for diverse plant species has presented an unprecedented opportunity for the study of transposable elements (TEs). Here, the dynamics of a CACTA-like family of DNA transposons, *Boc*-1, is analyzed in populations of *Brassica nigra*. The research was made possible by the previous identification of these recently active elements in searches of the TIGR genomic survey sequence of *B. oleracea*. The insertion polymorphism of *Boc*-1 elements was assayed by transposon display in four populations from southern California, and compared to the polymorphism of presumably neutral control markers. The neutrality of the TE insertions was tested using an approach based on the analysis of fixation indices, and using existing TE population genetic theory that has been applied to elements from *Drosophila*. Based on comparisons among fixation indices, most *Boc*-1 insertions appear to be neutral, but there was evidence of forces preventing individual insertions from reaching high population frequencies. Some explanations for these apparently conflicting results based on the biology of the elements are discussed.

# INTRODUCTION

Transposable elements (TEs) are ubiquitous components of plant genomes (Flavell *et al.* 1992; Kumar and Bennetzen 1999), and account for the majority of nuclear DNA in many species (Bennetzen 2002; Flavell *et al.* 1974; Kidwell 2002). The movement of these elements can create genetic polymorphisms within species (*e.g* Fu and Dooner 2002; Wright *et al.* 2001) and can lead to differences in genome structure and overall size between related species (*e.g* Joseph *et al.* 1990). Transposable element insertions have therefore been used as molecular markers in many agronomically important crops (Berenyi *et al.* 2002; Ellis *et al.* 1998; Gribbon *et al.* 1999; Pearce *et al.* 1996; Waugh *et al.* 1997). However, there have been relatively few studies investigating the polymorphism of transposable element insertions within natural populations of plants (Wright *et al.* 2001). Such data would provide information on the nature of the evolutionary forces acting on recent insertions and would be helpful in evaluating the utility of TEs as molecular markers for population studies.

The two major classes of eukaryotic mobile elements are defined by differences in their mechanism of transposition (Feschotte *et al.* 2002). Class I elements do not excise during transposition but produce a new DNA copy from an RNA template using the element-encoded enzyme, reverse transcriptase. Class I elements therefore increase in copy number during the process of transposition and can become very abundant. Markers based on the insertion polymorphism of these elements have proven to be phylogenetically informative in many groups of organisms because each insertion has a known ancestral state, the elements do not excise, and because independent insertions into the same site are unlikely (Cook and Tristem 1997; Tatout *et al.* 1999). Class II, or DNA elements, transpose using a "cut-and-paste" mechanism in which the element excises from one chromosomal location and moves to another. Some Class II elements,

particularly miniature inverted repeat transposable elements (MITEs), can also attain very high copy numbers through molecular mechanisms that are not completely resolved (Feschotte *et al.* 2002). Class II elements are also useful as molecular markers (*e.g* Casa *et al.* 2000).

Population studies of transposon polymorphism in *Drosophila melanogaster* have led to the general conclusion that euchromatic insertions in this species are deleterious and are eliminated from populations by the action of purifying selection (Charlesworth and Langley 1989), although the exact nature of the deleterious effects remains controversial (Biemont 1992; Biemont *et al.* 1997; Charlesworth *et al.* 1997). In other animals, including humans and other *Drosophila* species, many transposable element loci occur at high frequency or are fixed within a species and are therefore not likely to be deleterious (Eickbush and Furano 2002; Hey 1989). However, a minority of human insertions are clearly harmful and subject to purifying selection (*e.g* Boissinot *et al.* 2001). In plants, the accumulation of retrotransposon insertions in intergenic regions (San Miguel *et al.* 1996) suggests that these TE insertions are neutral, but insertion frequency data comparable to that available for fruit flies and humans have been collected for only a single family of Class II transposons (Wright *et al.* 2001).

The majority of the insertions of the human Class I element Alu likely occurred before the radiation of extant humans and are therefore monomorphic and not useful as markers (Roy-Engel *et al.* 2001). Analysis of Alu sequences in the public databases has led to the discovery of recently active Alu subfamilies which are polymorphic in human populations and informative as genetic markers (Roy-Engel *et al.* 2001). The approach taken in this study is similar. Sequence analysis of unassembled ~2.5 Kb genomic clones from *Brassica oleracea* was used to identify recently active subfamilies of TEs (Zhang and Wessler 2004) and to develop transposon display primers (Van den Broeck *et al.* 1998; Waugh *et al.* 1997) that could be used to score insertion

polymorphism in this and in other *Brassica* species. These markers are based on a modified amplified fragment length polymorphism (AFLP, Vos *et al.* 1995) protocol that uses a ligated adaptor primer and a primer specific for the terminal sequence of an element family to amplify fragments consisting of the element ends and adjacent genomic DNA.

*Brassia nigra* is a diploid, annual or biennial, self-incompatible, bee-pollinated species that is closely related to *B. oleracea* (Bateman 1955; Conner and Rush 1997; U 1935). Naturalized populations of *Brassica nigra* consisting of thousands of individuals are common in disturbed areas in southern California. A native of the Mediterranean region, there are anecdotal accounts of its introduction by Spanish missionaries as early as 1796. It is unlikely, however, that all of the *B. nigra* in California originated from a single introduction as numerous exotic plants, including several *Brassica* species, have been inadvertently introduced into the region. *B. nigra* was chosen for study because of the availability of these populations, and because of its close relationship to *B. oleracea*, for which genomic survey sequence is available.

The objective of this study was to investigate the polymorphism of TE insertion sites in natural populations of plants in order to make inferences regarding the selective consequences of transposition. To this end, transposon display was used to score the insertion polymorphism of the CACTA-like DNA transposon *Boc*-1 in four *B. nigra* populations from southern California. The polymorphism of *Boc*-1 was compared to the polymorphism of two sets of control markers that were generated using the same methodology, but were not anchored in the insertions of transposable elements. The fixation indices of the *Boc*-1 loci were consistent with their neutrality, but analysis of the probability distribution of insertion frequencies indicated the operation of some force or forces preventing individual insertions from drifting to high

frequency. Some aspects of the biology of CATA-like TEs are offered as an explanation for these apparently conflicting results.

## MATERIALS AND METHODS

**Sequence analysis:** The CACTA-like DNA transposon family, *Boc*-1, was identified as likely to be recently active following database searches and phylogenetic analysis (Zhang and Wessler 2004). A full length consensus sequence for the element was obtained using an iterative BLAST procedure. The nucleotide sequence encoding a fragment of the transposase domain was used as the initial query in BLASTn searches against the TIGR *B. oleracea* database. The nucleotide sequence of a hit with the highest identity to the initial query (at least 99% over >200 bp) was used as query in a second round of BLASTn searches. This process was carried out in both the 5' and 3' directions and repeated until the element termini were reached. A quasi consensus sequence was obtained by assembling these sequences and used as query in a BLASTn search against the TIGR *B. oleracea* database. Sequences of hits were compared to the quasi consensus to derive a consensus sequence by correcting mismatches or small insertion/deletions according to simple majority rule. Transposon display primers were designed based on the consensus sequence.

**Population sampling:** *Brassica nigra* plants were sampled from four populations in San Diego and Riverside Counties, California. Leaf tissue was collected from twenty individuals in each population directly into silica gel containing vials at approximate 5m intervals along a single linear transect. Although no estimates of population sizes were made, larger populations clearly consisting of thousands of individuals were deliberately selected to avoid issues associated with small effective population sizes ($N_e$). The populations were named for landmarks near their locations: Temecula (Tem), Penasquitos (Pen), El Camino (ElCam), and Guajome (Gua).

**Transposon display:** Nucleic acids were extracted from the dried tissue with a CTAB protocol (Doyle and Doyle 1987) and treated with RNAse A at room temperature. 10 µl of the extract was digested overnight at 37° with an excess of *Mse*I. Standard *Mse*I amplified fragment length polymorphism adapters (Vos *et al.* 1995) were ligated overnight at 25° using 5 units of T4 DNA ligase (Invitrogen) and the buffer supplied by the manufacturer. Transposon display was carried out as described (Casa *et al.* 2000) using the primers provided in Table 3.1. Amplification products were run on polyacrylamide sequencing gels and visualized by autoradiography. The gels were scored manually using a duplicate set of gels with 8 individuals from each population in adjacent lanes to facilitate the identification of loci across multiple gels. Bands of similar molecular weight were assumed to be homologous, after homology was confirmed for a subset of bands by sequencing.

Transposon display bands re-amplified from polyacrylamide gels were cloned using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA) according to the manufacturers instructions. All sequencing was performed by the Molecular Genetics Instrumentation Facility at the University of Georgia.

