In compliance with the
Canadian Privacy Legislation
some supporting forms
may have been removed from
this dissertation.

While these forms may be included
in the document page count,
their removal does not represent
any loss of content from the dissertation.

Diversity and mobility of transposons in *Arabidopsis thaliana*

by Quang Hien Le

Department of Biology
McGill University, Montreal
March 2002

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the degree of Doctor in Philosophy

© Quang Hien Le, 2002

National Library of Canada
Bibliothèque nationale du Canada

Acquisitions and Bibliographic Services
Acquisisitons et services bibliographiques

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-85719-0
Our file Notre référence
ISBN: 0-612-85719-0

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou aturement reproduits sans son autorisation.

Canada

Kính tặng bố mẹ.

(Xong rồi! Xong rồi!)

ABSTRACT

Transposons are a diverse collection of mobile genetic elements and are important components of nearly every genome. Because of their mobile and repetitive nature, transposons can have considerable effects on host gene expression, genome organization and evolution. The recent availability of genomic sequence information has expedited the discovery and study of transposons, as examplified in this thesis by the complete genome analysis of the model plant system *Arabidopsis thaliana*. Data mining in *Arabidopsis* has revealed a rich diversity of transposons, of which *Basho* and Terminal-repeat Retrotransposons In Miniature (TRIM) elements were previously unknown types. The identification of Related to Empty Sites (RESites) provide evidence for past transposition events. Examples of elements contributing to coding regions, acquiring cellular sequences, along with in-depth analysis of the insertions, their target sites and their distribution illustrate the impact of transposons on gene and genome structures. Computer-based searches of genomic sequences has also improved our understanding of previously identified transposon families, such as the origin, classification and mobilization of *Tourist* elements. In addition, information on transposons gathered from *in silico* analysis of genomic sequences has served to design *in vivo* experiments. In a whole genome strategy, Transposon Display was used to investigate transposition and regulation of mobility of *Tourist*-like elements in *A. thaliana* and in the nematode *Caenorhabditis elegans*.

RÉSUMÉ

Les transposons sont des éléments génétiques mobiles et une composante importante de presque tout génome. Parce qu'ils ont la capacité de changer de position et à cause de leur caractère répétitif, ils peuvent modifier l'expression des gènes ainsi que l'organisation et l'évolution du génome. L'imposante quantité de séquences génomiques disponibles depuis peu dans les bases de données a permis d'accélérer et de faciliter la découverte et l'étude des transposons. De fait, cette thèse présente une analyse approfondie des séquences génomiques de la plante modèle *Arabidopsis thaliana*. Malgré sa taille relativement petite, le génome d'*Arabidopsis* renferme néanmoins une grande diversité de transposons, comprenant notamment des types d'éléments jusqu'alors inconnus tels que les *Basho* et les Terminal-repeat Retrotransposons in Miniature (TRIM). L'identification de séquences Related to Empty Sites (RESites) indique que des événements de transposition ont eu lieu dans le passé. Des exemples où des éléments ont contribué à des régions codantes ou ont acquis des séquences cellulaires, ainsi qu'un examen détaillé des insertions, de leurs séquences cibles et de leur distribution, illustre l'impact des transposons sur la structure des gènes et du génome. L'analyse des séquences a aussi permis d'améliorer la compréhension actuelle de groupes d'éléments déjà connus, par exemple en ce qui concerne la classification et le mode de mobilisation des éléments de la famille *Tourist*. De plus, les données récoltées *in silico* ont permis de lancer des expériences *in vivo*. En utilisant une approche génomique, le Transposon Display a permis d'étudier les mécanismes de transposition et de régulation de la mobilité d'éléments *Tourist* chez *Arabidopsis* et le nématode *Caenorhabditis elegans*.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EST	Expressed Sequence Tag
gi	Geninfo Identifier
IS	Insertion Sequence
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
MITE	Miniature Inverted-repeat Transposable Element
MLE	*Mariner*-Like Element
MULE	*MUtator*-Like Element
NMD	Nonsense-Mediated mRNA Decay
ORF	Open Reading Frame
PPT	Poly-Purine Tract
PBS	Primer Binding Site
PTGS	Post-Transcriptional Gene Silencing
RESite	Related to Empty Site
RNAi	RNA interference
SINE	Short Interspersed Nuclear Element
TD	Transposon Display
TDR	Terminal Direct Repeat
TIR	Terminal Inverted Repeat
TGS	Transcriptional Gene Silencing
TRIM	Terminal-repeat Retrotransposons In Miniature
TSD	Target Site Duplication
UV-B	UltraViolet-B

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor, Dr. Thomas Bureau, for his guidance throughout my project and for his financial support. His dedication to science is inspiring and a source of motivation. I would also like to thank all past and present members of his lab, including Dr. Ruying Chang for his expertise; Dr. Janet George for all the advice, critical comments and good humour; my fellow graduate students Joël Savard, Lily Yu, Matthieu Cavey, Kime Turcotte and Stephen Wright alongside whom I worked or collaborated; my colleague Nabil Elrouby for technical suggestions, cultural exchanges and even the enlightening political discussions; Nikoleta Juretic for her invaluable help with sequencing and gel preparations; and thanks to Sylva Petrova, Hala Razi Khan, Chengbin Feng, Chris Olive, Boris-Antoine Legault and Newton Agrawal for their computer know-how.

I am very grateful to the members of my supervisory committee Dr. Rebecca Kellum, Dr. Daniel Schoen and Dr. Candace Waddell for lending me research equipment but mostly for their much appreciated counsel. I am also thankful to Dr. Rajinder Dhindsa, Dr. Richard Roy, Dr. Joseph Dent, Dr. Christian Hardtke and the members of their labs for advices and research equipment. I wish to recognize Dr. Claus-Peter Witte and Dr. Amar Kumar for their collaboration on the TRIM project. I am sending my thanks to Claire Cooney, Mark Romer and Frank Scopoletti for caring for my plants and to all the supporting staff at McGill. I value the help with protocols or simply the friendship of Dr. Andy Babwah, Dr. Aaron Windsor, Glen Cordner, Aura Navarro-Quezada, Joëlle Pérusse, Dominique Cloutier, Jacqui Brinkman, Dr. Heather Gray and Dr. Franz Oman. I would like to thank in particular my long-time colleague and friend (soon to be Dr.) Julie Poupart for her invaluable comments, for our lunch-time discussions and for her vast knowledge of *Arabidopsis*.

Finally, special thanks to all my family and friends, for their patience and encouragements, and to Caroline Marie, my greatest discovery.

PREFACE

CONTRIBUTIONS OF AUTHORS

Three of the chapters presented in this thesis have been co-authored in collaboration with colleagues at McGill University or at the Scottish Crop Research Institute, Scotland. I would like to acknowledge their contribution here.

In chapter 2 (Le *et al.*, 2000), Stephen Wright, Zhihui Yu and I have contributed equally to the mining process, summarized in Table 2.1, and to the identification of RESites shown in Figure 2.1. Logistically, Stephen Wright was responsible for the analysis of *Basho* elements, Zhihui Yu for the MULEs and I for the remaining transposon groups. I used the data that we have collectively generated to make Figures 2.2 to 2.6 inclusively. I would like to recognize the assistance of Chengbin Feng, Hala Razi Khan, Sylva Petrova and Chris Olive in building and maintaining the transposon database and for writing the Perl program for calculating AT-richness of insertion sites. The text submitted to *The Proceedings of the National Academy of Science USA* was prepared by Stephen and myself.

In chapter 3 (Witte *et al.*, 2001), Dr. Claus-Peter Witte and I have contributed equally to the work. I was responsible for the mining and analysis of all the TRIM elements within *Arabidopsis* whereas Claus-Peter Witte examined TRIMs in other plants.

In chapter 4 (Le *et al.*, 2001), phylogenetic analysis of the IS5/*Tourist* transposases was performed by Kime Turcotte.

References

LE, Q.H., WRIGHT, S., YU, Z. AND BUREAU, T. (2000). Transposon diversity in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **97,** 7376-7381.

LE, Q.H., TURCOTTE, K. AND BUREAU, T. (2001). Tc8, a *Tourist*-like transposon in *Caenorhabditis elegans*. *Genetics* **158,** 1081-1088.

WITTE, C.-P., LE, Q.H., BUREAU, T. AND KUMAR, A. (2001). Terminal-repeat Retrotransposons in Miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl. Acad. Sci. USA* **98,** 13778-13783.

INTRODUCTION

The genome can be compared to a book. The words in this book are genes that, when put together, tell a story. The story told by a genome is about the making of an organism. Transposons, like the genes in a book, can also be seen as words, but words that have the ability to multiply and change location in the text. Most of the time, inserting or moving a word will not impact much on the story but, in rare cases, it can change what the story was meant to tell.

This thesis is about the transposons in a book entitled *Arabidopsis thaliana* but that is written in a coded language. In the past decade, sequencing efforts have slowly deciphered and made accessible a growing number of these books. *Arabidopsis*, a small plant and model system, has a relatively small genome and is now one of the few eukaryotes completely sequenced. As a genome is being sequenced, all of the transposons that it contains are also being revealed and only await to be found and examined.

Thus, in the first phase of my project, I have taken advantage of the sequences available in public databases and, using computer-assisted methods, I have identified all the transposons contained within the genomic sequences of *Arabidopsis*. In this manner, I could acquire information on the positions and sequences of the full complement of transposons endogenous to *Arabidopsis*. The sequence data gathered would allow a glance at the origin and evolution of transposons and would contribute to the understanding of how elements have helped shape the genome.

Among the many transposons identified, each with their individual tales, the survey revealed additional data for Miniature Inverted-repeat Transposable Elements (MITEs), that helped elucidate how these elements might move. Some of the newly identified MITEs could potentially be mobile but are normally kept silent. Because transposon insertions can disrupt cellular functions, transposition is tightly regulated under ordinary circumstances, either by host- or by transposon-mediated mechanisms. In the second phase of my project, I focused on the questions regarding MITE activity. I also wanted to

test whether these elements could be stimulated by stress factors such as transformation, protoplast isolation and UV-B irradiation - all of which were previously reported to activate transposons in many systems. By exploiting the mutagenic character of transposons, stress-induced transposition is thought to be a way of promoting genetic variability within host organisms to help populations adapt to changing environments. The disruptive nature and the lack of obvious benefit to the host initially earned transposons the qualifiers of selfish and parasitic elements but another picture can be drawn from the mobile nature of transposons: a dynamic one in which they are constantly reshaping the genome.

My objectives here are first to identify all of the transposons present in the *Arabidopsis* genome and then, using the data gathered from the survey, to investigate transposon mobility and to study the relationship between environmental signals and transposon activation. My goal is to provide further insight into the significance of transposon activation by environmental factors and how this may affect the organization and evolution of genes and genomes. In other words, how transposons can reshape the story of a book.

CHAPTER 1

DATA MINING FOR TRANSPOSONS

Transposons are sequences that have the ability to change their genomic location. This mobile nature gives transposons a mutagenic potential and, since they have been found in nearly every living organisms studied to date, they are suggested to be a valuable source of genetic variability and important players in evolution. The recent availability of eukaryotic genome sequences has not only significantly increased the speed and efficiency of transposon discovery but has also allowed for the identification of major families of novel elements (LE *et al.* 2000). In addition, analysis of genomic information available in current sequence databases have provided us with a better understanding of the relationship between different element groups and how transposons may have affected gene and genome structures. This review will focus on three transposon families for which data mining was key in their discovery and in understanding their origins. We also discuss the possible evolutionary roles these elements may have in reshaping cellular genes and their potential impact on host genomes.

IMPORTANCE OF TRANSPOSONS

Transposons are important components of eukaryotic genomes where they can sometimes make up a large proportion of the genetic content. For example, they account for more than 10% of the *Arabidopsis* genome (THE ARABIDOPSIS GENOME INITIATIVE 2000), over 45% in humans (LANDER *et al.* 2001) and as much as 80% in maize is estimated to be composed of mobile elements (SANMIGUEL *et al.* 1996) (Table 1.1). This disparity in transposon content between genomes of different species can partly account

for what is known as the C-value paradox, which refers to the fact that genome sizes (*i.e.* C-value) between organisms can vary tremendously but with no correlation to gene content or biological complexity (LI 1997).

Transposons can cause genetic changes in a variety of ways. Insertions into exons or regulatory regions can disrupt gene sequences and in fact, this property has been widely exploited as a biotechnological tool for gene-tagging (BANCROFT *et al.* 1992), mapping (WILLIAMS *et al.* 1992) and enhancer/promoter traps (BELLEN *et al.* 1989; WILSON *et al.* 1989). Transposon excision is often an imprecise process which causes short deletions or generates small insertions. In addition, transposons can alter gene expression by providing regulatory sequences or alternative splice sites (BANVILLE and BOIE 1989; NORRIS *et al.* 1995; VARAGONA *et al.* 1992). They can also have a more global impact by affecting genome organization through chromosomal rearrangements. For example, non-homologous recombinations can occur between transposons because of their repetitive nature and aberrant transposition events can lead to deletions, translocations or inversions of large genomic regions (KIM *et al.* 1998). At one extreme, it has been hypothesized that gross chromosomal changes brought by transposition can lead to reproductive isolation of hybrids and potentially result in speciation (HURST and WERREN 2001). Parts of cellular gene sequences, such as exons, can also be mobilized along with transposons (MORAN *et al.* 1999; PICKERAL *et al.* 2000) and subsequently co-opted for novel cellular functions. In fact, transposon sequences themselves can assume functions for cellular processes (AGRAWAL *et al.* 1998; PARDUE *et al.* 1996). However, beneficial changes to the host genome resulting from transpositions are rare and not immediately obvious. Transposons are thus viewed at two extremes as either genomic parasites or as a molecular apparatus used to the advantage of host organisms (ORGEL and CRICK 1980; WESSLER 1996). Regardless, it is clear that there is a complex relationship between transposons and their host and that mobile elements are an important source of genetic variability upon which natural selection can act (CAPY *et al.* 2000; KIDWELL and LISCH 2001).

CLASSIFICATION

There has not been a common consensus that attempts to systematically name and classify the myriad of known elements identified from almost every living organism studied to date. Transposons which have been independently identified or related family members have often been given different designations which are sometimes more poetic than informative. However, transposons can primarily be classified as either class I or II elements depending on whether they move through an RNA or a DNA intermediate, respectively (Table 1.2). Structural features and sequence similarity are other general criteria used to further group and classify elements. Both class I and II transposons that encode for all the proteins required for their own mobilization process are said to be autonomous whereas non-autonomous elements rely on the mobility proteins provided in *trans* by an autonomous transposon located elsewhere in the genome.

Class I - Retroelements and retrotransposons

Class I elements move in a replicative manner through an RNA intermediate. A number of element-encoded proteins are involved in the retrotransposition process which has been described in detail elsewhere (FINNEGAN 1997; BOEKE and STOYE 1997). In brief, class I elements are transcribed into an RNA form and then reverse-transcribed into DNA before being integrated into a new location. Reverse-transcription from an RNA intermediate back to a DNA form is performed by a reverse-transcriptase (RT). For elements containing Long Terminal Repeats (LTRs), re-integration is performed by another element-encoded enzyme called integrase (IN) whereas re-integration for elements without LTRs proceed through a process involving nick-translation called target-primed reverse transcription (LUAN *et al.* 1993).

Based on the presence or the absence of LTR sequences, class I elements can be divided into LTR-retrotransposons or non-LTR retroelements. Similarity between RT sequences can also serve to differentiate class I elements. Non-LTR retrotransposons include the human Long and Short Interspersed Nuclear Elements (LINEs and SINEs) whereas examples of LTR-retrotransposons include the founding members *copia* and the retroviral-like *gypsy* elements identified from *Drosophila*. A group of short elements

with structural features reminiscent of LTR-retrotransposons, and collectively named Terminal-repeat Retrotransposons in Miniature (TRIM) (Table 1.2), have been recently identified from a number of plant genomes but their classification is presently uncertain because no member with coding capacity has been identified (WITTE *et al.* 2001).

Class II - DNA transposons

Class II elements, or DNA transposons, are directly mobilized as DNA through a conservative mode of transposition. Mobilization requires an element-encoded enzyme called transposase that has the ability to recognize, bind and excise the transposon in order to re-integrate it at a new location (PLASTERK 1995). Transposase recognizes the Terminal Inverted Repeats (TIRs) sequences typically delimiting DNA transposons. Along with the transposases, the length and sequence similarities between TIRs are used to further define transposon families such as the *En/Spm* elements from maize, *hobo* from *Drosophila* and Tc transposons from *Caenorhabditis elegans* (Table 1.2).

TSD AND THE DDE MOTIF

Except for a few rare cases, nearly all types of elements duplicate their target site upon re-insertion. This Target Site Duplication (TSD) is the result of the repair, or "filling in", of the staggered cut in the double-stranded DNA generated during element insertion. Because the size and sequence of a TSD is often characteristic of the target specificity of a given transposase or integrase, TSD is another feature often used to classify members of both classes of transposons.

The integration process carried out by transposase of DNA transposons and integrase of LTR-retrotransposons and retroviruses are actually quite similar (RICE and BAKER 2001). In fact, many prokaryotic and eukaryotic transposases and class I integrases contain a catalytic domain characterized by a DDE motif. This motif is also found in polynucleotidyl transferases, suggesting that all these proteins may have a common evolutionary origin. The DDE motif is composed of two aspartic acids separated by ~60-100 amino-acids, on average, and shortly (~35 residues) followed by a third

aspartic or a glutamic acid. Structures for transposases and integrases are available and co-crystal structures of the prokaryotic Tn5 transposase multimer bound to DNA show that, when folded into the catalytic site, these three acidic residues come in close proximity to coordinate divalent metal ions (such as Mg^{2+} or Mn^{2+}) to cleave bound DNA (RICE and BAKER 2001).

DATA MINING TO IDENTIFY TRANSPOSONS

Traditionally, transposition events were detected and studied through genetic screens for phenotypic variegation. For example, transposition into and out of genes in the anthocyanin pigmentation pathway can be detected as a spotted colouration pattern in maize kernels (FEDOROFF 1989). Cloning and molecular approaches have also served to identify and study elements but, recently, transposons have been more efficiently found by computer-based strategies. The repetitive nature of transposons and the large amount of information available in sequence databases now allows for the systematic identification of transposons using sequence similarity search tools. Many of the findings from database-searches have consolidated our knowledge of existing transposon groups. However, three transposon families of special interest here were discovered by data mining and their detailed analysis has forced a re-evaluation of our current understanding of transposon classification, origins and genomic impact.

From data mining I: MITEs

About a decade ago, the application of database search strategies led to the identification of a large group of novel transposons in economically important plant species, such as rice and maize (BUREAU *et al.* 1996; BUREAU and WESSLER 1992; BUREAU and WESSLER 1994). These transposons were all found to be short sequences (~80-500 bp), contained TIRs, could potentially form a DNA hairpin structure and did not appear to possess any coding capabilities. On the basis of the TIR sequence and the size and sequence of their TSD, they were either named *Stowaway* (generating a "TA" TSD) or *Tourist* (generating a "TAA" or "TTA" TSD) and were collectively known as

Miniature Inverted-repeat Transposable Elements (MITEs). A third group of MITEs that generate a "TA" TSD but have TIRs distinct from *Stowaway* were identified afterward and designated as *Emigrant* (CASACUBERTA *et al.* 1998). MITEs are abundantly represented in the genome (there are more than 10, 000 copies of the *Tourist-Zm1* subfamily in maize) where they were frequently associated with genes, sometimes contributing regulatory sequences such as TATA boxes or polyadenylation signals (BUREAU *et al.* 1996). Because of their ubiquity and abundance in plant genomes and their frequent proximity to genes, they were viewed as the plant counterparts of mammalian SINEs, short non-autonomous retroelements commonly found associated with mammalian genes but scarce in plants (WESSLER *et al.* 1995). For a long time after their discovery, a longer MITE with coding capacity for mobility-related proteins could not be found and, thus, the mechanism and regulation of MITE mobility was not known. However, the presence of TIR sequences suggested that they may be non-autonomous class II transposons.

Although searches in the limited number of *Arabidopsis* gene sequences available prior to1996 did not reveal the presence of any transposons (BUREAU *et al.* 1996), many were later identified in genomic sequences among which MITEs were numerously represented and mostly found distributed within pericentromeric regions (THE ARABIDOPSIS GENOME INITIATIVE 2000). From the analysis of soybean, *Arabidopsis* and rice sequences, longer MITE members harbouring Open Reading Frames (ORFs) coding for putative mobility-related proteins were finally found for *Tourist*, *Stowaway* and *Emigrant* elements (LE *et al.* 2000; TURCOTTE *et al.* 2001). For many of the elements with coding capacity, the predicted ORFs appeared to have suffered small insertions or deletions and thus these elements may not code for a functional protein. However, within *Tourist*, *Stowaway* and *Emigrant*, some members are highly similar to each other and the identification of insertion polymorphisms indicates recent activity. Regardless of the potential for activity, these hypothetical proteins have provided valuable information regarding the origins and classification of these transposons. Closer analysis of the predicted ORFs in all three MITEs revealed the presence of the catalytic DDE motif and sequence similarity to other class II elements. However, recent data indicate that *Tourist*, *Stowaway* and *Emigrant* have divergent evolutionary histories and form distinct families

(LE *et al.* 2001; TURCOTTE and BUREAU 2002). Their common grouping under the term MITE is not entirely accurate and this designation should simply be used to describe short non-autonomous TIR-containing elements (TURCOTTE and BUREAU 2002).

The putative transposase for *Tourist* elements is more similar to the transposases of a family of prokaryotic Insertion Sequences (IS) known as IS5, than to either *Stowaway* or *Emigrant*. IS5 elements are simple transposons commonly found in prokaryotes, that generally produce a 3-bp TSD (MAHILLON and CHANDLER 1998). *Tourist*-like elements are widespread as members were found by sequence similarity searches in a number of eukaryotes such as nematodes and fungi (LE *et al.* 2001; LE *et al.* 2000). In addition, sequences similar to the *Tourist* transposase in other plants, insects, fish, birds and mammals are also present in EST databases, suggesting an even more widespread distribution of *Tourist* and arguing for potential activity. *Tourist* elements that were thought at first to be restricted to the genomes of monocotyledonous plants, and consequently have a more recent origin, are now found to belong to a IS5/*Tourist* superfamily with members found in distantly related organisms.

The identification of full-length *Emigrant* and *Stowaway* and characterization of their corresponding tansposases has significantly contributed to the understanding of the life history of these elements. Alignment of the *Emigrant* and *Stowaway* transposases shows that these elements clearly belong to the IS630/Tc1/*mariner* superfamily of transposons, which are simple and widespread elements with members in bacteria, fungi, plants and animals. The terminal sequences and phylogenetic analysis of the *Emigrant* tranposase indicate that these elements are most related to the *pogo*-like family within IS630/Tc1/*mariner* (FESCHOTTE and MOUCHES 2000; LE *et al.* 2000; TURCOTTE and BUREAU 2002). *Stowaway* is as closely related to Tc1 or *mariner* as it is to *Emigrant* and *pogo*-like elements, suggesting that *Stowaway* and *Emigrant* form distinct families within the IS630/Tc1/*mariner* superfamily. The phylogeny of IS630/Tc1/*mariner* members does not always agree with the evolutionary history of their host and suggests horizontal transmission which may have facilitated their spread to diverse host genomes (ROBERTSON and LAMPE 1995). Horizontal transfers appear to be common within IS630/Tc1/*mariner* transposons (ROBERTSON and LAMPE 1995) and may explain the presence of *Emigrant* in *Arabidopsis* as this element appears to have emerged rather

recently from a common ancestor with *pogo*-like element found in animals (TURCOTTE and BUREAU 2002).

From data mining II: *Katydids*

Examination of genomic sequences from *Arabidopsis*, *Solanum* and a number of other monocotyledonous and dicotyledonous plants revealed the presence of short elements (~350 bp) that shared a surprisingly high level of nucleotide sequence similarity (THE ARABIDOPSIS GENOME INITIATIVE 2000; WITTE *et al.* 2001). These unique elements have collectively been named TRIM and exhibit structural features typical of LTR-retrotransposons: they are delimited by terminal direct repeats; they are flanked by a 5-bp TSD; they harbour *cis*-recognition sequences (*i.e.* a primer binding site and a poly-purine tract) essential for retrotransposition; and they can be found as solo-repeats analogous to solo-LTRs, which are retrotransposons lacking the central sequences as a result of a recombination event between the two LTRs (SHIRASU *et al.* 2000; VICIENT *et al.* 1999). However, TRIMs would be the smallest LTR-retrotransposons found to date, with sizes that are approximately 10-30 times smaller than the average LTR-retrotransposon and that preclude any coding capacity. Unlike retrotransposons that are found clustered in the pericentromeric regions of the chromosomes, TRIMs are dispersed and frequently found associated with genes. The small size of TRIMs may make them better tolerated in or near gene coding sequences and evidences suggest that they may have an active role in restructuring plant genes and genomes (WITTE *et al.* 2001).

In *Arabidopsis* where sequence for the complete genome is available, a larger TRIM coding for mobility-related proteins could not be found. Thus, in order to transpose, these elements would have to rely on mobility proteins from an autonomous element. This would mean that the integrase used in *trans* would have to be able to recognize and bind TRIM-specific sequences. In fact, selective pressure to maintain *cis* sequences for mobilization in *trans* may explain the unusually high conservation of terminal-direct-repeat sequences between TRIM members from divergent genomes. As a comparison, LTR sequences between retrotransposon families share very little sequence similarity within the same genome. In other words, TRIMs may represent a retrotransposon with the minimal sequences required for transposition. Alternatively,

high sequence conservation between transposons from divergent genomes could reflect horizontal transfer. Although unlikely, TRIMs may represent the progenitors of LTR-retrotransposons. In this scenario, TRIMs would have been the ancestral form and gave rise to LTR-retrotransposons by increasing in size and acquiring all the genes necessary for autonomous mobilization. The process of acquisition of cellular sequences by retrotransposons and retroelements is known as transduction and has been well documented for retroviruses and oncogenes (VARMUS 1984), for the mobilization and shuffling of exons by human LINE elements (PICKERAL *et al.* 2000), and for the successive acquisition of cellular genes by the maize retrotransposon *Bs1* (BUREAU *et al.* 1994; ELROUBY and BUREAU 2001; JIN and BENNETZEN 1989; PALMGREN 1994). In fact, TRIMs also appear to have transduced parts of a cellular gene coding for a putative non-sense mediated mRNA decay factor (WITTE *et al.* 2001).

To date, TRIMs have only been identified in plant genomes but it is expected that they will be found in other eukaryotic genomes as more sequences are made available. Additional TRIM members will allow a better understanding of the transposition and evolution of these peculiar elements, especially if a larger element with coding capacity is revealed. Larger elements may represent intermediate forms of TRIMs and may help resolve whether these elements accumulated coding sequences to give rise to LTR-retrotransposons or, on the contrary, they are the product of a streamlining process.

From Data Mining III: *Basho*

Probably the most intriguing group of transposons identified by systematic searches is *Basho*, also known as *AthE1* and *Helitron* (LE *et al.* 2000; KAPITONOV and JURKA 2001; SURZYCKI and BELKNAP 1999; THE ARABIDOPSIS GENOME INITIATIVE 2000; TURCOTTE *et al.* 2001). Distribution analysis of *Basho* elements in *Arabidopsis* reveals that they are the most abundant elements and are clustered within the (gene-poor) pericentromeric regions (THE ARABIDOPSIS GENOME INITIATIVE 2000). Despite variations in size (from 0.2 to over 3 kb) due to insertions or deletions of internal sequences, most members have an average length of ~1 kb and share overall high sequence similarity, especially at the 5'and 3' termini. Curiously, *Basho* elements do not exhibit any of the structural features characteristic to existing eukaryotic transposons

(LTRs, TIRs, poly-A tail) and because their mode of mobility is not known, their classification is not possible.

By analyzing a *Basho* inserted into another repetitive element, some authors have concluded that amplification occurs through an illegitimate recombination event that results in short deletions of the target site (SURZYCKI and BELKNAP 1999). Using a similar approach which consists in finding Related to Empty Sites (RESites), others have provided support that *Basho* transposition creates a 1-bp TSD (LE *et al.* 2000). RESites are sequences similar to the regions flanking an insertion but that do not have the insertion in question. They represent the ancestral sequences before element insertion and can be found in sequence databases through similarity searches or by PCR amplification of orthologous regions from a related population (LE *et al.* 2000). Recent findings suggest a mechanism for *Basho* transposition that is similar to prokaryotic elements that undergo rolling-circle transposition, a replicative form of mobilization that does not result in duplication of the target site. Consensus sequences were used to re-build and analyze the ancestral *Basho*-like element (Kapitonov & Jurka, 2001) which harboured predicted proteins similar to the ones carried by prokaryotic rolling-circle elements. Consequently, *Basho* would represent another group of elements related to prokaryotic elements and may also be related to geminiviruses, a group of plant viruses that also mobilizes using a rolling-circle mechanism (KAPITONOV and JURKA 2001). However, it has been argued by some that such a rolling-circle mechanism of amplification would not explain the presence of a large number of non-autonomous elements, thought to be generated by imperfect double-strand break repair of excised elements (FESCHOTTE and WESSLER 2001). It is clear that molecular and biochemical data are now required to prove, or disprove, hypotheses pertaining to *Basho* transposition. Although the mechanism through which *Basho* elements are mobilized is still being debated, there is little doubt that these repetitive sequences are *bona fide* transposons because of the large number of nearly identical elements and numerous RESites.

From data mining to molecular studies

RESites, high sequence similarity between family members and/or the fact that some elements harbour ORFs coding for mobility-related proteins suggest that there are

potentially active elements in *Arabidopsis*. Future challenges not only include the elucidation of the mechanisms underlying transposition of novel elements, such as *Basho* and TRIM, but also include the investigation of host-element dynamics that regulate transposition. Although database searches has identified evidence of past mobility and putative transposases for MITEs in *Arabidopsis*, none of these have actually been demonstrated to be currently active. Furthermore, nothing is known of the regulation of their mobility.

Genome sequencing and data mining will prove to be invaluable tools to address these questions. In addition, experimental evidences will be important in testing some of the hypotheses and predictions generated from data mining. Protocols to map transposon positions in organisms such as *C. elegans*, *Drosophila*, pea and petunia (KORSWAGEN *et al.* 1996, EGGERT *et al.* 1998, ELLIS *et al.* 1998, VAN DEN BROECK *et al.* 1999) can also be efficiently utilized to monitor transposition events. In these strategies, restriction fragments containing transposon along with flanking sequences are PCR-amplified and simultaneously displayed on polyacrylamide gels. A mobilization event would thus result in a change in fragment size detectable as a polymorphic band. These techniques have been referred to as Sequence-Specific Amplification Polymorphisms, Transposon Insertion Display or Vectorette-mediated PCR but are also more commonly known as Transposon Display (TD). Recently, TD was used to map and detect mobility of MITEs in maize (CASA *et al.* 2000, ZHANG *et al.* 2001), to investigate the factors that can activate Tnt1 retrotransposons in tobacco (MELAYAH *et al.* 2001), and to study the population dynamics of an *Ac*-like element in *Arabidopsis* (WRIGHT *et al.* 2001). Undoubtedly, TD will be key in detecting mobility for some of the many other transposons identified through data mining.

Conclusion

The recent accumulation of sequences available in public databases has proven to be an important source of information and has largely benefited the study of transposons. This is compounded by the fact that genome sequencing projects have suddenly made available non-genic regions traditionally considered as being "junk" DNA but that contained the majority of elements. Undoubtedly, as new sequences are being added and

other genomes are being deciphered, many more transposons will be found, some of which will contain important details to help our understanding of transposons. Computer-based sequence analysis can be efficiently used along with experimental approaches to address questions regarding mechanisms of mobility, origins and evolution of elements and their interactions with host genes and genomes.

Table 1.1. Transposon content in different genomes

Organism	Genome size (Mb)*	Transposon content (%)	References
Bacteria			
Escherichia coli	4.64	2	(BLATTNER *et al.* 1997)
Archaeabacteria			
Sulfolobus solfataricus	2.99	10	(BRUGGER *et al.* 2002; SHE *et al.* 2001)
Fungi			
Saccharomyces cerevisae	11.7	3	(GOFFEAU *et al.* 1996; KIM *et al.* 1998)
Plants			
Arabidopsis thaliana	116	10	(THE ARABIDOPSIS GENOME INITIATIVE 2000)
Oryza sativa	430	20	(TURCOTTE *et al.* 2001)
Animals			
Caenorhabditis elegans	97	~ 6	(THE *C. ELEGANS* SEQUENCING CONSORTIUM 1998)
Drosophila melanogaster	132.7	15	(ADAMS *et al.* 2000)
Homo sapiens	2.9×10^3	45	(LANDER *et al.* 2001)

* As of February 2002.

Table 1.2. Diversity of eukaryotic transposons

Class Sub-class Superfamily	Characteristic structural features*	Examples of families
Class I – RNA elements		
Retroviruses		
Retrovirus	TSDs; LTRs; *gag*, *pol*, and *env*; PBS, PPT	HIV, MLV
LTR-retrotransposons		
gypsy-like	TSDs; LTRs; *gag*, *pol***; PBS, PPT	*gypsy*, Ty3
copia-like	TSDs; LTRs; *gag*, *pol***; PBS, PPT	*copia*, Ty1, Tnt1
TRIMs	TSDs; TDRs, PBS, PPT	*katydid*
non-LTR-retroelements		
LINEs	TSDs; *gag*, *pol**; poly A/T-rich tail	LINEs
SINEs	TSDs; no ORFs; poly A/T-rich tail	*Alu*
Class II – DNA elements		
TIR transposons		
hAT	TSDs, TIRs, ORFs	*hobo*; Ac/*Ds*; *Tam3*
Mutator	TSDs, TIRs, ORFs	*Mutator*, MULEs
P	TSDs, TIRs, ORFs	*P*
CACTA	TSDs, TIRs, ORFs	*En/Spm*, *cacl*
IS630/Tc1/*mariner*	TSDs, TIRs, ORF	Tc1, *mariner*, *Emigrant*, *Stowaway*
IS5/*Tourist*	TSDs, TIRs, ORF	Tc8, *Tourist*
Unknown		
Basho	(TSDs), ORFs, short palindrome	*Basho*

* Autonomous elements code for the proteins necessary for their own mobilization. LINE, Long Interspersed Nuclear Element; LTR, Long Terminal Repeat; MITE, Miniature Inverted-repeat Transposable Element; MLE, *Mariner*-Like Element; MULEs,

MUtator-Like Element; ORF, Open Reading Frame; PPT, Poly-Purine Tract; PBS, Primer Binding Site; SINE, Short Interspersed Nuclear Element; TDR, Terminal Direct Repeat; TIR, Terminal Inverted Repeat; TRIM, Terminal-repeat Retrotransposons In Miniature; TSD, Target Site Duplication.

** Mainly, the order of the domains encoded by *pol* gene distinguishes *copia* (PR, IN, RT) from *gypsy* (PR, RT, IN) elements, the latter being similar to *pol* genes of retroviruses.

REFERENCES

ADAMS M. D. CELNIKER, S. E. HOLT, R. A. EVANS, C. A. GOCAYNE, J. D. *et al.*, (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185-2195.

AGRAWAL A., EASTMAN, Q. M. and SCHATZ, D. G., (1998). Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* **394,** 744-751.

BANCROFT I., BHATT, A. M., SJODIN, C., SCOFIELD, S., JONES, J. D. *et al.*, (1992). Development of an efficient two-element transposon tagging system in *Arabidopsis thaliana*. *Mol. Gen. Genet.* **233,** 449-461.

BANVILLE D. and BOIE, Y., (1989). Retroviral long terminal repeat is the promoter of the gene encoding the tumor-associated calcium-binding protein oncomodulin in the rat. *J. Mol. Biol.* **207,** 481-490.

BELLEN H. J., O'KANE, C. J., WILSON, C., GROSSNIKLAUS, U., PEARSON, R. K. *et al.*, (1989). *P*-element-mediated enhancer detection: a versatile method to study development in *Drosophila*. *Genes Dev.* **3,** 1288-12300.

BLATTNER F. R., PLUNKETT, G., 3RD, BLOCH, C. A., PERNA, N. T., BURLAND, V. *et al.*, (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453-1474.

BOEKE J. D. and STOYE, J. P., (1997). Retrotransposons, endogenous retroviruses and the evolution of retroelements, pp. 343-436 in *Retroviruses*, edited by J. M. Coffin, S. H. Hughes and H. E. Varmus. *Cold Spring Harbor Laboratory*, New York.

BRUGGER K., REDDER, P., SHE, Q., CONFALONIERI, F., ZIVANOVIC, Y. *et al.*, (2002). Mobile elements in archaeal genomes. *FEMS Microbiol. Lett.* **206,** 131-141.

BUREAU T. E., RONALD, P. C. and WESSLER, S. R., (1996). A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc. Natl. Acad. Sci. USA* **93,** 8524-8529.

BUREAU T. E. and WESSLER, S. R., (1992). *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell* **4,** 1283-1294.

BUREAU T. E. and WESSLER, S. R., (1994). *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6,** 907-916.

BUREAU T. E., WHITE, S. E. and WESSLER, S. R., (1994). Transduction of a cellular gene by a plant retroelement. *Cell* **77,** 479-480.

CAPY P., GASPERI, G., BIEMONT, C. and BAZIN, C., (2000). Stress and transposable elements: co-evolution or useful parasites? *Heredity* **85,** 101-106.

CASA A. M., BROUWER C., NAGEL A., WANG L., ZHANG Q., KRESOVICH S. and WESSLER S. R. (2000). The MITE family *Heartbreaker* (*Hbr*): molecular markers in maize. *Proc. Natl. Acad. Sci. USA* **97,** 10083-10089.

CASACUBERTA E., CASACUBERTA, J. M., PUIGDOMENECH, P. and AMPARO., M., (1998). Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: characterisation of the *Emigrant* family of elements. *The Plant Journal* **16,** 79-85.

ELLIS T. H. N., POYSER S. J., KNOX M. R., VERSHININ A. V. and AMBROSE M. J. (1998). Polymorphism of insertion sites of *Ty1-copia* class retrotransposons and its use for linkage and diversity analysis in pea. *Mol. Gen. Genet.* **260,** 9-19.

EGGERT H., BERGEMANN K. and SAUMWEBER H. (1998). Molecular screening for *P*-element insertions in a large genomic region of *Drosophila melanogaster* using polymerase chain reaction mediated by the vectorette. *Genetics* **149,** 1427-1434.

ELROUBY N. and BUREAU, T. E. (2001). A novel hybrid open reading frame formed by multiple cellular gene transductions by a plant long terminal repeat retroelement. *J. Biol. Chem.* **276,** 41963-41968.

FEDOROFF N. V., (1989). About maize transposable elements and development. *Cell* **56,** 181-191.

FESCHOTTE C. and MOUCHES, C., (2000). Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a *pogo*-like DNA transposon. *Mol. Biol. Evol.* **17,** 730-737.

FESCHOTTE C. and WESSLER, S. R., (2001). Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* **98,** 8923-8924.

FINNEGAN D. J., (1997). Transposable elements: how non-LTR retrotransposons do it. *Curr. Biol.* **7,** R245-248.

GOFFEAU A., BARRELL, B. G., BUSSEY, H., DAVIS, R. W., DUJON, B. *et al.*, (1996). Life with 6000 genes. *Science* **274,** 563-567.

HURST G. D. and WERREN, J. H., (2001). The role of selfish genetic elements in eukaryotic evolution. *Nat. Rev. Genet.* **2,** 597-606.

JIN Y. K. and BENNETZEN, J. L., (1989). Structure and coding properties of *Bs1*, a maize retrovirus-like transposon. *Proc. Natl. Acad. Sci. USA* **86,** 6235-6239.

KAPITONOV V. V. and JURKA, J., (2001). Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci. USA* **98,** 8714-8719.

KIDWELL M. G. and LISCH, D. R., (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution Int. J. Org. Evolution* **55,** 1-24.

KIM J. M., VANGURI, S., BOEKE, J. D., GABRIEL, A. and VOYTAS, D. F., (1998). Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* **8,** 464-478.

KORSWAGEN, H. C., R. M. DURBIN, M. T. SMITS and PLASTERK R. H. A. (1996) Transposon Tc1-derived, sequence-tagged sites in *Caenorhabditis elegans* as marker for gene mapping. *Proc. Natl. Acad. Sci. USA* **93,** 14680-14685.

LANDER E. S. LINTON, L. M. BIRREN, B. NUSBAUM, C. ZODY, M. C. *et al.*, (2001). Initial sequencing and analysis of the human genome. *Nature* **409,** 860-921.

LE Q. H., TURCOTTE, K. and BUREAU, T., (2001). Tc8, a *Tourist*-like transposon in *Caenorhabditis elegans*. *Genetics* **158,** 1081-1088.

LE Q. H., WRIGHT, S., YU, Z. and BUREAU, T., (2000). Transposon diversity in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **97,** 7376-7381.

LI W.-H., (1997). *Molecular Evolution. Sinauer Associates*, Sunderland, Massachussetts.

LUAN D. D., KORMAN, M. H., JAKUBCZAK, J. L. and EICKBUSH, T. H., (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72,** 595-605.

MAHILLON J. and CHANDLER, M., (1998). Insertion sequences. *Microbiol. Mol. Biol. Rev.* **62,** 725-774.

MELAYAH D., BONNIVARD, E., CHALHOUB, B., AUDEON, C. and GRANDBASTIEN, M. A., (2001). The mobility of the tobacco Tnt1 retrotransposon correlates with its transcriptional activation by fungal factors. *Plant J.* **28,** 159-168.

MORAN J. V., DEBERERDINIS, R. J. and KAZAZIAN JR, H. H., (1999). Exon shuffling by L1 retrotransposition. *Science* **283,** 1530-1534.

NORRIS J., FAN, D., ALEMAN, C., MARKS, J. R., FUTREAL, P. A. *et al.*, (1995). Identification of a new subclass of *Alu* DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *J. Biol. Chem.* **270,** 22777-22782.

ORGEL L. E. and CRICK, F. H., (1980). Selfish DNA: the ultimate parasite. *Nature* **284,** 604-607.

PALMGREN M. G., (1994). Capturing of host DNA by a plant retroelement: *Bs1* encodes plasma membrane H(+)-ATPase domains. *Plant Mol. Biol.* **25,** 137-140.

PARDUE M. L., DANILEVSKAYA, O. N., LOWENHAUPT, K., SLOT, F. and TRAVERSE, K. L., (1996). *Drosophila* telomeres: new views on chromosome evolution. *Trends Genet.* **12,** 48-52.

PICKERAL O. K., MAKALOWSKI, W., BOGUSKI, M. S. and BOEKE, J. D., (2000). Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10,** 411-415.

PLASTERK R. H., (1995). Mechanisms of DNA transposition, pp. 18-37 in *Mobile genetic elements*, edited by D. J. Sherratt. *Oxford University Press*, New York.

RICE P. A. and BAKER, T. A., (2001). Comparative architecture of transposase and integrase complexes. *Nat. Struct. Biol.* **8,** 302-307.

ROBERTSON H. M. and LAMPE, D. J., (1995). Recent horizontal transfer of a *mariner* transposable element among and between Diptera and Neuroptera. *Mol. Biol. Evol.* **12,** 850-862.

SANMIGUEL P., TIKHONOV, A., JIN, Y. K., MOTCHOULSKAIA, N., ZAKHAROV, D. *et al.*, (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274,** 765-768.

SHE Q., SINGH, R. K., CONFALONIERI, F., ZIVANOVIC, Y., ALLARD, G. *et al.*, (2001). The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *Proc. Natl. Acad. Sci. USA* **98,** 7835-7840.

SHIRASU K., SCHULMAN, A. H., LAHAYE, T. and SCHULZE-LEFERT, P., (2000). A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.* **10,** 908-915.

SURZYCKI S. A. and BELKNAP, W. R., (1999). Characterization of repetitive DNA elements in *Arabidopsis*. *J. Mol. Evol.* **48,** 684-691.

THE ARABIDOPSIS GENOME INITIATIVE, (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796-815.

THE *C. ELEGANS* SEQUENCING CONSORTIUM, (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012-2018.

TURCOTTE K. and BUREAU, T., (2002). Phylogenetic analysis reveals *Stowaway*-like elements may represent a fourth family of the family of the IS630-Tc1-*mariner* superfamily. *Genome* **45,** 82-90.

TURCOTTE K., SRINIVASAN, S. and BUREAU, T., (2001). Survey of transposable elements from rice genomic sequences. *Plant J.* **25,** 1-13.

VAN DEN BROECK D., MAES M., SAUER M., ZETHOF J., DE KEUKELEIRE P., D'HAUW M., VAN MONTAGU M. and GERATS T. (1998). Transposon Display identifies individual transposable element in high copy number lines. *Plant J.* **13,** 121-129.

VARAGONA M. J., PURUGGANAN, M. and WESSLER, S. R., (1992). Alternative splicing induced by insertion of retrotransposons into the maize *waxy* gene. *Plant Cell* **4,** 811-820.

VARMUS H. E., (1984). The molecular genetics of cellular oncogenes. *Annu. Rev. Genet.* **18,** 553-612.

VICIENT C. M., SUONIEMI A., ANAMTHAWAT-JOHNSON K., TANSKANEN J., BEHARAV A., *et al.*, (1999). Retrotransposon *BARE-1* and its role in genome evolution in the genus *Hordeum*. *Plant Cell* **11,** 1769-1784.

WESSLER S. R., (1996). Turned on by stress. Plant retrotransposons. *Curr. Biol.* **6,** 959-961.

WESSLER S. R., BUREAU, T. E. and WHITE, S. E., (1995). LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* **5,** 814-821.

WILLIAMS B. D., SCHRANK, B., HUYNH, C., SHOWNKEEN, R. and WATERSTON, R. H., (1992). A genetic mapping system in *Caenorhabditis elegans* based on polymorphic sequence-tagged sites. *Genetics* **131,** 609-624.

WILSON C., PEARSON, R. K., BELLEN, H. J., O'KANE, C. J., GROSSNIKLAUS, U. *et al.*, (1989). *P*-element-mediated enhancer detection: an efficient method for isolating and characterizing developmentally regulated genes in *Drosophila. Genes Dev.* **3,** 1301-1313.

WITTE C. P., LE, Q. H., BUREAU, T. and KUMAR, A., (2001). Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl. Acad. Sci. USA* **98,** 13778-13783.

WRIGHT, S.I., LE, Q.H., SCHOEN, D.J. AND BUREAU, T. (2001). Population dynamics of an *Ac*-like transposable element in self- and cross-pollinating *Arabidopsis*. *Genetics* **158,** 1279-1288.

ZHANG X., FESCHOTTE, C., ZHANG, Q., JIANG, N., EGGLESTON, W. B. *et al.*, (2001). *P* instability factor: an active maize transposon system associated with the amplification of *Tourist*-like MITEs and a new superfamily of transposases. *Proc. Natl. Acad. Sci. USA* **98,** 12572-12577.