**Data analysis:** Because transposon display produces dominant markers, the frequency of null (band absent) alleles was estimated by Zhivotovsky's (1999) Bayesian technique using the equations provided for a non-uniform prior distribution. Genetic diversity measures were computed for each locus, applying a correction for unequal sample sizes (Weir 1996, p. 167) because of the loss of one individual in the Tem and Gua populations. Due to the dominance of the markers, occupancy profiles (Charlesworth and Langley 1989) were generated for each population by multiplying the estimated element frequency at each locus by the number of haploid genomes in the sample. The parameters of the probability distribution of element

Table 3.1. Transposon display primers and selective bases.

| Element | Selective Bases | Pre-amp | Selective-amp |
|---|---|---|---|
| *Boc*-1 | +G, +T | TCCGACGACMTTGRTGTCCGT | CCGTCGGAACCTCCGTCGGAA |
| C-1 | +CATC, +CATG | CTCATTCTCTTTCTCCTCTTCTC | TCTCCTCTTCTCTTGAACTCC |
| C-2 | +GAG, +GTC | CYCTCYTTATGTYCTCAACTCC | ATGTYCTCAACTCCTTTGTGTC |

frequencies, $\alpha$ and $\beta$, were jointly estimated by the minimum $\chi^2$ method of Charlesworth and Charlesworth (1983, equation A9), which fits the observed occupancy profile to the expected distribution of element frequencies. However, the published version of equation A9 has a typographical error (B. Charlesworth personal communication) and the corrected equation used was

$$E\{n_i\} = \tfrac{1}{2}\hat{n}\beta \frac{(\alpha+1)(\alpha+2)...(\alpha+i-1)(\beta+1)...(\beta+m-i)}{(\alpha+\beta+1)(\alpha+\beta+2)...(\alpha+\beta+m-1)}\binom{m}{i} \qquad (1)$$

Where $E\{n_i\}$ is the expected number of chromosomal sites occupied by the element in $i$ genomes ($i$ = 1,2,3,...$m$), and $\frac{1}{2}\hat{n}$ is the mean number of occupied sites per haploid genome. Fits of the expected and observed distributions were tested by $\chi^2$, pooling classes to avoid expected values less than 5 (Charlesworth and Charlesworth 1983).

## RESULTS

***Boc*-1 and control polymorphisms:** The *Boc*-1 family of CACTA-like DNA elements (*Boc*1) was chosen for study based on analysis of the publicly available low-coverage *Brassica oleracea* genomic sequence. The high sequence similarity among the members of the family suggests that these TEs have been recently active in the *B. oleracea* genome (Zhang and Wessler 2004). As several families of CACTA-like transposons are present in *B. oleracea*, transposon display primers were designed that would selectively amplify only the members of the *Boc*-1 family. The two polymorphism control markers (C-1 and C-2, Table 3.1) were generated using the same templates and reaction conditions as for the *Boc*-1 markers, but are not anchored in the insertions of transposable elements.

The transposon display technique produces multiple bands that vary in size based on the distance between the primer site near the end of a transposon and a restriction site in the genomic sequence flanking the element (Van den Broeck *et al.* 1998; Waugh *et al.* 1997). When the

bands are visualized using autoradiography, individual bands can be recovered by cutting them from the gel followed by reamplification. The sequence of each band is expected to contain a short (~30bp) piece of element DNA flanking the primer and a longer sequence of genomic DNA flanking the transposon. This was observed for each of three *Boc*-1 loci sequenced, confirming that the bands resulted from priming within *Boc*-1 elements. Although originally intended to amplify retrotransposon insertions in *B. oleracea*, the sequencing of *B. nigra* bands from the two control primers revealed that these markers are not anchored TE sequences. Further experiments indicated that, in *B. nigra*, the targeted retrotransposons do not contain the primer sequences, and that both control primers produce many more bands than can be accounted for by the copy number of the elements (data not shown). The control markers are therefore similar to standard AFLP markers with one end of each band anchored in a random primer.

The insertion polymorphism of the markers generated by each of the three sets of primers was scored in each of the four *B. nigra* populations from southern California (Table 3.2). A total of 32 polymorphic bands were scored for *Boc*-1, 46 for C-1, and 39 for C-2. The banding patterns were highly reproducible, with independent runs producing identical results for each of the primer combinations assayed. Because transposon display produces dominant markers, the insertion frequency at each locus in each population was estimated using Zhivotovsky's (1999) Bayesian technique.

**Fixation indices:** Evolutionary biologists have long recognized that the comparison of fixation indices across loci can be used to detect the action of natural selection (Cavalli-Sforza 1966; Lewontin and Krakauer 1973). This is based on the principle that the effects of genetic drift will be uniform across all loci in the genome while natural selection can act independently at

Table 3.2. Genetic diversity measures.

| Marker | Tem $H_e$ | Pen $H_e$ | El Cam $H_e$ | Gua $H_e$ | $P$ (%) | $H_T$ | $F_{ST}$[a] |
|---|---|---|---|---|---|---|---|
| *Boc*-1 | 0.217 | 0.202 | 0.209 | 0.206 | 99.9 | 0.216 | 0.036* |
| C-1[b] | 0.260 | 0.252 | 0.271 | 0.277 | -- | 0.294 | 0.099* |
| C-1[c] | 0.265 | 0.260 | 0.281 | 0.286 | 40.0 | 0.285 | 0.041* |
| C-2[b] | 0.182 | 0.166 | 0.165 | 0.171 | -- | 0.197 | 0.130* |
| C-2[c] | 0.180 | 0.171 | 0.151 | 0.166 | 34.0 | 0.174 | 0.040* |

[a] Averaged across loci.

[b] Calculated with all loci.

[c] Calculated with outlier loci excluded.

* $P < 0.001$

individual loci because of recombination. To screen for transposable element loci that may be subject to selection (or linked to loci under selection) fixation indices were computed for each *Boc*-1 locus and for the markers generated by the C-1 and C-2 primers. The distribution of $F_{ST}$ for the polymorphic bands generated by each primer is presented in Figure 3.1. None of the *Boc*-1 loci had $F_{ST}$ values falling outside of the range of the numerous other loci studied. Two of the C-1 markers and three of the C-2 markers had very high $F_{ST}$ values. When these outliers are removed, the average $F_{ST}$ values across loci are nearly identical for the *Boc*-1 insertions and the two control markers (Table 3.2, Figure 3.1). Because of this result and for technical reasons discussed below, statistical tests for significant deviations of fixation indices among the *Boc*-1 loci were not conducted.

**Marker frequency distributions**: The insertion frequency distributions of TEs have been used to estimate the strength of forces removing elements from a population. In theory, the distribution of element frequencies across a large number of individual chromosomal sites is expected to follow a $\beta$-distribution. The parameters describing the shape of the distribution that best fits the observations can be used to estimate the strength of the forces acting on a family of elements (Charlesworth and Langley 1989; Langley *et al.* 1983). The parameter $\alpha$ measures the combined effects of drift and the probability of insertion into a given site, and the parameter $\beta$ estimates the effects of drift and the rate of removal of elements from the population by forces such as excision and selection (Charlesworth and Charlesworth 1983; Charlesworth and Langley 1989). Although the control markers are not anchored in TEs, under the mutation-drift hypothesis the insertion frequencies of neutral markers are expected to follow a $\beta$-distribution (Chakravarti *et al.* 1980; Crow and Kimura 1970; Wright 1937), so the parameters describing

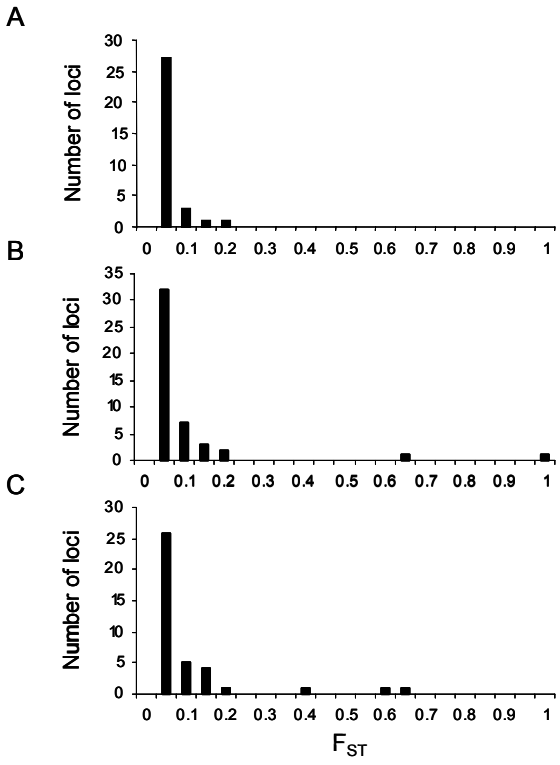



Figure 3.1. Frequency distributions of $F_{ST}$ across loci. (A) *Boc*-1 transposon display markers. (B) Markers generated by the C-1 primers. (C) Markers generated by the C-2 primers.

probability distributions of the frequencies of these markers were calculated to provide comparisons to *Boc*-1. The observed and expected frequency distributions for each of the three markers in each of the four *Brassica* populations surveyed is presented in Figure 3.2. The parameter estimates and, for *Boc*-1, the expected element frequency per occupiable band ( $\hat{x}$, estimated as $\alpha/(\alpha+\beta)$, Charlesworth and Charlesworth, 1983), are presented in Table 3.3. For each of the markers used, the parameter estimates were generally consistent across the four populations surveyed. Only loci polymorphic in at least one of the populations were included in the analysis to exclude potentially ancient fixed loci that cannot be removed from a population by selection (Wright *et al.* 2001). No fixed *Boc*-1 insertions were observed. The significance of the fit of the observed and expected distributions was tested using a $\chi^2$ test, pooling classes to avoid expected values < 5 (Figure 3.2). There was a significant lack of fit to the expected distribution in one of the populations for each of the three markers. Additionally, none of the marker distributions in the El Camino population seemed to be well described by a $\beta$-distribution (Figure 3.2), and the population may not be in equilibrium.

## DISCUSSION

Hypotheses on the evolutionary significance of mobile DNA range from regarding these sequences as merely parasitic (Doolittle and Sapienza 1980; Orgel and Crick 1980) to speculation that they could be important agents of evolutionary change (*e.g* Kidwell and Lisch 2001; McClintock 1984; McDonald 1998). The complete sequencing of the genomes of several model organisms has led to the general acceptance that the activities of TEs have an important influence on the composition and structure of genomes, but few studies in plants have attempted to investigate the link between "genome evolution" and the population-genetic forces that influence the accumulation of elements within genomes. While it is clear from genomic

Figure 3.2. Expected and observed marker frequency distributions for *Boc*-1 and the control markers. Observed and expected distributions are black and open bars, respectively. Expected values are the best fit $\beta$-distributions obtained using the minimum $\chi^2$ method. The presented $\chi^2$ values with degrees of freedom in parenthesis test the fit of the expected distribution with bins pooled to avoid low expected values. *P*-values less that 0.05 indicate a significant lack-of-fit to the observed distribution. (A) *Boc*-1 transposon display markers. (B) Markers generated by the C-1 primers. (C) Markers generated by the C-2 primers.

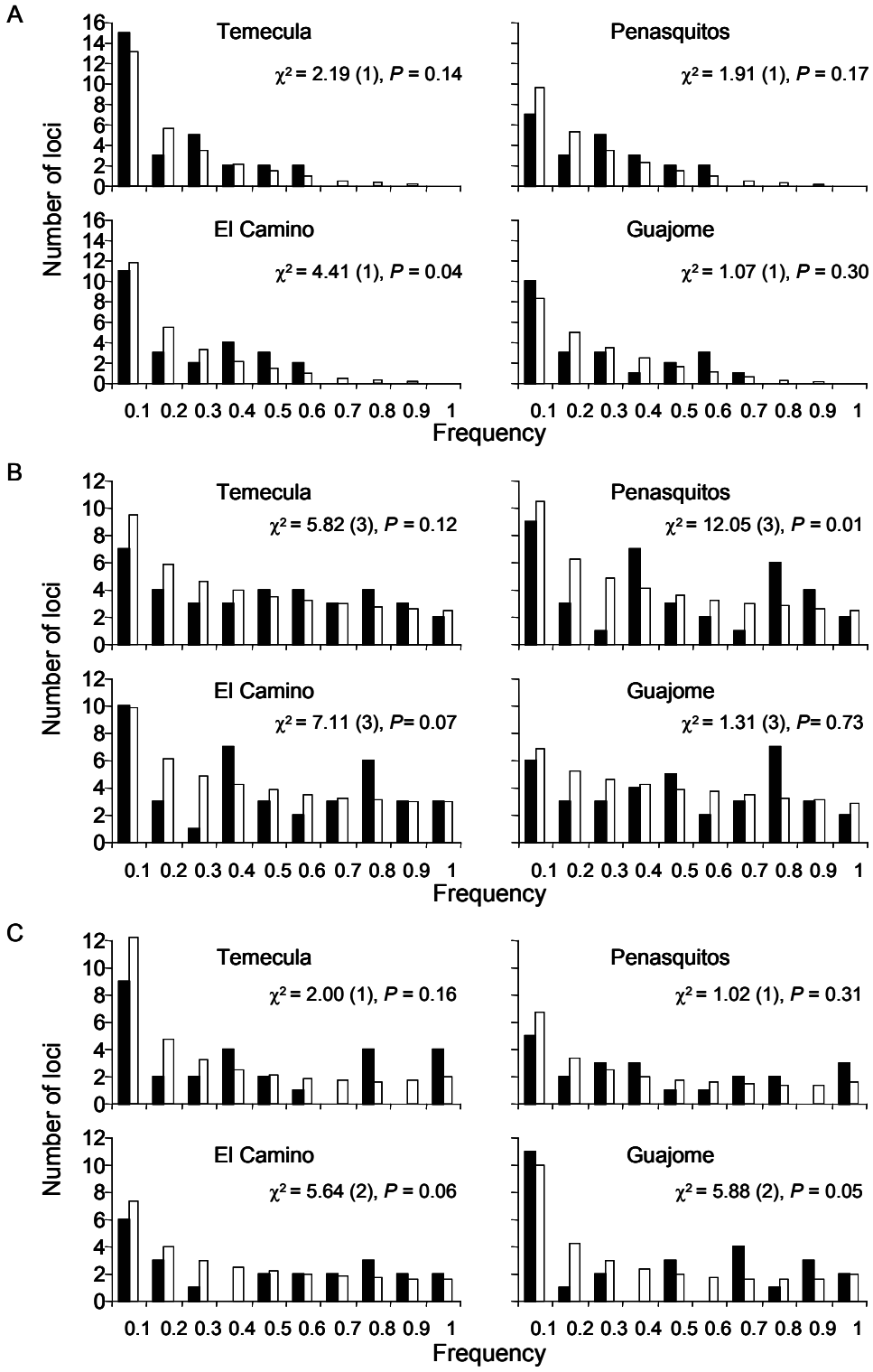

A
Temecula
χ² = 2.19 (1), P = 0.14
Penasquitos
χ² = 1.91 (1), P = 0.17
El Camino
χ² = 4.41 (1), P = 0.04
Guajome
χ² = 1.07 (1), P = 0.30
Number of loci
Frequency
B
Temecula
χ² = 5.82 (3), P = 0.12
Penasquitos
χ² = 12.05 (3), P = 0.01
El Camino
χ² = 7.11 (3), P= 0.07
Guajome
χ² = 1.31 (3), P= 0.73
Number of loci
Frequency
C
Temecula
χ² = 2.00 (1), P = 0.16
Penasquitos
χ² = 1.02 (1), P = 0.31
El Camino
χ² = 5.64 (2), P = 0.06
Guajome
χ² = 5.88 (2), P = 0.05
Number of loci
Frequency

Table 3.3. Parameters of the best-fit probability distributions of marker frequencies.

| Marker | Pop | $\alpha$ | $\beta$ | $\hat{x}$ |
|---|---|---|---|---|
| *Boc*-1 | Tem | 0.33 | 2.71 | 0.108 |
| | Pen | 0.65 | 3.16 | 0.170 |
| | El Cam | 0.44 | 2.71 | 0.139 |
| | Gua | 0.64 | 2.41 | 0.210 |
| C-1 | Tem | 0.54 | 1.04 | n/a |
| | Pen | 0.49 | 1.00 | n/a |
| | El Cam | 0.52 | 0.96 | n/a |
| | Gua | 0.74 | 1.00 | n/a |
| C-2 | Tem | 0.23 | 0.88 | n/a |
| | Pen | 0.32 | 0.83 | n/a |
| | El Cam | 0.41 | 0.95 | n/a |
| | Gua | 0.19 | 0.75 | n/a |

sequencing data that most TEs evolve as pseudogenes and therefore must be neutral with respect to fitness, it is not implausible that positive selection may act on some insertions, as it has been shown that TEs can influence the transcription of adjacent genes (*e.g* Kashkush *et al.* 2002), and that element copy number can be correlated with environmental conditions (Kalendar *et al.* 2000). This study was designed to investigate the relative contributions of selection and drift to the evolution of TEs in natural plant populations using two approaches. The first was to compare fixation indices across TE loci in order to screen for individual insertions that may be linked to loci under selection. The second was to use existing theory designed to quantify the average fitness effects of the insertions of a given family of elements based on the expected distribution of neutral alleles under the mutation-drift hypothesis (reviewed in Charlesworth and Langley 1989).

Fixation indices were computed for each polymorphic transposon and control locus based on Zhivotovsky's (1999) Bayesian estimators of allele frequencies using a non-uniform prior distribution. This approach was taken because it has been shown to give nearly unbiased estimates of heterozygosity and *F*-statistics (Zhivotovsky 1999), and to be superior to other methods in cases where polymorphism is low (Krauss 2000). There was little genetic differentiation among the four populations studied, and excellent agreement between the fixation indices calculated separately for *Boc*-1 and for the two control markers. This agreement suggests that the differentiation is due to the effects of genetic drift and biological processes acting at the population level (*e.g* Hamrick and Godt 1996), and indicates that the majority of *Boc*-1 insertions are likely to be neutral (Lewontin and Krakauer 1973). If the island model of population structure is assumed, the rate of gene flow among the populations (*Nm*) is about six migrants per generation. This seems biologically realistic because the *B. nigra* populations in southern

California often cover large areas, and many populations are interconnected by undeveloped canyons and streambeds, and by roadsides.