CHAPTER 2

DATA MINING I: TRANSPOSON DIVERSITY IN *ARABIDOPSIS THALIANA*

This chapter was published as a paper in the *Proceedings of the National Academy of Sciences* USA (Le *et al.*, 2000). It corresponds to the first phase of my project, which consisted in identifying transposons in the genome sequences of *Arabidopsis* in order to obtain a general overview of the transposon content. Even though *Arabidopsis* contains relatively few transposons when compared to other plants such as maize or barley, it nonetheless harbours a very diverse number of elements and this work has generated a collection of interesting individual stories to be followed up on. Because it was initially written when only ~14% of the *Arabidopsis* genome was completed, this chapter is merely a glimpse at the full set of transposons. Since then, a more complete data set has appeared together with the first report of the complete sequence of *Arabidopsis* published in *Nature* (The Arabidopsis Genome Initiative, 2000).

References

LE, Q.H., WRIGHT, S., YU, Z. AND BUREAU, T. (2000). Transposon diversity in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **97**, 7376-7381.

THE ARABIDOPSIS GENOME INITIATIVE (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815.

ABSTRACT

Recent availability of extensive genome sequence information offers new opportunities to analyze genome organization, including transposon diversity and accumulation, at a level of resolution that was previously unattainable. In this report, we used sequence similarity search and analysis protocols to perform a fine-scale analysis of a large sample (~17.2 Mb) of the *Arabidopsis thaliana* (Columbia) genome for transposons. Consistent with previous studies, we report that the *Arabidopsis thaliana* genome harbours diverse representatives of most known superfamilies of transposons. However, our survey reveals a higher density of transposons of which over one-fourth could be classified into a single novel transposon family designated as *Basho* which appears unrelated to any previously known superfamily. We have also identified putative transposase-coding ORFs for miniature inverted-repeat transposable elements (MITEs), providing clues into the mechanism of mobility and origins of the most abundant transposons associated with plant genes. In addition, we provide evidence that most mined transposons have a clear insertion preference for A+T-rich sequences and show that structural variation for many mined transposons is partly due to inter-element recombination. Taken together, these findings further underscore the complexity of transposons within the compact genome of *Arabidopsis thaliana*.

INTRODUCTION

Transposons are fundamental components of most eukaryotic genomes, with important contributions to their size, structure and variation. Based on mode of mobility, transposons are divided into two classes. Class I transposons move through an RNA intermediate and are reverse transcribed prior to their integration at another location in the genome. Retroelements can be further subdivided into retrotransposons (*e.g. copia*-like and *gypsy*-like) which are flanked by long terminal-repeats (LTRs) and non-LTR retroelements (*e.g.* long and short interspersed repetitive elements). Class II elements are characterized by terminal inverted-repeats (TIRs) and move directly through a DNA form by a "cut and paste" mechanism (*1*). Structural features, shared sequence similarity, and the size and sequence of the target site duplication (TSD) generated upon insertion, serve to further distinguish transposon superfamilies. As mutational agents, much of transposon impact may be deleterious to their hosts, although some insertions appear to play a significant role in adaptive evolution (*2-4*).

Historically, transposon discovery and analysis have been primarily conducted through the molecular genetic characterization of transposon-induced morphological mutations (*1*). While these studies have allowed for the characterization of many mobile element groups, they do not allow for fine-scale investigation into the extent of transposon diversity, and the forces driving this variability. As genome sequencing projects increase in scale and number, detailed analysis of the patterns of transposon diversity and abundance, and their contribution to genome organization becomes possible (*5*). The evidence thus far indicates that these patterns may be extremely variable among eukaryotic genomes (*6, 7*), suggesting the importance of such studies across diverse organisms.

The genome of the model plant *Arabidopsis thaliana* is small with a correspondingly low repetitive DNA content relative to other higher plants (*8*) and is currently targeted for complete sequencing. Previous studies using computer-based sequence similarity searches have revealed numerous repetitive elements within the *Arabidopsis* genome (*9, 10*). Similarly, the complete sequences of *Arabidopsis* chromosomes 2 and 4 have also allowed for the identification of transposon-related

sequence (*11, 12*). In this report, we performed a systematic and fine-scale search of *Arabidopsis* genome sequences to identify and characterize transposons. Our study differs from previous mining attempts in that we have not only compiled repetitive sequences but also provide evidence that many are in fact transposons by structural analysis and demonstration of past mobility. In this way, we provide insight into *Arabidopsis* transposon structure, mobility, distribution and diversity.

MATERIALS AND METHODS

Transposon Mining. *Arabidopsis* genomic sequences from 243 clones (approximately 17.2 Mb) with representation from all five linkage groups, were accessed from GenBank (National Center for Biotechnology Information, NCBI; http://www.ncbi.nlm.nih.gov/), between June and December 1998. Intergenic and intron sequences longer than 500 bp in length were used as BLAST search queries (version 2.0, http://www.ncbi.nlm.nih.gov/blast/) (*13*). Repetitive sequences with similarity to the query (score > 80) were compiled into groups and used as queries in additional database searches against GenBank entries released before December 1999. A total of 770 repetitive sequences belonging to 197 groups were identified. Of these, 444 mined sequences belonging to 135 groups were determined to be members of previously described transposon superfamilies by virtue of shared sequence composition and/or structural features such as TIRs, LTRs, a terminal poly A-rich sequence, coding capacity for mobility-related proteins, and flanking direct-repeats (*i.e.* TSDs). Another seven repetitive sequence groups representing 179 elements were determined to be mobile elements by documentation of a mobile history (see below). In many cases, some members of each mobile element group lacked one or both terminal sequences and were defined as truncated. Lastly, 147 repetitive sequences belonging to 55 groups could not be classified as transposons based on structural and sequence analysis. A detailed description of the mined mobile elements in our survey can be accessed from our relational database (http://soave.biol.mcgill.ca/clonebase/). A subset of the mined transposons were previously identified and/or annotated to be repetitive DNA, putative transposons, or transposon-related sequences by other researchers (*6, 9-12, 14-22*).

Data analysis. When necessary, further sequence analysis and alignments were performed using CLUSTALW (*23*), DIVERGE, TRANSLATE, PILEUP, BESTFIT and GAP from the University of Wisconsin Genetics Computing Group suite of programs (version 10) or additional BLAST search tools provided at NCBI (*13, 24*). Visualization of amino acid alignments was done using GeneDoc (*25*). A Perl program was written to compile sequences immediately flanking transposon insertion and to calculate the average G+C

content using a 20 bp sliding window (written by C. Olive). Only flanking sequences of intact, and not truncated, elements were included in the calculations. As a control, 50 positions within intergenic regions were randomly selected and submitted to the program. A copy of this program is available upon request.

Documentation of mobile history. Sequences immediately flanking the element were used as queries in database searches. Sequences sharing similarity to the queries typically represented either orthologous, or more frequently, paralogous regions (*e.g.*, multigene families, transposons, duplicated genomic regions or other repetitive sequences). In many cases, a pairwise comparison could be used to identify a gap corresponding to the absence of the insertion. This mined sequence with high nucleotide sequence similarity to the original query but lacking the insertion is referred to as a related to empty site or RESite. Examination of RESites also served to delimit element termini and to identify corresponding TSDs when this information could not be obtained from sequence and structural analysis. Alternatively, when a computer-assisted approach failed, RESites were amplified from other ecotypes of *Arabidopsis thaliana*, *A. lyrata*, or *Brassica* spp. using a previously described PCR approach (*26*; data not shown).

RESULTS AND DISCUSSION

In order to obtain a complete transposon profile of the *Arabidopsis* genome, we systematically surveyed a large sample (~17.2 Mb) of genomic sequences for transposon insertions using sequence similarity search algorithms (*13, 24*). Demonstration of past mobility of the mined elements was provided through the identification of RESites, which are sequences similar to the empty site of an insertion (Figure 2.1). This approach allows a rapid and efficient means to delimit element termini and highlight putative target site duplication events. We suggest that RESites provide convincing evidence that many mined interspersed repetitive sequences in *Arabidopsis* are in fact transposons. Together, mined repetitive sequences and their corresponding RESites strongly support a transposition-based insertion mechanism.

Based on shared sequence and structural similarities and the analysis of RESites, the majority of repetitive sequences mined in our survey (82%) could be compiled and categorized into 142 groups of putative transposons. The majority of the mined transposons corresponded to known superfamilies according to shared structural and sequence similarities (Table 2.1, Figure 2.1*A*). Among the groups of mined transposons, 28 have members that are structurally reminiscent to the maize *Mutator* element and/or harboured ORFs with significant similarity to the maize *Mutator* transposase (*i.e.* MURA). *Mutator*-like elements (MULEs) were in part defined as transposons with long TIRs (TIR-MULEs) as in the case of the maize *Mu* family of elements (*30*). However, some MULEs lack long TIRs (non-TIR-MULES) but have other features characterizing them as MULEs such as a 8-10 bp TSD and, for 5 elements, an ORF encoding a MURA-like transposase. In fact, sequence comparison and analysis of *Arabidopsis* TIR- and non-TIR- MULEs indicate that they share a common evolutionary history (Z. Yu, S. Wright and T. Bureau, unpublished data).

Interestingly, over one-fourth of the elements identified from 7 distinct groups did not belong to any of the previously described superfamilies but appeared related based on common structural features. The identification of 9 RESites indicate that these elements, which we have named *Basho* (after the nomadic Japanese haiku poet), have defined termini (5'-CHH....CTAG-3', where H = A, T, or C) and a target site preference for the

mononucleotide "T". *Basho*-like elements have been previously described as repetitive sequences and putative insertion sequences (*9, 16*) but no evidence (*e.g.*, RESites, defined element termini and target site sequence) was presented to indicate whether they are in fact mobile elements. Curiously, *Basho* elements appear to insert in a preferred orientation relative to the sequence context of the target site, namely 5'-AT-3' (Figure 2.1*B*). In addition, we have identified a *Basho*-like group in maize (data not shown) suggesting that *Basho* defines a new superfamily of transposons. Although high sequence similarity and RESites attest to past mobility, the mechanism by which *Basho* elements have transposed is presently unknown.

Our survey permitted an examination of the genomic patterns of transposons within *Arabidopsis*. For many eukaryotes there appears to be a relationship between the number of transposons, primarily class I elements (*i.e.* LTR and non-LTR retrotransposons), and genome size (*31*). In very large genomes, for instance, the transposon content can account for the majority of the genome (*32*). In agreement with the small size of the *Arabidopsis* genome, approximately 5% of the genomic sequences surveyed were composed of transposons of which only 2% were class I elements. Consistent with previous estimates of retrotransposon content in *Arabidopsis* (*10, 11, 18*), we found significantly fewer class I elements than reported in the genomes of other higher plants (*29*). This contrast in abundance of transposon type between various genomes may suggest a differential success depending on genomic environment.

The mined transposons were predominantly found in intergenic regions, with 5% located within introns of predicted genes and approximately 8% were nested insertions. The prevalence of mined transposons in non-coding regions is similar to the patterns which have been observed in other organisms; for example, heterochromatic regions in *Drosophila* and intergenic regions in maize typically contain numerous nested transposons (*32, 33*). A previous report also suggests that transposons are highly enriched within centromeric heterochromatin in *Arabidopsis* (*34*). In this report, only four of the clones surveyed in our study correspond to centromeric heterochromatin. Therefore, a meaningful comparison between these regions and euchromatic regions could not be achieved. While many insertions appeared to have a random target site sequence context, preferences for specific A+T-rich target sites were observed for three

of the most abundant types of elements, MITEs (TA, TAA, ATA), *Basho* (T), and *Mariner*-like elements (MLE; TA) (Figure 2.1). In addition to sequence-specific target sites, these elements appear to be distributed preferentially in A+T-rich regions. Up to 300 bp of sequences immediately flanking the site of insertion show a G+C content of ~25% (Figure 2.2), which is lower than previous estimations for the *Arabidopsis* genome (36.7% (*35*) and 35.8% (*11*)) and lower than observed in our control (Figure 2.2). *Arabidopsis* MULEs, which do not appear to have target sequence preference are also preferentially distributed in A+T-rich regions.

For a subset of mined transposons, the high level of shared nucleotide sequence similarity observed (>90%) suggests that elements have been recently active. Shared sequence identity between flanking regions of elements and matching RESites also attest to recent mobility (Figure 2.1). Although the majority of mined elements lacked any coding capacity, some members of mined groups were found harbouring corresponding ORFs such as reverse transcriptase, *Ac*-like transposase, CACTA-like transposase, or maize *Mutator* transposase (*29, 30, 36*). In addition, one group of mined elements (MLE I) not only shares structural features with the *Tc1/Mariner* transposon superfamily (Figure 2.3*A*), but also has at least one member located on chromosome 2 that harbours an ORF with up to 46% amino acid sequence similarity with the transposase of *Tc*1/*Mariner*-like elements, *Pogo*R11 and *Tigger*1 (Figure 2.3*B*). Furthermore, MLE I elements have the conserved terminal bases (5'-CAGT-3') necessary for the efficient transposition of other typical *Tc*1/*Mariner*-like elements (*37*). Some members of the MLE I have previously been reported to belong to a novel family of MITEs, referred to as *Emigrant* (*14*), based on their small size and target site preference for the dinucleotide TA. However, the MLE I elements clearly have more in common with transposons of the *Tc*1/*Mariner* superfamily (Figure 2.3) than to elements belonging to the MITE superfamily (*38*). The mined MLE I transposase shares no significant sequence similarity with two *Tc*1/*Mariner*-like transposases reported by Lin *et al.* (*11*) also on chromosome 2.

While the mechanism of MITE mobility was previously unclear we found evidence that some groups of MITEs have unusually large members which potentially encode for a transposase (Figure 2.4*A*). Specifically, putative ORFs found in members of two distinct MITE groups (X and XI) share 51% amino acid sequence similarity (Figure

2.4*B*). In addition, the ratio of synonymous to nonsynonymous nucleotide substitutions of 2.82 suggests the possibility of functional constraint on this coding sequence. These putative ORFs also share sequence similarity with the transposases of cyanobacterial and other prokaryotic insertion sequences (*IS*) (Figure 2.4*C*). Since MITE X and XI members are structurally similar to other MITE groups in *Arabidopsis*, as well as the *Tourist*-like elements in grasses (*43*), we suggest that the MITE X and XI ORFs represent the transposases characteristic of the *Tourist*-like family of elements.

We observe extensive diversity among mined elements, both within and among groups. The large number of mined transposon groups (Table 2.1) illustrates a high level of evolutionary divergence, and is indicative of long-term persistence or the result of horizontal transfer (*44*). Differences in size between related elements suggests that insertions and deletions may account for a significant proportion of variation within groups and is reflected by the high occurrence of apparently truncated elements (38%). In addition, inter-element recombination appears to generate sequence diversity among some of the mined elements. For example, the mosaic structure of members of some MULE groups is due to the exchange of terminal sequences (Figure 2.5). Such mosaic elements are apparently capable of transposition, as suggested by both their copy number, sequence similarity and presence of perfect TSDs. Another possible factor driving sequence diversity in MULEs is the acquisition of truncated cellular genes. For example, a member of MULE I harbours sequences which share ≈ 85% nucleotide sequence similarity to a region spanning exon 1, intron 1, and exon 2 of the *Arabidopsis* homeobox gene *Athb*-1 (Figure 2.6; *45*, Z. Yu, S. Wright and T. Bureau, unpublished data). Despite the sequence diversity observed, some conserved motifs among elements exist which suggest a common interaction with host factors (*e.g.* general transcription factors and gene regulatory proteins). We have identified a subterminal sequence motif (5'-TTTTCCCGCCAAAA-3') shared between MITE XI and *Ac*-like VII. This may be analogous to the motifs shared between the maize *Ac* and *Mutator* which are thought to be important in the regulation of mobility (*46*). Furthermore, a motif determined to be a matrix attachment region (5'-CAAATTTATTTTTATA-3') (*47*) within a member of MULE XVII is conserved among all group members.

Our report documents the features and characteristics of the surprisingly diverse forms of transposons residing in, for the most part, noncoding regions of the relatively small and compact genome of *Arabidopsis*. Lack of insertions into coding regions and prevalence in A+T-rich sequences/regions may reflect purifying selection against deleterious mutations and/or insertion preference of many mined elements. In either case, richness of elements in nongenic regions may represent merely genomic "junk" or, in some cases, elements may have a host function, as has been observed in other organisms (*2*). Clearly, the actual impact of transposons in genome evolution will require further molecular dissection of the factors important in gene expression, as well as analysis of population dynamics of element insertions. Our survey also provides additional clues as to the origin and evolution of transposons. Significant similarity between transposases of bacterial insertion sequences and the mobility-related proteins found in this (MITE X and XI) and other studies (*Mutator* or MULEs, *Tc1/Mariner* and retroelements) (*39-41, 48*) suggests that the majority of element superfamilies are ancient components of genomes. In addition, recombination appears to be important in giving rise to novel forms of MULEs and possibly other types of elements in *Arabidopsis*.

By providing a fine-scale approach in the mining of transposons from sequence databases, we have demonstrated that the abundance and variation of transposons in *Arabidopsis* is higher than previously reported. This obviously highlights the need for continued analysis of genomic sequence information in order to provide a high-resolution image of the *Arabidopsis* genome (*49*). Finally, information on transposon structures, mobile histories, and genomic coordinates presented in our report will expedite the development of novel biotechnologies for mapping purposes, systematic gene disruption and functional analysis.

ACKNOWLEDGMENTS

We like to thank Dr. Daniel J. Schoen for comments on the manuscript. We are grateful to Chris Olive, Chengbin Feng, Hala Razi Khan and Sylva Petrova for setting up and maintaining our complete transposon profile database which can be accessed at http://soave.biol.mcgill.ca/ clonebase/. This work was funded by a National Science and Engineering Research Council of Canada grant to T. E. B.

Table 2.1. Transposons in 17.2 Mb of the *Arabidopsis thaliana* (Columbia) genome

Type	Superfamily	Number of groups	Number of transposons
class I			
	SINEs *	3	14
	LINEs †	11	14
	copia–like retrotransposons	25	38
	gypsy-like retrotransposons	19	35
	undetermined ‡	29	38
class II			
	Ac-like	6	38
	CACTA-like	1	3
	MULEs §	28	108
	MITEs ¶	15	105
	MLEs ||	1	56
class?			
	Basho	8	179
total		147	630

* Short interspersed nuclear elements. † Long interspersed nuclear elements. ‡ These elements were determined to be class I based on sequence similarity to reverse transcriptase or were structurally reminiscent of solo-LTRs. However, informative coding regions were unavailable to group the elements into a given superfamily. § *Mutator*-like elements. ¶ Miniature inverted-repeat transposable elements. || *Mariner*-like elements.

FIGURE LEGENDS

Figure 2.1. RESites corresponding to mined *Arabidopsis* elements.

(*A*) Examples of RESites for different groups of mined elements. All of the RESites illustrated were identified by computer-assisted database searches. In total, 34 RESites were found by computer-assisted database searches and 13 were identified by cross-ecotype PCR analysis. The target sequences are underlined and the TSDs are shaded. GenBank gi numbers and nucleotide position on clones are indicated. * Inserted into a *Basho* III element; † Inserted into a *Basho* III element; ‡ Inserted into a MITE IX element. (*B*) RESites found for *Basho* insertions confirm mononucleotide TSD (shaded). § Inserted into a *Basho* V element.

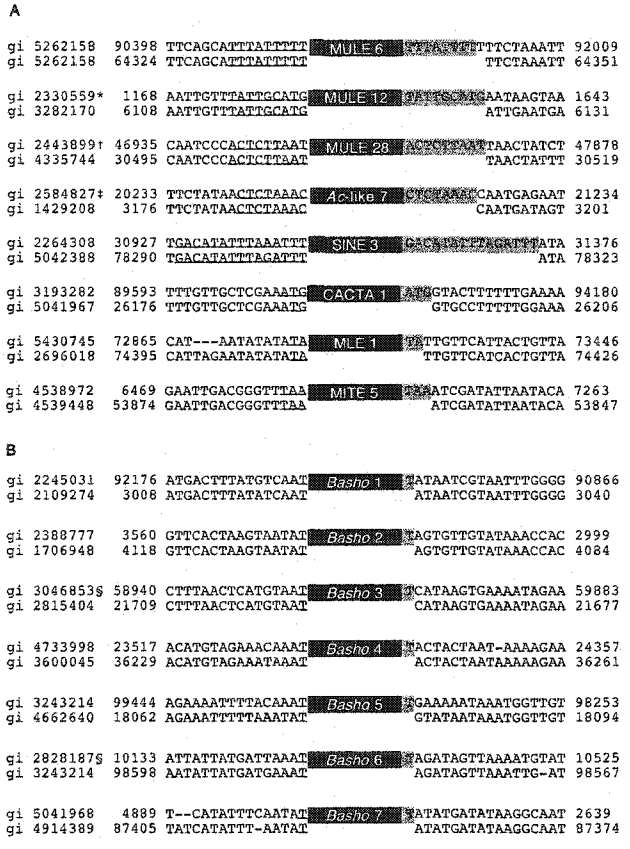

A
gi 5262158 90398 TTCAGCATTTATTTTT MULE 6 TTTCTAAATT 92009
gi 5262158 64324 TTCAGCATTTATTTTT TTCTAAATT 64351
gi 2330559* 1168 AATTGTTTATTGCATG MULE 12 AATAAGTAA 1643
gi 3282170 6108 AATTGTTTATTGCATG ATTGAATGA 6131
gi 2443899† 46935 CAATCCCACTCTTAAT MULE 28 TAACTATCT 47878
gi 4335744 30495 CAATCCCACTCTTAAT TAACTATTT 30519
gi 2584827‡ 20233 TTCTATAACTCTAAAC Ac-like 7 CAATGAGAAT 21234
gi 1429208 3176 TTCTATAACTCTAAAC CAATGATAGT 3201
gi 2264308 30927 TGACATATTTAAATTT SINE 3 ATA 31376
gi 5042388 78290 TGACATATTTAGATTT ATA 78323
gi 3193282 89593 TTTGTTGCTCGAAATG CACTA 1 GTACTTTTTTGAAAA 94180
gi 5041967 26176 TTTGTTGCTCGAAATG GTGCCTTTTTGGAAA 26206
gi 5430745 72865 CAT---AATATATATA MLE 1 TTGTTCATTACTGTTA 73446
gi 2696018 74395 CATTAGAATATATATA TTGTTCATCACTGTTA 74426
gi 4538972 6469 GAATTGACGGGTTTAA MITE 5 ATCGATATTAATACA 7263
gi 4539448 53874 GAATTGACGGGTTTAA ATCGATATTAATACA 53847
B
gi 2245031 92176 ATGACTTTATGTCAAT Basho 1 ATAATCGTAATTTGGGG 90866
gi 2109274 3008 ATGACTTTATATCAAT ATAATCGTAATTTGGGG 3040
gi 2388777 3560 GTTCACTAAGTAATAT Basho 2 AGTGTTGTATAAACCAC 2999
gi 1706948 4118 GTTCACTAAGTAATAT AGTGTTGTATAAACCAC 4084
gi 3046853§ 58940 CTTTAACTCATGTAAT Basho 3 CATAAGTGAAAATAGAA 59883
gi 2815404 21709 CTTTAACTCATGTAAT CATAAGTGAAAATAGAA 21677
gi 4733998 23517 ACATGTAGAAACAAAT Basho 4 ACTACTAAT-AAAAGAA 24357
gi 3600045 36229 ACATGTAGAAATAAAT ACTACTAATAAAAAGAA 36261
gi 3243214 99444 AGAAAATTTTACAAAT Basho 5 GAAAAATAAATGGTTGT 98253
gi 4662640 18062 AGAAATTTTTAAATAT GTATAATAAATGGTTGT 18094
gi 2828187§ 10133 ATTATTATGATTAAAT Basho 6 AGATAGTTAAAATGTAT 10525
gi 3243214 98598 AATATTATGATGAAAT AGATAGTTAAATTG-AT 98567
gi 5041968 4889 T--CATATTTCAATAT Basho 7 ATATGATATAAGGCAAT 2639
gi 4914389 87405 TATCATATTT-AATAT ATATGATATAAGGCAAT 87374

Figure 2.2. A majority of mined transposons show an insertion preference for A+T-rich regions.

The number of transposons used in the calculations is indicated in parentheses. The average G+C content was determined using a 20 bp sliding window over 2000 bp of sequences flanking the transposon insertion sites. Only two CACTA-like elements were used resulting in a low signal to noise ratio. Individual groups (see Table 1) of class I elements showed a similar profile (data not shown) as indicated for the entire class I average.

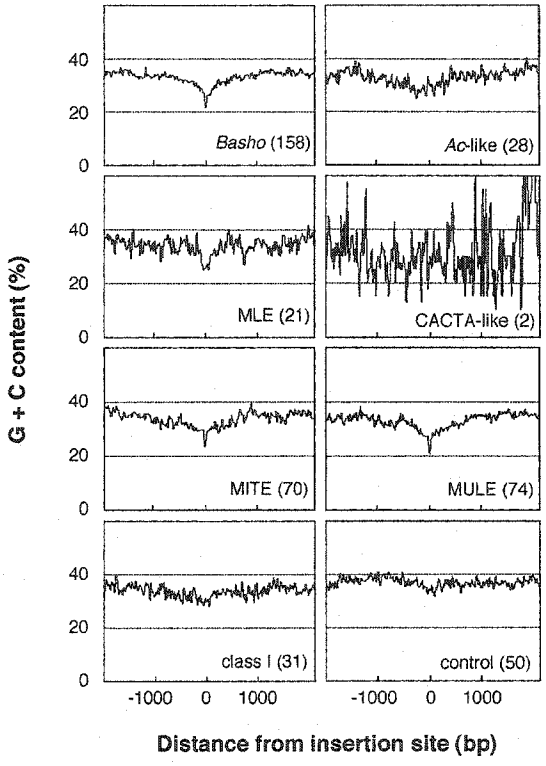

G + C content (%)
40
20
0
Basho (158)
Ac-like (28)
MLE (21)
CACTA-like (2)
MITE (70)
MULE (74)
class I (31)
control (50)
-1000
0
1000
Distance from insertion site (bp)

Figure 2.3. Structure of an *Arabidopsis* Tc1/*mariner*-like transposon.

(*A*) Similarities between TIRs and TSDs (underlined) of an *Arabidopsis* MLE I member and *Tc1/Mariner*-like elements *Pogo* (Drosophila, gi 8354) and *Tigger* (human, gi 2226003). (*B*) Putative transposase for the *Arabidopsis* MLE I (gi 4262216) aligned with transposases from *Drosophila melanogaster Pogo*R11 (gi 2133672) and from human *Tigger*1 (gi 2226004). Amino acid residues are shaded based on the level of structural and functional similarities. Residues conserved between three or two sequences are shaded black and grey, respectively. The arrow (⇑) indicates the predicted start of the *Arabidopsis* MLE I ORF as annotated in GenBank. The first methionine of the *Arabidopsis* MLE I transposase was inferred from the reading frame and sequence similarity with the human *Tigger*1 element. The stop (*) was introduced by a single nucleotide substitution (at position 85709 in gi 4262209) from GAG (glutamine) to TAG (stop).

A

```
MLE I      TA CAGTAAAACCTCTATAAATTAATA...TATTAATTTATAGAGGTTCTACTG TA
Tigger1    TA CAGGCATACCTCG.........................CGAGGTATGCCTG TA
PogoR11    TA CAGTATAATTCGCTTAGCTGC.........GCAGCTAAGCGAATTATACTG TA
```

B

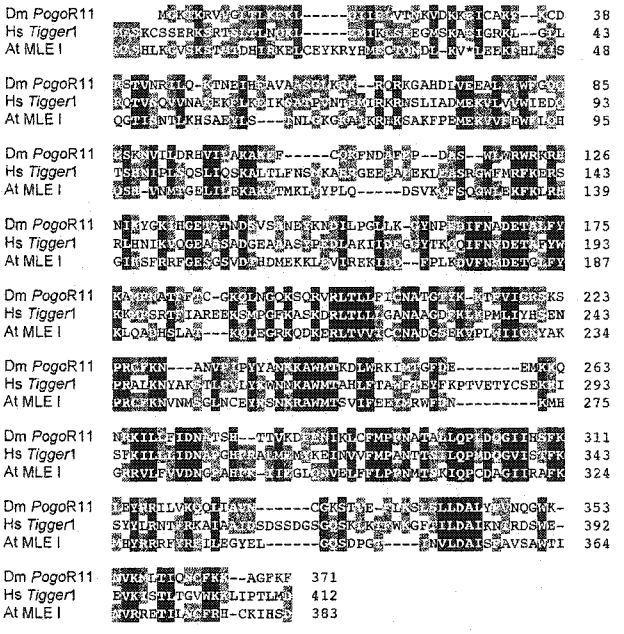
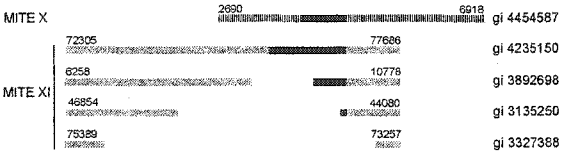


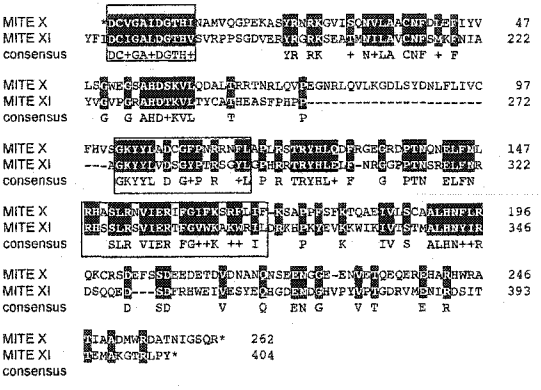
Figure 2.4. MITE transposases.

(*A*) Diagram depicting structural similarities between members of MITE XI elements (grey boxes) and a member of MITE X (striped box). Black boxes represent ORFs corresponding to a putative transposase: MITE XI ORFs share > 98% amino acid sequence similarity. MITE X and XI are distinct groups as no significant internal nucleotide sequence similarity is observed between MITE X and XI nor for the consensus sequence of their TIRs (5'-GG(G/T)GGTGTTATTGGTT-3' for MITE X and 5'-GGCCCTGTTTTGTTTG-3' for MITE XI). However, putative transposase ORFs, length of TIRs (16 bp) and TSD (5'-TTA-3') of MITE X and XI are similar indicating related mechanism of transposition and possibly a common origin for both groups. GenBank gi numbers from which elements were mined are indicated to the right. Nucleotide positions of transposon termini are indicated above each element. (*B*) Amino acid sequence similarity between the conceptual translation for MITE X (gi 4454587, nucleotide position 4650-5865) and the corresponding region of MITE XI ORF (gi 4585884). Identical amino acids and functionally or structurally related residues are shaded in black. Asterisks (*) represent translational stops. Boxed sequences indicate the region containing the DDE motif. (*C*) Alignment of conserved regions corresponding to the functionally important DDE motif found in transposases and integrases of many transposable elements (*39-42*). Transposases are from MITE X (gi 4454587, conceptual translation of nucleotides 4650-5865), MITE XI (gi 4585884), IS*5S* (gi 1256580) from *Synechocystis* sp., IS*493* (gi 1196467) and IS*903* (gi 136129) from *E. coli*. Amino acid residues are shaded based on the level of structural and functional similarities. Residues conserved between all, four or three sequences are shaded black, dark grey and light grey, respectively.

A


MITE X
2690
6918
gi 4454587
MITE XI
72305
77606
gi 4235150
6258
10778
gi 3892698
46854
44080
gi 3135250
75389
73257
gi 3327388

B


MITE X
MITE XI
consensus

C

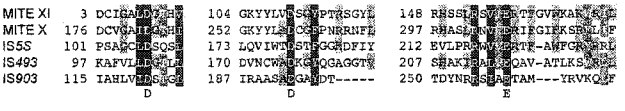
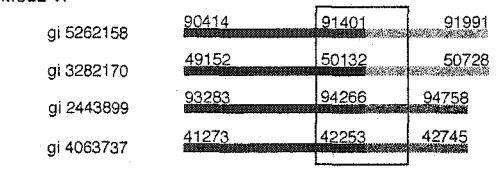

MITE XI
MITE X
IS55
IS493
IS903

Figure 2.5. Inter-element recombination in non-TIR MULEs generates mosaic transposons.

(*A*) Diagrammatic alignment of four members of MULE-VI. Black and grey boxes delimit regions of nucleotide sequence similarity (>90%) between members of MULE groups VI. GenBank gi numbers are marked to the left. Nucleotide positions on clones are noted above each element. (*B*) Expanded view (as boxed in *A*) indicating the sequence of the putative recombination region.

A


MULE VI
gi 5262158
90414
91401
91991
gi 3282170
49152
50132
50728
gi 2443899
93283
94266
94758
gi 4063737
41273
42253
42745

B

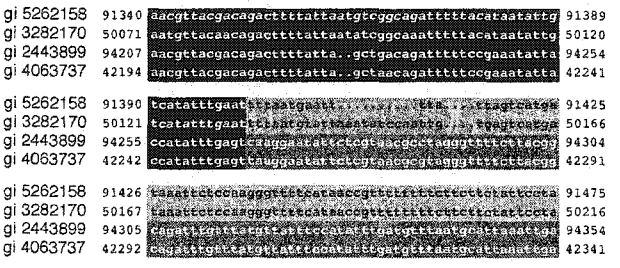
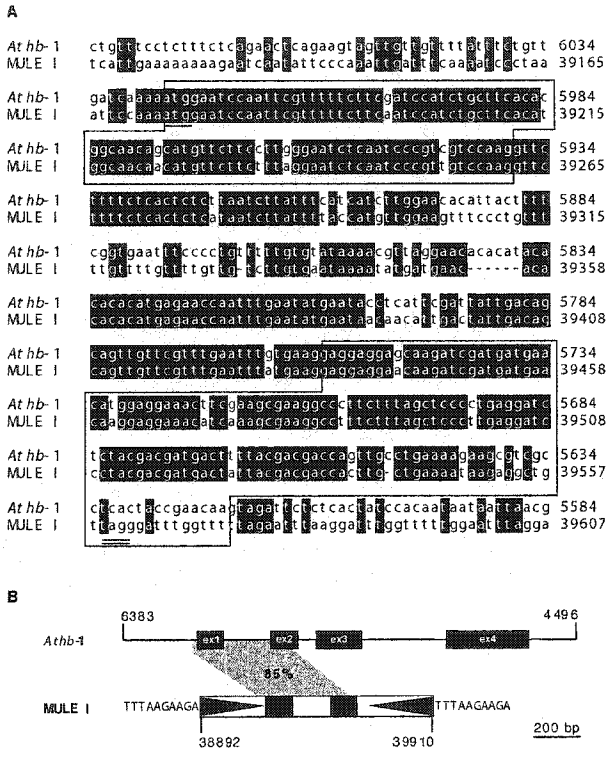

gi 5262158 91340 91389
gi 3282170 50071 50120
gi 2443899 94207 94254
gi 4063737 42194 42241
gi 5262158 91390 91425
gi 3282170 50121 50166
gi 2443899 94255 94304
gi 4063737 42242 42291
gi 5262158 91426 91475
gi 3282170 50167 50216
gi 2443899 94305 94354
gi 4063737 42292 42341

Figure 2.6. Acquisition of a truncated cellular gene by a member of MULE-I.
(*A*) MULE-I (gi 2182289) shares 85% nucleotide similarity with the first two exons and first intron of the homeobox gene *Athb*-1 (gi 6016704). Conserved nucleotides are shaded in black and position on clones are indicated on the right of the alignment. Sequences for the first two exons are boxed. The first ATG is underlined and an in frame stop codon of the MULE-I ORF is double underlined. (*B*) Diagram illustrating the regions of nucleotide similarity (85%) between the MULE-I and the genomic sequence of *Athb*-1 (shaded in grey). Positions for the *Athb*-1 gene and MULE-I element on their respective clones are indicated. MULE-I TIRs are represented by black triangles and the TSDs sequences are indicated flanking both termini.


A
Athb-1 ctgttcctctttctcgagcccagaagtaggtgtggtttatgtgtgtt 6034
MULE I tcattgaaaaaaaagagtcgaattcccaaattgaattgcaagatccctaa 39165
Athb-1 gatcaaaaatggaatccaattcgttttcttcgatccatctgcttcacac 5984
MULE I atccaaaatggaatccaattcgttttcttcaatccatctgcttcacat 39215
Athb-1 ggcaacagcatgttcttcccttgggaatctcaatcccgtcgtccaaggttc 5934
MULE I ggcaacaacatgttcttctttagggaatctcaatcccgttgtccaaggttc 39265
Athb-1 ttttctcactctctaaatcttatttcatcatcttggaacacattacttttt 5884
MULE I ttttctcactctcataatcttatttttaccatgttggaagtttccctgttt 39315
Athb-1 cggtgaatttccccctgttttgtgtataaaacgttaggaacacacataca 5834
MULE I ttgttttgttttgtttg-tcttgtgaataaaatatgatgaac------aca 39358
Athb-1 cacacatgagaaccaatttgaatatgaatacctcatccgattattgacag 5784
MULE I cacacatgagaaccaatttgaatatgaataacaacattgactattgacag 39408
Athb-1 cagttgttcgtttgaatttgtgaaggaggaggagcaagatcgatgatgaa 5734
MULE I cagttgttcgtttgaatttatgaaggaggaggaacaagatcgatgatgaa 39458
Athb-1 catggaggaaactcgaagcgaaggcccttctttagctcccctgaggatc 5684
MULE I caaggaggaaacatcaaagcgaaggccttctttagctcccttgaggatc 39508
Athb-1 tctacgacgatgacttttacgacgaccagttgcctgaaaagaagcgtcgc 5634
MULE I cctacgacgatgacttattacgacgaccacttc-ctgaaaataaggaggctg 39557
Athb-1 cccactaccgaacaagtagattcctcctcactaaccacaatgaatgaattaacg 5584
MULE I ttagggatttggttt-tagaatttaaggatttggttttgggaatttaggga 39607
B
6383
4496
Athb-1
ex1
ex2
ex3
ex4
85%
MULE I
TTTAAGAAGA
TTTAAGAAGA
38892
39910
200 bp

REFERENCES

1. Berg D. E. & Howe, M. M., eds. (1989) *Mobile DNA* (American Society for Microbiology, Washington, DC).
2. Britten, R. J. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 9374 -9377.
3. Kidwell, M.G. & Lisch, D. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 7704-7711.
4. McDonald, J. F. (1995) *Trends Ecol. Evol.* **10,** 123-126.
5. Jordan, I. K. & McDonald, J. F. (1999) *Genetics* **151,** 1341-1351.
6. Jurka, J. (1998) *Curr. Opin. Struct. Biol.* **8,** 333-337.
7. Smit, A. F. A. (1996) *Curr. Opin. Genet. Dev.* **6,** 743-748.
8. Leutwiler, L. S., Hough-Evans, B. R. & Meyerowitz, E. M. (1984) *Mol. Gen. Genet.* **194,** 15-23.
9. Surzycki, S. A. & Belknap, W. R. (1999) *J. Mol. Evol.* **48,** 684-691.
10. Wright, D. A. & Voytas, D. F. (1998) *Genetics* **149,** 703-715.
11. Lin, X., Kaul, S., Rounsley, S., Shea, T. P., Benito, M.-I., Town, C. D., Fujii, C. Y., Mason, T., Bowman, C. L., Barnstead, M. *et al.* (1999) *Nature (London)* **402,** 761-768.
12. Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, A., Stiekema, W., Entian, K.-D., Terryn, N. *et al.* (1999) *Nature (London)* **402,** 769-777.
13. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403-410.
14. Casacuberta, E., Casacuberta, J. M., Puigdomènech, P. & Monfort, A. (1998) *Plant J.* **16,** 79-85.
15. Bevan, M., Bancroft, I. Bent, E., Love, K., Goodman, H., Dean, C., Bergkamp, R., Dirkse, W., Van Staveren, M., Stiekema W. *et al.* (1998) *Nature (London)* **391,**485-488.
16. Doutriaux, M. P., Couteau, F., Bergounioux, C. & White, C. (1998) *Mol. Gen. Genet.* **257,** 283-291.
17. Chye, M.-L., Cheung, K.-Y. & Xu, J. (1997) *Plant Mol. Biol.* **35,** 893-904.

18. Konieczny, A., Voytas, D. F., Cummings, M. P. & Ausubel, F. M. (1991) *Genetics* **127,** 801-809.
19. Pélissier, T., Tutois, S., Deragon, J. M., Tourmente, S., Genestier, S. & Picard G. (1995) *Plant Mol. Biol.* **29,** 441-452.
20. Tsay, Y.-F., Frank, M. J., Page, T., Dean, C. & Crawford, N. M. (1993) *Science* **260,** 342-344.
21. Klimyuk, V. I. & Jones, J. D. (1997) *Plant J.* **11,** 1-14.
22. Wright, D. A., Ke, N., Smalle, J., Hauge, B. M., Goodman, H. M. & Voytas, D. F. (1996) *Genetics* **142,** 569-578.
23. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673-4680.
24. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389-3402.
25. Nicholas, K. B., Nicholas, H. B. Jr. & Deerfield, D. W. II (1997) *EMBNEW.NEWS* **4,** 14.
26. Bureau, T. E. & Wessler, S. R. (1992) *Plant Cell* **4,** 1283-1294.
27. Gilbert, N. & Labuda, D. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2869-2874.
28. Yoshioka, Y., Matsumoto, S., Kojima, S., Oshima, K., Okada, N. & Machida, Y. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 6562-6566.
29. Xiong, Y. & Eickbush, T. H. (1990) *EMBO J.* **9,** 3353-3362.
30. Walbot, V. (1991) in *Genetic Engineering*, ed. Setlow, J. K. (Plenum Press, New York), pp. 1-37.
31. Orgel, L. E. & Crick, F. H. C. (1980) *Nature (London)* **284,** 604-607.
32. SanMiguel, P., Tikhonov, A., Y.-K. Jin, Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z. *et al.* (1996) *Science* **274,** 765-767.
33. Vaury, C., Bucheton, A. & Pelisson, A. (1989) *Chromosoma* **98,** 215-224.
34. Copenhaver, G. P., Nickel, K., Kuromori, T., Benito, M.-I., Kaul, S., Lin, X., Bevan, M., Murphy, G., Harris, B., Parnell, L. D. *et al.* (1999) *Science* **286,** 2468-2474.
35. Barakat, A., Matassi, G. & Bernardi, G. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 10044-10049.

36. Fedoroff, N. V. (1989) *Cell* **56,** 181-191.
37. Fischer, S. E. J., van Luenen, H. G. A. M. & Plasterk, R. H. A. (1999) *Mol. Gen. Genet.* **262,** 268-274.
38. Wessler, S. R., Bureau, T. E. & White, S. E. (1995) *Curr. Opin. Genet. Dev.* **5,** 814-821.
39. Grindley, N. D. F., & Leschziner, A. E. (1995) *Cell* **83,** 1063-1066.
40. Lohe, A. R., De Aguiar, D. & Hartl, D. L. (1997) *Proc. Natl. Acad. Sci. USA.* **94,** 1293-1297.
41. Capy, P., Vitalis, R., Langin, T., Higuet, D. & Bazin, C. (1996) *J. Mol. Evol.* **42,** 359-368.
42. Tavakoli, N. P., DeVost, J. & Derbyshire, K. M. (1997) *J. Mol. Biol.* **274,** 491-504.
43. Bureau, T. E. & Wessler, S. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1411-1415.
44. Kidwell, M. (1992) *Curr. Opin. Genet. Dev.* **2,** 868-873.
45. Ruberti, I., Sessa, G., Lucchetti, S. & Morelli, G. (1991) *EMBO J.* **10,** 1787-1791.
46. Becker, H.-A. & R. Kunze, R. (1996) *Mol. Gen. Genet.* **251,** 428-435.
47. van Drunen, C. M., Oosterling, R. W., Keultjes, G. M., Weisbeek, P. J., van Driel, R. & Smeekens, S. C. M. (1997) *Nucleic Acids Res.* **25,** 3904 -3911.
48. Eisen, J. A., Benito, M.-I. & Walbot, V. (1994) *Nucleic Acids Res.* **22,** 2634-2636.
49. Karp, P. D. (1998) *Bioinformatics* **14,** 753-754.

CHAPTER 3

DATA MINING II: TERMINAL-REPEAT RETROTRANSPOSONS IN MINIATURE (TRIM) ARE INVOLVED IN RESTRUCTURING PLANT GENOMES

The data gathered from the initial survey of the *Arabidopsis* genome has revealed the presence of novel elements. Among these, *Basho* differs the most from known classes of elements but another group of repetitive elements, called *Katydid*, was also revealed to be quite unique (The Arabidopsis Genome Initiative, 2000). Structurally, *Katydid* elements are reminiscent of LTR-retrotransposons and may represent a "minimal-requirement" version of these elements. Closer examination of the members of this new group revealed that they are often closely associated with cellular genes and that they may have played a role in reshaping genes and genomes. At the same time, Dr. Witte and Dr. Kumar in the United Kingdom had isolated a similar element in potato. Realizing that we had stumbled upon a common group of plant transposons, we united our efforts in studying these new elements. I examined *Katydid* in *Arabidopsis* while the UK group focused on other plants. Following is the result of our collaboration, which was published in the *Proceedings of the National Academy of Sciences* USA (Witte *et al.*, 2001).

Reference

THE ARABIDOPSIS GENOME INITIATIVE (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815.

WITTE, C.-P.*, LE, Q.H.*, BUREAU, T. AND KUMAR, A. (2001). Terminal-repeat Retrotransposons in Miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl. Acad. Sci. USA* **98**, 13778-13783.

* C.-P.W. and Q.-H.L. contributed equally to this work.

ABSTRACT

A new group of long terminal repeats (LTR) retrotransposons; termed terminal-repeat retrotransposons in miniature (TRIM) are described, which are present in both monocotyledonous and dicotyledonous plant. TRIM elements have terminal direct repeat sequences between ≈100 to 250-bp in length that encompass an internal domain of ≈100 to 300-bp. The internal domain contains primer binding site (PBS) and polypurine tract (PPT) motifs, but lacks the coding domains required for mobility. Thus, TRIM elements are not capable of autonomous transposition and probably require the help of mobility-related proteins encoded by other retrotransposons. The structural organization of TRIM elements suggests an evolutionary relationship to either LTR retrotransposons or retroviruses. The past mobility of TRIM elements is indicated by the presence of flanking 5-bp direct repeats typically found at LTR retrotransposon insertion sites, by the high degree of sequence conservation between elements from different genomic locations, and by the identification of related to empty sites (RESites). TRIM elements appear to be actively involved in the restructuring of plant genomes, affecting the promoter, coding region and intron-exon structure of genes. In solanaceous species and in maize, TRIM elements provided target sites for further retrotransposon insertions. In *Arabidopsis thaliana,* evidence is provided that TRIM elements can also be involved in the transduction of host genes.

INTRODUCTION

Transposons, also known as mobile elements or transposable elements, are discrete genetic sequences found in nearly all eukaryotic genomes. As their name indicates, transposons have the ability to move from one location of the genome to another, generating target site duplications (TSD) upon insertion. Autonomous transposons encode the enzymes necessary for their mobilization whereas non-autonomous elements require these proteins to be provided in *trans* by an autonomous element (reviewed in ref.*1*).