Since the first test proposed by Lewontin and Krakauer (1973) there has been considerable interest in using $F_{ST}$ to test for the operation of selection at, or linked to, individual loci (*e.g* Bowcock *et al.* 1991; Taylor *et al.* 1995). In principle, loci under positive selection in a subset of populations should show elevated levels of differentiation, and loci under balancing selection should show a reduction in fixation index (Porter 2003). Linkage to a locus under negative selection in some populations ("background selection", Charlesworth *et al.* 1993) should also cause an increase in $F_{ST}$. However, the LK test as originally proposed has been shown to underestimate the variance of $F_{ST}$ for both statistical and biological reasons (Nei and Maruyama 1975; Robertson 1975a; Robertson 1975b). When applied to the *B. nigra* data, it fails to detect any of the extreme outliers generated by the C-1 and C-2 primers (Figure 3.1), because the degrees of freedom for the test is determined by the number of loci, which is so large for the present data that even fixation indices near 1.0 are not found to be significant. Recently, several alternative approaches have been proposed that solve many of the problems with the LK test (*e.g* Beaumont and Nichols 1996; Bowcock *et al.* 1991; Porter 2003; Vitalis *et al.* 2001). Unfortunately, these tests either require codominant markers, or additional parameters, such as independent estimation of the relationships among populations, for which data is not currently available. Nevertheless, none of the *Boc*-1 insertions had fixation indices outside of the range predicted by the 80 presumably neutral control marker loci (Figure 3.1), and it is therefore unlikely that any of the insertions experienced selection differentially the four populations studied.

In theory developed for analyzing the population genetics of *Drosophila* TEs, the observed distribution of insertion frequencies is fit to a $\beta$-distribution. A skew in the distribution towards low frequency alleles is used as evidence of selection against TE insertions. The estimates of the parameter $\beta$ are used to quantify the strength of the forces removing elements from the population, with stronger forces resulting in greater estimates of $\beta$ (reviewed in Charlesworth and Langley 1989). The forces can be in the form of selection against the deleterious effects of insertion into a given genomic site, ectopic recombination at that site (Langley,1988), or, in the case of DNA transposons, from the excision of elements during transposition. Distinguishing between these possibilities is not possible when only insertion frequency data is available.

The insertions of *Boc*-1 transposons at 32 loci in four populations rarely reach high frequencies, in contrast to the anonymous markers generated by the C-1 and C-2 primers, many of which segregate at frequencies greater than 0.8 (Figure 3.2). The lack of high frequency *Boc*-1 insertions therefore cannot be explained by demographic factors, or by variability in the *Mse*I restriction sites used to generate the polymorphisms, because the control reactions used identical templates. Consistent with the neutrality of the control markers, $\beta$ estimates for both are close to one in all of the populations (Charlesworth and Langley 1989, Table 3.3). For *Boc*-1, the average estimate of $\beta$ across the four populations studied was 2.75. This value is relatively low, suggesting that the forces removing these elements from the populations are weak. For example, $\beta$ estimates for *Drosophila* transposons range from 550 for the element *roo* on the X chromosome (Charlesworth and Langley 1989) to 4.0 for the element *2161* in a natural population from Maryland (Charlesworth *et al.* 1992), with the typical value falling in the range of 5 to 40. In plants, estimates are available for the Class II element *Ac*-III in *Arabidopsis*

*thaliana* ($\beta = 0.7$) and *Arabis lyrata* ($\beta = 9.7$), with the difference in values being attributed to the different mating systems, selfing and outcrossing, respectively, of these species (Wright *et al.* 2001).

The analysis of the fitness effects of *Boc*-1 insertions based on the comparison of $F_{ST}$ values across loci indicates that they are likely to be neutral, while analysis of the distributions of insertion frequencies indicates that some force or forces must be acting to prevent insertions from drifting to high frequency. While apparently in conflict, these observations can be reconciled by considering the biology of CACTA-like DNA transposons. One explanation could be that the *Boc*-1 insertions are typically neutral, but excise from their chromosomal sites at a rate that is sufficient to prevent any one site from drifting to fixation. Alternatively, it is known that nonautonomous *dSpm* elements in maize carry signals that allow them to be spliced out of the RNAs of genes into which they have inserted. These insertions therefore minimize their impact on host fitness, but only when proteins from autonomous elements are not present (reviewed in Kunze and Weil 2002). When present, these proteins prevent transcription of host genes carrying nonautonomous elements. It is conceivable that some of the *Boc*-1 insertions studied were anchored in nonautonomous elements, and that these may remain neutral until an autonomous element is expressed elsewhere in the genome. In such a scenario, both the autonomous and nonautonomous elements would be selected against upon the expression of *Boc*-1 proteins. This could explain how apparently neutral insertions never drift to high frequency, as they may be only episodically deleterious.

ACKNOLEDGEMENTS

We thank Edward G. Kentner for help in the field, Ryan Peeler for help in the laboratory, and Russell Malmburg, Daniel Promislow, and Jim Hamrick for help with the population genetics theory. EKK was supported by a training grant from the National Science Foundation (DBI 9602223) which funded this work.

LITERATURE CITED

Bateman, A. J., 1955 Self-incompatibility systems in angiosperms. III. *Cruciferae*. Heredity **9:** 52-68.

Beaumont, M. A., and R. A. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society of London Series B-Biological Sciences **263:** 1619-1626.

Bennetzen, J. L., 2002 Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica **115:** 29-36.

Berenyi, M., S. T. Gichuki, J. Schmidt and K. Burg, 2002 Ty1-copia retrotransposon-based S-SAP (sequence-specific amplified polymorphism) for genetic analysis of sweetpotato. Theoretical and Applied Genetics **105:** 862-869.

Biemont, C., 1992 Population-Genetics of Transposable DNA Elements - a Drosophila Point-of-View. Genetica **86:** 67-84.

Biemont, C., A. Tsitrone, C. Vieira and C. Hoogland, 1997 Transposable element distribution in Drosophila. Genetics **147:** 1997-1999.

Boissinot, S., A. Entezam and A. V. Furano, 2001 Selection against deleterious LINE-1-containing loci in the human lineage. Molecular Biology and Evolution **18:** 926-935.

Bowcock, A. M., J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto *et al.*, 1991 Drift, Admixture, and Selection in Human-Evolution - a Study with DNA Polymorphisms. Proceedings of the National Academy of Sciences of the United States of America **88:** 839-843.

Casa, A. M., C. Brouwer, A. Nagel, L. Wang, Q. Zhang *et al.*, 2000 The Mite family *Heartbreaker*: Molecular markers in maize. Proceedings of the National Academy of Sciences of the USA **97:** 10083-10089.

Cavalli-Sforza, L., 1966 Population structure and human evolution. Proceedings of the Royal Society of London Series B **164:** 362-379.

Chakravarti, A., P. A. Fuerst and M. Nei, 1980 Statistical studies of protein polymorphism in natural populations. III. Distribution of allele frequencies and the number of alleles per locus. Genetics **94:** 1039-1063.

Charlesworth, B., and D. Charlesworth, 1983 The Population-Dynamics of Transposable Elements. Genetical Research **42:** 1-27.

Charlesworth, B., and C. H. Langley, 1989 The population genetics of *Drosophila* transposable elements. Annual Review of Genetics **23:** 251-287.

Charlesworth, B., C. H. Langley and P. D. Sniegowski, 1997 Transposable element distributions in Drosophila. Genetics **147:** 1993-1995.

Charlesworth, B., A. Lapid and D. Canada, 1992 The Distribution of Transposable Elements within and between Chromosomes in a Population of Drosophila-Melanogaster .1. Element Frequencies and Distribution. Genetical Research **60:** 103-114.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The Effect of Deleterious Mutations on Neutral Molecular Variation. Genetics **134:** 1289-1303.

Conner, J. K., and S. Rush, 1997 Measurements of selection on floral traits in black mustard, Brassica nigra. Journal of Evolutionary Biology **10:** 327-335.

Cook, J. M., and M. Tristem, 1997 'SINEs of the times' - transposable elements as clade markers for their hosts. Trends in Ecology and Evolution **12:** 295-297.

Crow, J. F., and M. Kimura, 1970 *An introduction to population genetics theory*. Harper and Row, New York.

Doolittle, W. F., and C. Sapienza, 1980 Selfish genes, the phenotype paradigm and genome evolution. Nature **284:** 601-603.

Doyle, J. J., and J. L. Doyle, 1987 A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin **19:** 11-15.

Eickbush, T. H., and A. V. Furano, 2002 Fruit flies and humans respond differently to retrotransposons. Current Opinion in Genetics & Development **12:** 669-674.

Ellis, T. H. N., S. J. Poyser, M. R. Knox, A. V. Vershinin and M. J. Ambrose, 1998 Polymorphism of insertion sites of *Ty1-copia* class retrotransposons and its use for linkage and diversity analysis in pea. Molecular and General Genetics **260:** 9-19.

Feschotte, C., J. Jiang and S. R. Wessler, 2002 Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics **3:** 329-341.

Flavell, A., E. Dunbar, R. Anderson, S. Pearce, R. Hartley *et al.*, 1992 *Ty1-copia* group retrotransposons are ubiquitous and heterogeneous in higher plants. Nucleic Acids Research **20:** 3639-3644.

Flavell, R. B., M. D. Bennett, J. B. Smith and D. B. Smith, 1974 Genome Size and Proportion of Repeated Nucleotide-Sequence DNA in Plants. Biochemical Genetics **12:** 257-269.

Fu, H. H., and H. K. Dooner, 2002 Intraspecific violation of genetic colinearity and its implications in maize. Proceedings of the National Academy of Sciences of the United States of America **99:** 9573-9578.

Gribbon, B. M., S. R. Pearce, R. Kalendar, A. H. Shulman, L. Paulin *et al.*, 1999 Phylogeny and transpositional activity of Ty1-copia group retrotransposons in cereal genomes. Molecular and General Genetics **261:** 883-891.

Hamrick, J. L., and M. J. W. Godt, 1996 Effects of life history traits on genetic diversity in plant species. Philosophical Transactions of the Royal Society of London Series B **351:** 1291-1298.

Hey, J., 1989 The transposable portion of the genome of *Drosophila algonquin* is very different from that in *D. melanogaster*. Molecular Biology and Evolution **6:** 66-79.

Joseph, J. L., J. W. Sentry and D. R. Smyth, 1990 Interspecies distribution of abundant DNA sequences in *Lilium*. Journal of Molecular Evolution **30:** 146-154.

Kalendar, R., J. Tanskanen, S. Immonen, E. Nevo and A. H. Schulman, 2000 Genome evolution of wild barley (Hordeum spontaneum) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proceedings of the National Academy of Sciences of the USA **97:** 6603-6607.

Kashkush, K., M. Feldman and A. A. Levy, 2002 Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics **160:** 1651-1659.

Kidwell, M. G., 2002 Transposable elements and the evolution of genome size in eukaryotes. Genetica **115:** 49-63.

Kidwell, M. G., and D. R. Lisch, 2001 Perspective: Transposable elements, parasitic DNA, and genome evolution. Evolution **55:** 1-24.

Krauss, S. L., 2000 Accurate gene diversity estimates from amplified fragment length polymorphism (AFLP) markers. Molecular Ecology **9:** 1241-1245.

Kumar, A., and J. L. Bennetzen, 1999 Plant retrotransposons. Annual Review of Genetics **33:** 479-532.

Kunze, R., and C. Weil, 2002 The *h*AT and CACTA superfamilies of plant transposons, pp. 565-610 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. American Society for Microbiology Press, Washington, D.C.

Langley, C. H., J. F. Y. Brookfield and N. Kaplan, 1983 Transposable Elements in Mendelian Populations .1. A Theory. Genetics **104:** 457-471.

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74:** 175-195.

McClintock, B., 1984 The significance of responses of the genome to challenge. Science **226:** 792-801.

McDonald, J. F., 1998 Transposable elements, gene silencing and macroevolution. Tree **13:** 95-95.

Nei, M., and T. Maruyama, 1975 Lewontin-Krakauer test for neutral genes. Genetics **80:** 395.

Orgel, L. E., and F. H. C. Crick, 1980 Selfish DNA: the ultimate parasite. Nature **284:** 604-607.

Pearce, S., G. Harrison, L. Dongtao, J. S. Heslop-Harrison, A. Kumar *et al.*, 1996 The *Ty1-copia* group retrotransposons in *Vicia* species: copy number, sequence heterogeneity and chromosomal localisation. Molecular and General Genetics **250:** 305-315.

Porter, A. H., 2003 A test for deviation from island-model population structure. Molecular Ecology **12:** 903-915.

Robertson, A., 1975a Gene Frequency Distributions as a Test of Selective Neutrality. Genetics **81:** 775-785.

Robertson, A., 1975b Remarks on the Lewontin-Krakauer test. Genetics **80:** 396.

Roy-Engel, A. M., M. L. Carroll, E. Vogel, R. K. Garber, S. V. Nguyen *et al.*, 2001 Alu insertion polymorphisms for the study of human genomic diversity. Genetics **159:** 279-290.

San Miguel, P., A. Tikhonov, J. Young-Kwan, N. Motchoulakaia, P. S. Springer *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765-738.

Tatout, C., S. Warwick, A. Lenoir and J. M. Deragon, 1999 SINE insertions as clade markers for wild crucifer species. Molecular Biology and Evolution **16:** 1614-1621.

Taylor, M. F. J., Y. Shen and M. E. Kreitman, 1995 A Population Genetic Test of Selection at the Molecular-Level. Science **270:** 1497-1499.

U, N., 1935 Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. Japanese Journal of Botany **7:** 389-452.

Van den Broeck, D., T. Maes, M. Saur, J. Zetof, P. De Keukeleire *et al.*, 1998 *Transposon Display* identifies individual transposable elements in high copy number lines. The Plant Journal **13:** 121-129.

Vitalis, R., K. Dawson and P. Boursot, 2001 Interpretation of variation across marker loci as evidence of selection. Genetics **158:** 1811-1823.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee *et al.*, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research **23:** 4407-4414.

Waugh, R., K. McLean, A. J. Flavell, S. R. Pearce, A. Kumar *et al.*, 1997 Genetic distribution of BARE-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Molecular and General Genetics **253:** 687-694.

Weir, B. S., 1996 *Genetic data analysis II*. Sinauer Associates, Inc., Sunderland, MA.

Wright, S., 1937 The distribution of gene frequencies in populations. Proceedings of the National Academy of Sciences of the United States of America **23:** 307-320.

Wright, S. I., Q. H. Le, D. J. Schoen and T. E. Bureau, 2001 Population dynamics of an *Ac*-like transposable element in self- and cross-pollinating *Arabidopsis*. Genetics **158:** 1279-1288.

Zhang, X. Y., and S. R. Wessler, 2004 Genome-wide comparative analysis of transposable elements in the related speces *Arabidopsis thaliana* and *Brassica oleracea*. submitted.

Zhivotovsky, L. A., 1999 Estimating population structure in diploids with multilocus dominant DNA markers. Molecular Ecology **8:** 907-913.

# CHAPTER 4

# CONCLUSIONS

The research described in this dissertation was initiated in an effort to better understand the impact of transposable elements (TEs) on plant genome evolution. The motivation for much of this work was curiosity regarding the nature of polymorphisms generated by the activity of TEs. Very little information regarding the among plant variation in TE insertions was available when this research was initiated, and to date, only a few studies have addressed this question (*e.g.* Fu and Dooner 2002; Kalendar *et al.* 2000; Wright *et al.* 2001). Although TEs often compose the majority of DNA in plant genomes (Bennetzen 2002), basic questions concerning their dynamics remain unanswered. For example, it is known that stressful conditions can promote the activity of LTR retrotransposons (Grandbastien 1998; Wessler 1996), but the extent to which these elements move or cause mutations in natural plant populations has not been thoroughly investigated. Interestingly, the same chromosomal region in two maize lines contains completely different blocks of TEs, and even differs in gene content (Fu and Dooner 2002). Such unexpected results demonstrate the value of documenting the extent of the genetic variation due to the activities of TEs in plants. The study described in Chapter 2 has resulted in a retrotransposon-based marker system for *Iris* species that could be used to address this question, while Chapter 3 provides the first population-genetic data for CACTA-like DNA transposons.

The work on the Louisiana Iris species *I. brevicaulis* and *I. fulva*, describes a high copy group of Ty3/*gypsy* LTR retrotransposons and their use as the basis of a multilocus molecular

marker system (Kentner *et al.* 2003). These IRRE elements belong to an ancient clade of LTR retrotransposons that is characterized by the presence of a putative chromatin binding domain lacking in other retrotransposons (Malik and Eickbush 1999; Marin and Llorens 2000). Although the conservation of this domain in elements from monocots, dicots and gymnosperms suggests its importance to the life cycle of these elements, its function has not been elucidated. Similarly, the forces responsible for the variability documented among IRRE elements in the promoter region of the LTR remain subject to speculation. It could be that the variation is the result of a molecular arms race between the elements and silencing mechanisms present in the host genome (*e.g.* Hirochika *et al.* 2000). Interestingly, similar promoter variation among subfamilies of the tobacco retrotransposon Tnt1 is correlated with the differential expression of element subfamilies in response to specific stress-induced cues (Beguiristain *et al.* 2001). Accounting for nearly 10% of chromosomal DNA, IRRE elements have clearly been successful in their proliferation within the Louisiana Iris genome. The details of how this was accomplished and the consequences of this proliferation to *Iris* plants remain potential topics of further research.