Retrotransposons are a class of transposable element, which move via an RNA intermediate that is reverse transcribed into extrachromosomal DNA and inserted into the genome by their encoded reverse transcriptase, RNaseH and integrase enzymes (reviewed in *2, 3*). This replicative mode of transposition is similar to that used by retroviruses. However, unlike the retroviruses, which are usually found in animals, retrotransposons are found in all eukaryotes, where they are the most abundant class of mobile DNA (reviewed in *4, 5*). Retrotransposons are classified into two types, those with long terminal repeats (LTRs) and those without LTR (non-LTR retrotransposons). LTR retrotransposons are further sub-classified into the Ty1-*copia* and Ty3-*gypsy* groups that differ from each other in distinct sequence features and in the order of encoded gene products (*6, 7*). Retrotransposons make up a large portion of many eukaryotic genomes, including humans (*8, 9*) and a variety of plants (*4, 10,11*).

Many properties of transposable elements suggest that they are parasitic or selfish DNAs (*12, 13*). However, like any other component of a heritable genome, transposon DNA can serve as a genomic resource for mutation and natural selection. Indeed, a more recent paradigm suggests that transposable elements may play a central role in the evolution of gene function and genome structure in eukaryotic organisms (for reviews see *1, 4, 5, 14-16*). For example, transposons can contribute regulatory *cis*-factors (*17, 18*), alter transcript splicing (*19, 20*), promote exon shuffling (*21*) and facilitate chromosomal rearrangements or restructuring (*22, 23*) leading to changes in spatial and temporal expression patterns of host genes or even generation of novel genes. Furthermore, retrotransposons are involved in altering the genome size of eukaryotic organisms either

by increasing (*10, 11*) or by decreasing (*24-26*) their copy numbers within the host genome.

The properties and abundance of transposable elements make it important to identify and characterize the whole spectrum of mobile elements that inhabit the eukaryotic genomes and to investigate their origin, interactions with host genomes and their contributions to the evolution of host genes and genomes. In this report, a unique group of mobile retrotransposons is described. We provide evidence that these elements have some features typical of LTR retrotransposons but are unusually small and lack coding capacity for mobility related proteins. We also demonstrate that they are present in both dicotyledonous and monocotyledonous plants, and have been active in restructuring plant genomes.

MATERIALS AND METHODS

Data Analysis. TRIM elements were initially identified during sequence analysis of a genomic clone containing the *Solanum tuberosum* (potato) urease gene (Witte, Tiller, Davies & Kumar, unpublished) and during the analysis of the genome sequence of *Arabidopsis*, where they were referred to as *Katydid* (*27*). In potato, a DOTPLOT (*28*) revealed closely spaced direct repeats within intron 6 of the urease gene. Closer inspection indicated that the direct repeats were flanked by a 5-bp target site duplication (TSD) and contained a primer binding site (PBS) and a polypurine tract (PPT) within the region between the repeats. Initial DNA-database searches with the similarity search algorithm BLAST (*29*) using the complete element as query revealed the presence of similar sequences at several genomic locations in a number of solanaceous species. Independently, TRIM elements were identified during a survey of transposons in the Arabidopsis genome as part of the Arabidopsis Genome Initiative (AGI) annotation effort (*27*). Iterative database analyses (with BLAST) using TRIM sequences from previous searches as queries for new searches identified TRIM elements from several monocotyledonous and dicotyledonous plant species. In this search strategy all hits against the database were manually examined for typical features of TRIMs (*e.g.* a sequence of less than 540 bp containing direct repeats enclosing PBS and PPT motifs). Thereby, elements with some degree of truncation were also found for which the termini could not be accurately mapped due to deletions and/or sequence degeneracy. Sequence information of large insert clones were accessed from public sequence databases (GenBank, http://www.ncbi.nlm.nih.gov/Genbank/index.html and the European Molecular Biology Laboratory at http://www.ebi.ac.uk/embl/index.html) between October 2000 and May 2001. Sequence analysis was performed using programs of the UWGCG software package (University of Wisconsin Genetics Computer Group; version 10; http://www.gcg.com) or the EMBOSS program package (European Molecular Biology Open Software Suite; http://www.uk.embnet.org/Software/EMBOSS/). Multiple alignments were generated with PILEUP (UWGCG) or CLUSTALW (*30*). Graphical manipulations were carried out using Genedoc (*31*). TRIM elements were mapped onto

the assembled genomic sequences of Arabidopsis accessed at AGI (release of March 2001, http://www.*Arabidopsis*.org/agi.html).

RESite Analysis. Sequences which were similar to the region immediately flanking an insertion, but which did not harbour the insertion and had only one copy of the target site were defined as related to empty sites, or RESites (*32*). RESites can be used to delimit the termini of the insertion, confirm the size and sequence of the TSD and provide supporting evidence for past mobility. RESites were identified by using the sequences immediately flanking an insertion as a query in a BLAST search as previously described (*32*).

RESULTS

Identification, characterization and distribution of TRIM elements in plants

Database searches and sequence analyses revealed the presence of a family of repetitive sequences from a number of plant species, which share a unique set of characteristics. Specifically, these elements are short in size (mostly around 350 bp), have terminal direct repeats (TDRs) of 100-250 bp (on average, <140-bp), that flank an internal domain containing a primer binding site (PBS; located immediately downstream of the 5' terminal repeat and complementary to the methionine tRNA) and a polypurine tract (PPT; located immediately upstream of the 3' terminal repeat) (Fig.3.1*A*). None of the elements characterized to date encodes mobility related proteins in the internal domain (Fig.3.1). Intact elements are flanked by 5-bp direct repeats, representing the TSD that was generated upon insertion. These previously uncharacterized mobile elements have been named terminal-repeat retrotransposons in miniature or TRIM (Fig.3.1). TRIM elements are present in multiple copies and are dispersed within the host chromosomes (supplemental table 3.2). TRIM elements appear to be ubiquitous in the plant kingdom as they have been found in both dicotyledonous and monocotyledonous species (Table 3.1). To date, TRIM elements from the plant species examined have similar overall lengths and structures, and show some sequence similarity, especially in the TDRs (Fig.3.1 and 2). For example, there is over 80% and 90% sequence identity among the TDRs of TRIM elements in *Arabidopsis* and rice, respectively. Furthermore, TDRs of TRIM elements from distantly related species show significant similarity (Fig.3.2) ranging from 60-75% at the sequence identity level. In brief, TRIM elements are distinguished by their exceptionally small overall size (<540 bp) and by their presence in multiple copies in different locations within the host genome with highly conserved TDRs and short internal sequences (~100-300 bp). Moreover, despite the complete lack of coding capacity for mobility-related proteins (which makes it impossible to classify TRIM elements conventionally into Ty1-*copia* or Ty3-*gypsy* groups) they are able to transpose *in trans*, making them the smallest active LTR-retrotransposons known to date.

In *A. thaliana*, a total of 43 TRIM elements were identified (Table 3.1) and subdivided into 3 groups based on overall sequence similarity - *Katydid*-At1, 2, and 3.

Thirty-one copies of *Katydid*-At1 are present including 12 truncated elements (*i.e.* termini could not be accurately mapped). *Katydid*-At2 is present in 7 copies including 3 truncated elements while there are 5 copies of *Katydid*-At3, of which two are truncated. *Katydid*-At1 elements have TDRs of 116-bp (Figure 3.1*B*). The left and right TDRs are nearly identical in many elements. Because LTRs are initially identical upon element insertion, the high sequence identity between the TDRs suggests that *Katydid*-At1 elements were recently, and are perhaps still, active (*33*). The fact that three related to empty sites (RESites) were identified (Fig.3.3) and the high nucleotide similarity between some *Katydid*-At1 elements (up to 98%; Fig.3.1*B*) also attest to recent mobility. In contrast, similarity between members or between the TDRs of *Katydid*-At2 or 3 was lower, indicating that they may be ancient insertions (supplemental table 3.2). *Katydid*-At2 elements have TDRs similar to *Katydid*-At1 in size (~115-bp) but with only limited similarity in sequence (Fig.3.2). Although one *Katydid*-At3 element shared structural characteristics with other TRIM elements, two members had unusual features (supplemental table 3.2). For example, one element (gi 4263586) has terminal inverted repeats rather than TDRs. Because a perfect 5-bp flanking TSD can be identified, it is possible that this unique element was mobilized with this structure. The other element is truncated at one end yet still appears to have a 5 bp TSD. Structural rearrangements were also observed within other *Arabidopsis Katydid* elements. Although most *Katydid*-At1 elements have TDRs, "solo" and "triple" direct repeat structures were identified (supplemental figure 3.6). Solo and triple direct repeat elements are also immediately flanked by a 5-bp TSD. Solo direct repeats were most likely generated by illegitimate recombination between TDRs of an intact *Katydid*-At1 element (*10, 24, 25*). The middle repeat of the triple direct repeats containing element is unlikely to be a solo direct repeat as flanking TSDs are not present. Within the triple repeats the terminal direct repeat share 93% with each other but showed lower (89% and 90%) similarity with the middle direct repeat (gi 3150006). It is currently unclear as to the precise mechanisms that generated these unusual TRIM elements.

Unlike other *Arabidopsis* retrotransposons, which cluster in the pericentromeric and centromeric regions where gene density is low, *Katydid*-At1 is dispersed throughout the genome (supplemental table 3.2). In addition to the elements mined from nuclear

sequences, we have identified a *Katydid*-At1 insertion in the mitochondrial genome of *Arabidopsis* (gi 1785729). Like most elements found in the mitochondrial genome of *Arabidopsis* (*34*), this *Katydid*-At1 element is truncated and degenerate. Transposable elements make up approximately 4% of the total mitochondrial DNA and are predominately retrotransposon fragments, which originated from reverse transcripts of nuclear genomic elements (*34*). Similarly, we consider that the mitochondrial *Katydid*-At1 element is likely to be of nuclear origin since it shares sequence similarity with nuclear insertions and because no other such elements could be identified in mitochondria genomes.

In rice, TRIM elements from different genomic locations are highly similar, having nearly identical TDRs and highly conserved internal domains (Fig.3.1*C*). This attests to the recent activity of these elements in the rice genome. At the date of this analysis (March 2001), 16 rice TRIM elements could be found in the sequence database. There were 8 complete and 3 truncated elements, and 5 were only present as solo-LTRs (Table 3.1; supplemental table 3.2). Like *Arabidopsis Katydid* elements, the rice elements are also located on a variety of chromosomes and some of the insertions are near or within putative genes and overlap with predicted ORFs (*e.g.* gi 9711848 and 5777612 in supplemental table 3.2). However, only one expression sequence tag (EST; coding for a protein with similarity to leucine rich repeats (LRR)-like proteins; gi 4715547) with partial similarity to a terminal direct repeat of a rice TRIM was identified. In maize, only one complete TRIM element has been found to date. This element possesses unusually long TDRs of approximately 230-bp and a short internal domain of 72-bp. However, like TRIM elements from other plants, this element contains a PBS and a PPT motif and 5-bp TSDs.

In contrast to *A. thaliana* and rice, only limited genomic sequence data is available for members of the family solanaceous and leguminous species (Table 3.1). In both groups ancient TRIM insertions were found (supplemental table 3.2; gi 9392606), including truncated elements (supplemental table 3.2; gi 19200 and 170430), solo-LTRs (supplemental table 3.2; gi 21590 and 8747823) and more complex structures, probably generated by several insertion-deletion events involving more than one element (supplemental table 3.2; gi 9251206 and 10764221). However, TRIM elements with well-

conserved sequence were also identified (supplemental table 3.2; gi 9562370 and 19642), indicating that TRIM elements have a long evolutionary history in these plants and that some may be still active.

Retrotransposons nested within TRIM elements

The TRIM in intron 6 of one of the alleles of the urease gene from potato (*Solanum tuberosum*, cv. Désirée (gi 14599411) contains an insertion of a truncated LTR retrotransposon of the Ty1-*copia* group (supplemental table 3.2) . In the wild species *Solanum ochranthum,* the TRIM in intron 6 of urease sustained a LINE insertion, a non-LTR retrotransposon, 30-bp downstream of the Ty1-*copia* insertion found in potato (gi 14599445; supplemental table 3.2). In both cases the insertion was located within the internal sequence of the TRIM. Another case for TRIM elements being targeted by a Ty1-*copia* retrotransposons was found in the maize genome. In this case, a complete element of approximately 9-kb inserted within the 3' terminal direct repeat of a TRIM element (supplemental table 3.2).

Insertion within genes

TRIM elements contribute to coding and untranslated regions of plant genes. However, in many cases this notion is based on the observation that TRIM elements overlap with sequences predicted to contain a gene by computer-assisted intron/exon assignment. Although experimental evidence confirming the presence of a gene at predicted locations is often still lacking, the frequent presence of TRIM elements in ESTs from species of the Solanaceae and Fabaceae (Table 3.1; supplemental table 3.2) demonstrates that insertion into transcribed coding sequences is common at least in some species. In some cases, ESTs provided evidence for the insertion of TRIM elements into coding sequences with similarity to known genes, for example glycolate oxidase in tomato (gi 10902684), a nucleic acid binding domain (NBS)-LRR-like protein in potato (gi 9562370) or a P450-like protein in *M. trunculata* (gi 10520041). In solanaceous and leguminous species TRIM elements were also found in promoter regions and introns of genes (supplemental table 3.2). For example, in potato and *S. ochranthum* elements were found in intron 6 of the urease gene (gi 14599411, gi 14599414, and gi 14599445) in *M.*

sativa in intron 9 of the nodulin-25 gene (gi 19642), in potato in the promoter region of the proteinase inhibitor II gene (gi 21553), and in *P. vulgaris* a TRIM solo-terminal repeat was found in the promoter region of the chalcone isomerase gene (gi 20981; supplemental table 3.2). In contrast, TRIM elements are rarely found in ESTs from *Arabidopsis* and rice, although they often overlap with computer-predicted genes (supplemental table 3.2).

Two examples of *Katydid*-At1 elements altering host genes in *Arabidopsis* are shown in Fig.3.4 In the first case (Fig.3.4 *A* and *C*) a *Katydid*-At1 element in reverse orientation (gi 3510340, position 22675-22946) overlaps with the transcription and possible translation start site, and the first exon-intron boundary of a gene encoding a cytochrome P450-like protein (gi 3510340, position 19519-22694). That the gene is transcribed and correctly spliced is indicated by the existence of two ESTs corresponding to this gene (gi 5841198 and 1053954). The insertion of *Katydid*-At1, probably at the 5'-terminus of a pre-existing gene, introduced a transcription start site (a TATA motif within the terminal repeat is present about 40-bp upstream of transcription start). The element is truncated in the 3' terminal repeat about 20-bp downstream of the first exon-intron boundary. Thus, the first intron was either created *de novo* or the 5' boundary of an already existing intron was replaced due to the insertion of the element. Additionally, the potential use of a translation start site within the region between the two TDRs introduces 28 amino acids encoded by the element to the N-terminus of the protein. Fig.3.4 *B* shows a case where a *Katydid*-At1 element appears to have created a new intron within a gene, although no EST data to support this hypothesis is available.

Transduction of host genes

A *Katydid*-At1 element located in the *Arabidopsis* genome contains an ORF (gi 3241926) nearly identical to a putative cellular gene annotated as a nonsense-mediated mRNA decay trans-acting factor (NMD), but which is actually more similar to the yeast gene *sen1*. NMD is responsible for the degradation of prematurely terminated mRNAs (*35*) whereas SEN1 is involved in the endonucleolytic cleavage of pre-tRNAs (*36*). Both NMD and SEN1 contain a large domain involved in degrading ribonucleotides (*35*). The ORF linked to the *Katydid*-At1 elements lacks part of the first and second exons and does not contain intron sequences (Fig.3.5). This structure was most likely generated by the

recombination of a *Katydid* element with a mRNA of NMD during the transposition process. The presence of this ORF within *Katydid*-At1 is very reminiscent of the product of cellular gene transduction by retroviruses leading to the formation of oncoviruses (*37*). Previously, cellular gene transduction was documented for another plant retroelement, the maize *Bs1* element (*37-40*; Elrouby and Bureau, in press). Although the cellular (c-NMD) and transduced (r-NMD; where r- refers to retrotransposon) are both located on chromosome 5, they are not tightly linked. Nor are there other *Katydid*-At1 elements near c-NMD in the *A. thaliana* (Columbia ecotype) genomic sequence. This precludes the possibility that transduction resulted from a read-through transcript.

DISCUSSION

Computer-assisted data mining of transposable elements has proven to be the most efficient and informative route to the understanding of not only mobile element evolution but also their role in eukaryotic gene and genome evolution. Although many mined elements reinforce traditional paradigms concerning sequence- and structure-based transposon classification, unusual elements are occasionally mined that force a re-evaluation of element type definition. TRIM elements are one such type of element. Their size (<540-bp) is considerably less than that of typical LTR retrotransposons (5 to 15-kb), which makes them the smallest terminal direct repeat-containing mobile elements, discovered to date. Despite their small size and lack of coding capacity, TRIM elements are obviously mobile as indicated by several observations: (a) the high sequence similarity between elements from different genomic locations (Fig.3.1*B* and *C*), (b) the wide distribution within the individual genomes (supplemental table 3.2), and (c) the identification of RESites (Fig.3.3).

In spite of their small size, TRIM elements have some typical features of LTR retroelements: possessing TDRs, a 5-bp duplication of the target site sequence and PBS and PPT motifs, which are essential for the mobilization of LTR retrotransposons. Furthermore, the solo TDRs of TRIM elements are reminiscent to solo-LTRs and presumably arose by a similar mechanism (*10, 24, 25*). Unlike LTR retrotransposons, however, TRIM elements lack mobility related coding sequences such as *gag* and *pol* domains (Fig.3.1; -.3, -.4). A full-length TRIM with coding capacity for mobility-related proteins has not yet been discovered even though the sequencing of the *Arabidopsis* genome is nearly complete. Lack of an autonomous *Arabidopsis* TRIM element may merely reflect ecotype distribution or stochastic loss. In any case, TRIM elements are probably mobilized in *trans* by other retroelements. Mobility of elements lacking coding capacity is well documented. For example, defective or non-autonomous DNA transposons (*e.g.* class II elements) can be mobilized by functional autonomous element counterparts (*1*). The SINE *Alu* is the most abundant element in the human genome yet, like other SINEs, has no coding capacity (*41*). The *Bs1* elements of the Ty1-*copia* group retrotransposons in maize have been recently active but lack a *pol* domain (*42*). Recently

human LINEs were shown to mobilize transcripts not derived from LINE sequences (*43*). Furthermore, several retroviral vectors have been designed that lack coding capacity and only transpose in the presence of mobility-related proteins *in trans* (*44*). The abundance of streamlined non-autonomous elements, such as SINEs and TRIMs, in eukaryotic genomes may reflect a unique evolutionary relationship with their autonomous counterparts and the host genome.

A TRIM-like structure found in barley (gi 9623334) may have been generated by a combination of transposition and unequal recombination events among two identical LTR retrotransposons (*25*). Although this model would not necessarily account for the presence of PBS and PPT, it suggests a mechanism by which TRIM progenitor might have originally evolved from LTR retroelements. In any case, since retroelement classification relies on internal domains such as *gag* and *pol* and no TRIM-related retroelement containing these domains could be identified, we cannot be certain whether TRIM elements are related to Ty1-*copia* or Ty3-*gypsy*-like retrotransposons, retroviruses, or perhaps originated independently.

The presence of TRIM elements in dicotyledonous and monocotyledonous species indicates that these elements are probably ubiquitous in the plant kingdom. A comparison of the TDR sequences of TRIM elements from seven different plant species indicated significant sequence similarities (Fig.3.2; also see Results). This is surprising given that LTRs typically share little sequence similarity between members of Ty1-*copia* or Ty3-*gypsy* elements in different plant species (see references in *1, 4*). The relative conservation of TDRs of TRIM elements probably reflects selective constraints resulting in the TDRs only containing the minimal set of *cis*-factors necessary for the retrotransposition process. The high level of conservation in structure and size of TRIM elements from a wide range of plant species together with clear indications of mobility strongly suggests that TRIM elements are a novel form of streamlined LTR retrotransposon.

TRIM elements contribute to the restructuring of the host genomes

Retrotransposons have been reported to contribute to the coding regions of genes in plants (see references in *4, 14*). Recent data from the Human Genome Sequence has

also revealed many cases of *bona fide* and predicted proteins that contain L1 and *Alu* sequences (*9, 45*). Like other LTR-retrotransposons, there are indications that TRIM elements have been intimately involved in reshaping genomes and in gene evolution. TRIM elements are found dispersed throughout the genome in *Arabidopsis* and rice (supplemental table 3.2). Because TRIM elements are comparatively small in size, it can be argued that they are less likely to cause serious disruptions of promoter and intron integrity compared to insertions of the much larger LTR retrotransposons. Alternatively, some LTR retrotransposon have been suggested to sustain deletions immediately upon insertion near genes and thereby lessening the impact of transposition events (*14, 19, 46*). TRIM elements might also be able to contribute more effectively to ORFs of genes (Fig.4 and see supplemental data 3.2), especially if the element itself already contains long ORFs, (*e.g.* the *Katydid*-At1 element on gi 3241926). Apart from the potential direct influence of TRIM elements on genes, they appear to act indirectly as target sites for both LTR and non-LTR retrotransposons in *Solanum* plants and maize (supplemental table 3.2). Previously, it has been reported that LTR retrotransposons act as target sites for the insertions of other retrotransposons in maize (*10, 33*) and barley (*24, 25*). This may reflect the propagation of these elements into genomic regions where retrotransposons have already been established and, thus, may limit the deleterious effects of retrotransposon activity on host fitness (*4, 10*).

The ability of TRIM elements to transduce a host gene was documented for the *Katydid*-At1 element in *A. thaliana* (Fig.5). Together with the maize *Bs1* element, these are the only plant examples of retrotransposon-mediated transduction, a phenomenon thought to be limited to the acquisition of proto-oncogenes in viruses (*37-40*). Recent data from the human genome-sequencing project revealed many examples of LINEs, specifically L1 elements, with transduced genomic sequences located in their 3' flanking region (*21, 45, 47, 48*). Approximately 15-21% of these L1 elements harbour downstream sequences between 30 and 970-bp in length, meaning that approximately 1% of the human genome is the result of transduced sequences. Thus, transduction may account in part for the expansion of host genomes, may facilitate exon shuffling and amplified gene sequences may serve as the raw material for the evolution of new genes.

In conclusion, TRIM elements are the smallest known LTR-retrotransposons identified to date and are important players in plant genome evolution. Moreover, like other LTR-retrotransposons, they have been actively involved in the restructuring of plant genomes by acting as target sites for retrotransposon insertions, by altering host gene structure and by also being capable of transducing host genes. Although, only plant TRIM elements have been reported here, it is anticipated that TRIM-like elements would be found in other eukaryotic organisms.

ACKNOWLEDGEMENTS

We thank Nabil Elrouby and John Jones for critical reading of earlier version of this manuscript. AK and CPW acknowledge the funding of the Scottish Executive & Environmental Rural Affairs Department (SEERAD) and CPW also acknowledges the European Commission TMR (Training and Mobility of Researchers) program. This work was partly funded by grants from the National Science and Engineering Research Council (NSERC) of Canada to TB and a David Stewart McGill Majors Fellowship to QHL.

Table 3.1. TRIM elements found in plant sequences

Organism	¶Intact	Solo-LTR	§Trun-cated	Total	*EST
Brassicaceae					
Arabidopsis thaliana	21	5	17	43	1
Solanaceae					
Solanum tuberosum	3	2	3	8	4
S. ochrantum	1	0	0	1	0
Lycopersicon pennellii	0	0	1	1	1
L. esculentum	3	0	7	10	3
Nicotiana tabacum	1	0	0	1	0
Fabaceae					
Lotus japonicus	0	0	2	2	1
Medicago sativa	1	0	0	1	0
M. trunculata	0	7	5	11	9
Phaseolus vulgaris	1	1	0	2	0
Glycine max	0	0	1	1	1
Poaceae					
Oryza sativa	8	4	4	16	1
Zea mays	1	0	1	2	0

¶Elements with small internal deletions were counted as intact.

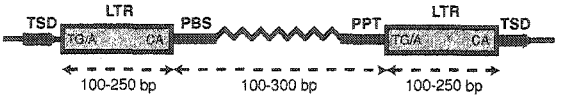
§Elements for which one or both termini could not be resolved.

*Each EST represents a TRIM insertion in a different gene (duplicate ESTs for the same gene were not counted)

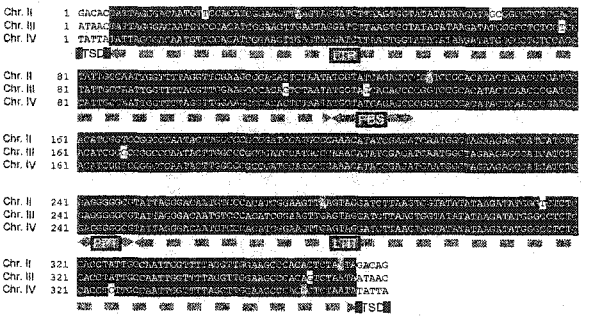
FIGURE LEGENDS

Figure 3.1. Schematic diagram of the general structure of TRIM elements.
(*A*) The archetypal TRIM contains the following sequence features: target site duplication (TSD), terminal direct repeat (TDRs, flanking shaded boxes); primer binding site (PBS); polypurine tract (PPT). Start and end bases of TDRs are given in the flanking boxes. The possible length of different elements is given underneath (maximum length <540 bp). (*B*) Multiple alignment of selected TRIM elements (*Katydid*-At1) elements from *Arabidopsis thaliana*. The alignment was created with CLUSTALW using default parameters. The BOXSHADE program was used for shading. Structural features (abbreviations as in *A*) are marked underneath. Elements in chromosome II (Chr. II, gi 6598569 position 33850-33481), chromosome III (Chr. III, gi 7629988 position 45624-45993), and in chromosome IV (Chr. IV, gi 3309259 position 95888-95519) are shown. (*C*) Multiple alignment of selected TRIM elements from rice (*Oryza sativa*). Alignment and abbreviations as above. Elements from chromosome I (Chr. I(a), gi 10800055 position 29385-29802), chromosome I (Chr. I (b), gi 9711848 position 74966-74558), and chromosome IV (Chr. IV, gi 5777612 position 77595-77183) are shown.

A


TSD
LTR
PBS
PPT
LTR
TSD
TG/A
CA
TG/A
CA
100-250 bp
100-300 bp
100-250 bp

B


Chr. II
Chr. III
Chr. IV
TSD
LTR
PBS
PPT
LTR
TSD

C

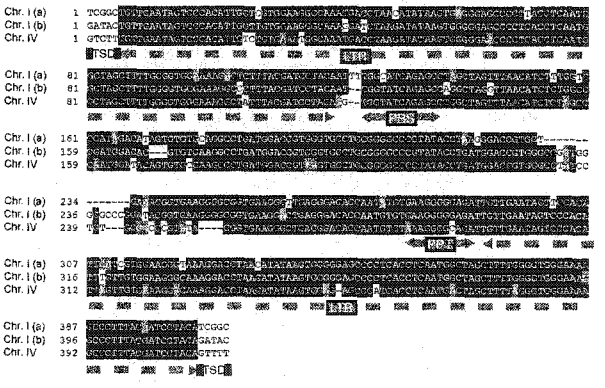

Chr. I (a)
Chr. I (b)
Chr. IV
TSD
LTR
PBS
PPT
LTR
TSD

Figure 3.2. Multiple alignment of selected 5' terminal direct repeat (TDR) sequences of TRIM elements from different species.

Sequences from *Nicotiana tabacum* (gi 9392606, position 3821-3947); *Lycopersicon esculentum* (gi 4220970, position 38-157); *Solanum tuberosum* (gi14599414, position 1522-1651); *Medicago sativa* (gi 19642, position 8088-8204); *Phaseolus vulgaris* (gi 2576326, position 1878-1761); *Oryza sativa* (gi 10800055, position 29390-29506); *Arabidopsis thaliana Katydid*-At2 (gi 7649355, position 68877-68994); *Arabidopsis thaliana Katydid*-At1 (gi 6598490, position 7722-7607). The alignment was created with CLUSTALW (gap opening parameter = 10; gap extension parameter = 1). The program BOXSHADE was used for shading.

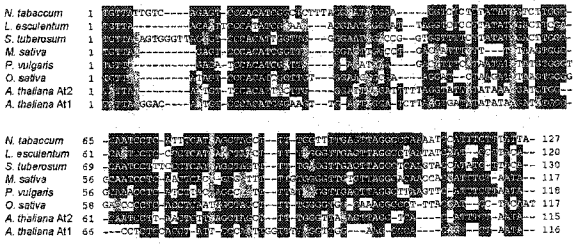

N. tabaccum
L. esculentum
S. tuberosum
M. sativa
P. vulgaris
O. sativa
A. thaliana At2
A. thaliana At1

Figure 3.3. Identification of RESites for TRIM elements (*Katydid-At1*).

Sequences harbouring the *Katydid*-At1 insertion (depicted by black box with arrowheads) are shown above the corresponding RESite. Positions on clone and gi numbers are indicated. Target sequences and TSDs are underlined. Insertion in gi 3702730 is a solo-LTR.


7209738 21706 ttccggtaattggtcatctt atcttcaccttcttctctct 22105
5672520 36998 ttcctatcattggtcatctt caccatctcctctct 37032
7629988 45609 aatctataaaattaaataac ttaactatttcaaaagaaaa 46008
7268832 4233 aatctataaaattaaataac tatttcaaaagaaca 4367
3702730 65214 ttttctcaattaattattat attctaacaaaagaatcatg 65469
12324309 6667 ttttctcaatcaattattat aacaa-aaaatcatg 6700

Figure 3.4. Examples of TRIM (*Katydid-At1*) contributions to gene structures.
Grey boxes with white arrowheads depict *Katydid*-At1 elements. Open boxes represent exons. Predicted start of translation (ATG) and TATA box are indicated. (*A*) *Katydid*-At1 element in gi 3510340 (position 22673-22946) contributes sequences for the promoter, first exon, and first exon-intron boundary of a gene encoding a cytochrome P-450-like gene. This insertion is truncated. Corresponding identical EST sequences (gi 5841198; gi 1053954) are represented by thick black bars. Thin lines connecting thick black bars represent a putative intron which is not found in the EST. (*B*) *Katydid*-At1 found on gi 7209738 contributes a splice site to a predicted cytochrome P-450 gene. Scale as in panel (*A*). (*C*) Sequence alignment of *Katydid*-At1 from panel (*A*) showing the putative translation start codon (boxed) and exon boundaries.

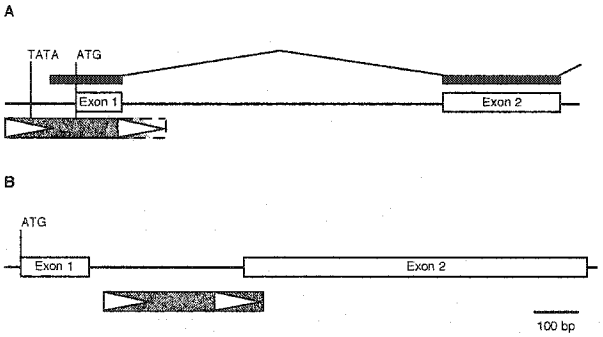
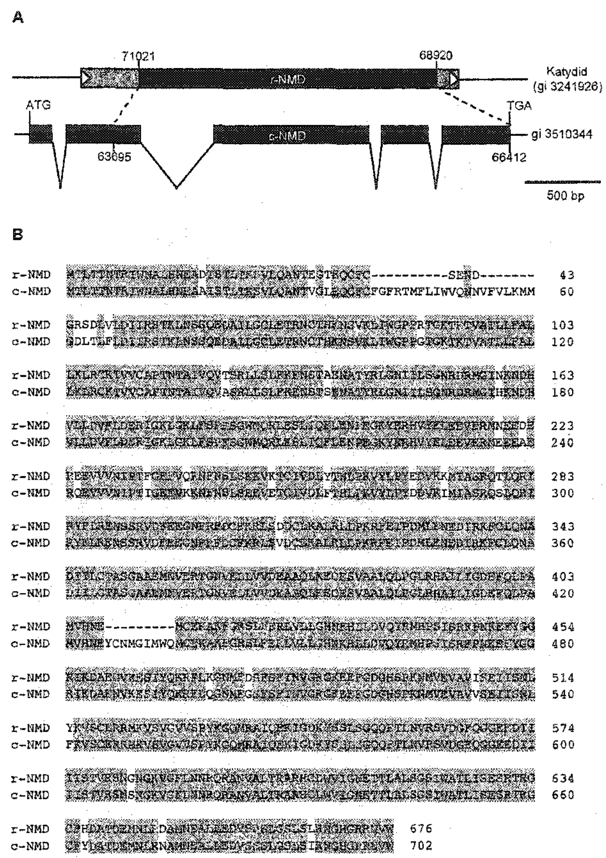

A
TATA
ATG
Exon 1
Exon 2
B
ATG
Exon 1
Exon 2
100 bp
C
gi 1053324 ccctaatangccccctcgagatgatggctcttctagccattgatctcgatatgtttgggc 60
gi 3510340 ccctaatacgccccctcgagatgatggctcttctagccattgatctcgatatgtttgggc 22779
Exon 1
gi 1053324 atcgatcgacggggcaagtattgggccggaccgatgtggatcgggttgagtatgtgcgga 120
gi 3510340 atggatcgacggggcaagtattgggccggaccgatgtggatcgggttgagtatgtgcgga 22719
Exon 1 | Exon 2
gi 1053324 tcgggctctgataccatattagagtgtaggttaagattattcaaacttcaaaggcaaaac 180
gi 3510340 tcgggctctgataccatattagagtgtaggttaagattattcaaacttcaaaggcaaaac 22777
gi 1053324 caaatgagcttcgtgtggtccgctgcggtgtgggtcatagctgtagccgctgttgtgatt 240
gi 3510340 caaatgagcttcgtgtggtccgctgcggtgtgggtcatagctgtagccgctgttgtgatt 21898

Figure 3.5. TRIM (*Katydid-At1*) involvement in the transduction of a cellular gene. (*A*) Diagram illustrating region of similarity (93%) between the cellular NMD gene (c-NMD) and the transduced version (r-NMD). r-NMD nucleotide sequence shared no significant similarity with the introns on c-NMD (thin lines). Exons are represented by black boxes whereas *Katydid*-At1 sequences and LTRs are represented by grey boxes and arrowheads, respectively. gi numbers and positions on clones are indicated. (*B*) Amino acid sequence alignment of r-NMD and c-NMD. Identical residues are shaded in grey.


A
71021
68920
r-NMD
Katydid
(gi 3241926)
ATG
TGA
c-NMD
gi 3510344
63695
66412
500 bp
B
r-NMD
c-NMD
43
60
103
120
163
180
223
240
283
300
343
360
403
420
454
480
514
540
574
600
634
660
676
702

REFERENCES

1. Kunze, R., Saedler, H. & Lonnig, W. E. (1997) *Adv. Bot. Res.* **27**, 331-470.
2. Boeke, J. D. & Corces, V. G. (1989) *Ann. Rev. Microbiol.* **43**, 403-434.
3. Boeke, J. D. & Stoye, J. P. (1997) in *Retroviruses*, eds. Varmus, H., Hughes, S., & Coffin, J. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 343-345.
4. Kumar, A. & Bennetzen, J. L. (1999) *Annu. Rev. Genet.* **33**, 479-532.
5. Kidwell, M. G. & Lisch, D. (2000) *Trends Ecol. Evol.* **15**, 95-99.
6. Doolittle, W. F., Feng, D. F., Johnson, M. S., & McClure M. A.. (1989) *Quart. Rev. Biol.* **64**,1-29.
7. Xiong, Y. & Eickbush, T. H. (1990) *EMBO J.* **9**, 3353-3362.
8. Smit, A. F. (1999) *Curr. Opin. Genet. Dev.* **9**, 657-663.
9. Li, W-H., Gu, Z., Wang, H. & Nekrutenko, A. (2001) *Nature* **409**, 847-849
10. SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., *et al.* (1996) *Science* **274**, 765-768.
11. Pearce, S. R., Harrison, G., Li, D., Heslop-Harrison, J. S., Kumar, A. & Flavell, A. J. (1996) *Mol. Gen. Genet.* **250**, 305-315
12. Doolittle, W. F. & Sapienza, C. (1980) *Nature* **284**, 601-603.
13. Orgel, L. E. & Crick, F. H. C. (1980) *Nature* **284**, 604-607.
14. Wessler, S. R., Bureau, T. E. & White, S. E. (1995) *Curr. Opin. Genet. Develop.* **5**, 814-821
15. McDonald, J. F., Matyunina, L. V., Wilson, S., Jordan, I. K., Brown, N. J. & Miller, W. J. (1997) *Genetica* **100**, 3-13.
16. Kazazin H.H. & Moran, J. V. (1998) *Nat. Genet.* **19**, 19-24.
17. Lister, C., Jackson, D. & Martin, C. (1993) *Plant Cell* **5**, 1541-1553.
18. Yang, Z., Boffelli, D., Boonmark, N., Schwartz, K. & Lawn, R. (1998) *J. Biol. Chem.* **273**, 891-897.
19. Marillonnet, S. & Wessler, S. R. (1997) *Plant Cell* **9**, 967-978.
20. Brodbeck, D., Amherd, R., Callearts, P., Hintermann, E., Mayer, U. A. & Affolter, M. (1998) *Cell Biol.* **17**, 621-633.

21. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. (1999) *Science* **283**, 1530-1534.

22. Agrawal, A., Eastman, Q. M. & Schatz, D. G. (1998) *Nature* **394**, 744-751.

23. Végh, Z., Vincze, E., Kadirov, R., Toth, G. & Kiss, G. B. (1990) *Plant Mol. Biol.* **15**, 295-306.

24. Vicient, C. M., Suoniemi, A., Anamthawat-Johnson, K., Tanskanen, J., Beharav, A., *et al.* (1999) *Plant Cell* **11**, 1769-1784.

25. Shirasu, K., Schulman, A. H., Lahaye, T. & Schulze-Lefert, P. (2000) *Genome Res.* **10**, 908-915.

26. Petrov, D. A. (2001) *Trends Genet.* **17**, 23-28.

27. *Arabidopsis* Genome Initiative (2000) *Nature* 408, 796-815

28. Maizel, J. V. & Lenk, R. P. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7665-7669.

29. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997) *Nucleic Acid Res.* **25**, 3389-3402.

30. Thomson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic. Acid Res.* **22**, 4673-4680.

31. Nicholas, K. B., Nicholas, H. B. Jr. & Deerfield, D. W. I. (1997) *EMBNEW. NEWS* **4**, -14.

32. Le, Q. H., Wright, S., Yu, Z. & Bureau, T. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 7376-7381.

33. SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. (1998) *Nat. Genet.* **20**, 43-45

34. Unseld, M., Marienfeld, J. R., Brandt, P. & Brenneicke, A. (1997) *Nat. Genet.* **15**, 57-61.

35. Jacobson A. & Peltz, S. W. (1996) *Annu. Rev. Biochem.* **65**, 693-739.

36. DeMarini, D. J., Winey, M., Ursic, D., Webb, F. & Culbertson, M. R. (1992) *Mol. Cell Biol.* **12**, 2154-2164.

37. Varmus, H. E. (1984) *Annu. Rev. Genet.* **18**, 553-612.

38. Bureau T.E., White, S. E. & Wessler, S. R. (1994) *Cell* **77**, 479-480.

39. Jin, Y.-K. & Bennetzen, J. L. (1994) *Plant Cell* **6**, 1177-1186.

40. Palmgren, M. G. (1994) *Plant Mol. Biol.* **25**, 137-140.

41. Deininger, P. L. (1989) in "Mobile DNA". Ed. Berg D.E. & Howe, M.M. American Society for Microbiology, Washington DC. pp 619-636.

42. Jin, Y.-K. & Bennetzen, J. L. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6235-6239.

43. Esnault, C., Maestre, J. & Heidmann, T. (2000) *Nat. Genet.* **24**, 363-367.

44. Kim, H. S., Kim, S. & Robbins, P. D. (2000) *Adv. Virus Res.* **55**, 545-663.

45. International Human Genome Sequencing Consortium (2001) *Nature* **409**, 860-921.

46. White, S. E., Habera, L. E., & Wesller, S. R. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 11792-11796.

47. Pickeral, O.K., Makalowski W., Boguski, M.S. & Boeke, J.D. (2000) *Genome Res.* **10**, 411-415.

48. Goodier, J.L., Ostertag, E.M. & Kazazian, H.H. (2000) *Hum. Mol. Genet.* **9**, 653-657.

CHAPTER 4

DATA MINING III: TC8, A *TOURIST*-LIKE TRANSPOSON IN *CAENORHABDITIS ELEGANS*

MITEs are another interesting story that has emerged from the initial survey of transposons in *Arabidopsis*. These transposons were first described nearly a decade ago as short elements widespread in the plant kingdom, flanked by either a 2- or 3-bp TSD and that had the potential to form a hairpin structure. Because the small size of these elements precluded any coding capacity, the mechanism underlying MITE mobility remained a mystery until the discovery of longer members with coding capacity during our initial survey of the *Arabidopsis* genome. Based on the size and sequence similarities between their target site duplication and terminal inverted repeat but, most importantly, their encoded transposase, MITEs are now known to be composed of three distinct families of elements, *Emigrant*, *Stowaway* and *Tourist*. This chapter, published in *Genetics* (Le, *et al.*, 2001), is an in-depth study of a *Tourist* transposon found in the nematode *C. elegans*. Once thought to be restricted to the plant kingdom, MITEs are now known to be widespread transposons, with representatives found in animals, fungi and bacteria.

Reference

LE, Q.H., TURCOTTE, K. AND BUREAU, T. (2001) Tc8, a *Tourist*-like transposon in *Caenorhabditis elegans*. *Genetics* **158**, 1081-1088.

ABSTRACT

Members of the *Tourist* family of *m*iniature *i*nverted-repeat *t*ransposable *e*lements (MITEs) are very abundant among a wide variety of plants, are frequently found associated with normal plant genes and thus, are thought to be important players in the organization and evolution of plant genomes. In Arabidopsis, the recent discovery of a *Tourist* member harbouring a putative transposase has shed new light on the mobility and evolution of MITEs. Here, we analyze a family of *Tourist* transposons endogenous to the genome of the nematode *Caenorhabditis elegans* (Bristol N2). One member of this large family is 7568 bp in length, harbours an ORF similar to the putative *Tourist* transposase from Arabidopsis and is related to the IS*5* family of bacterial *i*nsertion *s*equences (IS). Using database searches, we found *e*xpressed *s*equence *t*ags (ESTs) similar to the putative *Tourist* transposases in plants, insects and vertebrates. Taken together, our data suggest that *Tourist*-like and IS*5*-like transposons form a superfamily of potentially active elements ubiquitous to prokaryotic and eukaryotic genomes.

INTRODUCTION

Transposons are mobile genetic elements found in most, if not all, prokaryotic and eukaryotic genomes. Typically, transposons are defined as either class I, members of which move *via* an RNA intermediate (*e.g.* retrotransposons), or class II, members of which move directly as DNA (*e.g.* Tc1/*mariner*, *Ac/Ds*, *En/Spm*) (BERG and HOWE 1989). MITEs are transposons found in abundance among a wide variety of plant genomes. They are typically short (~100 to 500 bp), with conserved *t*erminal *i*nverted *r*epeats (TIRs), have a potential to form a stable DNA secondary structure (*i.e.* a hairpin structure), and generate a 2 or 3 bp *t*arget *s*ite *d*uplication (TSD) upon integration. Based on their TSD size and sequence, MITEs can primarily be divided into *Tourist*-like (5'-TAA-3') and *Stowaway*-like (5'-TA-3') families of elements (BUREAU and WESSLER 1992). MITEs are usually found in intimate association with genes and occasionally contributing *cis*-regulatory sequences (BUREAU and WESSLER 1994a; BUREAU and WESSLER 1994b). For these reasons, they are thought to play an important role in the evolution of plant genomes (WESSLER *et al.* 1995). Although ubiquitous in plants, there are only few reported MITEs from other eukaryotes such as fungi, insects, *C. elegans*, *Xenopus*, teleost fish, and humans (BESANSKY *et al.* 1996; FESCHOTTE and MOUCHES 2000; IZSVAK *et al.* 1999; OOSUMI *et al.* 1995; SMIT and RIGGS 1996; TU 1997; UNSAL and MORGAN 1995; YEADON and CATCHESIDE 1995). However, many of these MITEs are structurally similar to Tc1/*mariner*-like elements though members with ORFs are typically not available. The reason behind this difference in MITE abundance between plants and other organisms, and how this difference impacts host genome evolution, is not clear.

Since MITEs with coding capacity were previously unknown, the mechanism underlying their transposition remained elusive. Structurally, MITEs are reminiscent of non-autonomous deletion derivatives of class II transposons, presumably mobilized *in trans* by a transposase from a related autonomous element located elsewhere in the genome. Recently however, ORFs coding for putative *Tourist* transposases from *Arabidopsis thaliana* (Columbia) and *Stowaway* transposases from Arabidopsis and

Oryza sativa (domesticated rice) have been found, clearly defining MITEs as class II transposons (LE *et al.* 2000; TURCOTTE *et al.* 2001).

Further analysis of the putative Arabidopsis *Tourist* transposases indicates that *Tourist* elements are more related to specific bacterial IS, a large and heterogeneous group of simple inverted-repeat transposons widespread among prokaryotes (LE *et al.* 2000; MAHILLON and CHANDLER 1998), than they are to *Stowaway*. Detailed analysis of the putative transposases encoded by *Stowaway* suggests that these elements are related to members of the Tc1/*mariner* transposons (LE *et al.* 2000), Turcotte and Bureau, manuscript in preparation), a widespread superfamily of elements with representatives in vertebrates, invertebrates and fungi. Phylogenetic studies of transposases and integrases revealed that Tc1/*mariner* elements, members of diverse bacterial IS elements, retroviruses and retrotransposons share conserved catalytic residues (DOAK *et al.* 1994), known as the DDE motif, suggesting a common ancestry and even more widespread distribution.

Traditionally, transposons were identified and analyzed through genetic and molecular studies. However, the availability of sequence information and bioinformatic tools has now allowed for the identification and characterization of transposons through computer-assisted searches of sequence databases (BRITTEN 1995; BUREAU and WESSLER 1992; BUREAU and WESSLER 1994b; LE *et al.* 2000; OOSUMI *et al.* 1995; SURZYCKI and BELKNAP 1999; SURZYCKI and BELKNAP 2000). In *C. elegans*, molecular, genetic and sequence database search tools have revealed the presence of diverse representatives of both classes of transposons. Among these, Tc1 is probably one of the best characterized eukaryotic class II transposons and has been effectively used as a mapping, mutagenesis and gene-tagging tool (GREENWALD 1985; PLASTERK and VAN LUENEN 1997; RUSHFORTH *et al.* 1993; WILLIAMS *et al.* 1992). Tc1 transposition and regulation has been finely dissected at the molecular and biochemical levels (PLASTERK and VAN LUENEN 1997). Most Tc1 elements have a 54 bp TIR delimited by a 5'-TA-3' TSD. Autonomous Tc1 encodes a 343 amino acid transposase, which is the only protein required for transposition (PLASTERK and VAN LUENEN 1997; VOS *et al.* 1993; VOS *et al.* 1996). Tc1 copy number can vary from 10-15 fold depending on the *C. elegans* strain and element activity is cell-type dependent (EGILMEZ *et al.* 1995). In Bergerac, Tc1 excision

is detected in both somatic and germline cells whereas it is restricted to somatic cells in Bristol N2 (EGILMEZ *et al.* 1995; EIDE and ANDERSON 1988; PLASTERK and VAN LUENEN 1997; VOS *et al.* 1993; VOS *et al.* 1996). Recently, *mut-7* was found to encode a RNaseD homolog in *C. elegans* and may be involved in the regulation of transposon activity by post-transcriptional gene silencing (KETTING *et al.* 1999).