Hybridization between *I. fulva* and *I. brevicaulis* does not appear to result in the increased activity of IRRE retrotransposons, at least not on a scale similar to the dramatic increase that was observed in wallaby hybrids (Waugh O'Neil *et al.* 1998). However, the size of the *Iris* genome and the high copy number of the IRRE elements made the testing of activation in hybrids extremely difficult, and subtle changes in element activity would not have been detected with the techniques employed. In retrospect, the Louisiana Irises were not a good choice of system for studying the potential activation of TEs by interspecific hybridization in plants because of problems associated with their genome size, and because hybrid *Iris* do not display any

phenotypes or reduction in fitness that might suggest the activation of TEs (Burke *et al.* 1998). However, a beneficial outcome of this research has been the development of transposon display markers anchored in IRRE retrotransposon insertions. These markers have been used to create linkage maps for both *I fulva* and *I. brevicaulis* (A. Bouck unpublished data), and have many potential applications for studying the biology of these species and the polymorphism of retrotransposon insertion sites in natural populations.

The research on *Brassica* TEs described in Chapter 3 sought to evaluate the fitness effects of TE insertions in natural populations using two approaches. The first was to compare the fixation indexes across loci to screen for insertions that may be under selection (*e.g.* Lewontin and Krakauer 1973). The second was to use population genetic theory developed for the study of *Drosophila* TEs that estimates the average fitness effects of a family of TEs, while taking into account the effects of genetic drift on insertion frequencies (Charlesworth and Charlesworth 1983). Although the project was designed to make comparisons among three families of elements, only the primers for the CACTA-like DNA transposon *Boc*-1 produced markers anchored in TE insertions. The *Boc*-1 insertions were highly polymorphic among individuals sampled from four populations. Comparisons of fixations indexes at *Boc*-1 loci to those of a large number of presumably neutral markers revealed no insertions likely to be under differential selection among the populations. There was evidence that insertions are removed from populations, as no *B*oc-1 loci occurred at high frequency, but the magnitude of the forces removing the insertions was less than typically observed for *Drosophila* elements (Charlesworth and Langley 1989). Two explanations are offered for these apparently conflicting results. If elements remain capable of excision, this could prevent even neutral insertions from drifting to high frequency given a sufficiently high excision rate. Alternatively, it is known that

nonautonomous *dSpm* elements in maize carry signals that allow them to be spliced out of the RNAs of genes into which they have inserted. These insertions therefore minimize their impact on host fitness, but only when proteins from autonomous elements are not present (reviewed in Kunze and Weil 2002). If present, these proteins prevent the transcription of host genes carrying nonautonomous elements. It is conceivable that some of the *Boc*-1 insertions studied were anchored in nonautonomous elements, and that these may remain neutral until an autonomous element is expressed elsewhere in the genome. In such a scenario, both the autonomous and nonautonomous elements would be selected against upon the expression of *Boc*-1 proteins. This could explain how apparently neutral insertions never drift to high frequency

**Prospects for future studies:** During the period of time that elapsed while the work on this dissertation was being completed, the sequencing of the *Drosophila melanogaster*, *Caenorhabditis elegans*, *Homo sapiens*, *Arabidopsis thaliana*, and *Oryza sativa* genomes was completed. These sequencing projects have completely changed the manner in which TE research is conducted, as it is no longer necessary to devise genetic screens to identify active transposons. Instead, putatively active families can be identified by their high sequence similarity using database searches, and PCR primers allowing their study can be designed literally in a matter of hours or minutes. This approach has lead to the identification of the first active family of MITEs (Jiang *et al.* 2003) and has also identified active Alu elements in humans (Batzer *et al.* 1996; Roy-Engel *et al.* 2001). The power and speed of this approach is compelling.

While the power of whole-genome sequence data has already revolutionized the study of transposable elements, the sequence of a single individual of a species cannot address questions related to the polymorphism of TEs. However, database information can greatly facilitate the

collection of population data for species with sequenced genomes and even for related species, as illustrated by the *Brassica* work described in this dissertation. In this case, low coverage, unassembled shotgun sequence was sufficient to identify putatively active TEs in *Brassica oleracea.* As more of these genomic survey sequences (GSS) come on line, comparative analyses of diverse types of TEs in many plants will be possible and practical, and should lead to major advances in our understanding of the evolutionary forces governing their behavior.

## LITERATURE CITED


Batzer, M. A., S. S. Arcot, J. W. Phinney, M. Alegria-Hartman, D. H. Kass *et al.*, 1996 Genetic variation of recent Alu insertions in human populations. Journal of Molecular Evolution **42:** 22-29.

Beguiristain, T., M. A. Grandbastien, P. Puigdomenech and J. M. Casacuberta, 2001 ThreeTnt1 subfamilies show different stress-associated patterns of expression in tobacco. Consequences for retrotransposon control and evolution in plants. Plant Physiology **127:** 212-221.

Bennetzen, J. L., 2002 Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica **115:** 29-36.

Burke, J. M., S. E. Carney and M. L. Arnold, 1998 Hybrid fitness in the Louisiana irises: Analysis of parental and F1 performance. Evolution **52:** 37-43.

Charlesworth, B., and D. Charlesworth, 1983 The Population-Dynamics of Transposable Elements. Genetical Research **42:** 1-27.

Charlesworth, B., and C. H. Langley, 1989 The population genetics of *Drosophila* transposable elements. Annual Review of Genetics **23:** 251-287.

Fu, H. H., and H. K. Dooner, 2002 Intraspecific violation of genetic colinearity and its implications in maize. Proceedings of the National Academy of Sciences of the United States of America **99:** 9573-9578.

Grandbastien, M.-A., 1998 Activation of plant retrotransposons under stress conditions. Trends in Plant Science **3:** 181-187.

Hirochika, H., H. Okamoto and T. Kakutani, 2000 Silencing of retrotransposons in *Arabidopsis* and reactivation by the ddm1 mutation. The Plant Cell **12:** 357-368.

Jiang, N., Z. R. Bao, X. Y. Zhang, H. Hirochika, S. R. Eddy *et al.*, 2003 An active DNA transposon family in rice. Nature **421:** 163-167.

Kalendar, R., J. Tanskanen, S. Immonen, E. Nevo and A. H. Schulman, 2000 Genome evolution of wild barley (Hordeum spontaneum) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proceedings of the National Academy of Sciences of the USA **97:** 6603-6607.

Kentner, E. K., M. L. Arnold and S. R. Wessler, 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.

Kunze, R., and C. Weil, 2002 The *h*AT and CACTA superfamilies of plant transposons, pp. 565-610 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. American Society for Microbiology Press, Washington, D.C.

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74:** 175-195.

Malik, H. S., and T. H. Eickbush, 1999 Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. Journal of Virology **73:** 5186-5190.

Marin, I., and C. Llorens, 2000 Ty3/Gypsy retrotransposons: Description of new Arabidopsis thaliana elements and evolutionary perspectives derived from comparative genomic data. Molecular Biology and Evolution **17:** 1040-1049.

Roy-Engel, A. M., M. L. Carroll, E. Vogel, R. K. Garber, S. V. Nguyen *et al.*, 2001 Alu insertion polymorphisms for the study of human genomic diversity. Genetics **159:** 279-290.

Waugh O'Neil, R. J., M. J. O'Neil and J. A. Marshall Graves, 1998 Undermethylation associated with retroelement activation and chromosome remodeling in a interspecific mammalian hybrid. Nature **393:** 68-72.

Wessler, S. R., 1996 Plant retrotransposons: Turned on by stress. Current Biology **6:** 959-961.

Wright, S. I., Q. H. Le, D. J. Schoen and T. E. Bureau, 2001 Population dynamics of an *Ac*-like transposable element in self- and cross-pollinating *Arabidopsis*. Genetics **158:** 1279-1288.

# APPENDIX A

# IRRE RETROTRANSPOSONS IN THE PACIFIC COAST IRISES

## INTRODUCTION

The Pacific Coast Irises (*Iris* subgenus *Limniris*, section *Limniris*, series *Californicae*) are a group of 11 closely related species native to California Oregon, and Washington (Lenz 1958; Wilson 1998; Young 1998). Numerous hybrid zones between the members of this species complex have been described (Clarkson 1959; Lenz 1959; Young 1996), and the group has been used as a classic example of the concept of the syngameon (Grant 1971). Although reproductive isolation was once more complete among the members of the group, many species have come into secondary contact as the result of logging, road and powerline construction, grazing, mining and fire (Lenz 1959). However, the species are adapted to divergent ecological niches, and display divergent floral morphology (Lenz 1958) possibly reflecting their adaptations to different pollinators.

The members of the genus *Iris* typically have very large genomes (Bennett *et al.* 1998; Kentner *et al.* 2003). In the Louisiana Iris a group of high copy TY3/*gypsy* retrotransposons has been characterized that accounts for ~6-10% of the genome of these species (Kentner *et al.* 2003). These IRRE elements were used to develop a multilocus molecular marker system known as transposon display or S-SAP (Van den Broeck *et al.* 1998; Waugh *et al.* 1997) for these species. In *I. fulva* and *I. brevicaulis*, several subfamilies of IRRE elements were described using the sequence variation present in the U3 promoter region of the long terminal repeat

(LTR). Since the U3 domain occurs near the end of the LTR and is divergent among IRRE subfamilies, it was possible to design transposon display primers that are specific to the insertions of these subfamilies (Kentner *et al.* 2003).