Tc1, Tc2, Tc3, and Tc7 are members of the Tc1/*mariner* superfamily found in *C. elegans* (COLLINS *et al.* 1989; DREYFUS and EMMONS 1991; LEVITT and EMMONS 1989; REZSOHAZY *et al.* 1997). In addition, other types of transposons endogenous to *C. elegans* include LTR (*l*ong *t*erminal *r*epeat)-retrotransposons (BOWEN and MCDONALD 1999; BRITTEN 1995), non-LTR retrotransposons (or long interspersed nuclear elements) (MALIK and EICKBUSH 1998) and hAT (or *Ac*)-like elements (BIGOT *et al.* 1996). Elements designated as Tc4 and Tc5 were initially classified as foldback transposons (COLLINS and ANDERSON 1994; YUAN *et al.* 1991) but further analysis of their putative transposases revealed that they contain a DDE motif suggesting a distant relationship to the Tc1/*mariner* elements *pogo*, *Tigger1* and *Tigger2* (ROBERTSON 1996; SMIT and RIGGS 1996). Tc6 also has a foldback structure but its terminal sequences and TSDs are reminiscent of those of Tc1/*mariner* transposons (DREYFUS and EMMONS 1991). A number of MITEs have also been identified in *C. elegans* (OOSUMI *et al.* 1995; OOSUMI *et al.* 1996; SURZYCKI and BELKNAP 2000; THE *C. ELEGANS* SEQUENCING CONSORTIUM 1998) but these were not found to be related to *Tourist*. To date, *Tourist* transposons appear to be confined to plant genomes.

In this report, we analyze two predicted ORFs in *C. elegans* that share amino acid similarity to the putative Arabidopsis *Tourist* and bacterial IS transposases. The same ORFs were also identified independently by sequence similarity to an Arabidopsis element called *Harbinger* but no further attempts were made to characterize these putative *C. elegans* elements (KAPITONOV and JURKA 1999). We determine that one of the *C. elegans* ORFs is part of a *Tourist*-like element which belongs to a large group of nematode transposons, referred to as Tc8. The majority of Tc8 members are short (150 to 400 bp), can potentially form DNA-hairpins and have TIRs and TSDs similar to other *Tourist*-like elements. The Tc8 transposase shares similarity to a number of plant, insect and vertebrate EST sequences, indicating that the *Tourist* family of transposons are

potentially active in other organisms. In addition, several eukaryotic genomic sequences were identified with similarity to the Arabidopsis *Tourist* and *C. elegans* Tc8 transposases. Together our data confirm the presence of *Tourist*-like transposons beyond the plant kingdom and suggest that *Tourist*-like elements share a common ancestry with specific bacterial IS and are widely distributed in the prokaryote and eukaryote genomes.

MATERIALS AND METHODS

Transposon mining. All sequence information and BLAST search tools (ALTSCHUL *et al.* 1990; ALTSCHUL *et al.* 1997) were accessed through GenBank (http://www.ncbi.nlm.nih.gov/), the Genome Sequencing Center (http://www.genome.wustl.edu/gsc/index.shtml) or the Sanger Centre (http://www.sanger.ac.uk/Projects/C_elegans/). Tc8.1 was identified by using sequences corresponding to putative MITE transposases from *Arabidopsis* (LE *et al.* 2000) as queries in BLAST searches against the GenBank sequence database. Additional database searches were performed to identify other *C. elegans* elements related to Tc8.1 (BLAST score of > 80).

Identification of RESites. *R*elated *e*mpty *site*s (RESites) are sequences highly similar or nearly identical to the sequences flanking an insertion (Le *et al.* 2000). However, RESites do not harbour the transposon sequence and the target site of insertion is not duplicated. RESites may correspond to orthologous or paralogous sequences, serve to delimit the ends of the elements, determine the TSDs and suggest evidence of past mobility. The identification of RESites was performed as previously described (Le *et al.* 2000). In brief, RESites are identified using the region immediately flanking an insertion as queries in BLAST searches.

Data and phylogenetic analysis. Further sequence analysis and alignments were performed using TRANSLATE, PILEUP, BESTFIT and GAP as part of the University of Wisconsin Genetics Computing Group suite of programs (version 10.0) or additional BLAST search tools (BLASTP, BLASTX, TBLASTN, PSI-BLAST and BLAST 2 sequences) provided at NCBI (ALTSCHUL *et al.* 1997; TATUSOVA and MADDEN 1999). For phylogenetic analysis, the sequences surrounding the DDE motif of the transposase of 14 transposable elements were manually aligned, based on previous reports (CAPY *et al.* 1996; REZSOHAZY *et al.* 1993). Phylogenetic relationships were inferred with maximum parsimony (multiple trees heuristic search with Tree Bisection Reconnection) and neighbor joining methods from the PAUP program, version 4.0b4a (SWOFFORD 2000) from the amino acid sequence alignment shown in Figure 4.3. All amino acids and stop

codons were equally weighted. Unrooted trees were displayed using the midpoint rooting method. Bootstrap analysis of the data set was done with 200 replicates.

RESULTS AND DISCUSSION

Computer-based searches using the amino acid sequence of the putative *Tourist*-like transposases from Arabidopsis MITE X (conceptual translation of gi 4454587, positions 4650-5865) and MITE XI (gi 4585884) as queries revealed the presence of similar coding regions in several other plant, animal and eubacterial ESTs and/or genomic sequences. For the most part, no other full-length plant *Tourist*-like element could be resolved and likely reflects the presence of truncated or degenerate elements. Many of the eubacterial sequences, however, correspond to previously described mobile elements. Although, as in the case of plants, the animal sequences typically did not lead to the identification of full-length elements, a 7568 bp element was mined from the genomic sequence of *C. elegans*. This element, designated as Tc8.1, is located on chromosome 2 (clone CELF14D2, gi 2746790), has large imperfect TIRs, a putative 3 bp TSD (5'-TAA-3') and a predicted ORF coding for a hypothetical protein of 743 amino acids (gi 7499082). Overall, the presence of sequences with similarity to the *Tourist*-like element transposases in several eubacterial and eukaryotic genomes suggests a wide distribution.

Using the Tc8.1 nucleotide sequence as a query in computer-assisted database searches, we found 282 related sequences within the *C. elegans* genome. Of these sequences, 128 represent shorter versions of Tc8.1 complete with TIRs and a specific TSD sequence of 5'-TAA-3'. The remainder were truncated versions with only one discernable terminal end immediately flanked by the sequence 5'-TAA-3'. Together, Tc8.1 and the shorter intact and truncated elements comprise the Tc8 group of transposons within *C. elegans*. Even though Tc8 elements are highly abundant in *C. elegans*, the total nucleotide contribution represents only 0.05% of the genome.

Strikingly, 63 members (≈49%) of the Tc8 group are exactly 150 bp in length and share > 94% sequence similarity. Like plant *Tourist*-like members, these elements have the potential to form hairpin-like DNA secondary structures. Excluding Tc8.1, six members of the Tc8 group were > 1-kb in size but did not possess any coding capacity (Figure 4.1A). For four of these larger elements, the increase in size is the result of nested insertions of other types of transposons or repetitive elements. Evidence of past mobility for some members of Tc8 was provided through the identification of related

empty sites (RESites, Figure 4.2), which are sequences similar to the empty site of an insertion. Furthermore, BLASTN searches also revealed the presence of Tc8 within the genome of the related nematode *C. briggsae* (Figure 4.1B), which is thought to have diverged from *C. elegans* approximately 10-40 million years ago (BUTLER *et al.* 1981; HESCHL and BAILLIE 1990; KENNEDY *et al.* 1993). To date, eight complete and three truncated Tc8 elements were identified from *C. briggsae* genomic clones deposited in GenBank. Tc8 transposons in *C. briggsae* vary from 95 to 368 bp in length and are not as well conserved as are the Tc8 elements in *C. elegans* (data not shown). A complete list of the *C. elegans* and *C. briggsae* transposons mined in this survey is available at http://www.tebureau.mcgill.ca/. A second predicted ORF (gi 7504556) was identified in the *C. elegans* genome and shared 46.5% amino acid sequence similarity to the Arabidopsis *Tourist* transposase. However, neither TIRs nor TSD could be identified. This putative element did not share nucleotide sequence similarity to Tc8 members and was not found to be repetitive in the *C. elegans* genome.

In contrast to the 150 bp elements which appear to be recent insertions, Tc8.1 seems to be an ancient insertion since it has accumulated at least four nested insertions and a corresponding RESite appears degenerate (Fig.4.2). Alternatively, Tc8.1 mobilization occurred after all or some of the nested insertion events. An unusual feature of the nested insertions within Tc8.1 is the presence of another shorter member of the Tc8 group at the same position in each of the TIRs. These elements are not identical (99.3% similarity with a single nucleotide substitution and an indel). Therefore, the nested elements could be independent insertions which occurred at exactly the same location in the TIRs. However, the level of divergence between the nested insertions is similar to the level of divergence between both TIRs (data not shown). Thus, we cannot distinguish between the possibility that they represent independent insertions or that the actual structure of Tc8.1 TIRs is the result of a rearrangement, localized duplication, conversion or a combination of these events. The two other nested insertions within Tc8.1, are located at positions 26747-27109 and 29990-31095 (Figure 4.1) and are repetitive in the *C. elegans* genome. The insertion at position 26747-27109 shares > 85% nucleotide similarity with members of a group of putative inverted-repeat transposons previously identified as *Cele1* (OOSUMI *et al.* 1995). However, it appears to be a truncated insertion

as element boundaries or TSDs could not be clearly identified. The insertion located at position 29990-31095 is annotated in ACeDB as RepQ and, upon closer examination, displays structural characteristics of *IS630*/Tc*1*/*mariner*-like elements. It is not clear whether Tc8 or any of the other elements within Tc8.1 preferentially insert into mobile elements. Such a phenomena has been suggested to reflect a mechanism to minimize deleterious insertion events in other organisms (SANMIGUEL *et al.* 1996; VAURY *et al.* 1989).

The short hairpin-like Tc8 elements are most likely deletion derivatives of larger elements. As in the case of other DNA-based transposons (*i.e.* class II), the mobilization and spread of deletion derivatives would be facilitated by transposase provided in *trans* by an autonomous element located elsewhere in the genome. The high copy number of very similar Tc8 elements may reflect recent activity. Tc8.1 could have been the source of transposase which mobilized the 150 bp elements. It is unlikely, however, that Tc8.1 is presently active because the putative transposase ORF is disrupted by the *Cele1* insertion (Figure 4.1). Alternatively, a functional transposase may have been provided by a full length Tc8 element which has been lost from the Bristol N2 strain. It is also possible that Tc8 elements, similar to the case of Tc7 which can be mobilized by a Tc1 transposase (REZSOHAZY *et al.* 1997), are currently being mobilized in *trans* by the Tc4 and/or Tc5 transposases as these transposons have structural characteristics similar to Tc8 (see below). An abundance in shorter deletion derivatives is also observed for *Tourist*-like elements in Arabidopsis and other plant genomes. The selective spread of shorter elements may be a mechanism inherent in the transposition process of these elements or may reflect an active host mechanism to minimize element contribution to genome size.

Upon closer examination, the transposases from the Arabidopsis MITE X and XI elements share amino acid similarity to members of the bacterial IS*4* and IS*5* family of elements suggesting a common evolutionary history (KAPITONOV and JURKA 1999; LE *et al.* 2000). Likewise, PSI-BLAST searches (ALTSCHUL *et al.* 1997) indicate that the Tc8.1 transposase also shares similarity to the same IS transposases. This similarity is restricted to the N-terminal region which contains the DDE motif found in many transposases and integrases and is known to be essential in the catalysis of DNA integration step of transposition (REZSOHAZY *et al.* 1993). Members of the IS*4* and IS*5* families have

additional conserved residues proximal to the DDE motif, namely, the DX(G/A)(F/Y) and $YX_2RX_3EX_6K$, or YREK, motifs neighboring the last aspartic and glutamic acid residues (REZSOHAZY *et al.* 1993). These motifs are also present in the putative *Tourist* transposases (Figure 4.3). Although the YREK motif appears to be only partially conserved in the putative *Tourist* transposases from *C. elegans* and Arabidopsis, the invariant arginine and glutamic acid residues are conserved (Figure 4.3).

Phylogenetic analysis using the region corresponding to the conserved DDE motif shows that the Tc8 transposase clusters with the transposases from members of the IS*5* element family (Figure 4.4). According to both maximum parsimony and neighbor-joining analyses, Tc8 is more closely related to the other eukaryotic elements, Arabidopsis MITE-X and -XI, than they are to bacterial IS elements. Moreover, the eukaryotic *Tourist* elements are more related to bacterial IS5 members that generate a 3 bp TSD (typically, 5'-TAA-3'). This suggests that the Arabidopsis and *C. elegans* *Tourist*-like elements have evolved from a common ancestral IS sequence. We cannot rule out the possibility, however unlikely, that the results of our analysis reflect convergence of the *Tourist* and IS during evolution. With the identification of new *Tourist* transposases from other organisms, it will be interesting to test the hypothesis that eukaryotic *Tourist* emerged from a common ancestral IS*5* member. Nevertheless, these results suggest that *Tourist* and bacterial IS5 elements form an IS5/*Tourist* superfamily.

The terminal nucleotides of the Tc8 TIRs are not only similar to plant *Tourist*-like TIRs, but are also reminiscent to the TIRs of bacterial IS elements. The preference for insertion into the trinucleotide 5'-TAA-3' of Tc8, as confirmed by the RESite analysis, is also a feature of other *Tourist* and IS*5* elements (Figure 4.2, Table 4.1). Curiously, the 3 bp TSDs and terminal sequences of Tc8 TIRs are also reminiscent of two other well described active transposons endogenous to *C. elegans*, namely Tc4 and Tc5 (COLLINS and ANDERSON 1994; YUAN *et al.* 1991). Even though Tc4 and Tc5 putative tansposases also contain the DDE motif (ROBERTSON 1996; SMIT and RIGGS 1996), no significant nucleotide or amino acid sequence similarity could be detected beyond the TSDs and terminal nucleotides. The TIRs and TSDs are important in transposase-mediated recognition of element termini and catalytic cleavage of the target sequence, respectively (FISCHER *et al.* 1999; VAN LUENEN *et al.* 1994). Consequently, sequence similarity

between the TIRs and TSDs of different transposons suggests a common transposase specificity.

Although the identification of RESites provides evidence of past mobility, to date, no *Tourist* element has yet been shown to be currently active. Even though we could not identify ESTs similar to the putative *Tourist* transposases from *C. elegans* or *Arabidopsis*, we did mine ESTs from other plant species, insects and vertebrates (Table 4.2). Additional ESTs displayed high sequence similarity but only entries with similarities to the conserved motifs in the C-terminal end of the query (Figure 4.3) are shown in Table 4.2. These ESTs may correspond to the transposases of different *Tourist* element groups (Table 4.1). Alternatively, similarity to ESTs may not correspond to transposase expression but may simply represent the transcription of *Tourist* elements which have inserted into or near expressed genes. Unfortunately, genomic sequences corresponding to the EST clones were not found but TBLASTN searches using Tc8 amino acid sequences did reveal the presence of similar sequences in these organisms and in *C. briggsae* (data not shown); discernable ends of these putative elements could be identified. Despite these, it would appear that *Tourist* elements are present and possibly active in these other genomes. The lack of corresponding ESTs does not necessarily indicate lack of Tc8 activity. Transposon activity in *C. elegans* is highly strain dependent. For example, copy number of Tc1 elements can be 10-15 fold lower in Bristol N2 than in Bergerac BO (EGILMEZ *et al.* 1995). In Bergerac, Tc1 excision activity in somatic cells is higher than in germ line cells, indicating that element activity may be limited to specific tissues or developmental stages (EIDE and ANDERSON 1988). In some eukaryotes (*e.g.* higher plants), host-mediated regulation such as DNA methylation may play a role in silencing transposon activity (HIROCHIKA *et al.* 2000).

MITE insertions are often closely associated with normal plant genes (BUREAU and WESSLER 1992; BUREAU and WESSLER 1994a; BUREAU and WESSLER 1994b). This is in contrast to retrotransposons which are frequently found as nested insertions within heterochromatic regions (SANMIGUEL *et al.* 1996; VAURY *et al.* 1989). Although a more detailed examination is required to determine if Tc8 are preferentially found associated with genes or contribute *cis* regulatory sequences, there seems to be a negative correlation between Tc8 and gene distribution in *C. elegans*, where genes are clustered near the

centre of the chromosomes (Figure 4.5). However, this is not as obvious on the X chromosome where this gene clustering is less pronounced (BARNES *et al.* 1995; THE *C. ELEGANS* SEQUENCING CONSORTIUM 1998). Interestingly, Tc8 abundance and distribution relative to genes is similar to what has been previously observed for Tc1 elements in *C. elegans* and *Tourist*-like elements in Arabidopsis, which has a genome of approximately the same size as *C. elegans* (KORSWAGEN *et al.* 1996; THE *C. ELEGANS* SEQUENCING CONSORTIUM 1998).

Eukaryotic *Tourist* and bacterial IS elements appear to have emerged from a common ancestor. This is analogous to other eukaryotic elements, namely members of the Tc1/*mariner* superfamily, which are related to IS*630* family of bacterial transposons (CAPY *et al.* 1996). Furthermore, many members of the Tc1/*mariner* have been shown to be transpositionally active. Although active mobilization of a *Tourist* still needs to be demonstrated, the identification of ESTs similar to *Tourist* putative transposases suggests that this family is not merely a collection of transposon relics. Thus, *Tourist*-like transposons are potentially active elements related to the widespread bacterial IS*5* transposons, and together, members of the IS*5*/*Tourist* superfamily have successfully populated both prokaryotic and eukaryotic genomes.

ACKNOWLEDGMENTS

We thank Julie Poupart, Dr. Rick Roy and Dr. Joseph Dent for critical comments on our manuscript. We are grateful to Boris-Antoine Legault for providing computer-programming support. This work was funded by a National Science and Engineering Research Council (NSERC) grant to T. B. and a *Formation de Chercheurs et l'Aide à la Recherche* (FCAR) fellowship to K.T.

Table 4.1. TSD and TIR sequence similarity between insertion transposons

Element	Host Organism	TSD	TIR*	Nucleotide gi Nb.	Protein gi Nb.†
MITE X	*Arabidopsis thaliana*	TAA	GGGGGTGTTATTGGT	4454587	N/A‡
MITE XI	*Arabidopsis thaliana*	TAA	GGTCCTGTTTGTTTG	4235150	4585884
Tc8	*Caenorhabditis elegans*	TAA	TGGGGTTATTCAAGT	2746790	7499082
IS*5*Sa	*Synechocystis* sp.	TAA	GAGCGTGTTTGAAAA	1256581§	1256580
IS*1031*	*Acetobacter xylinum*	TAA	GAGCCTGATCCGAAA	349283	349284
IS*12528*	*Gluconobacter suboxydans*	TCA	GAGCCCTTTTGGAAA	2055292	2055293
IS*903*	*Escherichia coli*	9-bp¶	GGCTTTGTTGAATAA	43025	581236
ISH*11*	*Halobacterium halobium*	7-bp¶	GAGGGTGTCATAGAA	43507	43508
IS*6501*	*Brucella ovis*	TA¶	CTAGAGCTTGTCTGC	295859	454221
IS*5*	*E. coli* (lambda KH 100)	CWAR‖	GGAAGGTGCGAATAA	49080	455278
IS*112*	*Streptomyces albus*	TA‖	AGGGCTGTCCCGTAA	46594	581565
IS*493*	*Streptomyces lividans*	ANT‖	GAGCGTTTTTCAACC	1707869	1196467
IS*702*	*Calothrix* sp. PCC 7601	CGT‖	AGAACTCTTGCAAAA	40657	581004
ISL*2*	*Lactobacillus helveticus*	AAT‖	AGAGTTACTGCGAAA	763132	763133
Tc4	*Caenorhabditis elegans*	TNA	CTAGGGAATGACCAG	156456	**
Tc5	*Caenorhabditis elegans*	TNA	CAAGGGAAGGTTCTG	529006	**

* 15 bases of terminal end sequences are shown (left TIR, from 5' to 3'). The actual TIR may be longer or shorter. † Amino acid sequence used for phylogenetic analysis (Fig. 4). ‡ conceptual translation of gi 4454587 from position 4650-5865. § Sequence shown in table is the reverse-complement of the sequence found on clone in GenBank. ¶ No TSD consensus sequence. ‖ TSD sequence information extracted from the IS database (http://pc4.sisc.ucl.ac.be/is.html). W = A or T, R = A or G and N = any nucleotide. ** Not included in the phylogenetic analysis.

Table 4.2. dbEST entries with TBLASTN similarity to the putative *Tourist* transposases from *C. elegans* and *Arabidopsis*[*]

Query[†]	Organism	GenBank Accession Number[‡]
C. elegans (Tc8)	*Bombyx mori*	AV404936
	Anopheles gambiae	AJ281674, AJ284723
	Danio rerio	AI882971
	Zea mays	AW165620, AI783120, AW165636, BE345915, AW172080
	Lycopersicon hirsutum	AW616734, AW616736
	Sorghum bicolor	AW746190
	Oryza sativa	AU068684
	Glycine max	AW759312
	Bos taurus	AW353649
	Gallus gallus	AI981097
	Oryza sativa	AU068684
	Glycine max	AW759312
Arabidopsis (MITE-X and MITE-XI)	*Lycopersicon hirsutum*	AW616734, AW616736
	Lycopersicon esculentum	AW032471
	Zea mays	AW438057, AW438058, AW000374, AW172080, AI677510, AW165620, AI1783120, AW165636, BE345915, AI941809, AI947769, BE510064, BE509925, AW052790, AW352677, AW000051
	Triticum monococcum	BE492211
	Triticum aestivum	BE489993
	Sorghum bicolor	BE365933, BE593965, AW746190
	Oryza sativa	AU075345, AU068684, C25943
	Danio rerio	AI882971
	Gallus gallus	AI981097

Bos taurus	AW353649
Xenopus laevis	BE191894
Anopheles gambiae	AJ281674
Mus musculus	AA003110

---

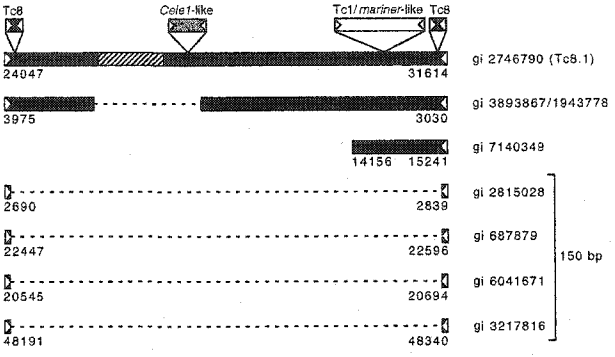
* Entries listed have similarity to the DDE motif of the query sequence. † Amino acid sequence used as query for *C. elegans* were Tc8.1 (gi 7499082) and the second putative *C. elegans Tourist* ORF (gi 7504556); for Arabidopsis MITE-X from gi 4585884 and MITE-XI from the conceptual translation of gi 4454587, nucleotide position 4650-5865. ‡ Alignments can be viewed at http://www.tebureau.mcgill.ca/.

FIGURE LEGENDS

Figure 4.1. *Tourist*-like elements in the *C. elegans* genome.

(*A*) Diagram of Tc8 members. The majority of the elements are 150 bp and share high sequence similarity. Additional longer members were found but were truncated. Dashes represent gaps and the hatched box indicates the region (~300 amino acids) of the annotated ORF which is similar to the C-terminal domain of other transposases. Boxes above Tc8.1 with inverted arrowheads represent nested insertions. The TIR and TSD sequences of the nested *Tourist*-like element are degenerate. GenBank gi numbers and positions of elements on corresponding clones are indicated to the right and below, respectively. (*B*) A Tc8-like group of elements in the genome of *C. briggsae*. Nucleotide alignment of *C. elegans* Tc8 from gi 2746790, position 24108-24275 and *C. briggsae Cb*-Tc8 from gi 11095049, position 34282-34399.

A



B

```
Tc8-Ce   1 tggggttattcaagtagtgtcggaaaattaaaaagtgtagaaaaattacg 50
           ||||||||||||| |||| |   |  |    || || |         |
Tc8-Cb   1 tggggttattcacgtagcgaaaatatgtattgaaatgcatttccgcgcct 50

Tc8-Ce  51 tcacaactgtatttaaaatacataaaaacatgtgttttaatacatttgtga 100
           ||  || |  || |    | |    | |||| ||| |||||||| |
Tc8-Cb  51 tcgaaattcaatgcatgagcgtcgcgaaatgtatttcaatacatat.... 96

Tc8-Ce 101 cgtcacaaatgtatttaaatacattttgctacattacttgaataacccca 150
                                      | |  ||| ||||||||||||
Tc8-Cb  97 ............................tttcgctacatgaataacccca 118
```

Figure 4.2. RESites corresponding to Tc8 element insertions.

The identification of RESites suggests evidence of past mobility. Black boxes with inverted arrowheads represent Tc8 insertions. GenBank gi numbers and position on clones are indicated. Target sites are underlined and TSDs are in bold.


3810698 67685-AAAGTGGCGAAAATTCGAGATTTTAGCTAA TAAAATCAGTGGAATTTTTTCGAAATTTTG-67220
7140353 4252-AAATTAGGCAAAATTCGAGATTTTAGC TAAAATTTGTGAAATTTTTTCGAAATTTTG-4308
1418595 13752-CTTTCAATAGAAAATGGGGTGCAGCACTAA TAATAGAGACTGCAGCACTATTTTTCGGAC-13543
727446 9634-CTTTCAATAGAAAATGGGGTGCAGCAC TAATAGAGACTGCAGCACTATTTTTCGGAC-9578
2746790 24305-AAATACATGTTTTTATTGTTACTTGAATAA TAAGGCAAAATATGTATTTAAATACACTTT-24078
3893867 4062-AAATACATGTTTTTATTGTTACTTGAA TAAGGCAAAATATGTATTTAAATACACTTT-4006
2746790 31374-AAATACATGTTTTTATTGTTACTTGAATAA TAAGGCAAAATATGTATTTAAATACACTTT-31583
1943778 2943-AAATACATGTTTTTATTGTTACTTGAA TAAGGCAAAATATGTATTTAAATACACTTT-2999
3869241 25810-TGAAAGTTGAACAATGGGGTGAAGTACTAA TAATAGAGGAAATACGGTAATTAACTTTAG-26019
2291239 17981-CCAAAGTTGAACAATGGGGTGCAATAC TAATAGAGGAAATACGGTAACTGTTATCAA-18037
2746790 24017-TAATGGTAAAAACCATTGATTTTTCACTAA TAAAAAGGTTTAATTTTCAGATAATACTGT-31644
2746790 18616-TAATGGTAAAAACTATTGATTTTTCACT GAAATAATAGTTTTTTTCAGATAATACTGT-18559

Figure 4.3. Similarity between putative MITE and bacterial IS transposases. GenBank gi numbers for amino acid sequences used in the alignment are indicated in Table 4.1. Amino acid coordinates are indicated to the left. The conserved DDE residues are indicated below the alignment. The DX(G/A)(Y/F) and YREK motifs found in other IS families (CAPY *et al.* 1996) are boxed.

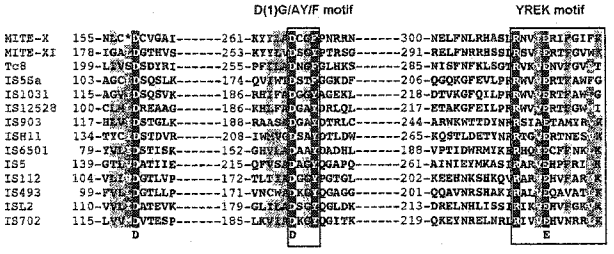

D(1)G/AY/F motif
YREK motif
MITE-X 155-NEC*LCVGAI------261-KYFLADCGFPNRRN------300-NELFNLRHASIRNVERIPGIFK
MITE-XI 178-IGALDGTHVS------253-KYFLVDSGFPTRSG------291-RELFNRRHSSIRSVERTFGVWK
Tc8 199-LAVSLSDYRI------255-PFLLADNGFLHKS------285-NISFNFKLSGTRVKLENVFGVLT
IS55a 103-AGCLLSQSLK------174-QVIWTDSTLGGKDF------206-QGQKGFEVLPRRNVLERTFAWFG
IS1031 115-AGVLLSQSVK------186-RRIFADCGLAGEKL------218-DTVKGFQILPRRNVLERTFAWFG
IS12528 100-CLALDREAAG------186-KHLFADGALDRLQL------217-ETAKGFEILPRRNVLERTFGWFI
IS903 117-HEALLSTGLK------188-RAASADGALDTRLC------244-ARWKWTTDINRRSIALTAMYRSK
ISH11 134-TYCLLSTDVR------208-IWSVGLSALDTLDW------265-KQSTLDETYNRRAGLERTNESLK
IS6501 79-YVLLDSTISK------152-GHVLADAAGADHL------188-VPTIDWRMIKERHQLCFFNKFLS
IS5 139-GTLLDATIIE------215-QFVSADLAGAPQ------261-AINIEYMKASIRARVEHPFRLLK
IS112 104-VELLDGTLVP------172-TLTIADCGLPGTGL------202-KEEHNKSHKQVRARLEHVFARLK
IS493 99-FVLLDGTLLP------171-VNCWADKGLQGAGG------201-QQAVNRSHAKIRALLEQAVATLK
ISL2 110-VVLLDATEVK------179-GLYLADSGLQGLDK------213-DRELNHLISSIRIKLEHVFGKLK
IS702 115-LVVLEVTESP------185-LKVIADKGLQGITK------219-QKEYNRELNRLRIVLEHVNRRLK
D
D
E

Figure 4.4. Evolutionary relationship between *Tourist* and bacterial IS5 elements. Phylogenetic analysis showing the relationship between the putative *Tourist* transposases from Arabidopsis (MITE-X and -XI) and from *C. elegans* to transposases from bacterial IS*5* elements. GenBank gi numbers for amino acid sequences used in the alignment are indicated in Table 4.1. Parsimony and the neighbor joining trees (not shown) are all concordant with the placement of Tc8 transposases with representatives of the bacterial IS*5* family. The unique and most parsimonious tree, based on the alignment in Figure 4.3, is shown (length = 335, CI = 0.818, RI = 0.564). Numbers at the base of nodes correspond to boostrap values. Typical size of TSDs are indicated to the right.

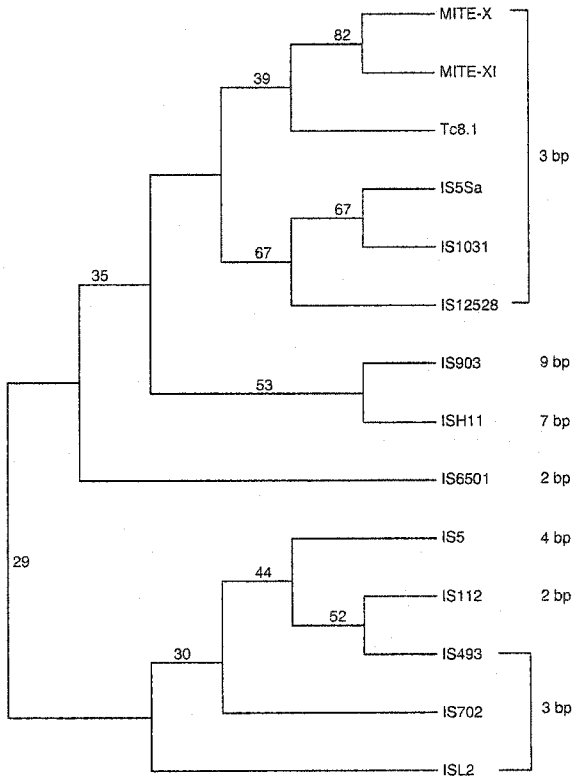
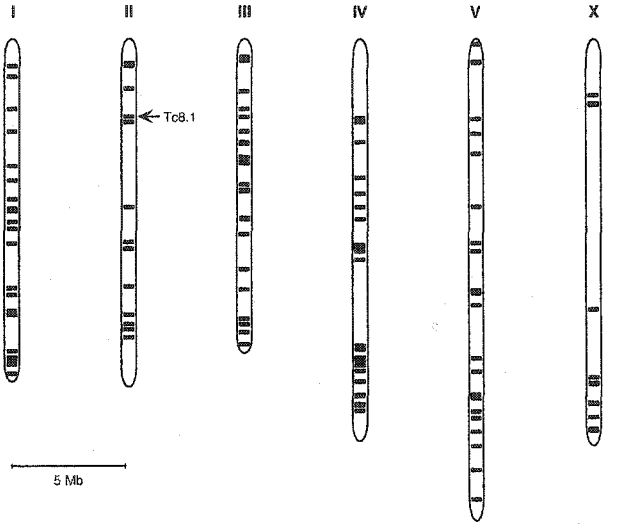

MITE-X
MITE-XI
Tc8.1
IS5Sa
IS1031
IS12528
IS903
ISH11
IS6501
IS5
IS112
IS493
IS702
ISL2
82
39
67
67
35
53
29
44
52
30
3 bp
9 bp
7 bp
2 bp
4 bp
2 bp
3 bp

Figure 4.5. Distribution of Tc8 elements in the *C. elegans* genome.
Each line represents a Tc8 insertion. Location of Tc8.1 is indicated. Scale is indicated at the bottom. The first nucleotide position on chromosome is located at the top.


I
II
III
IV
V
X
Tc8.1
5 Mb

REFERENCES

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic Local Alignment Search Tool. J. Mol. Biol. **215:** 403-410.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389-33402.

BARNES, T. M., Y. KOHARA, A. COULSON and S. HEKIMI, 1995 Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. Genetics **141:** 159-179.

BERG, D. E., and M. M. HOWE, 1989 *Mobile DNA*. American Society for Microbiology, Washington, D.C.

BESANSKY, N. J., O. MUKABAYIRE, J. A. BEDELL and H. LUSZ, 1996 *Pegasus*, a small terminal inverted repeat transposable element found in the white gene of *Anopheles gambiae*. Genetica **98:** 119-129.

BIGOT, Y., C. AUGE-GOUILLOU and G. PERIQUET, 1996 Computer analyses reveal a *hobo*-like element in the nematode *Caenorhabditis elegans*, which presents a conserved transposase domain common with the Tc1-*Mariner* transposon family. Gene **174:** 265-271.

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

BRITTEN, R. J., 1995 Active *gypsy/Ty3* retrotransposons or retroviruses in *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA **92:** 599-601.

BUREAU, T. E., and S. R. WESSLER, 1992 *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell **4:** 1283-1294.

BUREAU, T. E., and S. R. WESSLER, 1994a Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. Proc. Natl. Acad. Sci. U S A **91:** 1411-1415.

BUREAU, T. E., and S. R. WESSLER, 1994b *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell **6:** 907-916.

BUTLER, M. H., S. M. WALL, K. R. LUEHRSEN, G. E. FOX and R. M. HECHT, 1981 Molecular relationships between closely related strains and species of nematodes. J. Mol. Evol. **18:** 18-23.

CAPY, P., R. VITALIS, T. LANGIN, D. HIGUET and C. BAZIN, 1996 Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? J. Mol. Evol. **42:** 359-368.

COLLINS, J., E. FORBES and P. ANDERSON, 1989 The Tc3 family of transposable genetic elements in *Caenorhabditis elegans*. Genetics **121:** 47-55.

COLLINS, J. J., and P. ANDERSON, 1994 The Tc5 family of transposable elements in *Caenorhabditis elegans*. Genetics **137:** 771-781.

DOAK, T. G., F. P. DOERDER, C. L. JAHN and G. HERRICK, 1994 A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. Proc. Natl. Acad. Sci. USA **91:** 942-946.

DREYFUS, D. H., and S. W. EMMONS, 1991 A transposon-related palindromic repetitive sequence from *C. elegans*. Nucleic Acids Res. **19:** 1871-1877.

EGILMEZ, N. K., R. H. EBERT, 2ND and R. J. SHMOOKLER REIS, 1995 Strain evolution in *Caenorhabditis elegans*: transposable elements as markers of interstrain evolutionary history. J. Mol. Evol. **40:** 372-381.

EIDE, D., and P. ANDERSON, 1988 Insertion and excision of *Caenorhabditis elegans* transposable element Tc1. Mol. Cell. Biol. **8:** 737-746.

FESCHOTTE, C., and C. MOUCHES, 2000 Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a *pogo*-like DNA transposon. Mol. Biol. Evol. **17:** 730-737.

FISCHER, S. E., H. G. VAN LUENEN and R. H. PLASTERK, 1999 *Cis* requirements for transposition of Tc1-like transposons in *C. elegans*. Mol. Gen. Genet. **262:** 268-274.

GREENWALD, I., 1985 *lin-12*, a nematode homeotic gene, is homologous to a set of mammalian proteins that includes epidermal growth factor. Cell **43:** 583-590.

HESCHL, M. F., and D. L. BAILLIE, 1990 Functional elements and domains inferred from sequence comparisons of a heat shock gene in two nematodes. J. Mol. Evol. **31:** 3-9.

HIROCHIKA, H., H. OKAMOTO and T. KAKUTANI, 2000 Silencing of retrotransposons in *Arabidopsis* and reactivation by the *ddm1* mutation. Plant Cell **12:** 357-369.

IZSVAK, Z., Z. IVICS, N. SHIMODA, D. MOHN, H. OKAMOTO *et al.*, 1999 Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. J. Mol. Evol. **48:** 13-21.

KAPITONOV, V. V., and J. JURKA, 1999 Molecular paleontology of transposable elements from *Arabidopsis thaliana*. Genetica **107:** 27-37.

KENNEDY, B. P., E. J. AAMODT, F. L. ALLEN, M. A. CHUNG, M. F. HESCHL *et al.*, 1993 The gut esterase gene (*ges-1*) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. J. Mol. Biol. **229:** 890-908.

KETTING, R. F., T. H. HAVERKAMP, H. G. VAN LUENEN and R. H. PLASTERK, 1999 Mut-7 of *C. elegans*, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. Cell **99:** 133-141.

KORSWAGEN, H. C., R. M. DURBIN, M. T. SMITS and R. H. A. PLASTERK, 1996 Transposon Tc1-derived, sequence-tagged sites in *Caenorhabditis elegans* as marker for gene mapping. Proc. Natl. Acad. Sci. USA **93:** 14680-14685.

LE, Q. H., S. WRIGHT, Z. YU and T. BUREAU, 2000 Transposon diversity in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **97:** 7376-7381.

LEVITT, A., and S. W. EMMONS, 1989 The Tc2 transposon in *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA **86:** 3232-3236.

MAHILLON, J., and M. CHANDLER, 1998 Insertion sequences. Microbiol. Mol. Biol. Rev. **62:** 725-774.

MALIK, H. S., and T. H. EICKBUSH, 1998 The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Mol. Biol. Evol. **15:** 1123-1134.

OOSUMI, T., B. GARLICK and W. R. BELKNAP, 1995 Identification and characterization of putative transposable DNA elements in solanaceous plants and *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. U S A **92:** 8886-8890.

OOSUMI, T., B. GARLICK and W. R. BELKNAP, 1996 Identification of putative nonautonomous transposable elements associated with several transposon families in *Caenorhabditis elegans*. J. Mol. Evol. **43:** 11-18.

PLASTERK, R. H. A., and H. G. A. M. VAN LUENEN, 1997 Transposons, pp. 97-116 in *C. elegans II*, edited by D. L. Riddle, T. Blumenthal, B. J. Meyer and J. R. Priess. CSH laboratory Press, Cold Spring Harbor.

REZSOHAZY, R., B. HALLET, J. DELCOUR and J. MAHILLON, 1993 The IS*4* family of insertion sequences: evidence for a conserved transposase motif. Mol. Microbiol. **9:** 1283-1295.

REZSOHAZY, R., H. G. VAN LUENEN, R. M. DURBIN and R. H. PLASTERK, 1997 Tc7, a Tc1-hitch hiking transposon in *Caenorhabditis elegans*. Nucleic Acids Res. **25:** 4048-4054.

ROBERTSON, H. M., 1996 Members of the *pogo* superfamily of DNA-mediated transposons in the human genome. Mol. Gen. Genet. **252:** 761-766.

RUSHFORTH, A. M., B. SAARI and P. ANDERSON, 1993 Site-selected insertion of the transposon Tc1 into a *Caenorhabditis elegans* myosin light chain gene. Mol. Cell. Biol. **13:** 902-910.

SANMIGUEL, P., A. TIKHONOV, Y. K. JIN, N. MOTCHOULSKAIA, D. ZAKHAROV *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765-768.

SMIT, A. F., and A. D. RIGGS, 1996 *Tiggers* and DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. U S A **93:** 1443-1448.

SURZYCKI, S. A., and W. R. BELKNAP, 1999 Characterization of repetitive DNA elements in *Arabidopsis*. J. Mol. Evol. **48:** 684-691.

SURZYCKI, S. A., and W. R. BELKNAP, 2000 Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. Proc. Natl. Acad. Sci. USA **97:** 245-249.

SWOFFORD, D. L., 2000 PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), pp. Sinauer Associates, Sunderland, Massachusetts.

TATUSOVA, T. A., and T. L. MADDEN, 1999 BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. **174:** 247-250.

THE *C. ELEGANS* SEQUENCING CONSORTIUM, 1998 Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282:** 2012-2018.

TU, Z., 1997 Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, *Aedes aegypti*. Proc. Natl. Acad. Sci. USA **94:** 7475-7480.

TURCOTTE, K., S. SRINIVASAN and T. BUREAU, 2001 Survey of transposable elements from rice genomic sequences. Plant J. **25:** 1-13.

UNSAL, K., and G. T. MORGAN, 1995 A novel group of families of short interspersed repetitive elements (SINEs) in *Xenopus*: evidence of a specific target site for DNA-mediated transposition of inverted-repeat SINEs. J. Mol. Biol. **248:** 812-823.

VAN LUENEN, H. G., S. D. COLLOMS and R. H. PLASTERK, 1994 The mechanism of transposition of Tc3 in *C. elegans*. Cell **79:** 293-301.

VAURY, C., A. BUCHETON and A. PELISSON, 1989 The beta heterochromatic sequences flanking the I elements are themselves defective transposable elements. Chromosoma **98:** 215-224.

VOS, J. C., I. DE BAERE and R. H. PLASTERK, 1996 Transposase is the only nematode protein required for in vitro transposition of Tc1. Genes Dev. **10:** 755-761.

VOS, J. C., H. G. VAN LUENEN and R. H. PLASTERK, 1993 Characterization of the *Caenorhabditis elegans* Tc1 transposase *in vivo* and *in vitro*. Genes Dev. **7:** 1244-1253.

WESSLER, S. R., T. E. BUREAU and S. E. WHITE, 1995 LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr. Opin. Genet. Dev. **5:** 814-821.

WILLIAMS, B. D., B. SCHRANK, C. HUYNH, R. SHOWNKEEN and R. H. WATERSTON, 1992 A genetic mapping system in *Caenorhabditis elegans* based on polymorphic sequence-tagged sites. Genetics **131:** 609-624.

YEADON, P. J., and D. E. CATCHESIDE, 1995 *Guest*: a 98 bp inverted repeat transposable element in *Neurospora crassa*. Mol. Gen. Genet. **247:** 105-109.

YUAN, J. Y., M. FINNEY, N. TSUNG and H. R. HORVITZ, 1991 Tc4, a *Caenorhabditis elegans* transposable element with an unusual fold-back structure. Proc. Natl. Acad. Sci. USA **88:** 3334-3338.

CHAPTER 5

TOURIST MOBILITY IN *ARABIDOPSIS* AND *CAENORHABDITIS ELEGANS*

An obvious question that arises from the mining data concerns the mobility of the elements found. Are these transposons active or can they be activated? The identification of full-length *Tourist* members with coding capacity for putative transposases suggest potential activity in both *Arabidopsis* and *C. elegans*. Because uncontrolled transposition can be deleterious to the host, it is normally strictly regulated. There are known host- and element-encoded mechanisms that can repress transposon mobility. In addition, environmental stresses are also known to activate transposons. In this chapter, I have attempted to detect mobility of *Tourist* elements in *C. elegans* mutants for repression mechanisms and in *Arabidopsis* subjected to various stresses. But first, I needed a method to detect transposition therefore, in collaboration with Stephen Wright, I adapted a PCR-based mapping technique called Transposon Display (TD) to monitor mobilization events (Korswagen *et al.*, 1996; Wright *et al.*, 2001). The following chapter is an account and interpretation of the results I have observed for *Tourist* in *C. elegans* and *Arabidopsis*.

References

KORSWAGEN, H.C., DURBIN, R.M., SMITS, M.T. AND PLASTERK, R.H. (1996). Transposon Tc1-derived, sequence-tagged sites in *Caenorhabditis elegans* as markers for gene mapping. *Proc. Natl. Acad. Sci. USA* **93**, 14680-14685.

WRIGHT, S.I., LE, Q.H., SCHOEN, D.J. AND BUREAU, T. (2001). Population dynamics of an *Ac*-like transposable element in self- and cross-pollinating *Arabidopsis*. *Genetics* **158**, 1279-1288.

INTRODUCTION

Mobile genetic elements, or transposons, have the potential to reshape the genome. Transposon insertion or excision can alter gene expression by disrupting coding sequences or regulatory regions and their repetitive nature can promote chromosomal rearrangements (KIDWELL and LISCH 2001). Because they are found in nearly every living organism, and taking into account their mutagenic properties, transposons are thought to be an important source of genetic variability. However, mutations caused by transposons are mostly deleterious and uncontrolled activity can lead to accumulation of an excessive number of transposable elements. Therefore, the activity of endogenous transposable elements must be normally repressed, regulated by the elements, by the host or a combination of both.

Autoregulation of transposition activity can occur at different levels, including at the level of interactions between mobility proteins. Transposases are often composed of multiple protein subunits that bind to *cis*-recognition sequences within the element. A defective protein may retain its capability to interact with active sub-units causing a dominant-negative effect, rendering the entire transposase complex non-functional and diluting active proteins. This downregulation mechanism, termed overproduction inhibition, has been observed for *Mos*1 and other *mariner*-like elements in *Drosophila* (DE AGUIAR and HARTL 1999; HARTL *et al.* 1997a; HARTL *et al.* 1997b). Many different families of *mariner*-like elements co-exist within the *Drosophila* genome and these interactions between the related transposases, or "crosstalks", may drive active groups to rapidly diverge from non-functional ones or escape from the host genome (DE AGUIAR and HARTL 1999). Overproduction inhibition is one of the mechanisms thought to keep in check another *Drosophila* transposon, the *P*-element (COEN *et al.* 1994). Autoregulation can also act at the level of protein expression, as in the case of *Ac* elements in maize, where the transposase downregulates its own transcription by binding to its gene promoter (FRIDLENDER *et al.* 1998). Another level at which transposons can repress their own spread is through titration, where degenerate transposons compete for the transposase of functional elements (HARTL *et al.* 1997a; HARTL *et al.* 1997b; SIMMONS and BUCHOLZ 1985). Also, transposons have been observed to insert within

chromosomal regions that limit their deleterious effects such as heterochromatin or nested within other transposons (SANMIGUEL *et al.* 1996). However, it is not clear whether such a targeting mechanism is the result of element insertion preference or of purifying selection.

Several host mechanisms have been proposed to limit transposon activity including transcriptional silencing by methylation of cytosine residues (YODER *et al.* 1997), chromatin formation, RNA interference (RNAi) (KETTING *et al.* 1999; TABARA *et al.* 1999), and Post-Transcriptional Gene Silencing (PTGS). In marsupial hybrids that exhibit a genome-wide decrease in methylation levels, a considerable increase in the number of a retroviral-like element, KERV-1, was observed (O'NEILL *et al.* 1998). Similarly, in the *Arabidopsis* reduced-methylation mutant *ddm1*, the *cac1*, *Tar17* and *Mutator*-like elements show a sudden increase in activity levels (HIROCHIKA *et al.* 2000; MIURA *et al.* 2001; SINGER *et al.* 2001). Because the bulk of methylated cytosine residues are within transposons and other repeated sequences in many organisms, it is believed that the host heavily methylates transposon-specific sequences as a means to restrict their expression. Moreover, methylated elements become permanently inactivated over time due to transitions at methylated sites when 5-methylcytosines are deaminated to thymine (BESTOR 1999). In organisms that do not methylate their DNA, other mechanisms have been suggested to act as host defence against transposon proliferation. In *Drosophila* for example, chromatin formation is thought to act as transposon silencers. In fact, the true role of methylation may be to target sequences for condensation into compact organization (FEDOROFF 2000; HABU *et al.* 2001). The *Arabidopsis ddm1* mutant was recently shown to encode a SWI2/SNF2 chromosome remodeling homolog (JEDDELOH *et al.* 1999). In *Caenorhabditis*, where methylation of cytosine residues also does not occur, mutants resistant to sequence-specific post-transcriptional degradation of double-stranded RNA (dsRNA) show increased transposon activity, suggesting that interference by dsRNA (RNAi) may serve as a host defence against transposition (KETTING *et al.* 1999; TABARA *et al.* 1999). Targeted post-transcriptional degradation of dsRNA is actually a mechanism to inactivate foreign DNA (*e.g.* viruses and transgenes) that has been well studied in plants but is better known as co-suppression or PTGS (BAULCOMBE 2000; FAGARD *et al.* 2000; KETTING and PLASTERK 2000; VANCE and

VAUCHERET 2001). PTGS and RNAi may be widespread in eukaryotes as similar mechanisms have also been observed in fungi (COGONI and MACINO 1999), *Chlamydomonas* (WU-SCHARF *et al.* 2000), *Drosophila* (JENSEN *et al.* 1999a; JENSEN *et al.* 1999b) and mammals (WIANNY and ZERNICKA-GOETZ 2000).

In populations experiencing stress conditions, an increase in transposon activity suggests that environmental factors can also induce transposition. It is thought that transposition can cause genetic changes within a population that may lead to reproductive isolation, and possibly even result in speciation (BRITTEN 1996; CAPY *et al.* 2000; KIDWELL and LISCH 1997). Thus, an increase in transposon activity during periods of stress may generate genetic variability upon which selection can act, perhaps to benefit some individuals in the new environment. There are many examples of transposon activation in hosts submitted to adverse environments. Abiotic stresses were known to elicit mobilization since the early days of transposon studies when the activity of transposons in maize were induced by a "genomic shock" caused by gamma irradiation (MCCLINTOCK 1984). Recently, the activity of a maize element known as *Mutator* was shown to increase when pollen was exposed to ultraviolet-B (UV-B, 280-320 nm) doses equivalent to a 33% depletion of atmospheric ozone (WALBOT 1999). Biotic stresses such as fungal, bacterial and viral infections also appear to activate transposition (MOREAU-MHIRI *et al.* 1996). In fact, induction of elements in response to viral infection is proposed to be a mechanism by which horizontal transfer can occur (HOUCK *et al.* 1991). Protoplast formation, tissue culture, various biotic and abiotic elicitors can increase levels of retrotransposon transcripts in tobacco and potato (HIROCHIKA 1993; MHIRI *et al.* 1997; PEARCE *et al.* 1996; POUTEAU *et al.* 1991). In addition, the number of transposons in rice have been shown to significantly accumulate in tissue culture maintained over short (*i.e.* months) periods of time (HIROCHIKA *et al.* 1996). In tobacco, *Tnt1* activity was specifically increased in protoplast preparations but was nearly silent in cultures that had not undergone such treatment (MELAYAH *et al.* 2001). This observation lends support to the idea that microbial cell wall degrading enzymes used in protoplast isolation protocols may be the key elicitor, perhaps mimicking pathogen infection (MELAYAH *et al.* 2001). Interestingly, *cis*-acting sequences of retroelements known to be active have been found to share striking similarities with *cis*-regulatory sequences of plant

defence and stress activated genes (reviewed in GRANDBASTIEN 1998). This suggests that retroelements and stress-induced plant defence genes may be transcriptionally activated by the same environmental factors, and further demonstrates how environmental cues can regulate transposons. However, the exact nature of how stress signals can relieve host defence systems that represses transposition remains to be fully understood.

The *Tourist* family of transposons were first identified in monocotyledonous plants as unusually short elements that contained Terminal Inverted Repeats (TIRs). They were often found near genes, potentially contributing *cis*-regulatory sequences (BUREAU *et al.* 1996; BUREAU and WESSLER 1992; BUREAU and WESSLER 1994a; BUREAU and WESSLER 1994b). Until recently, the mechanism through which *Tourist* is mobilized and regulated was unknown because a family member with coding capacity had not been found. However, the presence of TIRs suggested that *Tourist* may belong to a class of transposons that move *via* a DNA form. This hypothesis was reinforced when *Arabidopsis* genomic sequences were made available and a putative *Tourist* transposase was finally found. Transposase sequence information also indicated that *Tourist* is related to a group of abundant elements in *C. elegans*. In fact, further searches using the transposase sequence reveals that *Tourist* belongs to a widespread group of elements now known as the IS5/*Tourist* superfamily whose transposase contain the catalytic DDE motif. This superfamily includes members not only in animals and plants, but also in fungi and bacteria. Recently, a maize IS5/*Tourist* element called *PIFα* was shown to be mobile and necessary for the transposition of shorter non-autonomous *mPIF* elements by using Transposon Display (TD), a PCR-based transposon mapping technique to detect mobility by looking for polymorphic patterns in band sizes (ZHANG *et al.* 2001).

To this day, *PIFα* is the only IS5/*Tourist* element shown to be currently active. The genomes of *C. elegans* and *Arabidopsis*, two eukaryotes for which the complete genomic sequence is available, also contain members of the IS5/*Tourist* transposons (LE *et al.* 2001; LE *et al.* 2000). Most are short and probably defective sequences, but some larger elements with coding capacity may potentially be active. In order to further investigate the mobility of *Tourist*-like elements and their regulation by environmental factors, we used a modified TD protocol to monitor mobility of two IS5/*Tourist* elements in *Arabidopsis* undergoing tissue culture, protoplast isolation or irradiated with UV-B.

We have also examined the mobility of Tc8, an IS5/*Tourist* member, in *C. elegans* defective for silencing by dsRNA. We did not observe polymorphisms for either IS5/*Tourist* elements in *Arabidopsis* or *C. elegans* and our findings have led us to doubt the potential for activity of these elements. Taken together, our results suggest that the IS5/*Tourist* elements examined may no longer be active and represent families of transposon relics from a past explosive invasion.

MATERIALS AND METHODS

Growth conditions. All nematode strains were maintained at room temperature (23 ± 1°C) on NGM plates seeded with *Escherichia coli* OP50 as described (LEWIS and FLEMING 1995). *C. elegans* strains were obtained from *Caenorhabditis* Genetics Center (http://biosci.umn.edu/CGC/CGChomepage.htm) except for Bristol N2 which was obtained from Dr. Richard Roy at McGill University.

UV-B treatment. Plants were grown in soil and maintained in growth chambers (Conviron) at 25°C under constant light conditions. One to four-week old plants were submitted to 14 kJ/m^2/day for 10 days in a growth chamber (Conviron) retrofitted with UV-B 313 lamps (ChemRoy Canada). To selectively filter-out wavelengths < 290 nm, a cellulose di-acetate membrane (ComPlast Plastics Inc.) was placed between the light source and the plants whereas a mylar (ComPlast Plastics Inc.) membrane was used to shield control plants from UV-B wavelengths. A Solar Light Co. radiometer and detector (models PMA 2100 and PMA 2101, respectively) were used to calibrate UV irradiance. Plants were allowed to recover under normal growth conditions as described above and DNA from the progeny was extracted for transposon display (see below). *Arabidopsis* seeds were concurrently irradiated in Petri dishes covered with either cellulose di-acetate or mylar membranes and plants grown from these seeds were examined for transposition events.

Tissue culture and protoplast isolation. Calli regenerated from *Arabidopsis* (Columbia) roots were prepared and maintained according to the protocol described by Mathur and Koncz (1998). Protoplasts from Arabidopsis were prepared and regenerated as described (MATHUR and KONCZ 1998) except that they were embedded by adding 0.4 M sucrose containing 1% Bacto-agar as described by Wenck and Marton (1995). Calli from root explants and a protoplast regenerant were maintained on Petri dishes as described (MATHUR and KONCZ 1998) and periodically sampled and tested for transposition events.

DNA isolation. Nematode genomic DNA was extracted as described (SULSTON and HODGKIN 1988) with slight modifications. In brief, organisms grown for three days on two 10-cm NGM plates, seeded with *E. coli* strain OP50 over a thin layer of 2% agarose, were collected using 5 mL of M9 media and pooled into 15-mL tubes. Nematodes were washed once with 1M sucrose and three times with M9 media by centrifuging and resuspending and then stored at -70 until extraction. Thawed pellets were incubated in 0.5 mL of extraction buffer (100 mM NaCl; 100 mM Tris-HCl, pH8.5; 50 mM EDTA, pH 7.4; 1% SDS, 1% β-mercaptoethanol; 100 μ/mL proteinase K) at 65°C for 30 minutes, then ethanol precipitated. Resuspended samples were incubated with 20 μg/mL RNaseA (Gibco/BRL), extracted with phenol:chloroform before ethanol precipitation and resuspension in elution buffer (QIAgen). *Arabidopsis* genomic DNA was extracted using the Plant DNA Miniprep Kit (QIAgen) following the protocol provided by the manufacturer.

Primer Design. Adaptor-specific and Tc1-specific pre-selective and selective primers for TD (Table 5.1 and see below) were described by Korswagen *et al.* (1996). Information on transposons previously identified is available at http://www.tebureau.mcgill.ca/ (LE *et al.* 2001; LE *et al.* 2000) and MITE X and XI reported therein are hereafter referred to as *Tourist-10* and *-11*, respectively. Nucleotide sequences alignments and determination of *BfaI* restriction sites for elements of each group were performed with the PILEUP and MAP functions from the suite of programs of the University of Wisconsin Genetics Computing Group (version 10). Primers were designed against conserved regions of the aligned elements sequences.

Transposon Display. TD as described by Korswagen *et al.* (1996) was modified for the Li-Cor fluorescence detection system (WRIGHT *et al.* 2001). In brief, this method makes use of a primer targeted within a transposon and another located in an adaptor ligated to genomic sequences to PCR amplify transposon sequences along with their flanking sequences (Figure 5.1). A novel insertion or excision event would result in a change in the length of the flanking sequence and would be represented as a polymorphic band. Approximately 100 ng of genomic DNA was digested with 2.5 units of the restriction

enzyme *Bfa*I (New England Biolabs) and ligated to 15 pmol adaptor cassettes (5'-TAG CAAGGAGAGGACGCTGTCTGTCGAAGGTAAGGAACGGACGAGAGAAGGGAG A-3' and 5'-TCTTCCCTTCTCGAATCGTAACCGTTCGTACGAGAATCGCTGTCTC TCCTTGC-3') with T4 DNA ligase (Invitrogen/GibcoBRL) as specified by the manufacturer. Digested genomic DNA ligated to adaptor cassettes were diluted 4-fold before a pre-selective amplification step using a pair of element- and adaptor -specific primers. Pre-selective amplification products were diluted 100-fold and re-amplified using a second set of nested element- and adaptor-specific primers, the latter being fluorescent labeled to allow detection. We used the *AmpliTaq* PCR system (Perkin-Elmer) in a MJ Research (PTC-225) thermocycler and the conditions for both amplifications were as follows: one cycle at 94°C for 10 minutes, 20 cycles of 94°C for 1 minute, 55°C or 60°C for 1 minute, and 72°C for 1 minute followed by a final step of one cycle at 72°C for 10 minutes. Annealing temperatures and sequences for element-specific primers used are indicated in Table 5.1. The adaptors are not perfectly complementary and the adaptor-specific primers are specially designed against the region of mismatch. This strategy allows amplification from the transposon-targeted primer only after first strand synthesis from the adaptor-specific primer has occurred (Figure 5.1, KORSWAGEN *et al.* 1996). Following the second amplification, 6 µL of loading dye (95% formamide, 10 mM EDTA, 0.1% bromophenol blue) were added to the final amplification products which were then separated by size and visualized on a 5.5% denaturing polyacrylamide gel (BioShop) using a Li-Cor DNA4200 automated sequencer. A Perl program was used to calculate and confirm the sizes of bands observed on TD based on BLAST search results of element-specific selective primer sequences against the *Arabidopsis* genome, allowing for 15-20% of primer nucleotide mismatches (see Appendix 2).

RESULTS

Tc8 activity in *C. elegans*

A family of *C. elegans* IS5/*Tourist* elements, Tc8, has been previously described and includes a full-length member with coding capacity for a potentially functional transposase. In order to test whether Tc8 transposons are active, we examined for mobilization in *C. elegans* mutants *mut-7*, *rde-1*, *rde-2*, *rde-3* and *rde-4* in which endogenous elements, including Tc1, have previously been shown to be activated (KETTING *et al.* 1999). These mutants are all resistant to the silencing mechanism mediated by dsRNA which is thought to function as a host defence against transposons. We used the transposon mapping technique described by Korswagen *et al.* (1996) to detect transposition events (see *Materials and Methods).* Although high rates of mobility were observed for Tc1, Tc8 did not show any difference in banding patterns when compared to wild-type Bristol N2 strains (Figure 5.2).

Tourist activity in *Arabidopsis*

We also examined the activity of *Tourist-10* and *Tourist-11* in *Arabidopsis* by subjecting plants to situations that were previously reported to activate other types of transposons in other systems (MELAYAH *et al.* 2001; WALBOT 1999). As a control, we monitored the activity of *cac1*, a TIR-containing CACTA-like transposon known to be active in *Arabidopsis* (MIURA *et al.* 2001). We first examined the effect of UV-B by subjecting plants at different growth stages (seeds and one- to four-week old plants) to irradiation levels equal to twice the estimated normal dose observed for a sunny day at temperate latitudes (~14 kJ/m^2/day) (LI *et al.* 1993). This level of UV-B was previously shown to cause 50% decrease in normal plant growth. A total of 69 plants were examined and, as shown in Figure 5.3*A* and 5.3*B*, we did not observe polymorphic bands for either *Tourist-10* or *Tourist-11* elements when compared to non-irradiated control plants. However, *cac1* was also not activated in these same plants.

We then tested for *Tourist* activity in *Arabidopsis* from tissue culture and also in regenerants from isolated protoplasts. The passage of plants through tissue culture has been reported to activate some transposons (Hirochika, 1993; Hirochilka, 1996). It is

currently believed that the cell wall degradation enzymes of fungal source used when isolating protoplasts is the transposon activating agent (MELAYAH *et al.* 2001). Calli kept in tissue culture for two months were regularly examined for signs of transposition. In total, 56 calli were tested but no activity was detected for both *Tourist* elements (Figure 5.3*C*). Also, *cac1* was also not active.

One callus was successfully regenerated from isolated *Arabidopsis* protoplast and maintained. The initial callus and sub-cultures derived after 4 and 8 months were maintained independently and examined but none showed signs of *Tourist* transposition. Again, no *cac1* activity was detected in these same calli (data not shown).

DISCUSSION

Tourist activity in *Arabidopsis*

UV-B irradiation, tissue culture and protoplast isolation are conditions capable of inducing the activity of endogenous transposons. Tissue culture is believed to activate transposons by reducing genome-wide levels of methylation, which is used by the host to silence transposon expression (KAEPPLER and PHILLIPS 1993; PLANCKAERT and WALBOT 1989). Protoplast isolation activates endogenous elements probably because the fungal extract used for digesting the cell wall during preparation elicits a cellular response similar to that of pathogen infection (MELAYAH *et al.* 2001; POUTEAU *et al.* 1991). It is less clear how the "genomic shock" generated by UV-B can elicit transposon activity but it may be that the activation of transposition proteins also follows de-methylation of silent elements (WALBOT 1999).

In order to activate *Tourist* transposons in *Arabidopsis* and to study their response to environmental factors, we submitted plants to the stress conditions mentioned above. We did not observe mobility for *Arabidopsis Tourist* elements submitted to either of these conditions. However, the endogenous element *cac1*, which is known to be active, also did not transpose under the same stresses. Because *cac1* did not show signs of mobility in our experiments, it is possible that the intensity of the stresses used were inadequate to induce mobility.

In calli from regenerated protoplasts, digestion with methylation sensitive restriction enzymes indirectly suggests that genome wide levels of methylation may be reduced (data not shown). Mobility for *Tourist-10*, *Tourist-11* and *cac1* was not observed probably because our sample size was too small to detect low rates of transposition under these conditions. In fact, we were only able to regenerate one callus from isolated protoplast and were not successful in regenerating mature plants. Protoplasts can be easily prepared and efficiently regenerated from some *Arabidopsis thaliana* ecotypes (MATHUR and KONCZ 1998). In our hands, we found it very difficult to obtain regenerants from protoplasts for the Columbia ecotype.

In plants exposed to UV-B, irradiation levels were sufficient to cause physiological changes such as growth retardation and bleaching. Irradiated plants also

produced fewer seeds than untreated plants and some perished before reaching maturity. On the other hand, it is possible that the UV-B dosage was too high, causing plants that would have exhibited mobilization of *Tourist* or *cac1* to suffer other lethal mutations. The sample size used would have been sufficient to detect *Tourist* mobility in at least 1.5% of the UV-B treated individuals (n = 69) and in at least 1.8% for the explants (n = 56). However nothing is known about basal transposition rates for *Tourist*-like elements. For comparison, transposition rates for the active *Arabidopsis* element *Tag1* in stress-induced plants were estimated to be ~30% compared to a transposition rate of < 0.8 % in wild-type (BHATT 1998). For *Tnt1* in tobacco, 3% of individuals from tissue culture showed new insertions compared to 25% for isolated protoplasts (MELAYAH *et al.* 2001).

Alternatively, *Tourist* and *cac1* elements were not mobilized after tissue culture, protoplast isolation or UV-B irradiation treatments because they are insensitive to these stress conditions. This would suggest that different stress conditions can affect different element types. It would be interesting to compare the activity of these and other types of elements in plants exposed to other abiotic and biotic stresses (*e.g.* heat shock, wounding and exposure to various pathogens).

Tc8 activity in *C. elegans*

RNA interference (RNAi) in nematodes is thought to be a host defence mechanism against the transposition of endogenous elements based on the evidence that endogenous transposons, including Tc1, were found to be activated in *C. elegans* defective for RNAi. Here, we tested if Tc8 was activated in *mut-7*, *rde-1*, *-2*, *-3* and *-4*. Although, the different components underlying the interference process have not been completely elucidated, *mut-7* was determined to code for a RNAseD homolog (KETTING *et al.* 1999) whereas, RDE-1 contains a PAZ/Piwi domain also found in AGO-1, which is involved in RNA silencing in *Arabidopsis* (FAGARD *et al.* 2000; KETTING *et al.* 1999). Our results indicate that Tc1 is activated in the mutants examined but not in the wild-type strain Bristol N2. However, we did not observe mobility for the *Tourist*-like element Tc8 in these same lines.

Although *C. elegans* and *Arabidopsis Tourist* elements examined here did not show mobility, we did not examine the expression levels of the putative transposases for

these elements. Transcription of transposases may be inhibited because promoters are inaccessible due to methylation or to the binding of repressors. The *Tourist* transposase itself may be such a repressor. This would be similar to the case of maize *Ac* elements where the transposase bound to its own promoter prevents transcription (Fridlender, 1996). Transposons can also insert into chromosomal regions that later become condensed into chromatin. Consequently, they are physically inaccessible to the transcription apparatus or to the transposase. In fact, the majority of *Arabidopsis* elements are clustered within heterochromatin (THE ARABIDOPSIS GENOME INITIATIVE 2000). Although Tc8 insertions appear to be random (LE *et al.* 2001), we do not know whether there is a position-effect exerted on these elements. Even if the transposase is expressed, mobility may be repressed by mechanisms such as overproduction inhibition, similar to the case of the distantly related *mariner*-like elements, or titration, where functional transposases are competed for by defective transposons (HARTL *et al.* 1997a; HARTL *et al.* 1997b; SIMMONS and BUCHOLZ 1985). Thus, examining transposon activity in *Arabidopsis* and additional *C. elegans* mutants for transcriptional and post-transcriptional silencing and gene regulation mechanisms may be informative. However, it is also conceivable that *Tourist-10*, *Tourist-11* and Tc8 activity was not detected simply because these elements may be defective transposons.

C. elegans and *Arabidopsis Tourist* elements are possibly defective transposons

We re-examined Tc8, *Tourist-10* and *Tourist-11* sequences in order to elucidate why we did not detect mobility for these transposons. First, Tc8, *Tourist-10* and *Tourist-11* are highly AT-rich (64%, 75% and 66%, respectively). For reference, the active *PIF*α in maize has an AT-content of 59% (gi 15987054). Repeat induced point mutations such as C to T transitions are thought to be a mechanism by which the host can permanently destroy transposable elements in the long term (BESTOR 1999). Thus, *Tourist* sequences have had the time to accumulate enough point mutations to lose the capacity to code for functional transposases or because members have lost *cis*-sequences required for transposition. Consequently, the AT-richness may be a reflection of their actual old age. Despite the fact that few members within a family still share high nucleotide sequence similarities (>95%), many other members have significantly diverged. In addition, the

mere presence of *Tourist* members in a wide range of hosts (LE *et al.* 2001) indicates that these elements are ancient components of these genomes. This case is similar to the IS630/Tc1/*mariner* transposons which are also found in a wide range of hosts, often times with many families cohabitating a genome but only a handful are known to be currently active (DE AGUIAR and HARTL 1999). Second, the majority of Tc8 elements found in *C. elegans* are short imperfect hairpins that do not carry any coding capacity. A larger and apparently full-length element does harbour an ORF coding for the putative transposase though this coding sequence may have suffered insertions of other transposon types (LE *et al.* 2001). Another ORF similar to the *Tourist* transposase is present in *C. elegans* but it is not associated with a full-length element (LE *et al.* 2001). Thus, the putative *Tourist* transposases may all be defective; meaning that every Tc8 is a non-autonomous element that would require the assistance of a related protein from another *Tourist* family member for mobility. However, database searches using the Tc8 transposase as a query did not reveal the presence of another *Tourist* family within the *C. elegans* genome.

In *Arabidopsis*, among the many putative IS5/*Tourist* ORFs found, only one *Tourist-10* and three *Tourist-11* predicted ORF actually correspond to full-length elements. Other members longer in length either do not have coding capacity or their predicted proteins show no similarities with IS5/*Tourist* transposases. Of the four putative *Arabidopsis Tourist* transposases, only one *Tourist-10* (gi 2191187) and one *Tourist-11* transposase (gi 4585884) have an intact catalytic DDE motif along with conservation of the DXG/A/YF and YREK motifs (LE *et al.* 2001). The *Tourist-10* and *-11* proteins share 34% and 38% amino acid identity with the active maize *PIFα*. Thus, only two *Tourist* transposases appear to be intact. However, our results suggest that they are non-functional, possibly due to deleterious mutations at other essential, but uncharacterized, residues in the IS5/*Tourist* transposases.

Relics of the past

The fact that 49% of Tc8 elements are all short hairpins of exactly 150-bp is intriguing. Deletion derivatives of transposons are believed to be generated by incomplete gap repair of excised elements using another allele or a homologous sequence as template (PLASTERK 1991). However, this process cannot account for the exact and

constant size of the large number of elements. One explanation is that the 150-bp Tc8 elements are non-autonomous transposons that were preferentially amplified by the transposase of the full-length Tc8 before it became inactive. The high sequence similarity between the many 150-bp sequences may also be the result of gene conversion events. An alternative, but not necessarily exclusive, explanation is that the smaller Tc8 represent transposition footprints, short sequences left behind by excised elements. The transposase of the *Drosophila P*-element was recently shown to generate a 17-bp staggered cut in the TIR during excision that results in a 34-bp excision footprint after end-joining repair (BEALL and RIO 1997). Although experimental confirmation is required, the transposition process of Tc8 may involve a large 75-bp staggered cut that would subsequently be repaired to generate a 150-bp palindrome. From the few longer Tc8 found in *C. elegans*, we know that this element has unusually long TIRs (~1300 bp). If the short elements truly represent excision footprints, then this would mean that Tc8 elements have a high rate of excision without novel insertions since nearly half of Tc8 sequences found are exactly 150 bp in length. Many *Tourist* and other miniature inverted-repeat transposable element families (better known as MITEs) are often found as short elements capable of forming perfect hairpins (BUREAU and WESSLER 1992). Unfortunately, information on the excision footprint for *PIFα*, the only active *Tourist* known to date, is not yet available.

Some authors have drawn attention to the fact that the abundance of certain transposons is phylum specific. For example, *P* elements have only been found in *Drosophila*. Long and Short Interspersed Nuclear Elements are the most abundant types of transposons found in mammalian genomes whereas *Tourist* is prevalent in plant genomes (WESSLER 1996). This discrepancy in the abundance of transposon type between different genomes reflects a complex transposon-host relationship resulting from the combination of factors that include the invasive success of an element, the responsiveness of the host genome and the selective pressures that come into play. Transposons can be considered as more or less aggressive parasites and their intrusion into a new genome occurs in three stages (for a review, see KIDWELL and LISCH 2001). Following the initial invasion, there is a rapid amplification phase of transposons before the host can react and mount its defence. Depending on the strength of the invasion and

the defensive response, a balance is eventually reached where the rate of transposition and element loss is at an equilibrium. Finally, when all autonomous transposons have been lost, elements slowly degenerate or are deleted. The *Tourist* transposons studied here may represent elements that have reached this final stage.

Conclusion

In summary, further experimentation is required to establish the effects of stress on transposition. As indicated by the results from *C. elegans*, examination of mutants for mechanisms thought to repress transposition may be a more efficient strategy to detect mobility. A re-evaluation of *Tourist* sequences suggests that these are ancient components of the *Arabidopsis* and *C. elegans* genomes. Although the maize *Tourist PIFα* is the only active element known to be currently active, IS5/*Tourist* elements were once active and very prolific considering the numerous and large families that can be found in very different genomes, but especially within plants. Taken together, our results seem to point to the fact that IS5/*Tourist* elements in *Arabidopsis* and in *C. elegans* are degenerate transposons, relics of a past invasion and ensuing explosive spread but that may be no longer capable of transposition.

ACKNOWLEDGMENTS

We are grateful to Dr. Richard Roy for providing us with *C. elegans* Bristol N2 strain and to Dr. Candace Waddell and Dr. Daniel J. Schoen for the use of equipment and comments in experiments involving *Arabidopsis*. Some nematode strains used in this work were provided by the *Caenorhabditis* Genetics Center, which is funded by the NIH National Center for Research Resources.

Table 5.1. Primers used in TD.

Primer	Amplification step**	Sequence (5' to 3')	T_m* (°C)
Adaptor	ap1	CGAATCGTAACCGTTCGTACGAGAATCGCT	var.
	ap2	GTACGAGAATCGCTGTCCTC	var.
Tourist-10	ep1***	CAGTGGATTAGTGTTGGATTTCAATTC	60
		CAGTGGATTCGGCTTGGATTTCAATTC	
	ep2***	CCATTAACTAATAACACCCCC	60
		(C/T)CATTAACGAATAACACCCCC	
Tourist-11	ep1	CATTGAGATGGTTCATCCAAATGGAC	60
	ep2	GGTTCATCCAAATGGACAAACAAACAGG	60
cac1	ep1	(T/C)TTTCGTAATGCTATGGTTGAAACACCTAAC	60
	ep2	CATACAATTCTGACGCTATC	60
Tc8	ep1	ATNTACTTGAATA(C/A)AGTTGTGAC	55
	ep2	CCGACACTACTTGAATAACCCC	55
Tc1	ep1	GCTGATCGACTCGATGCCACGTCG	55
	ep2	GATTTTGTGAACACTGTGGTGAAG	55

* Different T_m were tested for each adaptor- and element-specific primer combination but the one generating a banding pattern that was in best agreement with the predicted band sizes were used in the experiments.

** Adaptor- (ap1) and element-primer (ep1) for pre-selective amplification. Adaptor- (ap2) and element-primer (ep2) for selective amplification.

*** To minimize the number of non-specific primer sequences resulting from the different permutations of multiple degenerate positions, two separate primers were synthesized and mixed in equal parts for the PCR reactions.

FIGURE LEGENDS

Figure 5.1. Diagram summarizing the steps involved in the TD strategy.

Black shading represents genomic sequences, white shading represents transposon sequences and grey shading represents adaptor sequences. (1) Genomic DNA is digested using a restriction enzyme (*BfaI*) that cuts within the transposon and the flanking sequence. (2) A specially designed double-stranded adaptor that is not completely complementary in sequence is ligated to the digested product. (3) A pre-selective PCR step using a pair of primers targeted against the adaptor and a transposon family amplifies transposon termini with flanking sequences. The adaptor is designed such that, in the initial amplification round (3.1), elongation from the adaptor-specific primer (ap1) can only occur after first strand synthesis from the transposon-specific primer (ep1, 3.2), thus selecting for transposon-specific products. (4) A selective PCR amplification using nested primers is performed to enrich for transposon-specific products. The adaptor-specific primer (ap2) is fluorescent labeled (*) to allow detection on polyacrylamide gel electrophoresis (5). A transposition event results in a change in the flanking sequence that can be visualized as a novel (insertion) or a missing band (excision).

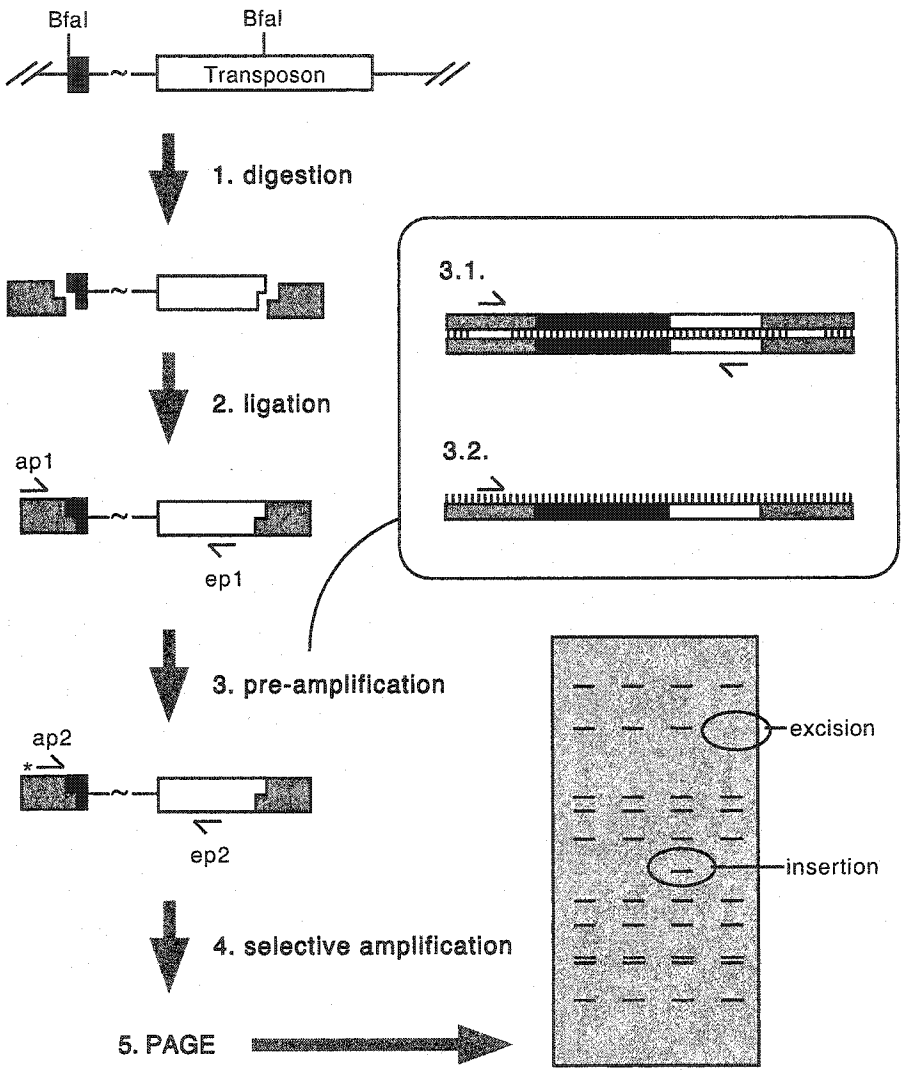

BfaI
BfaI
Transposon
1. digestion
2. ligation
ap1
ep1
3. pre-amplification
3.1.
3.2.
ap2
*
ep2
4. selective amplification
5. PAGE
excision
insertion

Figure 5.2. Tc8 is not active in *C. elegans* lines resistant to RNAi.

Transposon Display showing Tc8 (right panel) and Tc1 (left panel) banding pattern in the different mutants tested. Mutant strains are indicated above TDs; (N2) wild type Bristol N2, (M.W.) Li-Cor 50-700 fluorescent molecular weight marker, (predicted) bands predicted by computer analysis (see Appendix 2). Examples of polymorphic bands are indicated (*).

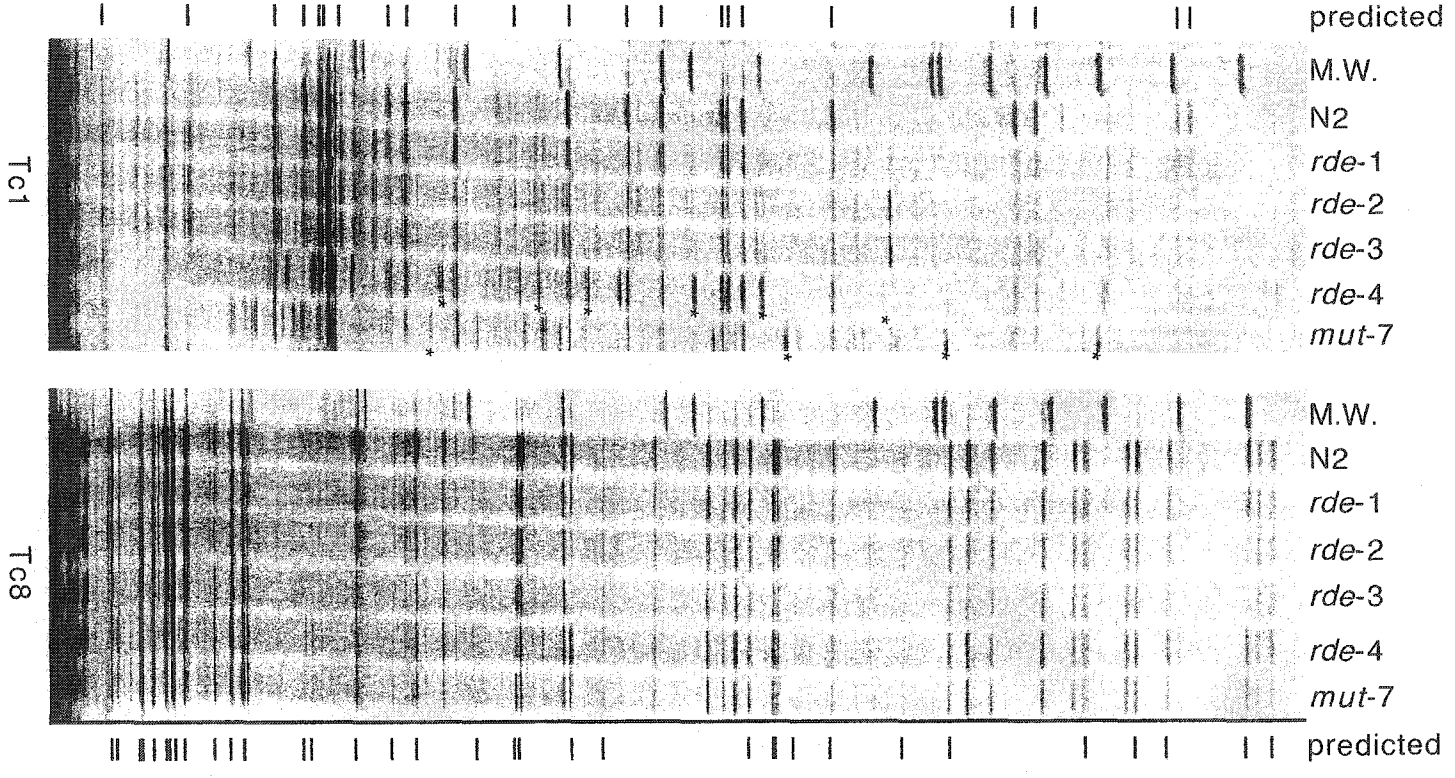
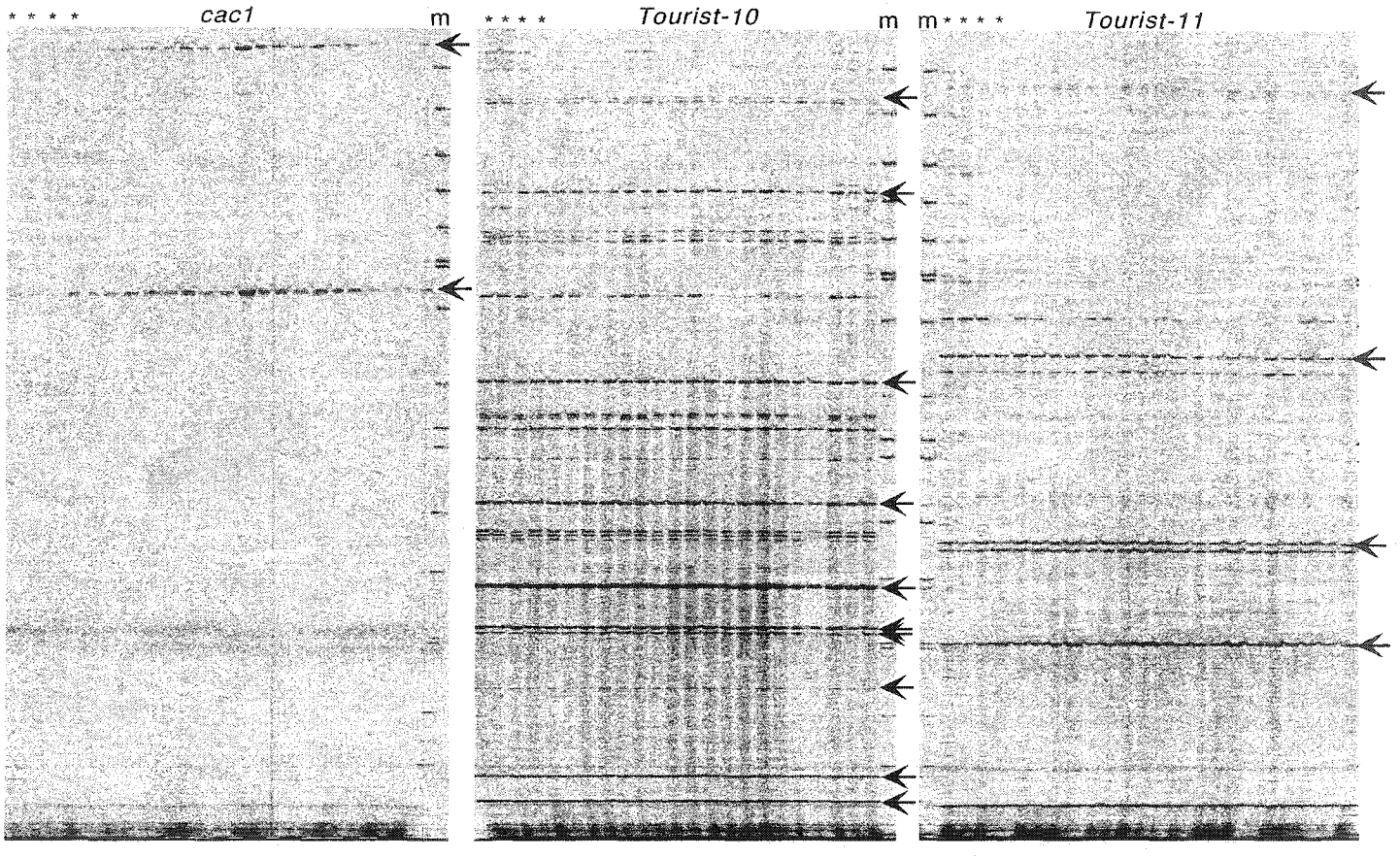

predicted
M.W.
N2
rde-1
rde-2
rde-3
rde-4
mut-7
Tc1
M.W.
N2
rde-1
rde-2
rde-3
rde-4
mut-7
Tc8
predicted

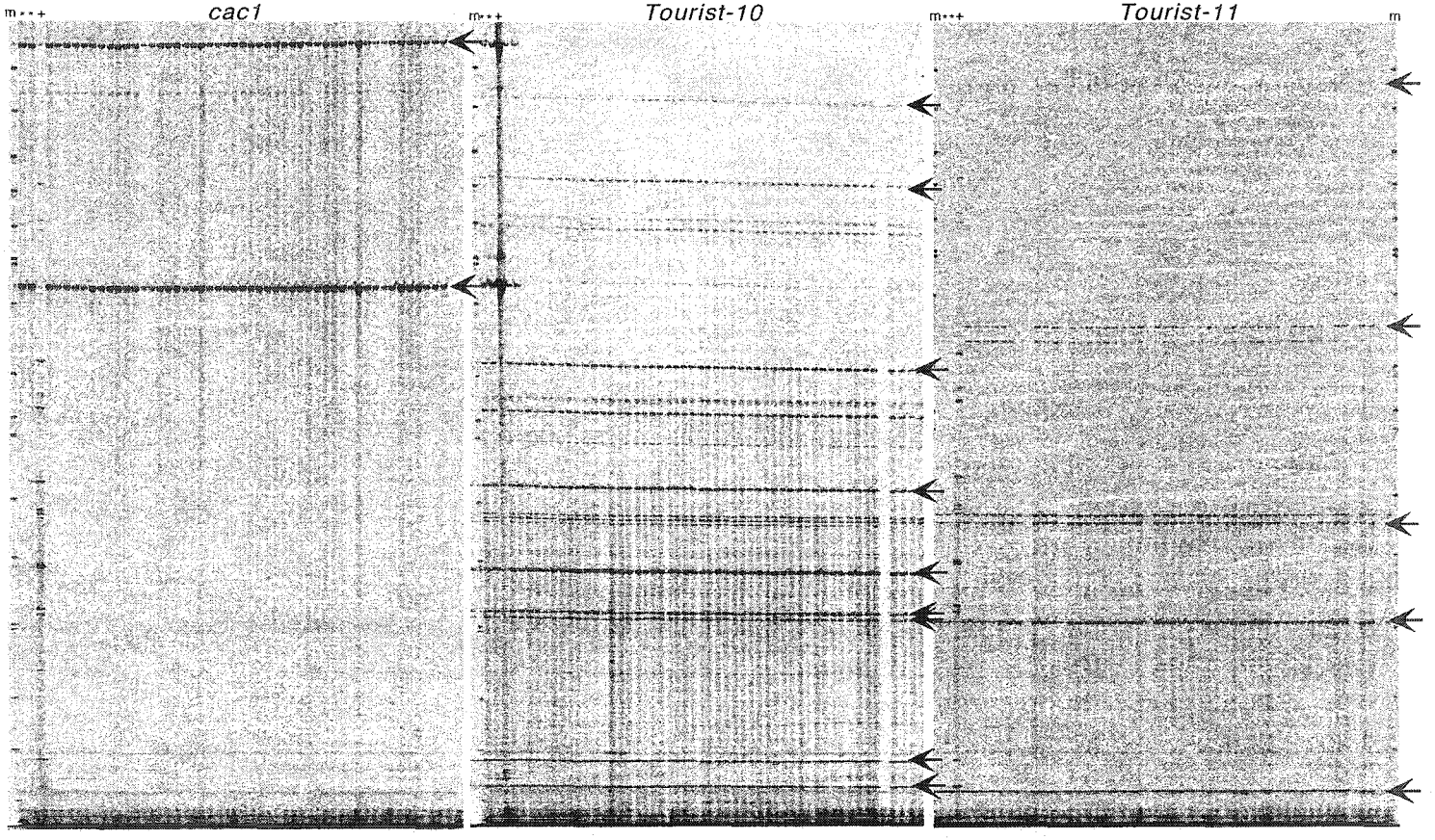
Figure 5.3. *Tourist* and *cac1* are not mobile in *Arabidopsis* submitted to UV-B or tissue culture stresses.

TD of *cac1*, *Tourist-10* and *-11* activity in the progeny of UV-B treated plants (*A*), in plants irradiated as seeds (*B*) or in plants undergoing tissue culture (*C*). Lanes marked with (*) indicates *Arabidopsis* plants that have not been submitted to stress; (m) indicates molecular weight marker (Li-Cor 50-700-bp marker); and (+) indicates additional controls for PCR amplification. Unpredicted bands that appear throughout the samples may correspond to mispriming or elements located in repetitive regions of the genome where the sequence has not been completely resolved.