The work described here was undertaken with the objective of using the insertions of IRRE retrotransposons as genetic markers to aid in the study of hybridization and speciation in the Pacific Coast Irises. To this end IRRE LTRs were cloned and sequenced from the genomes of *I. douglasiana*, *I. bractiata*, and *I. inominata*. Two major groups of IRRE elements were found in the genomes of these species. One group is closely related to the IRRE elements found in the Louisiana Irises, and the other group was found only in the Pacific Coast Iris. Subfamilies within each of the two major groups were defined by phylogenetic analysis and transposon display primers specific to several of these subfamilies were designed and tested.

## MATERIALS AND METHODS

**Plant materials:** Individuals of *Iris douglasiana*, *I. bracteata* and *I. inominata* were collected from natural populations in California using the site information provided by Lenz (1958). For all species, young leaf tips were collected directly into silica gel-containing vials, desiccated, and stored at room temperature.

**Nucleic acid analysis:** Nucleic acids were extracted using a standard CTAB protocol and the resulting extracts treated with RNase A for one hour at room temperature. Fragments of IRRE LTR retrotransposons were amplified by PCR using the LTRSCREEN primers described in Kentner *et al* (2003). The resulting fragments were cloned using the TopoTA cloning kit (Invitrogen, Carlsbad, CA) using the manufacturers instructions. All sequencing was performed using the Big Dye terminator sequencing kit (Perkin Elmer/Applied Biosystems, Foster City, CA) on an ABI 377 automated DNA sequencer (Perkin Elmer/Applied Biosystems, Foster City,

CA ). Transposon display was carried out as described (Kentner *et al.* 2003) using primers designed to be specific for individual subfamilies of elements. The primers are available upon request.

**Phylogenetic analysis:** DNA sequences were aligned with the ClustalW Service at the European Bioinformatics Institute (http://www2.ebi.ac.uk/clustalw) using the default parameters, and GeneDoc (www.psc.edu/biomed/genedoc) was used to manually edit and box-shade the alignments. Neighbor-joining trees were constructed using MEGA 2.1 (http://www.megasoftware.net/) using the Kimura 2 parameter substitution model with the pairwise gap deletion option, and tested by 1000 bootstrap replications.

## RESULTS AND DISCUSSION

Fragments of IRRE retrotransposons were readily amplified from the genomes of the Pacific Coast Irises using degenerate primers designed to amplify IRRE sequences in the Louisiana Iris species (Kentner *et al.* 2003). Seventy seven of these fragments were cloned and sequenced. The fragments ranged in size from 632bp to 463bp and were aligned with 61 IRRE sequences from the Louisiana Iris species *I. fulva* and *I brevicaulis*. This alignment was the basis for the neighbor joining (NJ) tree presented in Figure A.1. Two major groups of IRRE elements in the Pacific Coast Irises were resolved. The groups can be conveniently identified by the first six base pairs of the LTR, which compose the inverted repeat at the ends of the LTRs (Kentner *et al.* 2003). In the IRRE elements from the Louisiana Iris, and the group of Pacific Coast Iris IRRE most closely related to them, these six bases are TGTCACT. A second major group of IRRE elements from the Pacific Coast Iris is defined by LTRs that begin with the sequence TGTGAGA. This group was given the designation IRRE3.

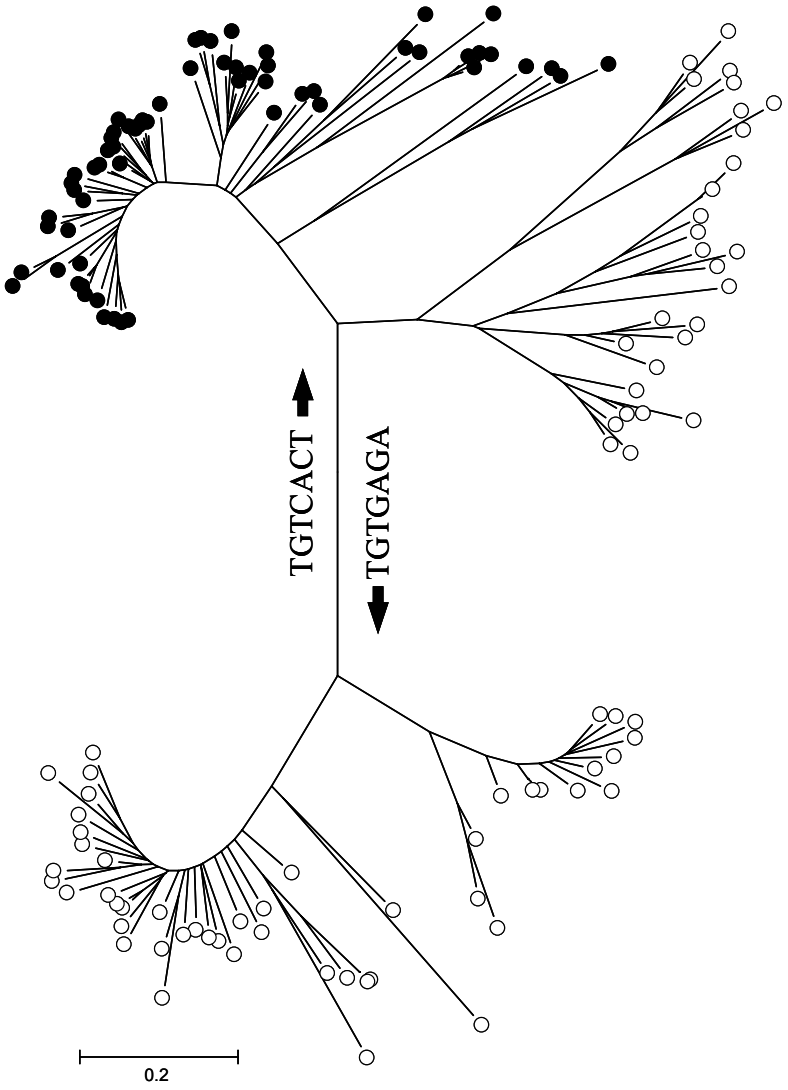



Figure A.1. Neighbor joining tree of Louisiana and Pacific Coast Iris IRRE retrotransposon fragments. Open and black circles represent Pacific Coast Iris, and Louisiana Iris sequences, respectively. The presented first six bases of the LTR are diagnostic of the two major groups resolved by the tree. The scale bar represents genetic distance, see methods.

The LTRs of IRRE elements are highly variable and contain many small insertion/deletion polymorphisms that make the alignment of sequences difficult and subject to uncertainty when members of divergent lineages are included in the same alignment. The NJ tree presented in Figure A.2 is based on an alignment containing only Pacific Coast Iris sequences and better represents the relationships among the IRRE elements found in these plants. Subgroups were given somewhat arbitrary names and used as the basis for the design of transposon display primers specific to each group.

In spite of numerous attempts and much troubleshooting, the transposon display reactions produced disappointing results in the Pacific Coast Irises. Several sets of primers specific for each subfamily shown in Figure A.2 invariably displayed relatively few bands on top of a background smear (ugly data not shown). Altering the number of selective bases failed to influence the number of bands resolved. However, several bands were cloned and confirmed to be anchored in IRRE insertions (Figure A.3). The sequence of the cloned bands confirmed that the IRRE LTR ends had been properly identified, and that the transposon display primers produced bands anchored in TE insertions. However, the reactions were judged to be too unreliable to score polymorphisms in natural populations. The reason for the poor performance of the transposon display in these species was not determined. It could be that the standard *Eco*RI adapters used in these experiments anneal to some unknown repetitive sequence in the Pacific Coast Iris genome and foul the reactions. However, experiments with different adaptors and/or enzymes were not attempted.

Figure A.2. Neighbor joining tree of Pacific Coast Iris IRRE retrotransposon fragments. Numbers to the left of nodes represent percentage support in 1000 bootstrap replicates, with values < 60 not displayed. Names of phylogenetic groups for which transposon display primers were designed are to the right of the figure. Not all groups were given names. The scale bar represents genetic distance, see methods.