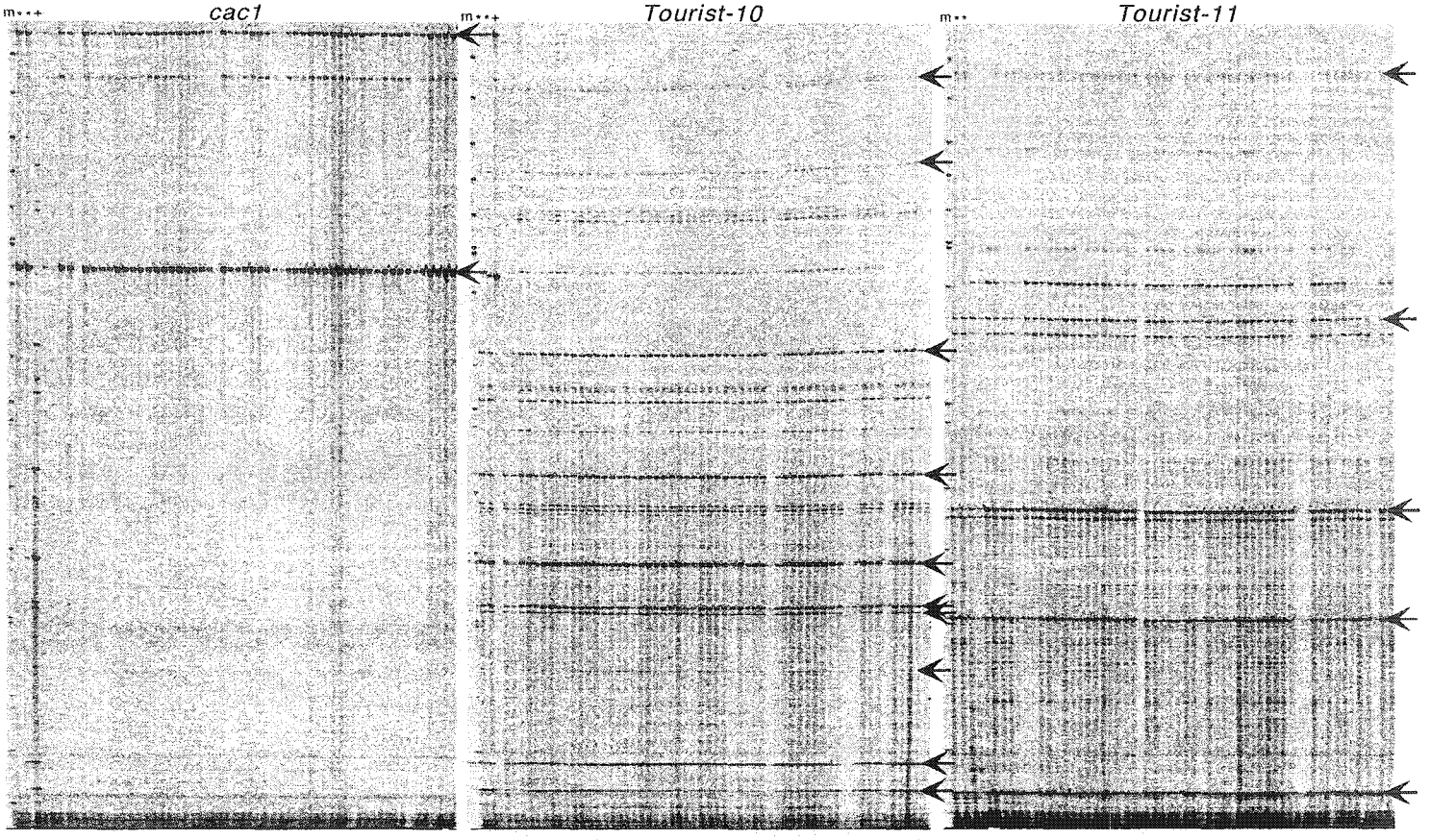
A



B



C



REFERENCES

BAULCOMBE D. C., (2000). Molecular biology. Unwinding RNA silencing. *Science* **290,** 1108-1109.

BEALL E. L. and RIO, D. C., (1997). *Drosophila P*-element transposase is a novel site-specific endonuclease. *Genes Dev.* **11,** 2137-2151.

BESTOR T. H., (1999). Sex brings transposons and genomes into conflict. *Genetica* **107,** 289-295.

BHATT A. M., LISTER, CLARE, CRAWFORD, NIGEL, DEAN, CAROLINE, (1998). The transposition frequency of *Tag1* elements is increased in transgenic *Arabidopsis* lines. *Plant Cell* **10,** 427-434.

BRITTEN R. J., (1996). Cases of ancient mobile element DNA insertions that now affect gene regulation. *Molecular phylogenetics and evolution* **5,** 13-17.

BUREAU T. E., RONALD, P. C. and WESSLER, S. R., (1996). A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc. Natl. Acad. Sci. USA* **93,** 8524-8529.

BUREAU T. E. and WESSLER, S. R., (1992). *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell* **4,** 1283-1294.

BUREAU T. E. and WESSLER, S. R., (1994a). Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. *Proc. Natl. Acad. Sci. U S A* **91,** 1411-1415.

BUREAU T. E. and WESSLER, S. R., (1994b). *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6,** 907-916.

CAPY P., GASPERI, G., BIEMONT, C. and BAZIN, C., (2000). Stress and transposable elements: co-evolution or useful parasites? *Heredity* **85,** 101-106.

COEN D., LEMAITRE, B., DELATTRE, M., QUESNEVILLE, H., RONSSERAY, S. *et al.*, (1994). *Drosophila P* element: transposition, regulation and evolution. *Genetica* **93,** 61-78.

COGONI C. AND MACINO G. (1999). Gene silencing in *Neurospora crassa* requires a protein homologous to RNA-dependent RNA polymerase. *Nature* **399,** 166-169.

DE AGUIAR D. and HARTL, D. L., (1999). Regulatory potential of nonautonomous mariner elements and subfamily crosstalk. *Genetica* **107,** 79-85.

FAGARD M., BOUTET, S., MOREL, J. B., BELLINI, C. and VAUCHERET, H., (2000). AGO1, QDE-2, and RDE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals. *Proc. Natl. Acad. Sci. USA* **97,** 11650-11654.

FEDOROFF N., (2000). Transposons and genome evolution in plants. *Proc. Natl. Acad. Sci. USA* **97,** 7002-7007.

FRIDLENDER M., SITRIT, Y., SHAUL, O., GILEADI, O. and LEVY, A. A., (1998). Analysis of the *Ac* promoter: structure and regulation. *Mol. Gen. Genet.* **258,** 306-14.

GRANDBASTIEN M.-A., (1998). Activation of plant retrotransposons under stress conditions. *Trends Plant Sci.* **3,** 181-187.

HABU Y., KAKUTANI, T. and PASZKOWSKI, J., (2001). Epigenetic developmental mechanisms in plants: molecules and targets of plant epigenetic regulation. *Curr. Opin. Genet. Dev.* **11,** 215-220.

HARTL D. L., LOHE, A. R. and LOZOVSKAYA, E. R., (1997a). Regulation of the transposable element *mariner*. *Genetica* **100,** 177-184.

HARTL D. L., LOZOVSKAYA, E. R., NURMINSKY, D. I. and LOHE, A. R., (1997b). What restricts the activity of *mariner*-like transposable elements. *Trends Genet.* **13,** 197-201.

HIROCHIKA H., (1993). Activation of tobacco retrotransposons during tissue culture. *EMBO J.* **12,** 2521-2528.

HIROCHIKA H., OKAMOTO, H. and KAKUTANI, T., (2000). Silencing of retrotransposons in *Arabidopsis* and reactivation by the *ddm1* mutation. *Plant Cell* **12,** 357-369.

HIROCHIKA H., SUGIMOTO, K., OTSUKI, Y., TSUGAWA, H. and KANDA, M., (1996). Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl. Acad. Sci. USA* **93,** 7783-7788.

HOUCK M. A., CLARK, J. B., PETERSON, K. R. and KIDWELL, M. G., (1991). Possible horizontal transfer of *Drosophila* genes by the mite *Proctolaelaps regalis*. *Science* **253,** 1125-1128.

JEDDELOH J. A., STOKES, T. L. and RICHARDS, E. J., (1999). Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nat. Genet.* **22,** 94-97.

JENSEN S., GASSAMA, M. P. and HEIDMANN, T., (1999a). Cosuppression of *I* transposon activity in *Drosophila* by *I*-containing sense and antisense transgenes. *Genetics* **153,** 1767-1774.

JENSEN S., GASSAMA, M. P. and HEIDMANN, T., (1999b). Taming of transposable elements by homology-dependent gene silencing. *Nat. Genet.* **21,** 209-212.

KAEPPLER S. M. and PHILLIPS, R. L., (1993). Tissue culture-induced DNA methylation variation in maize. *Proc. Natl. Acad. Sci. USA* **90,** 8773-8776.

KETTING R. F., HAVERKAMP, T. H., VAN LUENEN, H. G. and PLASTERK, R. H., (1999). *mut-7* of *C. elegans*, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. *Cell* **99,** 133-141.

KETTING R. F. and PLASTERK, R. H., (2000). A genetic link between co-suppression and RNA interference in *C. elegans*. *Nature* **404,** 296-298.

KIDWELL M. G. and LISCH, D., (1997). Transposable elements as sources of variation in animals and plants. *Proc. Natl Acad. Sci. USA* **94,** 7704-7711.

KIDWELL M. G. and LISCH, D. R., (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution Int. J. Org. Evolution* **55,** 1-24.

KORSWAGEN H. C., DURBIN, R. M., SMITS, M. T. and PLASTERK, R. H. A., (1996). Transposon Tc1-derived, sequence-tagged sites in *Caenorhabditis elegans* as marker for gene mapping. *Proc. Natl. Acad. Sci. USA* **93,** 14680-14685.

LE Q. H., TURCOTTE, K. and BUREAU, T., (2001). Tc8, a Tourist-like transposon in *Caenorhabditis elegans*. *Genetics* **158,** 1081-1088.

LE Q. H., WRIGHT, S., YU, Z. and BUREAU, T., (2000). Transposon diversity in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **97,** 7376-7381.

LEWIS J. A. and FLEMING, J. T., (1995). Basic culture methods, pp. 3-29 in *Caenorhabditis elegans: Model Biological Analysis of an Organism*, edited by D. C. Epstein H.F. and Shakes. *Academic Press*, New York.

LI J., OU-LEE, T.-M., RABA, R., AMUNDSON, R. G. and LAST, R., (1993). Arabidopsis flavonoid mutants are hypersensitive to UV-B irradiation. *Plant Cell* **5,** 171-179.

MATHUR J. and KONCZ, C., (1998). Protoplast Isolation, Culture, and Regeneration, pp. 35-42 in *Arabidopsis Protocols*, edited by J. M. Martinez-Zapater and J. Salinas. *Humana Press Inc.*, Totowa, NJ.

MCCLINTOCK B., (1984). The significance of response of the genome to challenge. *Science* **226,** 792-801.

MELAYAH D., BONNIVARD, E., CHALHOUB, B., AUDEON, C. and GRANDBASTIEN, M. A., (2001). The mobility of the tobacco Tnt1 retrotransposon correlates with its transcriptional activation by fungal factors. *Plant J.* **28,** 159-168.

MHIRI C., MOREL, J. B., VERNHETTES, S., CASACUBERTA, J. M., LUCAS, H. *et al.*, (1997). The promoter of the tobacco *Tnt1* retrotransposon is induced by wounding and by abiotic stress. *Plant. Mol. Biol.* **33,** 257-266.

MIURA A., YONEBAYASHI, S., WATANABE, K., TOYAMA, T., SHIMADA, H. *et al.*, (2001). Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature* **411,** 212-214.

MOREAU-MHIRI C., MOREL, J.-B., AUDEON, C., FERAULT, M., GRANDBASTIEN, M.-A. *et al.*, (1996). Regulation of expression of the tobacco *Tnt1* retrotransposon in heterologous species following pathogen-related stresses. *Plant J.* **9,** 409-419.

O'NEILL R. J., O'NEILL, M. J. and GRAVES, J. A., (1998). Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393,** 68-72.

PEARCE S. R., KUMAR, A. and FLAVELL, A. J., (1996). Activation of the Ty1-copia group retrotransposons of potato (Solanum tuberosum) during protoplast isolation. *Plant Cell Reports* **15,** 949-953.

PLANCKAERT F. and WALBOT, V., (1989). Molecular and genetic characterization of *Mu* transposable elements in *Zea mays*: behavior in callus culture and regenerated plants. *Genetics* **123,** 567-578.

PLASTERK R. H., (1991). The origin of footprints of the Tc1 transposon of *Caenorhabditis elegans*. *EMBO J.* **10,** 1919-1925.

POUTEAU S., HUTTNER, E., GRANDBASTIEN, M. A. and CABOCHE, M., (1991). Specific expression of the tobacco *Tnt1* retrotransposon in protoplasts. *EMBO J.* **10,** 1911-1918.

SANMIGUEL P., TIKHONOV, A., JIN, Y. K., MOTCHOULSKAIA, N., ZAKHAROV, D. *et al.*, (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274,** 765-768.

SIMMONS M. J. and BUCHOLZ, L. M., (1985). Transposase titration in *Drosophila melanogaster*: a model of cytotype in the P-M system of hybrid dysgenesis. *Proc. Natl. Acad. Sci. USA* **82,** 8119-8123.

SINGER T., YORDAN, C. and MARTIENSSEN, R. A., (2001). Robertson's *Mutator* transposons in *A. thaliana* are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). *Genes Dev.* **15,** 591-602.

SULSTON J. and HODGKIN, J., (1988). Methods, pp. 587-595 in *The nematode Caenorhabditis elegans*, edited by W. B. Wood. *Cold Spring Harbor Laboratory*, New York.

TABARA H., SARKISSIAN, M., KELLY, W. G., FLEENOR, J., GRISHOK, A. *et al.*, (1999). The *rde-1* gene, RNA interference, and transposon silencing in *C. elegans*. *Cell* **99,** 123-132.

THE ARABIDOPSIS GENOME INITIATIVE, (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796-815.

VANCE V. and VAUCHERET, H., (2001). RNA silencing in plants--defense and counterdefense. *Science* **292,** 2277-2280.

WALBOT V., (1999). UV-B damage amplified by transposons in maize. *Nature* **397,** 398-399.

WENCK A. R. and MARTON, L., (1995). Large-scale protoplast isolation and regeneration of *Arabidopsis thaliana*. *Biotechniques* **18,** 640-643.

WESSLER S. R., (1996). Turned on by stress. Plant retrotransposons. *Curr. Biol.* **6,** 959-961.

WIANNY F. and ZERNICKA-GOETZ, M., (2000). Specific interference with gene function by double-stranded RNA in early mouse development. *Nat. Cell Biol.* **2,** 70-75.

WRIGHT S. I., LE, Q. H., SCHOEN, D. J. and BUREAU, T. E., (2001). Population dynamics of an *Ac*-like transposable element in self- and cross-pollinating *Arabidopsis*. *Genetics* **158,** 1279-1288.

WU-SCHARF D., JEONG, B., ZHANG, C. and CERUTTI, H., (2000). Transgene and transposon silencing in *Chlamydomonas reinhardtii* by a DEAH-box RNA helicase. *Science* **290,** 1159-1162.

YODER J. A., WALSH, C. P. and BESTOR, T. H., (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13,** 335-340.

ZHANG X., FESCHOTTE, C., ZHANG, Q., JIANG, N., EGGLESTON, W. B. *et al.*, (2001). *P* instability factor: an active maize transposon system associated with the amplification of *Tourist*-like MITEs and a new superfamily of transposases. *Proc. Natl. Acad. Sci. USA* **98,** 12572-12577.

SUMMARY AND CONCLUSION

Although the *Arabidopsis* genome is relatively small with few repetitive DNA, data mining revealed that it nonetheless contains a great diversity of transposons. The identification of RESites suggest that elements found through data mining are *bona fide* transposons by presenting evidence of past mobility. Many of the elements identified belonged to previously known groups and served to consolidate our knowledge of existing families. Some, however, like *Basho* and *Katydid*, are novel types with unique structural features. Others, like *Tourist*, *Emigrant* and *Stowaway*, were known but the discovery of longer members coding for putative transposase led to a re-evaluation of our understanding of their mobility and their origins. Furthermore, examples of transposons that contributed sequences to cellular genes, acquired cellular sequences or are involved in recombination highlight the potential impact of transposons on the evolution of gene and genome structure in *Arabidopsis*.

The discovery of full-length members and putative transposases for *Arabidopsis Tourist* clarified the origins and life histories of this family of transposons. *Tourist* elements were once thought to be restricted to plants but database searches extended to other organisms revealed their presence in *C. elegans* as well as in many distant and diverse genomes. Along with bacterial IS5 transposons, *Tourist* is found to form a superfamily of elements widespread in eukaryotes and prokaryotes. The data gathered from the transposon survey also served to initiate experiments to investigate *Tourist* activity in *Arabidopsis* and *C. elegans*. Using TD to detect transposition events, attempts were made to mobilize *Tourist* elements in *C. elegans* mutants resistant to RNAi and in *Arabidopsis* submitted to UV-B and tissue culture stresses. Altogether, results from transposon mining and experimental approaches suggest that, despite a prolific past, *Tourist* appears to be no longer active in these two organisms.

Transposon mining in *Arabidopsis* has generated critical information for the analysis of *Tourist* elements. A large number of other transposon families were

uncovered from the survey but have not been studied in-depth here. We may have only scratched the surface as many questions still need to be answered. For example, the mode of mobility of *Katydid* and *Basho* remains unclear. With the full complement of *Arabidopsis* transposon now available, we can establish the evolutionary history of elements. Systematic identification of transposable elements can be done in other systems as many more genome sequences will be made available. Transposon abundance and organization within the genome of *Arabidopsis* can also be compared with close relatives, such as other ecotypes or species of the sister group *Arabis*. Mining for elements in other systems promises to be interesting since most eukaryotic organisms have a greater transposon content than *Arabidopsis*. Comparative studies between phyla will help to better understand why there are often large differences in the abundance of transposon type between organisms (*e.g.* the low representation of SINEs in plants compared to their numerous mammalian counterparts).

While *in silico* strategies accelerates element discovery and characterization, TD offers a fast and efficient method to monitor transposition in small populations. Although the *Tourist* elements identified and examined here may be transposon relics incapable of transposition, many other families of transposons were found during data mining. Some of these families included members sharing high sequence similarity or harbouring ORFs for mobility-related proteins, which suggest potential activity. Considering the unusual structures of *Basho* and *Katydid* elements, additional information regarding their transposition process will be most revealing. TD combined with genetic approaches may also help to appreciate the host mechanisms thought to regulate transposition. Even though we were not able to detect mobility for Tc8, this and the previous analysis of Tc1 activity have demonstrated that host regulation of transposition can readily be examined in different RNAi-deficient *C. elegans*. This can be extended to other systems, such as in *Arabidopsis*, where large collections of mutants are available and characterized.

In addition, data mining and TD are tools well-suited to examine transposon activation by environmental stresses. At the moment, the pathways and processes linking stress perception and element activation are not fully understood. From an ecological perspective, and in light of such phenomena as global warming and thinning of the ozone

layer, it may be imperative for us to better comprehend the implications of stress activation and the underlying mechanisms.

In my thesis, the example with *Tourist* showed how the analysis of genomic sequence information offers an invaluable tool for the better understanding of transposons and how *in silico* data can serve as a platform to design and initiate experiments using molecular tools. These combined approaches can be efficiently applied to study the multitude of elements that will be identified from ongoing genome sequencing projects, and to provide further insight into the evolutionary impact of transposons.

APPENDIX 1

SUPPORTING INFORMATION

Supporting information for chapter 3 (Table 3.2 and Figure 3.6).

WITTE, C.-P., LE, Q.H., BUREAU, T. AND KUMAR, A. (2001). Terminal-repeat Retrotransposons in Miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl. Acad. Sci. USA* **98,** 13778-13783.

Table 3.2. Supporting Table.

GenBank accession	gi number	plant	position[¶]	length	TSD	comments
Solanaceae						
BE341675.1	9251206	potato	1-555	555	-	EST; two elements joined in LTR region
BF053887.1	10807699	potato	~160-380	221	-	EST; 5'LTR truncated
BE471879.1	9562370	potato	196-546	351	-	EST; translation is similar to NBS-LRR proteins; available sequence ends in 5'LTR
BE922581.1	10448657	potato	292-409	118	GTTGT/ ATTGT	EST; solo LTR; translation is similar to protein kinases; overlaps with kinase ORF
Z11883.1	21590	potato	4292-4413	122	GAAGT/ GAGGA	solo LTR
Z12753.1	21553	potato	25-379	355	ATAAG/ GTACG	in promoter region of gene for proteinase inhibitor II
AJ276865.1	14599414	potato	1522-1856	334	GAGCT/ AAGCT	in intron 6 of gene for urease (allelic form 1)
AJ276864.1	14599411	potato	7601-7737 + 9493-9695	340	GAGCT/ AAGCT	in intron 6 of gene for urease (allelic form 2); interrupted by Ty1-*copia* insertion
AJ409162.1	14599445	*Solanum ochran-thum*	468-640 + 4244-4407	337	GAGTT/ AAGCT	in intron 6 of gene for urease; interrupted by LINE insertion

AW650487.1	7411725	tomato	60-385	326	CTATG/CTACC	EST
BF096974.1	10902684	tomato	1-202	202	-	EST; translation is similar to glycolate oxidase; available sequence ends in 3'LTR
AI781248.1	5279289	tomato	~210-507	298	-	EST; 5'LTR truncated
AF259793.1	10764221	tomato	1922-2391	470	-	in intron of aldehyde oxidase gene; three LTRs in complex truncated structure
AR110591.1	12826867	tomato	~1471-1812	342	-	truncated
AF198177.1	7688741	tomato	1-149	149	-	in promoter region of Asc1 gene; available sequence ends in 3'LTR
M20241.1	170430	tomato	2528-2781	254	GATGC/GATGC	in intron 3 of Cab-7 gene; deletions in the LTRs
AF001000.1	2459810	tomato	~4627-4846	220	-	downstream of polygalacturonase 1 gene; complex truncated structure
X53043.1	19200	tomato	6- ~201	196	-	upstream of gene for elongation factor 1-alpha; truncated (one LTR missing).
AQ367378.1 |	4220970	tomato	38-378	341	TATCT/TGTCT	GSS
AW399723.1	6918193	Lycopersicon *pennellii*	1-297	297	-	EST; available sequence ends in 3'LTR

AF231351.1	9392606	tobacco	3821-4182	362	GTTAT/GTCAT	in intron 4 of glucose-6-phosphate dehydrogenase
Fabaceae						
AV412624.1	7741792	*Lotus japonicus*	99-400	302	-	EST; translation similar to nodulin-like proteins; available sequence ends in 5'LTR
Y12859.1	2073447	*Lotus japonicus*	19770-19960	191	-	in intron 2 of *krm* gene; truncated
BE204536.1	8747823	*Medicago trunculata*	244-363	119	ATTTC/ATTTC	EST; solo LTR
AW773783.1	7717696	*Medicago trunculata*	~422-681	259	-	EST; truncated complex structure involving three LTRs; available sequence ends in LTR.
BE942282.1	10520041	*Medicago trunculata*	344-458	114	CTAGT/ATAGT	EST; translation is similar to cytochrome P450-like proteins; solo LTR
AJ388733.1	6604001	*Medicago trunculata*	~269-364	96	-	EST; solo LTR; truncated
BF005269.1	10705544	*Medicago trunculata*	370-~522	153	-	EST; truncated
BF635034.1	11899192	*Medicago trunculata*	137-253 458-574	117 117	TAATC/GAAAC GTGTG/GCGGN	EST; two solo LTRs
BG452435.1 BG451390.1	13371229 13370184	*Medicago trunculata*	304-422	119	ATTTC/TATTC	EST; solo LTR
BG450464.1	13369162	*Medicago trunculata*	67-354	288	-	EST; truncated element

BG453410.1	13372204	*Medicago trunculata*	~209-323 362-~510	115 149	-	EST; solo LTR and truncated element
X13287.1	19642	*Medicago sativa*	8088-8488	401	TTACT/ TTACT	in intron 9 of nodulin-25 gene
Y15080.1	2576326	*Phaseolus vulgaris*	1504-1878	375	AACAC/ AACAC	between genes of tRNA-Pro
Z15046.1	20981	*Phaseolus vulgaris*	310-432	123	ATATA/ ATATG	in promoter region of gene for chalcone isomerase; solo LTR
AW509076.1	7147154	soybean	1-138	138	-	EST; available sequence ends in internal region
Poaceae						
AP003381.1	13365598	rice	167032-167421	390	ATCTT/ ATTTT	chromosome 1
AP003052.1	11967924	rice	10961-11367	407	ATTAT/ CCGTG	chromosome 1; next to gene of 14-3-3-like protein (5'); directly flanked by truncated Ty1-*copia Opie-2*-like element (3').
AP002816.1	9558485	rice	23864-~24136	273	-	chromosome 1; truncated, only 3'LTR and partial internal region; spans exon 2 of hypothetical protein (BAB03408.1)
AP002861.2l	10800055	rice	29390-29797	408	GATAC/ GATAC	chromosome 1; overlaps with the coding region of hypothetical protein (BAB16482.1)

AP002839.1 AP002744.2	9711848 11761068	rice	74563-74961 147478-147876	399	GCCGA/GCCGA	chromosome 1; overlaps with last exon of hypothetical protein (BAB19113.1)
AJ245900.1	5777612	rice	77188-77590	403	AAAAC AAGAC	chromosome 4; covers first exon and promoter of hypothetical protein (CAB53487.1)
AC027657.2 (A)	13185063	rice	62292-62782	491	ATTAT/ ATTAT	chromosome 10; longer internal region
AC027657.2 (B)	13185063	rice	65213-~65423	211	-	chromosome 10; truncated (3'LTR missing)
AC027657.2 (C)	13185063	rice	50812-50926	115	CATTT/ CATTT	chromosome 10; solo LTR
AP003023.2	13872907	rice	90750-91137	388	AAAAT/ ATAAG	chromosome 1
AP002869.1 (A)	10257386	rice	119158-119279	122	CCCTG/ CCCTG	chromosome 1; solo LTR; in intron 2 of hypothetical protein (BAB39248.1)
AP002869.1 (B)	10257386	rice	~107449-~107666	218	-	chromosome 1; truncated internal region + truncated 3'LTR; overlaps with the coding region of a hypothetical protein (BAB39245.1)
AP002483.1	8468045	rice	29267-29382	116	TATTA/ TTCCT	chromosome 1; solo LTR; in intron 4 of a hypothetical protein (BAB16449.1)

AC079685.2	12597733	rice	82951-83060	110	GGCTT/GGCTT	chromosome 10; solo LTR
AC074283.3	11527448	rice	78491-79003	513	CATGT/AGTCT	chromosome 10; longer internal region; between genes
AU056663.1	4715547	rice	~356-456	100	-	EST; truncated LTR; translation is similar to LRR-protein from *A. thaliana* (AAF82144.1)
AF090447.2	13606087	maize	230428-230567 + 239490-239883 and 263504-263642 + 272564-272957	533	AATAA/AATAA	near alpha zein gene cluster within region present in two tandem repeats (a larger genomic region was duplicated including the TRIM). A Ty1-copia retrotransposon is inserted in the 3' LTR of the TRIM.
AZ921375.1l	13392948	maize	191-310	120	-	partial LTR
Brassicaceae						
AB015478.1	3241926	Arabidopsis	68600-71176	2577	TCGAG	Katydid-At1; contains helicase, Chromosome 5
AF072897.1	3264774	Arabidopsis	69113-69460	348	TATAC	Katydid-At1, Chromosome 4
AF074021.1	3309259	Arabidopsis	95524-95883	360	TAATA	Katydid-At1; same as 7267212, Chromosome 4
AC005359.1	3367500	Arabidopsis	60901-61248	348	GTATG/GTATA	Katydid-At1, Chromosome 4

AB020754.1	3985957	Arabidopsis	8598-8928	331		Katydid-At1, Chromosome 5
AL049746.1	4741184	Arabidopsis	80497-80741	245	CTAAA	Katydid-At1, Chromosome 3
AB026657.1	4757413	Arabidopsis	42243-42565	323		Katydid-At1, Chromosome 3
AP000388.1	5672588	Arabidopsis	2302-2606	305		Katydid-At1, Chromosome 3
AL133298.1	6522607	Arabidopsis	21548-21907	360	TTTAG	Katydid-At1, Chromosome 3
AL133298.1	6522607	Arabidopsis	44123-44422	300		Katydid-At1, Chromosome 3
AC005824.2	6598490	Arabidopsis	7363-7722	360	CATAC	Katydid-At1, Chromosome 2
AC006439.3	6598569	Arabidopsis	33486-33845	360	CTGTC	Katydid-At1, Chromosome 2
AC007069.5	6598658	Arabidopsis	48904-49263	360	GATAA	Katydid-At1, Chromosome 2
AP001304.1	7209738	Arabidopsis	21726-22085	360	ATCTT	Katydid-At1; with RESite; part of cyt P-450, Chromosome 3
AL161501.2	7267212	Arabidopsis	62195-62554	360	TAATA	Katydid-At1; same as 3309259. Chromosome 4
AL161514.2	7267546	Arabidopsis	20927-21274	348	GTATG/GTATA	Katydid-At1, Chromosome 4
AL163832.1	7573477	Arabidopsis	31290-31497	208		Katydid-At1, Chromosome 3
AL132969.2	7629988	Arabidopsis	45629-45988	360	ATAAC	Katydid-At1; with RESite, Chromosome 3
AL021749.1	2842474	Arabidopsis	23191-23272	82	ATGTC	Katydid-At1, Chromosome 4
AB016873.1	3449314	Arabidopsis	50050-50165	116	ATAAA	Katydid-At1, Chromosome 5
AB016880.1	3449321	Arabidopsis	69279-	116	ATTGA	Katydid-At1,

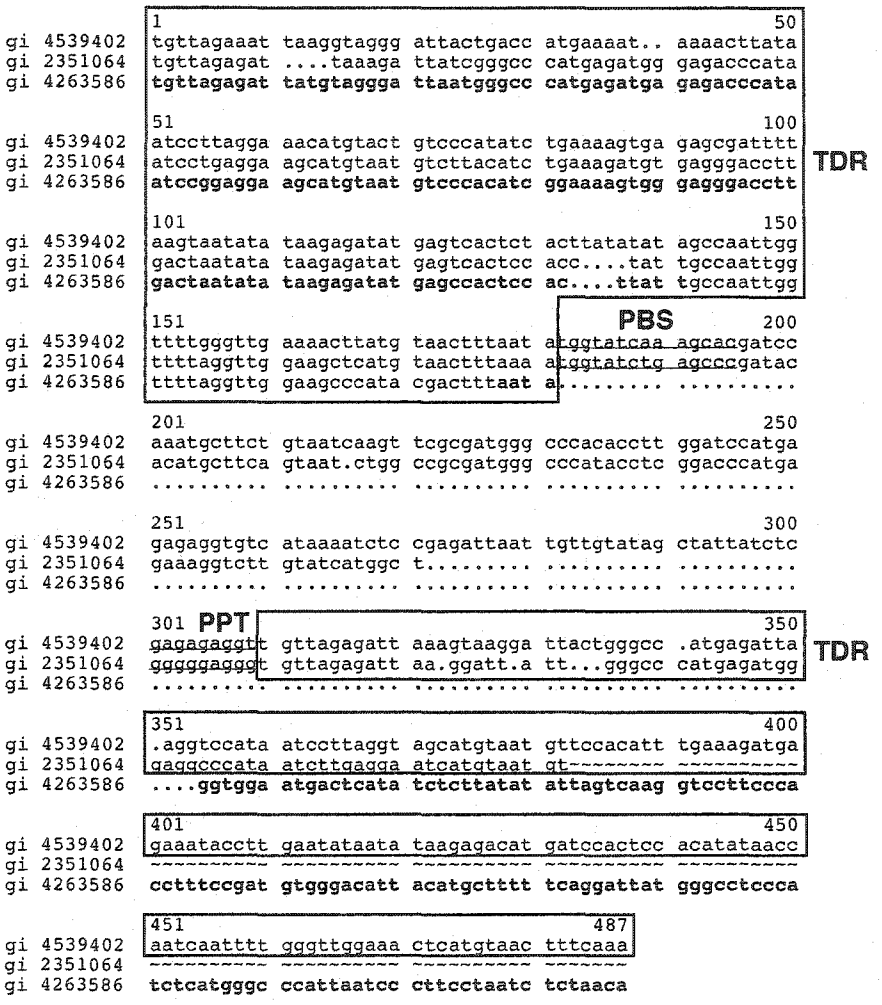
			69394			Chromosome 5
AB017064.1	3510340	Arabidopsis	22673-22946	274		Katydid-At1; part of cyt-P450, Chromosome 5
AB018112.1	3702730	Arabidopsis	65334-65449	116	ATTAT/ ATTCT	Katydid-At1, Chromosome 5
AC008148.2	5686694	Arabidopsis	39930-40045	116	GTTCC	Katydid-At1, Chromosome 1
AC006284.3	6598551	Arabidopsis	78366-78566	201		Katydid-At1, Chromosome2
AC006304.3	6598556	Arabidopsis	64383-64494	112	TCTTG	Katydid-At1, Chromosome 2
AC007171.4	6598695	Arabidopsis	40187-40282	96	ATATA	Katydid-At1, Chromosome 2
AL161572.2	7269643	Arabidopsis	161108-161189	82		Katydid-At1, Chromosome 4
AC068809.1	7767732	Arabidopsis	65033-65392	360	AATAT	Katydid-At1, Chromosome 5
AC007152.9	9581921	Arabidopsis	35330-35433	104	AAGGG	Katydid-At1, Chromosome 1
AC015446.4	9828652	Arabidopsis	65909-66000	92		Katydid-At1, Chromosome 1
AF296835.1	9885848	Arabidopsis	11166-11224	59		Katydid-At1, Chromosome 5
AC004708.1	3150006	Arabidopsis	101444-102047	604	GAGTT	Katydid-At1, Chromosome 2
AF069299.1	3193311	Arabidopsis	70895-71498	604	GAGTT	Katydid-At1, Chromosome 4
AL161471.2	7267087	Arabidopsis	76033-76636	604	GAGTT	Katydid-At1; Triple LTR elements; another copy on gi 3193311, Chromosome 4
AL353871.1	7649355	Arabidopsis	68877-69262	386	ATATT	Katydid-At2, Chromosome 3
AL138644.1	6899913	Arabidopsis	50623-52206	1584	AGGTC/ TGGTA	Katydid-At2, Chromosome 3

AL138644.1	6899913	Arabidopsis	57019-58484	1466		Katydid-At2, Chromosome 3
AP000415.1	5832736	Arabidopsis	28562-28932	371	AATTA/GTAAT	Katydid-At2; From a duplication? Pretty sure about the ends but TSD not identical, Chromosome 3
AC007190.4	5019262	Arabidopsis	59904-60265	362	TAAAT/TAAAA	Katydid-At2, Chromosome 1
AC007047.6	6598653	Arabidopsis	51022-51354	333	ATCCT/ACCCT	Katydid-At2, Chromosome 2
AC007210.5	6598715	Arabidopsis	9434-9764	331		Katydid-At2, Chromosome 2
AF118222.1	4115912	Arabidopsis	86667-87149	483	CAAAG/CAAAA	Katydid-At3; Same as 7267723 and 4539402, Chromosome 4
AL161517.2	7267723	Arabidopsis	119477-119959	483	CAAAG/CAAAA	Katydid-At3; Same as 4115912 and 4539402, Chromosome 4
AL049524.1	4539402	Arabidopsis	59514-59996	483	CAAAG/CAAAA	Katydid-At3; Same as 4115912 and 7267723, Chromosome 4
AB006699.1	2351064	Arabidopsis	798-1136	339	GGATA/GGAGA	Katydid-At3, Chromosome 5
AC004793.2	4263586	Arabidopsis	40126-40435	310	ATAGA	Katydid-At3; LTRs are TIRs, Chromosome 1
AB028605.1	5041958	Arabidopsis	22239-22400	162		Katydid-At3, Chromosome 5
AF296835.1	9885848	Arabidopsis	11068-11250	183		Katydid-At3, Chromosome 5

[9]truncated element positions are approximations (~) if the limits are unclear.

Figure 3.6. Alignment of *Katydid-At3* sequences.

gi numbers are indicated to the left. TDR sequences are boxed. PBS and PPT sequences are underlined and indicated. Sequences in bold represent the inverted repeat structure in gi 4263586.


1 50
gi 4539402 tgttagaaat taaggtaggg attactgacc atgaaaat.. aaaacttata
gi 2351064 tgttagagat ....taaaga ttatcgggcc catgagatgg gagacccata
gi 4263586 tgttagagat tatgtaggga ttaatgggcc catgagatga gagacccata
51 100
gi 4539402 atccttagga aacatgtact gtcccatatc tgaaaagtga gagcgatttt
gi 2351064 atcctgagga agcatgtaat gtcttacatc tgaaagatgt gagggacctt
gi 4263586 atccggagga agcatgtaat gtcccacatc ggaaaagtgg gagggacctt
101 150
gi 4539402 aagtaatata taagagatat gagtcactct acttatatat agccaattgg
gi 2351064 gactaatata taagagatat gagtcactcc acc....tat tgccaattgg
gi 4263586 gactaatata taagagatat gagccactcc ac....ttat tgccaattgg
151 PBS 200
gi 4539402 ttttgggttg aaaacttatg taactttaat atggtatcaa agcacgatcc
gi 2351064 ttttaggttg gaagctcatg taactttaaa atggtatctg agcccgatac
gi 4263586 ttttaggttg gaagcccata cgactttaat a..... .....
201 250
gi 4539402 aaatgcttct gtaatcaagt tcgcgatggg cccacacctt ggatccatga
gi 2351064 acatgcttca gtaat.ctgg ccgcgatggg cccatacctc ggacccatga
gi 4263586 ....
251 300
gi 4539402 gagaggtgtc ataaaatctc cgagattaat tgttgtatag ctattatctc
gi 2351064 gaaaggtctt gtatcatggc t..... ....
gi 4263586 ....
301 PPT 350
gi 4539402 gagagaggtt gttagagatt aaagtaagga ttactgggcc .atgagatta
gi 2351064 gggggaggg t gttagagatt aa.ggatt.a tt...gggcc catgagatgg
gi 4263586 ....
351 400
gi 4539402 .aggtccata atccttaggt agcatgtaat gttccacatt tgaaagatga
gi 2351064 gaggcccata atcttgagga atcatgtaat gt~~~~~~~~ ~~~~~~~~~~
gi 4263586 ....ggtgga atgactcata tctcttatat attagtcaag gtccttccca
401 450
gi 4539402 gaaatacctt gaatataata taagagacat gatccactcc acatataacc
gi 2351064 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
gi 4263586 cctttccgat gtgggacatt acatgctttt tcaggattat gggcctccca
451 487
gi 4539402 aatcaatttt gggttggaaa ctcatgtaac tttcaaa
gi 2351064 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~
gi 4263586 tctcatgggc ccattaatcc cttcctaatc tctaaca
TDR
TDR

APPENDIX 2

PROGRAMS

Following are the codes for three of my Perl programs, their accompanying documentation and the results I obtained from them.

`countBfa1.pl`

Calculates the sizes of *BfaI* fragments in a given sequence. Results for the *Arabidopsis* genome sequence are included.

`readblast.pl` and `virtualTD.pl`

For predicting the number and sizes of bands seen on TD based on BLAST results using element-specific primers as queries. A MacPerl version is also available but is not included.

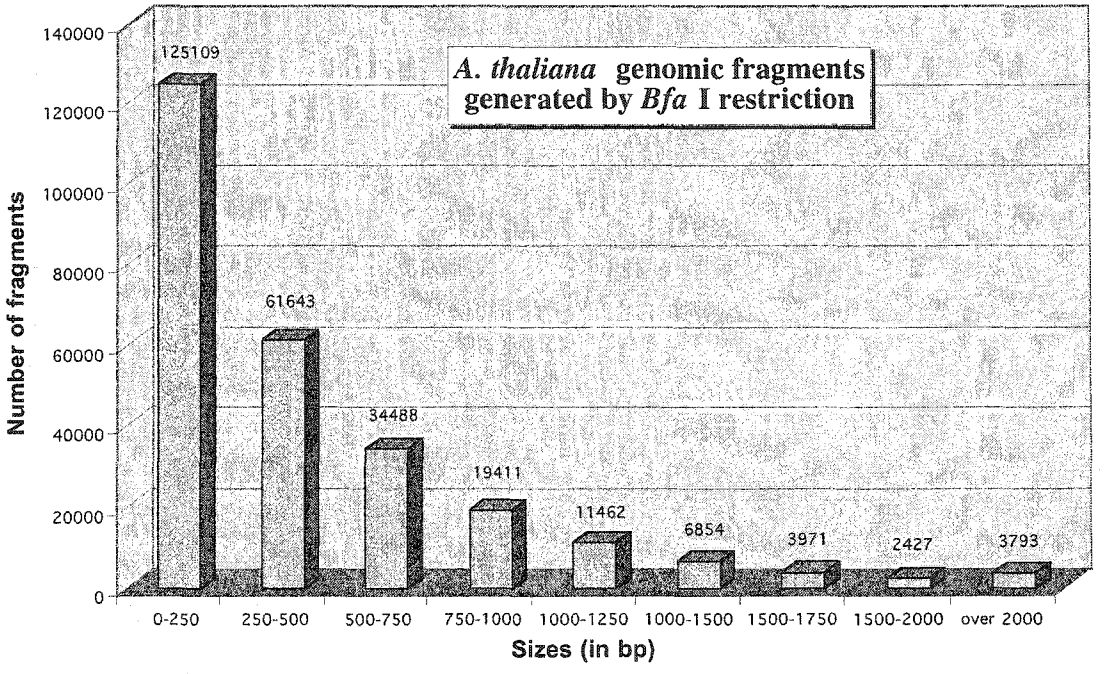
```
#!/usr/bin/perl -w
# please report bugs to hien.le@mail.mcgill.ca
# SYNTAX: cutBfa1.pl [InputFile] [OutputFile]
# INPUT:  chromosome in FASTA format
# OUTPUT: List of fragments and size in TAB file
#         followed by number of fragments smaller and greater than 1 kb

# open input file
open( InputFile, $ARGV[0] ) or die( "could not open file!\n" );
print( "\nOpening inputfile $ARGV[0]...\n" );

# delete existing output file with same name
unlink $ARGV[1];

# open output file for appending
open( OutputFile, ">>$ARGV[1]" );
print( "\nOpening output file $ARGV[1]...\n" );

# read from file, one line at a time and concatenate sequence
print ( "Concatenating sequence...\n" );
while( $line = <InputFile> ) {
      if( $line =~ /^([actgn]+)/i ) {
            $sequence .= $1;
      }
}

# print OutputFile "$sequence\n";    # for testing

# split at BfaI pattern (ctag)
print ( "Finding BfaI sites...\n" );
@fragment = split( /ctag/i,$sequence);

# Calcualte size of each fragment and group by size
print( "\nInitializing fragment counts...\n" );
$under = 0;
$twofifty = 0;
$fivehundred = 0;
$sevenfifty = 0;
$onethousand = 0;
$twelvefifty = 0;
$fifteenhundred = 0;
$seventeenfifty = 0;
$twothousand = 0;
$over = 0;
foreach $fragment (@fragment) {
    $FragmentLength = length( $fragment ) + 4;   # +4 for RE pattern
    print ( "$FragmentLength, " );
    if( $FragmentLength <= 2000 ) {
      if( $FragmentLength <= 250 ) {
                  $under++;
                  $twofifty++;
            }
            elsif( $FragmentLength <= 500 ) {
                  $under++;
                  $fivehundred++;
            }
            elsif( $FragmentLength <= 750 ) {
                  $under++;
```

```
                $sevenfifty++;
            }
            elsif( $FragmentLength <= 1000 ) {
                $under++;
                $onethousand++;
            }
            elsif( $FragmentLength <= 1250 ) {
                $twelvefifty++;
            }
            elsif( $FragmentLength <= 1500 ) {
                $fifteenhundred++;
            }
            elsif( $FragmentLength <= 1750 ) {
                $seventeenfifty++;
            }
            elsif( $FragmentLength <= 2000 ) {
                $twothousand++;
            }
    }
    else {
      $over++;
    }
}

$ArraySize = @fragment;
print ( "no more fragments!\nGenerated $ArraySize fragments\n" );
print ( "\nWARNING! Sizes of first and last fragment may be off by 3\n"
);

print OutputFile "   0 to  250 bp\t$twofifty\n";
print OutputFile " 250 to  500 bp\t$fivehundred\n";
print OutputFile " 500 to  750 bp\t$sevenfifty\n";
print OutputFile " 750 to 1000 bp\t$onethousand\n";
print OutputFile "1000 to 1250 bp\t$twelvefifty\n";
print OutputFile "1000 to 1500 bp\t$fifteenhundred\n";
print OutputFile "1500 to 1750 bp\t$seventeenfifty\n";
print OutputFile "1500 to 2000 bp\t$twothousand\n";
print OutputFile "over 2000 bp\t$over\n";

print OutputFile "\n$ArraySize Bfa1 fragments from input file
$ARGV[0]\n";
print OutputFile "$under fragments under 1 kb\n";

close( InputFile );
close( OutputFile );

__END__
```

Results from `countBfa1.pl` analysis of *Arabidopsis* genome sequences (gi NC_003070, NC_003071, NC_003072, NC_003073 and NC_003074)

```
0 to  250 bp      30863
 250 to  500 bp   15782
 500 to  750 bp   8643
 750 to 1000 bp   4848
1000 to 1250 bp   2936
1000 to 1500 bp   1765
1500 to 1750 bp   1038
1500 to 2000 bp   615
over 2000 bp      975

67465 Bfa1 fragments from input file NC_003070.fna
60136 fragments under 1 kb

   0 to  250 bp   21726
 250 to  500 bp   10248
 500 to  750 bp   5922
 750 to 1000 bp   3271
1000 to 1250 bp   1927
1000 to 1500 bp   1139
1500 to 1750 bp   660
1500 to 2000 bp   425
over 2000 bp      649

45967 Bfa1 fragments from input file NC_003071.fna
41167 fragments under 1 kb

   0 to  250 bp   25441
 250 to  500 bp   12668
 500 to  750 bp   6982
 750 to 1000 bp   3911
1000 to 1250 bp   2299
1000 to 1500 bp   1326
1500 to 1750 bp   814
1500 to 2000 bp   482
over 2000 bp      699

54622 Bfa1 fragments from input file NC_003072.fna
49002 fragments under 1 kb

   0 to  250 bp   19312
 250 to  500 bp   9294
 500 to  750 bp   5151
 750 to 1000 bp   2956
1000 to 1250 bp   1703
1000 to 1500 bp   1074
1500 to 1750 bp   576
1500 to 2000 bp   355
over 2000 bp      564

40985 Bfa1 fragments from input file NC_003073.fna
36713 fragments under 1 kb

   0 to  250 bp   27767
```

```
 250 to  500 bp    13651
 500 to  750 bp    7790
 750 to 1000 bp    4425
1000 to 1250 bp    2597
1000 to 1500 bp    1550
1500 to 1750 bp    883
1500 to 2000 bp    550
over 2000 bp       906

60119 Bfa1 fragments from input file NC_003074.fna
53633 fragments under 1 kb
```