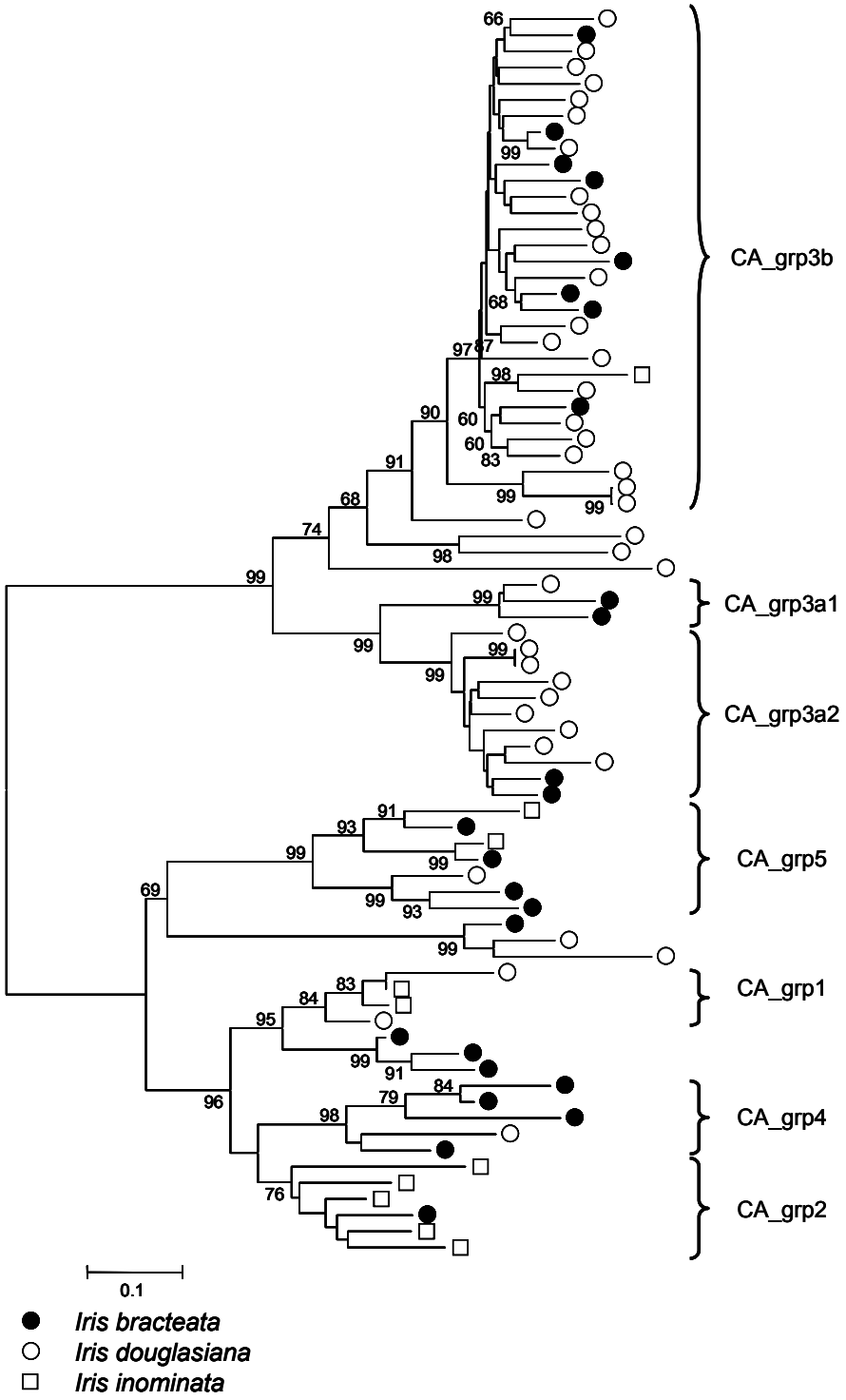
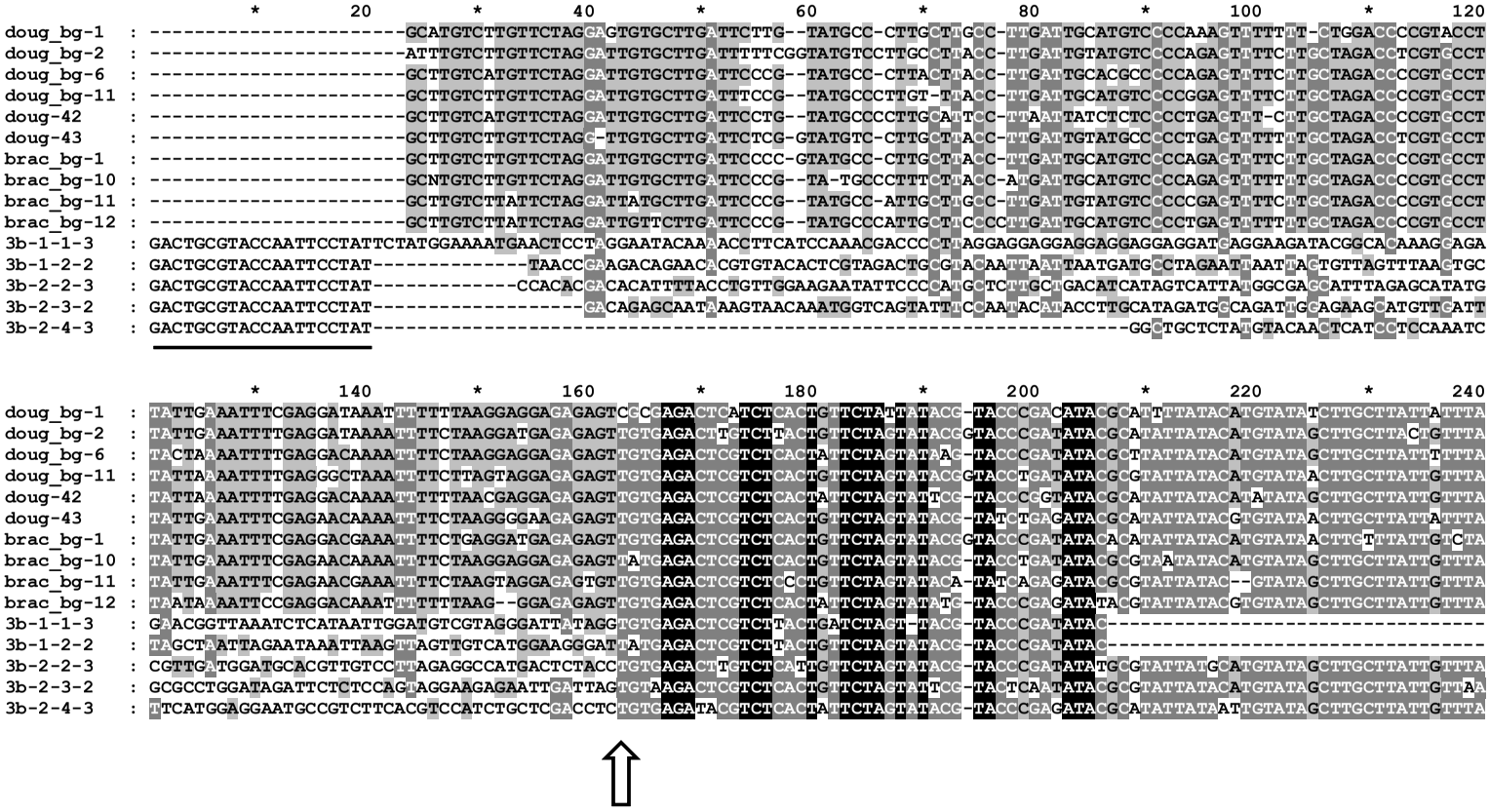

CA_grp3b
CA_grp3a1
CA_grp3a2
CA_grp5
CA_grp1
CA_grp4
CA_grp2
0.1
Iris bracteata
Iris douglasiana
Iris inominata

Figure A.3. Alignment of IRRE transposon display PCR products and some representative IRRE fragments from the Pacific Coast Irises. Arrow indicates the 3' end of the LTR, and the underscore identifies the *Eco*RI adapter primer sequences.


doug_bg-1
doug_bg-2
doug_bg-6
doug_bg-11
doug-42
doug-43
brac_bg-1
brac_bg-10
brac_bg-11
brac_bg-12
3b-1-1-3
3b-1-2-2
3b-2-2-3
3b-2-3-2
3b-2-4-3

LITERATURE CITED

Bennett, M. D., A. V. Cox and I. J. Leitch, 1998 Angiosperm DNA C-values database. http://www.rbgkew.org.uk/cval/database1.html

Clarkson, Q. D., 1959 Field studies of natural hybridization in the Oregon species of *Iris* L. subsection *Californicae* Diels. Madrono **15:** 115-122.

Grant, V., 1971 *Plant Speciation*. Columbia University Press, New York.

Kentner, E. K., M. L. Arnold and S. R. Wessler, 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.

Lenz, L. W., 1958 A revision of the Pacific coast irises. Aliso **4:** 1-71.

Lenz, L. W., 1959 Hybridization and speciation in the Pacific coast irises. Aliso **4:** 237-309.

Van den Broeck, D., T. Maes, M. Saur, J. Zetof, P. De Keukeleire *et al.*, 1998 *Transposon Display* identifies individual transposable elements in high copy number lines. The Plant Journal **13:** 121-129.

Waugh, R., K. McLean, A. J. Flavell, S. R. Pearce, A. Kumar *et al.*, 1997 Genetic distribution of BARE-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Molecular and General Genetics **253:** 687-694.

Wilson, C. A., 1998 A cladistic analysis of *Iris* series *Californicae* based on morphological data. Systematic Botany **23:** 73-88.

Young, N. D., 1996 Concordance and discordance: A tale of two hybrid zones in the Pacific coast irises (Iridaceae). American Journal of Botany **83:** 1623-1629.

Young, N. D., 1998 Pacific coast *Iris* species delimitation using three species definitions: biological, phylogenetic, and genealogical. Biological Journal of the Linnean Society **63:** 99-120.