A. thaliana genomic fragments generated by Bfa I restriction
Number of fragments
140000
120000
100000
80000
60000
40000
20000
0
125109
61643
34488
19411
11462
6854
3971
2427
3793
0-250
250-500
500-750
750-1000
1000-1250
1000-1500
1500-1750
1500-2000
over 2000
Sizes (in bp)

```
"readblast.readme"

SYNTAX:    perl readblast.pl [filename] [threshold_match]
INPUT :    [filename_]filenumber.blastn
           ex: [Basho2_2_]1.blastn
           [threshold_match] sets the primer mismatch allowed (0 to 1)
           Tested optimum = 0.85. For perfect match, use 1.
           BLAST search result as flat-anchored query
           !!USE INPUT WITH UNIX LINEBREAKS!!
OUTPUT:    [filename].tab
           For use as INPUT in retrieve_gi_sequences.pl
           Descriptions appended with _sens indicate same orientation as
clone
           whereas _rev indicate reverse orientation on clone.
SYNOPSIS: Reads BLAST search results files (of primer sequences) one
after
           the other and generates a tab delimited file of position of
primers
           on clones plus 1000 nucleotides of flanking sequences.

010923
Because of problem with regular expression in virtualTD, had to change
description lines of elements. Descriptions appended with "_rev" now
changed to "rev" to indicate reverse orientation on clone. Same
orientation as on clone now appended with "sens".
```

```
#!/usr/bin/perl -w
# please report bugs to hien.le@mail.mcgill.ca
# SYNTAX:   perl readblast.pl [filename] [threshold_match]
# INPUT :   [filename_]filenumber.blastn
#           ex: [Basho2_2_]1.blastn
#           [threshold_match] sets the primer mismatch allowed (0 to 1)
#           Tested optimum = 0.85. For perfect match, use 1.
#           BLAST search result as flat-anchored query
#           !!USE INPUT WITH UNIX LINEBREAKS!!
# OUTPUT:   [filename].tab
#           For use as INPUT in retrieve_gi_sequences.pl
#           Descriptions appended with _sens indicate same orientation
as clone
#           whereas _rev indicate reverse orientation on clone.
# SYNOPSIS: Reads BLAST search results files (of primer sequences) one
after
#           the other and generates a tab delimited file of position of
primers
#           on clones plus 1000 nucleotides of flanking sequences.


$element = $ARGV[0];
$element_rev = "$element.rev";
$element_sens = "$element.sens";
$threshold_match = $ARGV[1];

for( $filenumber = 1; ; $filenumber++ ) {

# open input file
open( InputFile, "$element$filenumber.blastn" ) or die( "could not open
file!\n"
);
print( "\nOpening inputfile $element$filenumber.blastn...\n" );

# open output file for appending
open( OutputFile, ">>$element.tab" );
print( "\nOpening output file...\n" );

# read from the file, one line at a time
while( $line = <InputFile> ) {

  # set the threshold
  if( $line =~ /^(\d)\s+(\d+)\s+\w+\s+(\d+)/ ) {
  $threshold = $3 * $threshold_match;
  print "threshold set to $threshold";
  }

  if( $line =~ /^(\d\d+)\s+(\d+)\s+\w+\s+(\d+)/ ) {
    # check orientation of hit and extend by 1000 nucleotides
    # for right orientation on clone
    if( ($3 - $2) > 0) {
        if( ($3 - $2) + 1 >= $threshold ) {
        $rightend = $3 + 1000;
        print OutputFile "$element_sens\t$1\t$2\t$rightend\n";
      }
    }
    # for reverse orientation on clone
    else{
```

```
        if( ($2 - $3) + 1 >= $threshold ) {
          $leftend = $3 - 1000;
          # avoid negative position on clone
          if( $leftend < 0 ){
            print OutputFile "$element_rev\t$1\t1\t$2\n";
          }
          else {
            print OutputFile "$element_rev\t$1\t$leftend\t$2\n";
          }
        }
      }
    }

}

close( InputFile );
close( OutputFile );

}
```

"virtualTD.readme"

```
!/usr/local/bin/perl
For use in Transposon Display analysis.
Use with Boris' retrieve_gi_sequences.pl
INPUT: output from Boris' retrieve_gi_sequences.pl
       IMPORTANT!
       FASTA sequence descriptor line must be formatted as such:
       >GroupNumber-Primer(_rev) gi:GInumber Start End OtherStuff
   Ex: >Ac6-2_rev gi:6598480 57514 58530 size:1017 Arabidops...
       >Tag2-2A gi:3046853 3796 4814 size:1019 Arabidopsis th...
       where _rev indicates orientation of primers on clone
OUTPUT: list of Nla3 fragment sizes


Known bugs and fixes

010923
Macperl does not remove *.tmp file when program and input files are not
in the same directory.
NOT FIXED!

010923
Macperl does not recognize UNIX new lines and vice versa.
NOT FIXED!

010923
Two occurences of lines
  elsif( $line =~ /^>([A-z]+\d+-\d+)( gi:)(\d+ \d+ \d+)/ ) {
causes version 3.1 to not recognize element names such as "MITE10_2A".
Changed the regular expression in both of those lines:
  elsif( $line =~ /^>(.+_sens)( gi:)(\d+ \d+ \d+)/ ) {
Version 3.2 now requires to have either the tag _rev, for reverse
orientation on clones, or _sens, for same orientation as clones,
appended to the description. Make changes to readblast.pl.
```

```
#!/usr/local/bin/perl
# please report bugs to hien.le@mail.mcgill.ca
# For use in Transposon Display analysis.
# Use with Boris' retrieve_gi_sequences.pl
# SYNTAX: perl virtualTD [InputFileName] [RestrictionPattern]
# INPUT: output from Boris' retrieve_gi_sequences.pl
#        IMPORTANT!
#        FASTA sequence descriptor line must be formatted as such:
#        >GroupNumber-Primer(_rev) gi:GInumber Start End OtherStuff
#    Ex: >Ac6-2_rev gi:6598480 57514 58530 size:1017 Arabidops...
#        >Tag2-2_snes gi:3046853 3796 4814 size:1019 Arabidopsis th...
#        where _rev indicates reverse orientation of primers on clone
#        and _sens indicates same orientation as clone.
# OUTPUT: list of Nla3 fragment sizes in "InputFileName.vtd"

# PARTI - Concatenate the sequence output from Boris'
retrieve_gi_sequences.pl
# open input file
open( InputFile, "$ARGV[0]" ) or die( "could not open $ARGV[0] file!\n"
);
print( "\nOpening inputfile $ARGV[0]...\n" );

# open output file for appending
$OutputFileName = $ARGV[0];
open( OutputFile, ">>$OutputFileName.tmp" );
print( "\nCreating temporary output file $OutputFileName.tmp...\n" );

# read from the file, one line at a time
while( $line = <InputFile> ) {

  # For reverse orientation sequences
  if( $line =~ /^>(.+rev)( gi:)(\d+ \d+ \d+)/ ) {
    $fragment = $1;
    $identifier = $3;
                          print OutputFile "\n$fragment ($identifier)\n";
  }

  # For right orientation sequences
  elsif( $line =~ /^>(.+sens)( gi:)(\d+ \d+ \d+)/ ) {
    $fragment = $1;
    $identifier = $3;
    print OutputFile "\n$fragment ($identifier)\n";
  }

             elsif( $line =~ /([actgn\s]+\n)/ ) {
     $line =~ s/\n//g;
     $line =~ s/\s//g;
     print OutputFile "$line";
  }

}

close( InputFile );
close( OutputFile );
print( "\nFormatted sequences successfully!" );

# PART II - Finds Nla3 restriction site and counts size of fragments
```

```
# choose a restriction pattern
$REpattern = $ARGV[1];

# open input file
open( InputFile, "$ARGV[0].tmp" ) or die( "could not open file
$ARGV[0].tmp!\n" );
print( "\nOpening temporary file\n$OutputFileName.tmp...\n" );

# open output file for appending
open( OutputFile, ">>$ARGV[0].vtd" );
print( "\nOpening output file\n$OutputFileName.vtd...\n" );
print OutputFile "Perl virtualTD v3.2\n$OutputFileName.vtd\n";
print OutputFile "Searching for $REpattern pattern\n";

# read from the file, one line at a time
while( $line = <InputFile> ) {

  # For reverse orientation sequences
  if( $line =~ /^(.+rev)\s+(.+)/ ) {
    $TEtype = $1;
    $identifier = $2;
    $orientation = -1;
                        print OutputFile "\n$TEtype\t$identifier";
  }

  # For right orientation sequences
  elsif( $line =~ /^(.+sens)\s+(.+)/ ) {
    $TEtype = $1;
    $identifier = $2;
    $orientation = 0;
    print OutputFile "\n$TEtype\t\t$identifier";
  }

      elsif( $line =~ /([actgn]+)/ ) {
    # split at REpattern
    @fragment = split( /$REpattern/, $1);
    $FragmentLength = length( $fragment[$orientation] ) + 30;
    print OutputFile "\t$FragmentLength";
    push @SizeArray, $FragmentLength;
  }

}

print OutputFile "\n\nOrdered by size:\n";
@SizeArray = sort {$a <=> $b} @SizeArray;
print OutputFile "@SizeArray\n";

$now_string = localtime;
print OutputFile "\n\nJob completed $now_string";

unlink <*.tmp>;
print( "Deleting temporary file $ARGV" );

close( InputFile );
close( OutputFile );
print( "\nJob completed!\n" );

__END__
```

```
Perl virtualTD v3.1
/Users/hien/documents/transposondisplay/virtualtd/mite10/mite10_2A.out.v
td
Searching for catg pattern

MITE10_2A_rev     (10086525 65690 66710)  1828
MITE10_2A_rev     (7270176 15929 16949)   191
MITE10_2A_rev     (6434225 9710 10728)    108
MITE10_2A_rev     (6598446 81596 82615)   435
MITE10_2A_rev     (12324933 15764 16784)  138
MITE10_2A_rev     (6598638 6698 7718)     213
MITE10_2A_rev     (6598529 54075 55095)   105
MITE10_2A_rev     (5225383 61832 62852)   105
MITE10_2A_rev     (7269318 165661 166681) 167
MITE10_2A_rev     (7269318 42187 43206)   566
MITE10_2A_rev     (7267319 107832 108852) 196
MITE10_2A_rev     (7267247 123357 124377) 502
MITE10_2A_rev     (4455229 83913 84933)   167
MITE10_2A_rev     (4454004 23780 24800)   246
MITE10_2A_rev     (2760169 73040 74060)   315
MITE10_2A_rev     (4678258 12763 13782)   58
MITE10_2A_rev     (12324933 15764 16784)  138
MITE10_2A_rev     (6598638 6698 7718)     213
MITE10_2A_rev     (6598529 54075 55095)   105
MITE10_2A_rev     (5225383 61832 62852)   105
MITE10_2A_rev     (7269318 165661 166681) 167
MITE10_2A_rev     (7269318 42187 43206)   566
MITE10_2A_rev     (7267319 107832 108852) 196
MITE10_2A_rev     (7267247 123357 124377) 502
MITE10_2A_rev     (4455229 83913 84933)   167
MITE10_2A_rev     (4454004 23780 24800)   246
MITE10_2A_rev     (2760169 73040 74060)   315
MITE10_2A_rev     (4678258 12763 13782)   58
MITE10_2A_rev     (6434225 9710 10728)    191
MITE10_2A_rev     (12323462 67004 68021)  196
MITE10_2A_rev     (12324388 45573 46593)  131
MITE10_2A_rev     (10086525 65690 66710)  1828
MITE10_2A_rev     (7270176 15929 16949)   1051
MITE10_2A_rev     (6598446 81596 82615)   435
MITE10_2A_rev     (3869063 32152 33169)   61
MITE10_2A_rev     (7269318 42187 43207)   567
MITE10_2A_rev     (7269318 165661 166680) 166
MITE10_2A_rev     (4678258 12763 13783)   58
MITE10_2A_rev     (12324933 15764 16783)  138
MITE10_2A_rev     (6598638 6698 7717)     212
MITE10_2A_rev     (6598529 54075 55094)   105
MITE10_2A_rev     (5225383 61832 62851)   105
MITE10_2A_rev     (7267319 107832 108851) 195
MITE10_2A_rev     (7267247 123357 124376) 501
MITE10_2A_rev     (4455229 83913 84932)   166
MITE10_2A_rev     (4454004 23780 24799)   246
MITE10_2A_rev     (2760169 73040 74059)   314
MITE10_2A_rev     (6434225 9710 10728)    191
MITE10_2A_rev     (12323462 67004 68020)  195
MITE10_2A_rev     (6692719 21457 22477)   139
MITE10_2A_rev     (6598446 81596 82616)   613
MITE10_2A_rev     (7267761 69401 70421)   613
MITE10_2A_rev     (5672588 6693 7713)     74
```

```
MITE10_2A_rev      (10086525 65690 66709)   661
MITE10_2A_rev      (7270176 15929 16948)    106
MITE10_2A_rev      (6598686 71027 72044)    75

Ordered by size:
58 58 58 61 74 75 105 105 105 105 105 105 106 108 131 138 138 138 139
166 166 167 167 167 167 191 191 191 195 195 196 196 196 212 213 213 246
246 246 314 315 315 435 435 501 502 502 566 566 567 613 613 661 1051
1828 1828
```

```
Perl virtualTD v3.1
mite11.out.bfa1.vtd
Searching for ctag pattern

Mite11-2_rev      (6598388 43672 44693)   1043
Mite11-2_rev      (7229443 21681 22702)   366
Mite11-2_rev      (9502397 17210 18231)   136
Mite11-2_rev      (6598388 43802 44825)   1054
Mite11-2_rev      (6598422 72299 73322)   667
Mite11-2          (5738368 26393 27414)   790
Mite11-2          (5738368 26523 27544)   660
Mite11-2          (7269071 75201 76222)   790
Mite11-2          (7269071 75331 76352)   660
Mite11-2          (9971623 74180 75201)   136
Mite11-2          (4235150 77519 78542)   1054
Mite11-2_rev      (4235150 71308 72335)   928
Mite11-2_rev      (5738368 21030 22057)   204
Mite11-2_rev      (7269071 69838 70865)   204
Mite11-2_rev      (9971623 69371 70398)   279
Mite11-2          (6598388 47413 48440)   423
Mite11-2          (6598422 75257 76284)   60
Mite11-2          (7229443 25085 26112)   1058
Mite11-2          (9502397 22013 23040)   279

Ordered by size:
60 136 136 204 204 279 279 366 423 660 660 667 790 790 928 1043 1054
1054 1058


Job completed Mon Sep 10 12:05:35 2001
```

```
Perl virtualTD v3.2
cac1/cac1.bfa1.vtd
Searching for ctag pattern

cac1_2_sens       (12322938 38659 39678)  478
cac1_2__rev (12322530 63965 64984)  849
cac1_2_sens       (12597825 77110 78129)  849
cac1_2__rev (6598564 76060 77079)   836
cac1_2_sens       (6598495 60644 61663)   735
cac1_2_sens       (14017746 4034 5053)    162
cac1_2_sens       (14017747 8326 9345)    161
cac1_2_sens       (14017748 7920 8939)    157
cac1_2_sens       (14017745 8349 9368)    161

Ordered by size:
157 161 161 162 478 735 836 849 849


Job completed Wed Sep 26 13:38:40 2001
```

```
Perl virtualTD v3.2
tc8_2.fna.vtd
Searching for ctag pattern

Tc8_2_.sens       (4262641 11077 12098)   753
Tc8_2_.rev  (17402823 21449 22470)  1052
Tc8_2_.rev  (1216305 25180 26201)   589
Tc8_2_.sens       (10800384 136635 137656)      236
Tc8_2_.rev  (10800384 159201 160220)      80
Tc8_2_.sens       (7140379 241294 242315) 207
Tc8_2_.rev  (7140347 40955 41976)   910
Tc8_2_.sens       (13775519 16025 17046)  281
Tc8_2_.sens       (4263177 5286 6307)     759
Tc8_2_.rev  (16950465 16299 17320)  1052
Tc8_2_.rev  (16950449 1658 2679)    76
Tc8_2_.rev  (14550376 23561 24582)  313
Tc8_2_.sens       (7105633 31423 32444)   168
Tc8_2_.rev  (3319447 14236 15257)   126
Tc8_2_.rev  (13435314 18076 19097)  144
Tc8_2_.rev  (2105490 1 376)   406
Tc8_2_.rev  (14578209 27147 28168)  258
Tc8_2_.sens       (1943778 3008 4029)     162
Tc8_2_.rev  (2873378 29070 30091)   477
Tc8_2_.sens       (7140370 56538 57559)   670
Tc8_2_.sens       (14625195 6508 7529)    90
Tc8_2_.sens       (532814 34944 35965)    997
Tc8_2_.sens       (1947136 15745 16766)   302
Tc8_2_.sens       (14625161 17613 18634)  1052
Tc8_2_.rev  (14574129 13951 14972)  83
Tc8_2_.rev  (156211 13843 14864)    1052
Tc8_2_.rev  (4966265 14625 15646)   840
Tc8_2_.sens       (4966265 8334 9355)     698
Tc8_2_.sens       (4966265 25420 26440)   917
Tc8_2_.sens       (485156 32582 33603)    1052
Tc8_2_.sens       (1072159 12717 13738)   405
Tc8_2_.sens       (14625203 31832 32853)  423
Tc8_2_.sens       (6041671 20672 21693)   326
Tc8_2_.rev  (1465850 19172 20193)   833
Tc8_2_.sens       (3165551 7998 9019)     604
Tc8_2_.rev  (2315795 10741 11762)   442
Tc8_2_.rev  (3168936 37327 38348)   1052
Tc8_2_.rev  (2773201 21416 22437)   882
Tc8_2_.rev  (2429531 25774 26795)   110
Tc8_2_.sens       (6671781 24218 25239)   279
Tc8_2_.rev  (6671781 12399 13420)   505
Tc8_2_.sens       (1326347 17115 18136)   1052
Tc8_2_.sens       (1326347 15179 16200)   64
Tc8_2_.rev  (10801501 57318 58339)  1052
Tc8_2_.sens       (10801501 54333 55354)  1052
Tc8_2_.rev  (4262597 21747 22768)   388
Tc8_2_.sens       (2746781 7469 8490)     102
Tc8_2_.sens       (10801502 50605 51626)  1052
Tc8_2_.sens       (10801502 34989 36007)  68
Tc8_2_.sens       (2429519 32223 33244)   68
Tc8_2_.rev  (1109795 2981 4002)     130
Tc8_2_.sens       (13606082 9527 10548)   75
Tc8_2_.sens       (4263302 6923 7944)     1052
Tc8_2_.rev  (4263302 1 603)   633
```

```
Tc8_2_.rev  (1403163 15103 16124)   193
Tc8_2_.sens       (927391 32227 33248)    645
Tc8_2_.sens       (7140424 31423 32444)   168
Tc8_2_.rev  (5824723 58544 59565)   765

Ordered by size:
64 68 68 75 76 80 83 90 102 110 126 130 144 162 168 168 193 207 236 258
279 281 302 313 326 388 405 406 423 442 477 505 589 604 633 645 670 698
753 759 765 833 840 882 910 917 997 1052 1052 1052 1052 1052 1052 1052
1052 1052 1052 1052


Job completed Sat Dec 15 23:23:47 2001
```

```
Perl virtualTD v3.2
tc1_2.fna.vtd
Searching for ctag pattern

Tc1_2_.sens       (7140354 40836 41859)    271
Tc1_2_.rev  (1049461 14926 15949)    440
Tc1_2_.rev  (10800383 280 1303)      305
Tc1_2_.rev  (2315759 20824 21847)    561
Tc1_2_.sens       (14625349 4044 5067)     900
Tc1_2_.rev  (14625291 9070 10093)    1054
Tc1_2_.sens       (4966308 1170 2193)      1054
Tc1_2_.rev  (14625314 10286 11309)   182
Tc1_2_.sens       (5921978 12629 13652)    657
Tc1_2_.sens       (4263265 651 1674)       190
Tc1_2_.sens       (14574228 3105 4128)     182
Tc1_2_.rev  (1051304 16396 17419)    1054
Tc1_2_.rev  (3810688 47328 48351)    548
Tc1_2_.sens       (17065911 71608 72631)   384
Tc1_2_.sens       (1122846 15544 16567)    250
Tc1_2_.sens       (2653095 4653 5676)      700
Tc1_2_.rev  (1627677 29851 30874)    393
Tc1_2_.sens       (4226095 10937 11960)    61
Tc1_2_.rev  (6425438 256838 257861)  183
Tc1_2_.rev  (1515135 22390 23413)    664
Tc1_2_.sens       (1402970 5573 6596)      159
Tc1_2_.sens       (3218082 5006 6029)      854
Tc1_2_.rev  (2815198 140 1163)       175
Tc1_2_.sens       (2653063 20583 21606)    351
Tc1_2_.sens       (3218092 17752 18775)    217
Tc1_2_.sens       (1628209 32142 33165)    997
Tc1_2_.rev  (2814863 10208 11231)    226
Tc1_2_.sens       (1729486 20069 21092)    334
Tc1_2_.sens       (6887 1511 2534)   130
Tc1_2_.sens       (156449 1556 2579)       271
Tc1_2_.rev  (861394 23508 24528)     378
Tc1_2_.rev  (726429 9705 10725)      1051
Tc1_2_.rev  (1263486 9965 10988)     1054

Ordered by size:
61 130 159 175 182 182 183 190 217 226 250 271 271 305 334 351 378 384
393 440 548 561 657 664 700 854 900 997 1051 1054 1054 1054 1054


Job completed Sat Dec 15 23:23:22 2001
```

APPENDIX 3

REPRINTS AND MANUSCRIPTS

CAVEY, M., LE, Q.H. AND BUREAU, T. (2002). Hybridization stringency correlates with BLAST score in determining copy number. Unpublished.

WRIGHT, S.I., LE, Q.H., SCHOEN, D.J. AND BUREAU, T. (2001). Population dynamics of an *Ac*-like transposable element in self- and cross-pollinating *Arabidopsis*. *Genetics* **158,** 1279-1288.

THE ARABIDOPSIS GENOME INITIATIVE (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796-815.

Title: Hybridization stringency correlates with BLAST score in determining copy number

Authors: Matthieu Cavey, Quang Hien Le and Thomas Bureau

Affiliation: Department of Biology, McGill University, 1205 Docteur Penfield Avenue, Montréal, Quebec H3A 1B1, Canada

Address correspondence to: Thomas E. Bureau, Department of Biology, McGill University, 1205 Ave. Docteur Penfield, Montréal, QC H3A 1B1 CANADA. e-mail: thomas_bureau@maclan.mcgill.ca

Keywords: copy number, BLAST score, hybridization, stringency, Southern

ABSTRACT

Since its development in 1975, the Southern blot technique has proved to be an extremely powerful and convenient tool to determine gene copy number. More recently, sequence similarity algorithms, notably Basic Local Alignment Search Tool (BLAST), can also be used to determine copy number of genes sequenced from model organisms such as *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae and Drosophila melanogaster*. Here, the results from 13 studies that have probed Southern blots in these model organisms to determine gene or transposon copy number are compared with the results found by BLAST searches using the same "probe sequences" as queries. Overall, our data indicate that fragments resolved by using high stringency conditions (wash temperature less than 18°C below the T_m) in Southern blots correspond well to similar fragments with scores greater than 200 bits as determined by BLAST searches. More precisely, the scores of all the fragments detected are consistently greater than their length. We also describe discrepancies between the predicted (BLAST) and observed (Southern blot) number of gene/transposons of certain families.

INTRODUCTION

Since its publication in 1975 by E.M. Southern, the Southern blot technique has been extensively used to detect the presence of specific sequences in the genomes of organisms and is an invaluable tool for understanding gene structure, genome organization and control of gene expression (1). In particular, it is especially useful when determining copy number for multigene families as well as transposons, using heterologous probes. With the availability of whole genome sequences, it is now possible to reproduce such experiments virtually using sequence similarity search algorithms such as the Basic Local Alignment Search Tool (BLAST) on genome sequence databases.

Because of the inherent variability of laboratory experiments, and the critical choices of the probe specificity (*e.g.* the portion of the gene that it spans) as well as the stringency levels, the Southern blotting technique can be imprecise when determining copy number. Thus, the reported copy number might be under- or overestimated. Sequence similarity searches provide the possibility to determine copy number at high resolution. Our study attempts to virtually reproduce selected studies reported in the literature that use Southern hybridization to estimate copy number. The basic strategy consists of performing BLAST searches with the "probe sequences" used in Southern hybridization protocols as queries and comparing stringency levels and BLAST scores to obtain an estimate of the reliability of the Southern hybridization approach. The complete genome sequence information for four eukaryotic model organisms, namely *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Drosophila melanogaster* are exploited for our comparative analyses.

MATERIALS AND METHODS

Case studies. A literature search was performed to identify relevant studies on the four model organisms. Twelve studies were selected; two reporting on a multigene family and one on a transposable element family for each organism. As the BLAST parameters remained constant, consistent stringency levels between studies were required for comparisons to be relevant. The studies were thus selected when they met the following requirements: the estimated number of genes/transposons detected for the family was clearly stated, the sequence of the probe could be unambiguously accessed and the stringency conditions used for the Southern hybridizations were comparable to the other studies (for most, wash temperature of 60-68°C, and SSC concentrations of 0.1-0.5 x SSC). Among the 12 cases considered, 3 (*i.e.* studies on *copia*, *apx-1b* and *act-2*) with lower stringency conditions (50°C to 65°C and 2 SSC to 3 SSC) were included to see if they differed from the other 9 (Table 1). In addition, one study that did not detect any transposon in a related genome was included in our analysis (*i.e. Tag-1*).

BLAST searches. In most cases, the restriction sites used to design the probes experimentally had to be located on the clones to obtain the exact sequences used in the studies (Table 1). Genomic sequences (assembled chromosomes) were downloaded from the NCBI web site (http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/org.html), and BLAST searches were performed locally using the default settings of the BLAST program (2; http://www.ncbi.nlm.nih.gov/BLAST/). The function for filtering out highly repetitive sequences was disabled as it would have yielded underestimates of the total number of hits, especially in the case of transposable elements. The percent identity of the sequences included in the alignments and the expect-value were not considered in our analysis as the normalized score reflects these parameters (the expect-value is a measure of the probability that the sequence found in the database be due to chance alone, considering its complexity and length relative to the whole database size).

Analysis of BLAST output. Hits with scores <30 bits were considered to be non-significant and were thus ignored from further analyses. The bands detected on Southern

blots were assumed to represent the High Scoring Pair (HSP) fragments (*i.e.* BLAST hits with the highest scores) for each case. The sets of HSP fragments separated by a distance smaller than the total size of the gene/transposon or overlapping fragments were considered as a single gene/transposon (clusters were likely to consist of only exons when cDNA probes were used). Thus, for each of these clusters of fragments, only the highest HSP fragment was kept as representing one copy of the gene/transposon for further analyses. The limitations of this assumption will be discussed.

Stringency assessment. BLAST results were classified according to score cutoffs determined on the basis of the NCBI classification scheme: all fragments with scores ≥200 bits were grouped as a set, ≥80 bits as another set and so on for fragments with scores ≥50, ≥40 and ≥30 bits. These cutoffs are equivalent to different levels of stringency in our "virtual" Southern blots. In addition, we determined the minimal score of the hits that corresponded to the exact number of genes/transposons detected in the studies. The results were then classified according to the stringency level of the Southern blots, based on the melting temperature (T_m) of the hybrids which was calculated from $T_m = 81.5°C + 16.6 \times Log_{10}[Na^+] + 0.41 \times (\%G+C) - 500/n$, where $[Na^+]$ is the concentration of salt (mol/l) given by SSC, %G+C is the G-C content of the hybrids, and n is the length of the hybrids in nucleotides (3). The difference between the T_m and the most stringent wash temperature was computed. Studies with wash temperatures very close to the T_m were considered as very stringent and studies with wash temperatures much lower than the T_m were considered as less stringent.

RESULTS

For *tif-1*, *ub3-d*, *Rte-1*, *Ta1-1*, *gpd-1*, *pau-1*, *his-10*, *Ty5* and *s7*, the fragments that were detected by Southern blot generally correspond to BLAST hits with scores greater than 200 bits (Table 2). However, it is worth noting that studies on *Rte-1* and *pau-1* seem to have missed a large number of copies (12 retrotransposons and 11 genes, respectively). This is also the case for *gpd-1* and *s7*, but to a lesser extent: 2 and 1 fragment(s), respectively, were not detected on the Southern blots even though their scores are greater than 200 bits. The study on *histones* in *C. elegans* (*i.e. his-10*) appears somewhat less stringent, as all the genes detected on the Southern blot have scores in the range of 80 to 200 bits.

A second group (*i.e. copia*, *apx-1b* and *act-2*) can be defined from our study which show more variable results. The expected trend was to observe decreasing scores as the stringency conditions were lowered. Lower scores are observed, but increasing BLAST scores are found as the stringency level is decreased.

Regarding *Tag-1*, a transposon present in *A. thaliana* ecotype Landsberg but not detected in ecotype Columbia (4), the BLAST results agree well with the Southern blot data; the unique fragments that show similarities to the probe have scores smaller than 40 bits (Table 2). It is interesting to note however that the probe spanned only part of the internal sequence of the transposon, while many fragments corresponding to the terminal repeats were detected when using the whole *Tag-1* sequence as a query (data not shown). *Tag-1*-like elements are present in *A. thaliana* ecotype Columbia, but only their termini are highly conserved, and a probe spanning the less-conserved internal sequence is unable to detect them at this stringency level. A similar pattern was observed for the LTR-retrotransposon *Ty5* in *S. cerevisiae*, where both Southern blots and BLAST searches revealed many more fragments when LTR probes were used instead of internal probes (5; data not shown).

The minimal BLAST score of the genes/transposons that corresponded to bands detected by Southern blots show a clear positive correlation between the magnitude of the scores and stringency levels (Table 2). However, there are some exceptions, such as the relatively low score obtained for *ub3-d* (383 bits), when compared to the high scores for

tif-1 and *Rte-1* (2307 and 3172 bits, respectively). The ratios of score to probe-length also show a positive correlation with stringency level, but the trend is variable, with studies such as *ub3-d*, *Ta1-1*, *pau-1*, *Ty5* and *apx-1b* that show higher ratios than the more stringent hybridizations that precede (Table 2).

DISCUSSION

Two major qualitative conclusions can be drawn from our results. First, we have shown that overall, higher scores correspond to higher hybridization stringencies while lower scores are associated with lower hybridization stringencies. This was the expected trend, and we may argue that the variability observed is in part attributable to differences in the sizes of the probes that were used. In fact, the BLAST scores of the corresponding fragments that were detected by stringent Southern blots are very high relative to the size of their respective probes. As mentioned in the results, the ratios of score to probe-length are maximal at high stringency and decrease at lower stringencies. This confirms our expectation that the fragments should have higher proportions of mismatches when the stringency is lowered, even though there is some variation in the trend. Our results thus show that in most cases, stringency conditions of 60-68 °C and 0.1-0.5 x SSC will allow fragments with scores greater than 200 bits to be detected on the Southern blots, and that the score to probe-length ratio will be greater than 1.

Second, we found conflicting evidence between BLAST and Southern blots in determining gene copy number. From our results, several studies appear to have missed some genes/transposons in the genome such as in the case of *Rte-1* and *pau-1*. Other studies have not detected genes/transposons because of the choice of the probe even though these sequences might be highly divergent but still contain stretches of similarity, such as in the case of *Tag-1* and *Ty5*.

Limitations

Although we attempted to use studies with similar stringency conditions for comparisons, there is some variation. Our calculation was based on the final wash of the blot (*i.e.* most stringent), which is critical for the amount and specificity of hybrids that will be retained and detected by autoradiography (6). However, the preceding washes were variable among the studies and could have skewed our comparison.

More importantly, the formula used for calculating the T_m of the hybrids is inaccurate for probes greater than about 100 nucleotides (3). In this study, the T_m calculations provided in Table 2 only serve as ratings for relating the level of stringency

used in Southern blots to the BLAST search stringency (*i.e.* score). Thus, the level of stringency used in the different case studies must be considered over broader ranges of temperature and SSC concentrations.

Moreover, the calculation of the scores by the BLAST algorithm does not take into account the position of the mismatches in the probe-target sequence hybrid. Similarly, gaps are being given a certain weight by the algorithm, but their position is not. The occurrence of mismatches in the middle of the hybrid are much more likely to be unstable than those at the ends (3). The specificity of the mismatches, that is whether they concern G-C or A-T base pairs, also influences the stability of the hybrids. In addition, the length of the hybrids plays a role in their stability. The longer the hybrid, the more stable it is and the smaller the impact of mismatches (1). Unfortunately, these parameters were not taken into account in the present study, mainly to limit the complexity of the analysis. We did not modify BLAST default settings for the values of penalties and rewards for the calculation of the score as it would not have resulted in a better simulation of Southern blots. Gaps could have been given a greater weight but not their position, nor the specificity of the mismatches.

Another limitation in the present study is imposed by our strategy and the BLAST algorithm. As explained in the methodology, in many cases, clusters of fragments similar to different parts of the probe were found, but only the highest HSP fragment was kept for each gene/transposon. This was a convenient way to directly analyze the results, however it does not reflect the actual similarity of the probe to its target sequence as in the case of global alignments (2). Hence, it can be argued that in some cases, the overall similarity between the probe and a given gene/transposon could be higher than for another gene/transposon of the family, even though the two HSP fragments representing the two genes/transposons in Table 2 had a similar score and thus were grouped using the same BLAST score cutoff. In other words, there could be some genes/transposons consisting of many HSP fragments and others consisting of only a few HSP fragments that would not be distinguished in our analysis. As differentiating genes/transposons from one another is also difficult, fragments separated by less than the whole gene size were considered to represent only one gene/transposon. However, tandemly-repeated genes can sometimes be separated by only a few hundreds base pairs, which leads to an underestimation of the

total number of genes/transposons when our strategy is adopted. We also have to consider the possibility that some of our HSP fragments may have comigrated on the Southern blots. If so, our estimate of the minimal score for the fragments detected on the Southern blots would be over-estimated.

The different labeling methods used to design the probes in the studies also have to be considered. Most studies used the random priming method while a few used nick-translation. The former often produces probes that do not span the whole sequence corresponding to the queries in our searches. However, we could not explain the discrepancies observed in our results by differences in the labeling methods alone. For example, probes produced by random priming do not hybridize with lower HSP fragments (data not shown).

A quantitative comparison of Southern blots and BLAST searches (*i.e.* equating stringency conditions to BLAST scores) is thus severely limited by the variability and uncertainties that we have outlined above. Despite these limitations, it is clear from our results that BLAST scores greater than the length of the query sequence (or more generally, scores greater than 200 bits) can be assumed as representing true genes/transposons when searching for multicopy genes/transposons. We also found that BLAST searches can provide higher resolution than some Southern hybridization protocols. In light of the recent release of the human genome as well as additional model eukaryotic genomes, this re-affirms the potential of combining experimental and computer-based approaches in genomics.

ACKNOWLEDGEMENTS

We thank Sylvia Levine, Julie Poupart and Dr. Beat Suter for critical reading of our manuscript and Boris-Antoine Legault for facilitating the computer analyses. This work was supported by a grant from the National Science and Engineering Research Council (NSERC) of Canada to T. B. Correspondences should be sent to Dr. Thomas Bureau, Department of Biology, McGill University, 1205 Ave. Docteur Penfield, Montréal, QC H3A 1B1, Canada. E-mail: thomas_bureau@maclan.mcgill.ca.

REFERENCES

Meinkoth,J., and Wahl,G. (1984) Review: Hybridization of Nucleic Acids Immobilized on Solid Supports. *Analyt. Biochem.*, **138**, 267-284.

Arribas,C., Sampedro,J. and Izquierdo,M. (1986) The ubiquitin genes in *D. melanogaster*: transcription and polymorphism. *Biochim. Biophys. Acta.*, **868**, 119-127.

Sambrook,J., and Russell,D.W. (2001) *Molecular Cloning: A Laboratory Manual.* 3rd Ed. Cold Spring Harbor Laboratory Press, NY.

Frank,M.J., Preuss,D., Mack,A., Kuhlmann,T.C., and Crawford,N.M. (1998) The *Arabidopsis* transposable element *Tag1* is widely distributed among *Arabidopsis* ecotypes. *Mol. Gen. Genet.*, **257**, 478-484.

Zou,S., Wright,D.A., and Voytas,D.F. (1995) The *Saccharomyces* Ty5 retrotransposon family is associated with origins of DNA replication at the telomeres and the silent mating locus *HMR*. *Proc. Natl. Acad. Sci. USA*, **92**, 920-924.

Ausubel,F.M., Brent,R., Kingston,R., Moore,D., Seidman,J., Smith,J.A., and Struhl,K. (1999) Current Protocols in Molecular Biology. John Wiley & Sons, NY.

Santos,M., Gousseau,H., Lister,C., Foyer,C., Creissen,G., and Mullineaux,P. (1996) Cytosolic ascorbate peroxidase from *Arabidopsis thaliana* L. is encoded by a small multigene family. *Planta*, **198**, 64-69.

McDowell,J.M., Huang,S., McKinney,E.C., An,Y.Q., and Meagher,R.B. (1996) Structure and Evolution of the Actin Gene Family in *Arabidopsis thaliana*. *Genetics*, **142**, 587-602.

Voytas,D.F., Konieczny,A., Cummings,M.P., and Ausubel,F.M. (1990) The Structure, Distribution and Evolution of the Ta*1* Retrotransposable Element Family of *Arabidopsis thaliana*. *Genetics*, **126**, 713-721.

Boseman Roberts,S., Sanicola,M., Emmons,S.W., and Childs,G. (1987) Molecular Characterization of the Histone Gene Family of *Caenorhabditis elegans*. *J. Mol. Biol.*, **196**, 27-38.

Huang,X.Y., Barrios,L.A.M., Vonkhorporn,P., Honda,S., Albertson,D.G., and Hecht,R.M. (1996) Genomic Organization of the Glyceraldehyde-3-Phosphate

Dehydrogenase Gene Family of *Caenorhabditis elegans*. *J. Mol. Biol.*, **206**, 411-424.

Youngman,S., van Luenen,H.G.A.M., and Plasterk,R.H.A. (1996) Rte-1, a retrotransposon-like element in *Caenorhabditis elegans*. *FEBS Letters*, **380**, 1-7.

Viswanathan,M., Muthukumar,G., Cong,Y.S., and Lenard,J. (1994) Seripauperins of *Saccharomyces cerevisiae*: a new multigene family encoding serine-poor relatives of serine-rich proteins. *Gene*, **148**, 149-153.

Linder,P., and Slonimski,P.P. (1989) An essential yeast protein, encoded by duplicated genes *TIF*1 and *TIF*2 and homologous to the mammalian translation initiation factor eIF - 4A, can suppress a mitochondrial missense mutation. *Proc. Natl. Acad. USA*, **86**, 2286-2290.

Lee,H., Simon,J.A., and Lis,J.T. (1988) Structure and Expression of Ubiquitin Genes of *Drosophila melanogaster*. *Mol. Cell. Biol.*, **8**, 4727-4735.

Arribas,C., Sampedro,J., and Izquierdo,M. (1986) The ubiquitin genes in *D. melanogaster*: transcription and polymorphism. *Biochim. Biophys. Acta.*, **868**, 119-127.

Di Franco,C., Pisano,C., Dimitri,P., Gigliotti,S., and Junakovic,N. (1989) Genomic distribution of copia-like transposable elements in somatic tissues and during development of *Drosophila melanogaster*. *Chromosoma*, **98**, 402-410.

Table 1: **Studies and Probes used in the present study**

Organism	Gene/Transposable Element Family	Study	Gene/ Transposon Length (nt)	Probe Length (nt)	GenBank Accession #	Coordinates[a] on the clone	Reference
A. thaliana							
	Ascorbate peroxidase	*apx-1b*	3321	1023	X80036.1	all	7
	Actins	*act-2*	3172	1134	U41998.1	cds	8
	Ta1-1	*Ta1-1*	519	241	X53972.1	64-304	9
	Tag-1	*Tag-1*	3295	1322	L12220.1	1098-2420	4
C. elegans							
	Histone (H4)	*his-10*	311	92	X15634.1	402-493	10
	Glyceraldehyde-3-Phosphate Dehydrogenase	*gpd-1*	1622	1500	Z49070.1	7771-9270	11
	Rte	*Rte-1*	3297	2177	AF054983.1	all	12
S. cerevisiae							
	Seripauperins	*pau-1*	1125	247	L25123.1	280-526	13
	Tif	*tif-1*	2363	2056	X12813.1	303-2358	14
	Ty5	*Ty5*	897	839	NC_001135.2	1179-4322	5
D. melanogaster							
	Ubiquitin (1)	*ub3-d*	2640	193	M22536.1	238-430	15
	Ubiquitin (2)	*s7*	not found	228	M33122.1	all	16
	Copia	*copia*	5144	2005	AE003411.1	63104-65108	17

[a] Coordinates of the probe on the clone from GenBank; all: complete clone, cds: coding sequence

Table 2: Comparison of the Number of Elements found by Southern blot and BLAST

Gene/ Transposable Element	SOUTHERN BLOT					BLAST						
	Stringency Conditions				Number of Genes/ Transposons found	Score Cutoffs[b] (bits)					Minimal Score[c]	Score:Probe Length Ratio
	T_{wash} (°C)	SSC	%(G+C) of probe	T_m-T_{wash} [a] (°C)		≥200	≥80	≥50	≥40	≥30		
tif-1	68	0.1	36.7	-2.0	2	2	2	4	5	92	2307	1.12
ub3-d	68	0.1	51.8	1.9	1	1	3	5	7	40	383	1.98
Rte-1	68	0.1	47.3	2.4	15	27	34	39	46	224	3172	1.46
Ta1-1	65	0.2	39.0	5.1	1	1	1	1	5	855	478	1.98
gpd-1	60	0.1	42.1	8.2	2	4	4	6	37	984	1818	1.21
pau-1	60	0.1	49.8	9.6	13	24	24	27	27	31	363	1.47
his-10	65	0.5	52.2	13.8	12	0	16	17	19	89	125	1.36
Ty5	65	0.5	42.0	14.5	1	1	1	1	2	36	1161	1.38
s7	50	0.1	47.8	18.7	2	3	5	6	6	18	254	1.11
copia	65	2	32.6	20.9	25	21	22	24	26	90	123	0.06
apx-1b	65	2	47.4	26.7	2	1	1	2	4	53	327	0.32
act-2	50	3	47.4	44.7	10	6	11	12	14	28	212	0.19
Tag-1	60	0.2	37.8	11.3	0	0	0	0	0	68	-	-

[a] Most probes are greater than 100 nt (see Table 1).

[b] Areas are shaded when the number of genes/transposons predicted by BLAST match with the number found by Southern blot.

[c] Minimal score of the corresponding genes/transposons found by Southern blot.

Copyright © 2001 by the Genetics Society of America

Population Dynamics of an *Ac*-like Transposable Element in Self- and Cross-Pollinating Arabidopsis

Stephen I. Wright, Quang Hien Le, Daniel J. Schoen and Thomas E. Bureau

Department of Biology, McGill University, Montreal, Quebec H3A 1B1, Canada

Manuscript received November 24, 2000
Accepted for publication April 11, 2001

ABSTRACT

Theoretical models predict that the mating system should be an important factor driving the dynamics of transposable elements in natural populations due to differences in selective pressure on both element and host. We used a PCR-based approach to examine the abundance and levels of insertion polymorphism of *Ac*-III, a recently identified *Ac*-like transposon family, in natural populations of the selfing plant *Arabidopsis thaliana* and its close outcrossing relative, *Arabidopsis lyrata*. Although several insertions appeared to be ancient and shared between species, there is strong evidence for recent activity of this element family in both species. Sequences of the regions flanking insertions indicate that all *Ac*-III transposons segregating in natural populations are in noncoding regions and provide no evidence for local transposition events. Transposon display analysis suggests the presence of slightly higher numbers of insertion sites per individual but fewer total polymorphic insertions in the self-pollinating *A. thaliana* than *A. lyrata*. Element insertions appear to be segregating at significantly lower frequencies in *A. lyrata* than *A. thaliana*, which is consistent with a reduction in transposition rate, reduction in effective population size, or reduced efficacy of natural selection against element insertions in selfing populations.

TRANSPOSABLE elements (TEs) are mobile self-replicating segments of DNA. Their ability to self-replicate and increase in abundance makes them an important source of spontaneous mutation and cause of genome evolution (KIDWELL and LISCH 2000). Population genetics theory predicts that the abundance of TEs in natural populations may be controlled by a balance between the forces of transposition increasing copy number and the action of purifying selection removing insertions from populations (CHARLESWORTH and LANGLEY 1989). However, the general importance of purifying selection in controlling TE dynamics and the nature of their deleterious effects in natural populations remain uncertain.

Evidence for the role of weak purifying selection in controlling transposon abundance in euchromatic regions of the genome was documented primarily in natural populations of *Drosophila melanogaster*, using the techniques of *in situ* hybridization (MONTGOMERY and LANGLEY 1983; CHARLESWORTH and LAPID 1989) and RFLP analysis (reviewed in CHARLESWORTH and LANGLEY 1989). From these studies TE families have been shown to be represented typically by a small number of copies per genome, despite evidence for much higher rates of transposition than excision. In addition, individual element insertions appear to be maintained at very low frequencies, their presence restricted to only one to a few individuals in population samples. This skewed frequency distribution shows a significant departure from neutral expectation (GOLDING *et al.* 1986; TAJIMA 1989) and is consistent with a model of transposition-selection balance, where new insertions are introduced through the transposition process and weak selection prevents them from rising to high frequencies (CHARLESWORTH and CHARLESWORTH 1983; CHARLESWORTH and LANGLEY 1989; CHARLESWORTH *et al.* 1992). While this model of transposition-selection balance has become well accepted for many families of transposons in *D. melanogaster*, the exact nature of purifying selection and copy number control on TEs remains an issue of contention (BIEMONT *et al.* 1997; CHARLESWORTH *et al.* 1997).

Although there is strong evidence indicating selective regulation of TE copy number in the large random-mating populations of *D. melanogaster*, very few population studies are available to examine its relevance in other eukaryotic genomes. In fact, preliminary analysis of retrotransposon families in several other animal species, including other species in the genus Drosophila (HEY 1989; BATZER *et al.* 1996; TAKASAKI *et al.* 1997), indicates that transposon insertions may often rise to high frequencies, and the accumulation of many ancient insertions in several taxa (*e.g.*, MURATA *et al.* 1996; SANMIGUEL *et al.* 1998) is suggestive of a neutral process of fixation or loss of TE insertions over evolutionary time. This suggests the importance of detailed comparisons of the patterns of transposon insertion polymorphism in related species with contrasting life histories.

Corresponding author: Stephen I. Wright, Institute of Cell, Animal and Population Biology, University of Edinburgh, Ashworth Laboratories, King's Bldgs., W. Mains Rd., University of Edinburgh, Edinburgh, EH9 3JT Scotland. E-mail: stephen.wright@ed.ac.uk

There has been recent theoretical interest in the role of the host breeding system in driving evolutionary changes in TE dynamics, particularly in plant populations (Charlesworth and Charlesworth 1995; Wright and Schoen 1999; Morgan 2001). If many transposon insertions cause recessive deleterious mutations (Biemont *et al.* 1997), then greater expression of these deleterious effects due to higher levels of homozygosity may drive transposon abundance and frequency to lower levels in selfing populations (Wright and Schoen 1999; Morgan 2001). In contrast, if element abundance is regulated by the dominant effects of ectopic recombination between elements (Montgomery *et al.* 1991), the presence of most insertions in a homozygous state may reduce chances for ectopic pairing (Montgomery *et al.* 1991), leading to a relaxation of selection on TE copy number and a potential for rapid increase in element abundance (Charlesworth and Charlesworth 1995; Wright and Schoen 1999; Morgan 2001). Reductions in effective population size in selfing populations, through frequent bottleneck events and/or strong selection at linked loci (Maynard Smith and Haigh 1974; Charlesworth *et al.* 1993; Braverman *et al.* 1995), could also drive an increase in TE frequency and abundance (Brookfield and Badge 1997; Wright and Schoen 1999; Morgan 2001), but frequent stochastic loss of element families in selfing lineages is also a possible outcome (Wright and Schoen 1999). Finally, differences in the breeding system are also expected to cause changes in the selective pressure on TEs themselves, leading to conditions favoring self regulation in genomes with low recombination (Charlesworth and Langley 1986), which may also cause reduced levels of insertion polymorphism.

Arabidopsis thaliana is a highly selfing (Abbot and Gomes 1989) species with one of the smallest known genomes in higher plants (Leutwiler *et al.* 1984). Recent surveys of the diversity and abundance of transposable elements in the Arabidopsis genome have revealed the presence of an extremely diverse array of transposon families, with most present at low copy number (Surzycki and Belknap 1999; Le *et al.* 2000). This low abundance of elements is a strong contrast to many plant species, for which element families may comprise as much as 50% of the genome (SanMiguel *et al.* 1996). Despite the low abundance, the high sequence similarity observed in several families of elements (Surzycki and Belknap 1999; Le *et al.* 2000), and the identification of putative empty insertion sites in related ecotypes, (Le *et al.* 2000) suggests that many of these families may have been recently active.

The *hobo/Ac/Tam3* (*h*AT) superfamily of transposons is a widespread group of class II elements, which are known to be responsible for diverse morphological (Coen and Carpenter 1986) and chromosomal (Dooner and Belachew 1991; Shalev and Levy 1997; Zhang and Peterson 1999) mutations. Several families of *Ac*-like elements have been identified in Arabidopsis, many of which show evidence for current activity (Tsay *et al.* 1993; Frank *et al.* 1997) and/or recent historical transposition events (Frank *et al.* 1998; Henk *et al.* 1999; Le *et al.* 2000). The Arabidopsis *Ac*-like III family (hereafter *Ac*-III), was recently identified in a survey of TE diversity in the *A. thaliana* (Columbia) genome project (Le *et al.* 2000). It was classified within the *h*AT superfamily on the basis of shared sequence and structural similarity of the terminal inverted repeats, the presence of several copies of a putative *Ac* transposase binding motif (TGGGC), and an 8-bp target site duplication (see Henk *et al.* 1999). Amplification of putative "empty" *Ac*-III insertion sites in several ecotypes related to Columbia suggested the possibility of recent mobility (Le *et al.* 2000).

In this study, we utilized an amplified fragment length polymorphism (AFLP)-based (Vos *et al.* 1995) technique, transposon display (Korswagen *et al.* 1996; Waugh *et al.* 1997; Van den Broeck *et al.* 1998), to examine the distribution, abundance, and levels of insertion polymorphism of the *Ac*-III transposon family in natural populations of *A. thaliana* and its self-incompatible, highly outcrossing (Kärkkäinen *et al.* 1999) relative, *A. lyrata*.

MATERIALS AND METHODS

Transposon display—element family and primer design: The technique of transposon display involves the digestion of total genomic DNA with a frequent-cutting restriction enzyme, the ligation of adaptors to the digested fragments, and labeled PCR amplification of this digestion-ligation template using a TE-specific primer and an adaptor-specific primer. Since the restriction sites flanking transposon insertions are located at variable positions, this allows for the visualization of individual element insertions on a polyacrylamide gel, and variation among individuals in the sizes of these bands allows for an assessment of the patterns of transposon-based polymorphism in natural populations.

Sequence information for the *Ac*-III elements was accessed from the *A. thaliana* transposable element database at McGill University (http://www.tebureau.mcgill.ca; see Le *et al.* 2000), and several additional elements from recently released genomic clones were also identified through BLAST searches of GenBank (National Center for Biotechnology Information (NCBI); http://www.ncbi.nlm.nih.gov/blast/). All nine elements identified to date from the genome sequence data were aligned using the program Pileup (University of Wisconsin Genetics Computing Group, Version 10.0), and conserved subterminal regions were identified. Two overlapping degenerate primers were designed, on the basis of sequence alignments, for nested PCR using transposon display [ep1, 5′ GGTTCGGTTA(A/T)TCGGTTAGC(G/T)G 3′; ep2, 5′ G(C/A)TTCGGTTCGGTTA(A/T)TCGGTTAG 3′]. By examining the frequent-cutter restriction map of the *Ac*-III elements, the four-cutter restriction endonuclease *Nla*III was selected for use in transposon display, since no sites were observed between the primer sequence and the end of the element. Restriction mapping of *A. thaliana* chromosomes 2 and 4 suggested that the size distribution of *Nla*III restriction fragments was suitable for transposon display (97% of restriction frag-

TABLE 1

Source of population samples

Species	Population	No. of individuals	Origin
A. lyrata ssp. petraea	MJ	6	Mjallom, Sweden[a]
A. lyrata ssp. petraea	KM	1	Karhumaki, Russia[a]
A. lyrata ssp. lyrata	LR	2	Michigan[b]
A. lyrata ssp. lyrata	GM	11	Michigan[c]
A. lyrata ssp. lyrata	BB	1	USA[d]
A. thaliana	PS	7	Tennessee[e]
A. thaliana	SI	1	Tennessee[c]
A. thaliana	WF	1	Tennessee[b]
A. thaliana	Y2	6	North Carolina[b]
A. thaliana	W2	3	North Carolina[b]
A. thaliana	SG	1	Tennessee[e]
A. thaliana	No-O	1	Nossen, Germany[f]
A. thaliana	Ws	1	Wassilewskija, Russia[f]
A. thaliana	Nd-1	1	Niederzenz, Germany[f]
A. thaliana	Sn-1	1	Czechoslovakia[f]
A. thaliana	Tsu-1	1	Tsu, Japan[f]
A. thaliana	RLD-1	1	Rschew, Russia[f]
A. thaliana	DiG	1	Dijon, France[f]
A. thaliana	S96	1	Netherlands[f]
A. thaliana	Tol-O	1	Ohio[f]
A. thaliana	Be-O	1	Bensheim, Germany[g]

[a] Seed material obtained from O. Savolainen.
[b] Seed material obtained from R. Mauricio.
[c] Seed material obtained from C. Langley.
[d] Seed material obtained from T. Mitchell-Olds.
[e] Seed material obtained from M. Pigliucci.
[f] Seed material obtained from ABRC.
[g] Seed material obtained from NASC.

ments fall between 4 and 1000 bp). An *Nla*III adaptor was designed on the basis of the vectorette used by KORSWAGEN *et al.* (1996; *Nla*III 503, 5′ CAAGGAGAGGACGCTGTCTGT CGAAGGTAAGGAACGGACGAGAGAAGGGAGA 3′; *Nla*III 504, 5′ TCTCCCTTCTCGAATCGTAACCGTTCGTACGAGA ATCGCTGTCCTCTCCTTGCATG 3′), and two nested adaptor-specific primers were also constructed (ap1, 5′ CGAATCG TAACCGTTCGTACGAGAATCGCT 3′; ap2, 5′ GTACGAGAA TCGCTGTCCTC 3′; KORSWAGEN *et al.* 1996). Using the available *A. thaliana* genomic sequence (~93% of the total genome), the expected sizes of transposon display bands within the *A. thaliana* Columbia ecotype could be predicted by identifying the nearest *Nla*III site to element insertions.

Population samples: A total of 21 *A. lyrata* individuals and 29 *A. thaliana* individuals from a wide range of geographical locations were analyzed in this study (Table 1). The standard *A. thaliana* ecotype Columbia was also used as a positive control to confirm specificity of the technique. All *A. lyrata* samples were derived from maternal seed families that were field collected by various researchers. *A. thaliana* individuals were either obtained from field-collected seed, the progeny of field-collected seed that were selfed for one generation, or from the Arabidopsis Biological Resource Center (ABRC) or the Nottingham Arabidopsis Stock Center (NASC; see Table 1). One individual per seed family was analyzed for both species.

Transposon display analysis: Genomic DNA was extracted using a modified version of DELLAPORTA *et al.* (1983), with the addition of a phenol:chloroform:IAA extraction after the first precipitation. Genomic DNA (20–50 ng) was digested with 2 units of *Nla*III, 1× NEB buffer 4 (New England BioLabs, Beverly, MA), and 1 mg/ml bovine serum albumin in a total volume of 10 μl for at least 2 hr. Digested DNA was then ligated to the vectorette overnight at 16°, using 5 units of T4 DNA ligase, 15 pmol adaptor, and 1× T4 DNA ligase buffer in a total volume of 25 μl. The ligation reaction was then diluted to a total volume of 100 μl for use as template in PCR. Preselective PCR amplification was carried out in a total volume of 25 μl, using 3 μl of the digestion-ligation mixture, 10 pmol of *Nla*III adaptor-specific primer (ap1), and 10 pmol of element-specific primer (ep2), 1.5 mM $MgCl_2$, 0.2 mM dNTPs, and 1 unit Amplitaq DNA polymerase (Perkin-Elmer, Foster City, CA). PCR reactions were run on a Perkin-Elmer PCR model 9700 device for 20 cycles consisting of 1 min denaturing at 94°, 1 min annealing at 65°, and 1 min extension at 72°. This PCR reaction was diluted 1/100 and then used as template for selective PCR amplification under the same reaction conditions, except that 10 pmol of the nested element-specific primer and 2 pmol of the IRD700-labeled (LiCor, Lincoln, NB) nested adaptor-specific primer were used. Selective amplification products were then denatured for 5 min at 95°, rapidly cooled on ice, and run on a 66 cm, 5.5% denaturing polyacrylimide gel (3.75% acrylimide, 7 M urea, 1× Tris-borate EDTA), and bands were read using the LiCor DNA sequencer Long ReadIR4200 system. Bands were sized using the LiCor IRD700-labeled 50 to 700-bp molecular weight ladder, and their presence or absence was scored for each individual. Band scoring was checked with the assistance of the program Cross Checker (version 2.91, Dr. J. B. Buntjier, 1999). For a subset of samples, banding patterns were rechecked by repeating the process of digestion, ligation, and PCR. All amplifications were conducted twice from the same digestion-ligation reaction, and inconsistencies were exam-

ined. Several band sizes that were inconsistently present were excluded from subsequent analysis.

To assess the relative importance of insertion polymorphism *vs.* nucleotide and indel variation in producing polymorphic bands, selective amplifications were also run using the unlabeled adaptor-specific primer, and reaction products were cloned into the pCR 2.1 vector using the TA cloning kit (Invitrogen, Carlsbad, CA). Clones were then sequenced using the LiCor DNA sequencer Long ReadIR4200 according to the manufacturer's protocol. Sequences were analyzed by BLAST searches against GenBank (NCBI; http://www.ncbi.nlm.nih.gov/blast/) for sequence similarity and through multiple alignments using Pileup (University of Wisconsin Genetics Computing Group, version 10.0).

Data analysis: A table of presence/absence for each insertion site was generated for analysis of patterns of insertion polymorphism. For all analyses, bands of the same size were assumed to be identical, while bands of distinct size were assumed to be independent insertions. Evidence supporting these assumptions, and the implications of any violations, are addressed in the DISCUSSION.

To test for a role for purifying selection in driving patterns of insertion polymorphism, the minimum χ^2 method of CHARLESWORTH and CHARLESWORTH (1983; Appendix 3) was applied. This method estimates the parameters of the probability distribution for element frequency x, using the β-distribution

$$\phi(x) \approx \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}, \quad (1)$$

where $\alpha = 4N_e u\hat{n}/(T - \hat{n})$, $\beta = 4N_e(\upsilon + s)$, N_e is the effective population size, u is the rate of transposition, T is the total number of insertion sites, $\hat{n}$ is the equilibrium element copy number, υ is the rate of excision, and s is the strength of selection against element insertions. α measures the effects of repeated transposition into the same insertion site, and β provides an estimate of the strength of forces removing insertions from natural populations (CHARLESWORTH and LANGLEY 1989). If the number of sites is effectively infinite, Equation 1 reduces to

$$\phi(x) \approx \beta x^{-1}(1 - x)^{\beta-1} \quad (2)$$

(CHARLESWORTH and CHARLESWORTH 1983). The parameter estimation method uses the estimated frequency distribution of individual insertion sites per haploid genome and finds the values of α and β, which minimize the deviation between observed and expected insertion frequencies. From *in situ* hybridization studies of Drosophila, insertion frequencies were estimated directly from haploid chromosomes (CHARLESWORTH and LANGLEY 1989). However, because the current study examines diploid chromosomes, and transposon display does not distinguish between heterozygous and homozygous insertions, estimates of allele frequencies, element copy number per haploid genome, and the insertion site frequency profile cannot be obtained directly from the banding patterns. In *A. thaliana*, estimates of outcrossing rates from natural populations indicate that populations are almost 100% selfing (ABBOT and GOMES 1989), and all insertions were thus assumed to be homozygous for the purposes of allele frequency and copy number estimation. Estimates of insertion site frequencies per haploid genome could then be estimated directly from transposon display. For *A. lyrata*, a modified version of Equation A9 in CHARLESWORTH and CHARLESWORTH (1983) was used to generate expected numbers of insertions present in m diploid genomes using the assumption of Hardy-Weinberg equilibrium and was compared to the observed values. Specifically, let g_i be the number of insertion sites present in i diploid individuals ($i = 1, 2, \ldots m$). The expected value of g_i is, assuming Hardy-Weinberg equilibrium and from CHARLESWORTH and CHARLESWORTH (1983),

$$E[g_i] = \frac{1}{2} T \binom{m}{i} \int_0^1 (x^2 + 2x(1 - x))^i (1 - (x^2 + 2x(1 - x)))^{m-i} \phi(x)\, dx. \quad (3)$$

T is then estimated in the usual way (CHARLESWORTH and CHARLESWORTH 1983) as $T = \hat{n}(\alpha + \beta)/\alpha$, where $\hat{n}$ is estimated by the sum of the Hardy-Weinberg estimates of insertion frequency,

$$\hat{n} = \sum_{i=1}^{m} \left(1 - \sqrt{1 - z_i}\right), \quad (4)$$

and z_i is equal to the frequency of the "null" allele for the ith insertion site.

RESULTS

Consistency and specificity of transposon display in Arabidopsis: Several lines of evidence indicate that the technique of transposon display provides a specific and repeatable visualization of *Ac*-III transposon insertions in Arabidopsis. First, virtually all insertion site sizes predicted from the available Arabidopsis genomic sequence were observed in analysis of the Columbia ecotype genomic DNA (Figure 1a). Only the largest predicted insertion site, 1444 bp, was not observed, suggesting the presence of an upper size limitation on the ability to amplify and visualize insertion sites. Estimates of numbers of insertion sites per individual should thus be considered as minimum estimates. However, given that analysis of *Nla*III restriction fragment sizes in the Arabidopsis genome suggests that <5% of digested fragments should be >1 kb, this should not result in extreme underestimates. Second, replicated digestion-ligations from the same genomic DNA and repeated amplifications from the same digestion-ligation reaction resulted in consistent banding patterns (Figure 1a). Sequencing of a subset of amplification products from both species also allowed for confirmation that many observable bands were derived from *Ac*-III transposon insertions, through sequencing of the termini of the elements (*e.g.*, Figure 1b). Although several amplification products that did not correspond to element insertions were also sequenced, these fragments did not appear on transposon display gels. Finally, in some cases, sequence similarity between the regions flanking the insertion and regions of the Columbia genome lacking the element provides evidence that novel bands correspond to new transposition events (*e.g.*, Figure 1b).

***Ac*-III transposon insertions in Arabidopsis:** Amplification of the *Ac*-III element family was successful in all populations of both *A. thaliana* and *A. lyrata* analyzed in this study (Figure 2, Table 2). The vast majority of insertion sites identified in both species appeared to be polymorphic, although several bands of identical size appeared to be shared between both species, possibly representing ancient fixed insertions (Figure 2, Table

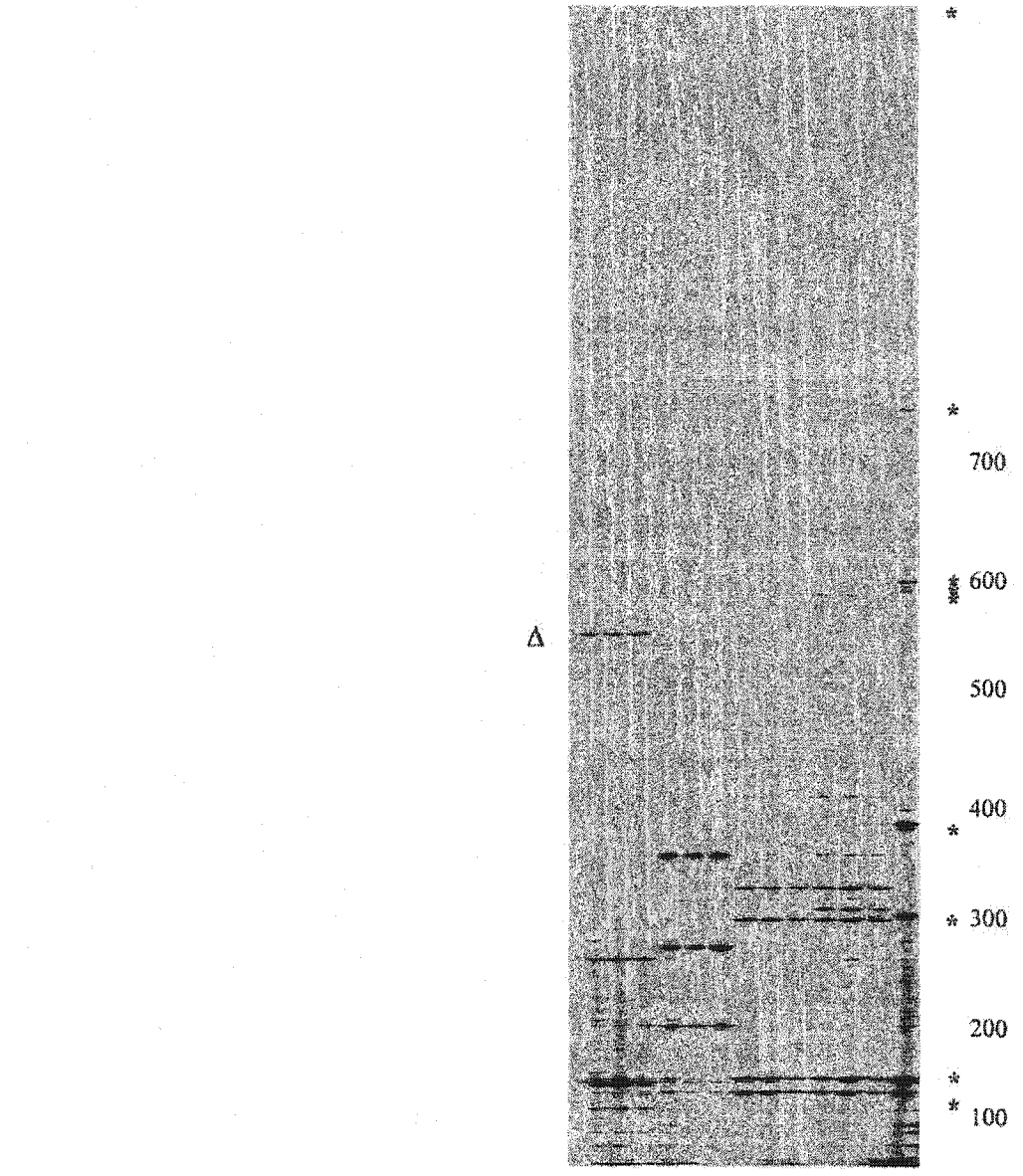


FIGURE 1.—Characteristic results from the application of transposon display to the *Ac*-III family in Arabidopsis. (a) Transposon display patterns for four individuals from natural populations and the *A. thaliana* ecotype Columbia. Three replicates for each individual from natural populations are shown; the first two represent two different amplifications from the same digestion-ligation, and the third is derived from an independent digestion-ligation reaction. Lanes 1–3, *A. lyrata* GM-1; 4–6, *A. lyrata* GM-2; 7–9, *A. thaliana* Y2-1; 10–12, *A. thaliana* Y2-2. Lane 13, *A. thaliana* ecotype Columbia. Asterisks indicate the sizes of insertion sites predicted from the Columbia genome sequencing project. The triangle indicates the size of the insertion site for which the sequence is shown in b. Numbers represent the molecular weight, in base pairs. (b) Sequence analysis of cloned transposon display fragment. Top sequence: 5′ end of cloned fragment. Middle sequence: 3′ end of *Ac*-III transposon. Bottom sequence: region of *Arabidopsis thaliana* (Columbia) genome with high sequence similarity to cloned flanking region, with the accession number for the genomic clone given. ep1, element-specific primer used in transposon display; IR, inverted repeat.

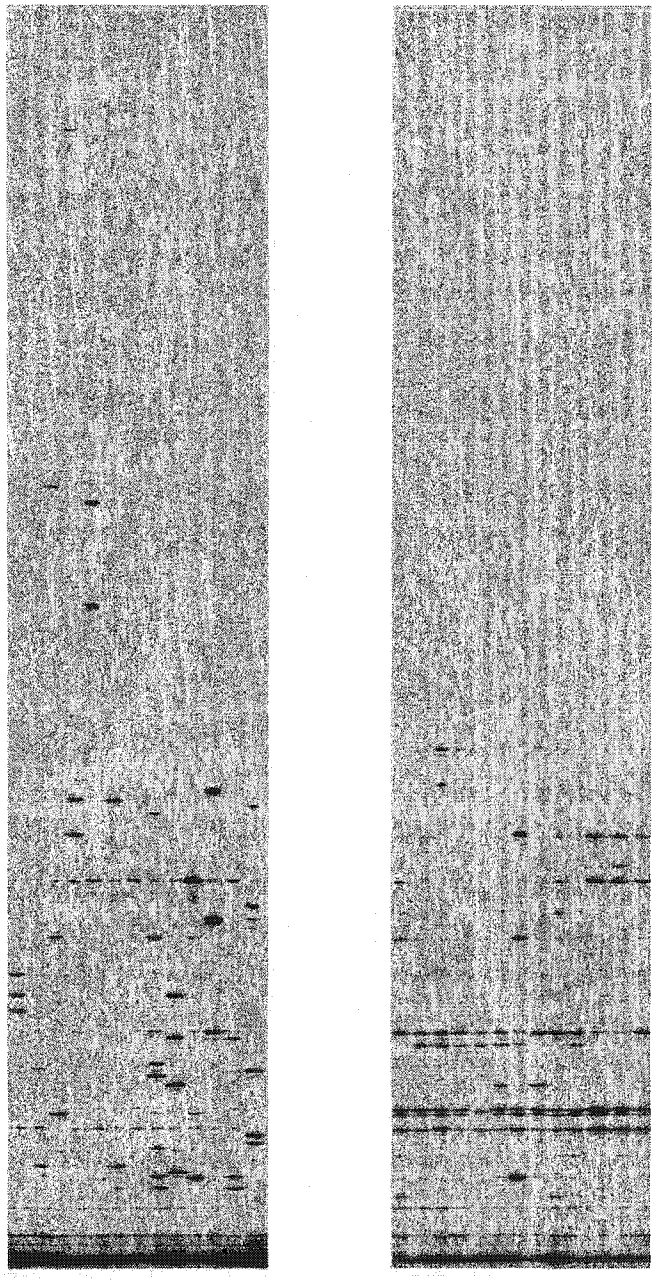


FIGURE 2.—Example of among-population variability of transposon display patterns. (a) *Arabidopsis lyrata.* (b) *A. thaliana.*

2). These latter sites were present only faintly and somewhat inconsistently in *A. lyrata*, while they were predominantly present as intense bands in *A. thaliana.* While a total of 46 polymorphic insertion sites were identified in *A. lyrata*, only 24 were identified in *A. thaliana* (Table 2). Despite the lower total number of polymorphic insertion sites, *A. thaliana* populations exhibited a slightly higher average number of insertion sites per individual and a higher estimated equilibrium number of insertions per individual, $\hat{n}$ (Table 2).

In total the sequences of 15 (30%) insertion sites in *A. lyrata* and 6 (22%) in *A. thaliana* populations were determined, through the cloning and sequencing of amplification products from several individuals and from shared insertion sites to elements present in the Columbia genome project (Table 3). Nine of the sequenced flanking regions (57%) showed very high (90%) sequence similarity to genomic regions that were sequenced in *A. thaliana.* All of these insertion sites were in noncoding regions, including both introns and intergenic regions, with several inserted into other repetitive sequences, including a non-LTR retrotransposon (Table 3). All identified flanking regions appeared to be unique, suggesting that a significant proportion of distinct bands were the result of independent insertion events. Insertion sites showed similarity to physically distant regions across all five *A. thaliana* chromosomes, providing no evidence for local transposition events, which were observed for maize *Ac*-like transposons (*e.g.*, BANCROFT and DEAN 1993).

Patterns of polymorphism: In *A. thaliana*, the majority of insertion sites were present at intermediate to high frequency (Figures 2 and 3). In contrast, a large fraction of the insertion sites were segregating at low frequency in *A. lyrata*, with many insertion sites present in only one or a few individuals (Figures 2 and 3). To test for a significant difference among species in the insertion site frequency distribution, a 2×2 contingency table was constructed, similar to the approach of SAWYER *et al.* (1987). This table compares the numbers of insertion sites unique to a single individual ("singletons") to higher frequency insertions between the two species and can be used to test the null hypothesis of an equal proportion of singleton insertions in each species. In *A. lyrata*, the number of singletons is 23, compared to 26 nonsingletons. In contrast, in *A. thaliana*, there are 3 singletons compared with 24 nonsingletons. Under Fisher's exact test, the contingency analysis shows a highly significant deviation from random expectation ($P < 0.01$), reflecting the much higher number of low-frequency insertions in the *A. lyrata* sample in comparison to *A. thaliana.*

To test for departures from neutral expectations of insertion frequency distributions, the parameters α and β were estimated for each species. For both data sets, α was not significantly different from 0, and χ^2 values increased as α was increased from 0. This suggests that the number of insertion sites are effectively infinite, and estimates of β were subsequently made under the assumption that $\alpha = 0$. Table 4 shows the estimates of β and the corresponding minimum χ^2 values for both species under two sets of conditions: (a) including and (b) excluding the potentially ancient high frequency insertions. Excluding these insertions is justified if they are in fact ancient and fixed, since they do not contribute to current population dynamics in each species. χ^2 values were obtained after pooling classes for which

TABLE 2

Summary of transposon display patterns in *A. thaliana* and *A. lyrata*

Species	Sample size	Total no. of insertion sites	No. of segregating sites (S)	Average no. of insertion sites per individual	$\hat{n}^a$
A. lyrata	21	49	46	8.2	11.63
A. thaliana	29	27	24	10.7	21.7

[a] Estimate of the equilibrium element copy number based on transposon display.

expected values were <5. In all cases, the insertion site profiles do not show a significant deviation from expected by the minimum χ^2 parameter estimates. Because of the small total number of insertion sites and the error associated with the estimation of the occupancy distribution from dominant markers, the estimates are subject to a high variance. However, under either scenario, the *A. lyrata* population data show a significant deviation from the hypothesis that $\alpha = 0$ and $\beta = 1$, which is consistent with a significant role for purifying selection in driving patterns of insertion frequency. In contrast, for the *A. thaliana* insertion site distribution, the estimated minimum χ^2 values of β are <1, suggesting that the effects of genetic drift dominate over selection. Estimates of β for *A. lyrata* when the potentially ancient insertions are excluded are of the same order of magnitude as was observed for several element families in *D. melanogaster* (CHARLESWORTH and LANGLEY 1989), suggesting that selection coefficients may be similar. Figure 3 shows a comparison of the expected insertion site occupancy profile under the estimated parameters to that calculated from the data.

DISCUSSION

Our study represents the first detailed investigation of the levels and patterns of insertion polymorphism of a class II transposon family in natural plant populations. The high levels of polymorphism of element insertions and the diversity of sequenced flanking regions strongly suggest that the *Ac*-III transposon family has been active in both species since their divergence. This allows for a detailed comparison of the population dynamics of the same transposon family in related selfing and outcrossing species.

The insertion site frequency distributions are consistent with the importance of the action of purifying selection on transposon insertions in outcrossing populations; segregating insertions in *A. lyrata* were present at lower frequencies than expected on the basis of neutral expectations. The few high-frequency bands identified in *A. lyrata* were of similar size to insertions identified in *A. thaliana* and may thus be potentially ancient. In contrast to the patterns observed in the outcrossing *A. lyrata*, most insertions were segregating at intermediate to high frequencies in the selfing *A. thaliana* and did not show evidence of a strong departure from neutral expectations. Virtually all insertions identified in a given population in *A. thaliana* were shared with other populations and present in other individuals from the same population. Although differences in element abundance between the species were less pronounced, there was a slightly higher number of insertion sites per individual in *A. thaliana*.

One limitation of our study concerns the population sampling: *A. thaliana* collections represent a more extensive worldwide sample, while *A. lyrata* is more concentrated, with more individuals sampled from fewer popu-

TABLE 3

Locations and properties of naturally occurring insertion sites of *Ac*-like III in *A. thaliana* and *A. lyrata*

Insertion site (bp)	Population[a]	Region of similarity to *A. thaliana*	Location of insertion site
549	*A. lyrata* GM	Chromosome 1	Intergenic region
204	*A. lyrata* MJ	Chromosome 3	Intron 3 of hypothetical protein
135	*A. lyrata* GM	Chromosome 3	Intergenic region
188	*A. lyrata* GM	Chromosome 1	Intergenic region
295	*A. thaliana* PS		5′ flanking region of glycoprotein gene family
353	*A. lyrata* GM	Chromosome 5	NA
271	*A. lyrata* GM	Chromosome 5	Intergenic region
323	*A. thaliana* PS		Non-LTR retrotransposon
193	*A. thaliana* PS	Chromosome 4	Intergenic region

[a] Populations refer to those listed in Table 1. NA, not available.

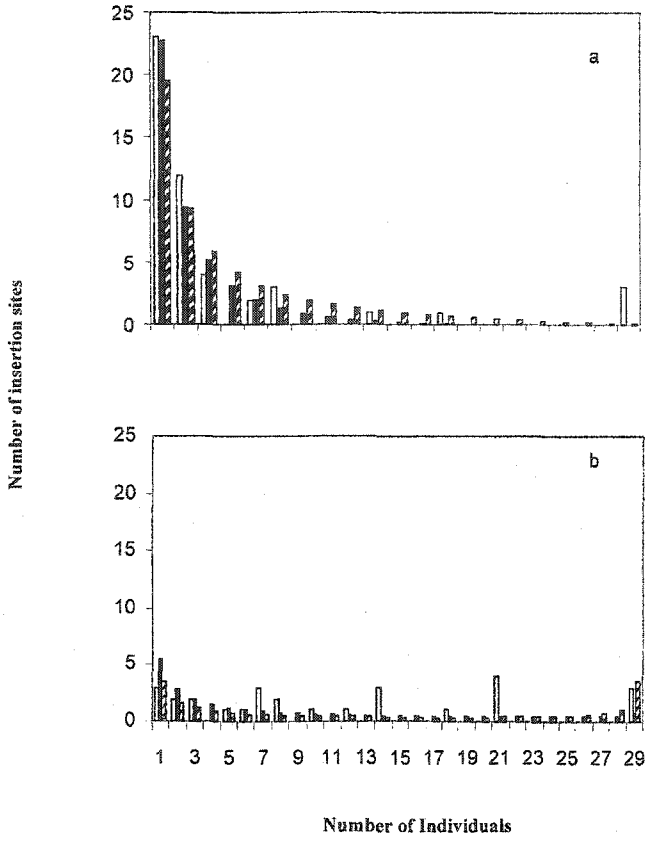


FIGURE 3.—Frequency distributions of element insertions in (a) *A. lyrata* and (b) *A. thaliana*. Open bars, estimated insertion site frequency distribution using the transposon display patterns. Solid bars, predicted site occupancy profile assuming a β-distribution, where the fixed insertion site class is excluded in the estimation. Hatched bars, predicted site occupancy profile assuming a β-distribution, where the fixed insertion class is included. Estimations of the parameters of the β-distribution are as described in the text.

lations. This approach was taken because diversity is structured primarily between selfing populations, making it likely for most insertions to be fixed within populations. However, this contrasting population structure between selfers and outcrossers makes a direct test of population genetics models difficult, since demographic differences between the species may be substantial, and a worldwide sample from *A. thaliana* may represent a population that is out of equilibrium (SAVOLAINEN *et al.* 2000). Nevertheless, more extensive sampling of *A. thaliana* would be expected to inflate levels of transposon diversity and lower their average frequency, while the opposite pattern was observed in this study. Similarly, an alternative explanation to selection for the abundance of low-frequency insertions in *A. lyrata* could be the presence of independent transposition events among diverged populations. Detailed analysis of insertion polymorphism for a large within-population sample would provide an important confirmation of the hypothesis of purifying selection in *A. lyrata*. Once again, however, given that overall nucleotide divergence between

TABLE 4

Parameter estimates based on patterns of insertion polymorphism

Species	β	χ^2(d.f.)
A. lyrata		
a.	3.5*	7.77 (5)
b.	9.7*	1.53 (3)
A. thaliana		
a.	0.5	5.12 (3)
b.	0.7	3.59 (3)

β estimates $4N_e(\nu + s)$, where N_e is the effective population size, ν is the excision rate, and s is the selection coefficient. (a) Including potentially ancient, fixed insertions; (b) excluding fixed insertions. *, estimates of β that are significantly >1, $P < 0.05$.

A. thaliana populations appears to be as great as divergence in *A. lyrata* (SAVOLAINEN *et al.* 2000), the presence of population structure would be unlikely to explain the contrast between species.

The contrasting patterns of insertion polymorphism between the two species are consistent with the hypothesis that the breeding system contributes to the control of transposon dynamics. Several possible forces may be involved in driving this difference. First, a decrease in the abundance of the low-frequency class of insertions is consistent with a reduction in the effectiveness of purifying selection in selfing lineages (WRIGHT and SCHOEN 1999). Alternatively, the differences may be caused by a reduction in the transposition rate in *A. thaliana*. The population data indicate that *A. thaliana* populations show a reduced number of low-frequency insertions relative to *A. lyrata*, but there is not a significant increase in the number of high-frequency insertions, and the total number of insertion sites per individual does not differ dramatically between species (Table 2). Transposon display patterns also suggest a reduction in the species-wide number of polymorphic insertions in *A. thaliana* (Table 2), which has not been observed for nucleotide substitutions (SAVOLAINEN *et al.* 2000). Many or all of the *Ac*-III elements may lack coding capacity, and their transposition rates may thus be subject to control by other autonomous *Ac*-like elements. Stochastic loss of some of these autonomous elements in *A. thaliana* may also be an important force that might reduce transposition rates and drive down levels of insertion polymorphism. For nonautonomous elements, the assumption of a simple linear relationship between the number of elements and transposition rate may be violated.

The use of the technique of transposon display for surveying levels of transposon polymorphism has several weaknesses that may affect interpretation of the results. First, the inability to discriminate between heterozygous and homozygous insertions prevents direct estimation

of the relevant population parameters. However, the use of Hardy-Weinberg assumptions for *A. lyrata* is conservative with respect to the hypothesis of the action of purifying selection, since deleterious selection would be expected to generate a lower frequency of homozygous insertions than predicted. The contrast between the species is also not an artifact of the assumption of complete homozygosity in Arabidopsis; even if the Hardy-Weinberg estimation of allele frequencies is made for the data from *A. thaliana*, no significant departure from neutral expectation is found (data not shown), and thus the contrast between species remains.

The second limitation of the transposon display technique is the possibility that some novel bands arise not from transposition events but from nucleotide substitution leading to restriction site variation or from indel variation. Some of the rare insertion sites may thus not derive from recent transposition events but from polymorphism generated by other classes of mutation. This potential problem is of particular concern with the *A. lyrata* data, where most insertions are present at low frequencies. However, given the sequence analysis of a significant fraction of the identified insertion sites, the small sizes of the regions amplified, and the relatively small number of insertion sites overall, variability in banding pattern is likely to arise primarily from insertion polymorphism. PCR amplification of many individual insertion sites, using primers specific to flanking regions, could allow for confirmation of this assumption as well as for a more direct examination of the levels of heterozygosity of insertion sites.

Our results provide support for the hypothesis of an effect of the selfing rate on driving transposon dynamics in natural populations. This suggests the importance of examining more element families and other related selfing and outcrossing species to investigate whether such patterns are in fact general and consistently associated with differences in the rate of self-fertilization. Comparisons of the transposition properties of element families is also useful to tease apart the role of selection on elements *vs.* their hosts in driving changes in population dynamics. Detailed comparisons of the within-*vs.* between-population patterns of insertion polymorphism are also important for distinguishing the importance of purifying selection and transposition rate. The Arabidopsis genomic sequencing project, as well as the one for rice, provides useful sources of sequence information on transposable elements, and this will allow for further investigation into plant transposon dynamics and the impact of transposons on mutation and evolution.

We thank Brian Charlesworth and Isabel Gordo for detailed advice and assistance with the data analysis; Chuck Langley for helpful discussion; Brian Charlesworth, Deborah Charlesworth, and Martin Morgan for comments on the manuscript; and Outi Savolainen, Chuck Langley, Massimo Pigliucci, Rodney Mauricio, and Thomas Mitchell-Olds for their generous contributions of seed material. This research was supported by National Sciences and Engineering Research Council (NSERC) operating grants to T.E.B. and D.J.S. and by an NSERC PGSA fellowship to S.I.W.

LITERATURE CITED

ABBOT, R. J., and M. F. GOMES, 1989 Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity **62:** 411–418.

BANCROFT, I., and C. DEAN, 1993 Transposition pattern of the maize element Ds in *Arabidopsis thaliana*. Genetics **134:** 1221–1229.

BATZER, M., S. S. ARCOT, J. W. PHINNEY, M. ALEGRIA-HARTMAN, D. H. KASS *et al.*, 1996 Genetic variation of recent Alu insertions in human populations. J. Mol. Evol. **42:** 22–29.

BIEMONT, C., A. TSITRONE, C. VIEIRA and C. HOOGLAND, 1997 Transposable element distribution in *Drosophila*. Genetics **147:** 1997–1999.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

BROOKFIELD, J. F. Y., and R. M. BADGE, 1997 Population genetics models of transposable elements. Genetics **109:** 281–294.

CHARLESWORTH, B., and D. CHARLESWORTH, 1983 The population dynamics of transposable elements. Genet. Res. **42:** 1–27.

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

CHARLESWORTH, B., and C. H. LANGLEY, 1986 The evolution of self-regulated transposition of transposable elements. Genetics **112:** 359–383.

CHARLESWORTH, B., and C. H. LANGLEY, 1989 The population genetics of *Drosophila* transposable elements. Annu. Rev. Genet. **23:** 251–287.

CHARLESWORTH, B., and A. LAPID, 1989 A study of ten families of transposable elements on X chromosomes from a population of *Drosophila melanogaster*. Genet. Res. **54:** 113–125.

CHARLESWORTH, B., A. LAPID and D. CANADA, 1992 The distribution of transposable elements within and between chromosomes in a population of Drosophila melanogaster. I. Element frequencies and distribution. Genet. Res. **60(2):** 103–114.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

CHARLESWORTH, B., C. H. LANGLEY and P. SNIEGOWSKI, 1997 Transposable element distributions in *Drosophila*. Genetics **147:** 1993–1995.

COEN, E. S., and R. CARPENTER, 1986 Transposable elements in Antirrhinum majus: generators of genetic diversity. Trends. Genet. **2:** 292–296.

DELLAPORTA, S. L., J. WOOD and J. B. HICKS, 1983 A plant DNA minipreparation: version II. Plant Mol. Biol. Rep. **1:** 19–21.

DOONER, H. K., and A. BELACHEW, 1991 Chromosome breakage by pairs of closely linked transposable elements of the *Ac/Ds* family in maize. Genetics **129:** 855–862.

FRANK, M. J., D. LIU, Y.-F. TSAY, C. USTACH and N. M. CRAWFORD, 1997 *Tag1* is an autonomous transposable element that shows somatic excision in both Arabidopsis and Tobacco. Plant Cell **9:** 1745–1756.

FRANK, M. J., D. PREUSS, A. MACK, T. C. KUHLMANN and N. M. CRAWFORD, 1998 The *Arabidopsis thaliana* transposable element Tag1 is widely distributed among Arabidopsis ecotypes. Mol. Gen. Genet. **257:** 478–484.

GOLDING, G. B., C. F. AQUADRO and C. H. LANGLEY, 1986 Sequence evolution within populations under multiple types of mutation. Proc. Natl. Acad. Sci. USA **83:** 427–431.

HENK, A. D., R. F. WARREN and R. W. INNES, 1999 A new Ac-like transposon of Arabidopsis is associated with a deletion of the *RPS5* disease resistance gene. Genetics **151:** 1581–1589.

HEY, J., 1989 The transposable portion of the genome of *Drosophila algonquin* is very different from that in *D. melanogaster*. Mol. Biol. Evol. **6:** 66–79.

KÄRKKÄINEN, K., H. KUITTINEN, R. VAN TREUREN, C. VOGL, S. OIKARINEN *et al.*, 1999 Genetics basis of inbreeding depression in Arabis petraea. Evolution **53:** 1354–1365.

KIDWELL, M. G., and D. R. LISCH, 2000 Transposable elements and host genome evolution. Trends Ecol. Evol. **15:** 95–99.

KORSWAGEN, H. C., R. M. DURBIN, M. T. SMITS and R. H. A. PLASTERK, 1996 Transposon Tc1-derived, sequence tagged sites in Caenorhabditis elegans as markers for gene mapping. Proc. Natl. Acad. Sci. USA **93:** 14680–14685.

LE, Q. H., S. WRIGHT, Z. YU and T. BUREAU, 2000 Transposon diversity in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **97:** 7376–7381.

LEUTWILER, L. S., B. R. HOUGH-EVANS and E. M. MEYEROWITZ, 1984 Mol. Gen. Genet. **194:** 15–23.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **231:** 1114–1116.

MONTGOMERY, E. A., and C. H. LANGLEY, 1983 Transposable elements in Mendelian populations. II. Distribution of three copia-like elements in a natural population. Genetics **104:** 473–483.

MONTGOMERY, E. A., S.-M. HUANG, C. H. LANGLEY, and B. H. JUDD, 1991 Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: genome structure and evolution. Genetics **129:** 1085–1098.

MORGAN, M. T., 2001 Transposable element number in mixed mating populations. Genet. Res. (in press).

MURATA, S. N., N. TAKASAKI, M. SAITOH, H. TACHIDA and N. OKADA, 1996 Details of retrotranspositional genome dynamics that provide a rationale for a genetic division: the distinct branching of all the pacific salmon and trout (*Oncorhynchus*) from the atlantic salmon and trout (*Salmo*). Genetics **142:** 915–926.

SANMIGUEL, P., A. TIKHONOV, J. YOUNG-KWAN, N. MOTCHOULSKAIA, D. ZAKHAROV *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765–768.

SANMIGUEL, P., B. S. GAUT, A. TIKHONOV, Y. NAKAJIMA and J. L. BENNETZEN, 1998 The paleontology of intergene retrotransposons of maize. Nat. Genet. **20:** 43–45.

SAVOLAINEN, O., C. H. LANGLEY, B. P. LAZZARO and H. FREVILLE, 2000 Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. Mol. Biol. Evol. **17:** 645–655.

SAWYER, S. A., D. E. DYKHUIZEN and D. L. HARTL, 1987 Confidence interval for the number of selectively neutral amino acid polymorphisms. Proc. Natl. Acad. Sci. USA **84:** 6225–6228.

SHALEV, G., and A. A. LEVY, 1997 The maize transposable element *Ac* induces recombination between the donor site and an homologous ectopic sequence. Genetics **146:** 1143–1151.

SURZYCKI, S. A., and W. R. BELKNAP, 1999 Characterization of repetitive DNA elements in Arabidopsis. J. Mol. Evol. **48:** 684–691.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–593.

TAKASAKI, N., T. YAMAKI, M. HAMADA, L. PARK and N. OKADA, 1997 The salmon *SmaI* family of short interspersed repetitive elements (SINEs): interspecific and intraspecific variation of the insertions of SINEs in the genomes of chum and pink salmon. Genetics **146:** 369–380.

TSAY, Y.-F., M. J. FRANK, T. PAGE, C. DEAN and N. M. CRAWFORD, 1993 Identification of a mobile endogenous transposon in *Arabidopsis thaliana*. Science **260:** 342–344.

VAN DEN BROECK, D., T. MAES, M. SAUER, J. ZETHO, P. DE KEUKELEIRE *et al.*, 1998 Transposon display identifies individual transposable elements in high copy number lines. Plant J. **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 Res. **23:** 4407–4414.

WAUGH, R., K. MCLEAN, A. J. FLAVELL, S. R. PEARCE, A. KUMAR *et al.*, 1997 Genetic distribution of BARE-1 retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms. Mol. Gen. Genet. **253:** 687–694.

WRIGHT, S. I., and D. J. SCHOEN, 1999 Transposon dynamics and the breeding system. Genetica **1/3:** 139–148.

ZHANG, J., and T. PETERSON, 1999 Genome rearrangements by nonlinear transposons in maize. Genetics **153:** 1403–1410.

Communicating editor: O. SAVOLAINEN

Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*

The Arabidopsis Genome Initiative*

* *Authorship of this paper should be cited as 'The Arabidopsis Genome Iniative'. A full list of contributors appears at the end of this paper*

The flowering plant *Arabidopsis thaliana* is an important model system for identifying genes and determining their functions. Here we report the analysis of the genomic sequence of *Arabidopsis*. The sequenced regions cover 115.4 megabases of the 125-megabase genome and extend into centromeric regions. The evolution of *Arabidopsis* involved a whole-genome duplication, followed by subsequent gene loss and extensive local gene duplications, giving rise to a dynamic genome enriched by lateral gene transfer from a cyanobacterial-like ancestor of the plastid. The genome contains 25,498 genes encoding proteins from 11,000 families, similar to the functional diversity of *Drosophila* and *Caenorhabditis elegans*— the other sequenced multicellular eukaryotes. *Arabidopsis* has many families of new proteins but also lacks several common protein families, indicating that the sets of common proteins have undergone differential expansion and contraction in the three multicellular eukaryotes. This is the first complete genome sequence of a plant and provides the foundations for more comprehensive comparison of conserved processes in all eukaryotes, identifying a wide range of plant-specific gene functions and establishing rapid systematic ways to identify genes for crop improvement.

The plant and animal kingdoms evolved independently from unicellular eukaryotes and represent highly contrasting life forms. The genome sequences of *C. elegans*[1] and *Drosophila*[2] reveal that metazoans share a great deal of genetic information required for developmental and physiological processes, but these genome sequences represent a limited survey of multicellular organisms. Flowering plants have unique organizational and physiological properties in addition to ancestral features conserved between plants and animals. The genome sequence of a plant provides a means for understanding the genetic basis of differences between plants and other eukaryotes, and provides the foundation for detailed functional characterization of plant genes.

Arabidopsis thaliana has many advantages for genome analysis, including a short generation time, small size, large number of offspring, and a relatively small nuclear genome. These advantages promoted the growth of a scientific community that has investigated the biological processes of *Arabidopsis* and has characterized many genes[3]. To support these activities, an international collaboration (the Arabidopsis Genome Initiative, AGI) began sequencing the genome in 1996. The sequences of chromosomes 2 and 4 have been reported[4,5], and the accompanying Letters describe the sequences of chromosomes 1 (ref. 6), 3 (ref. 7) and 5 (ref. 8).

Here we report analysis of the completed *Arabidopsis* genome sequence, including annotation of predicted genes and assignment of functional categories. We also describe chromosome dynamics and architecture, the distribution of transposable elements and other repeats, the extent of lateral gene transfer from organelles, and the comparison of the genome sequence and structure to that of other *Arabidopsis* accessions (distinctive lines maintained by single-seed descent) and plant species. This report is the summation of work by experts interested in many biological processes selected to illuminate plant-specific functions including defence, photomorphogenesis, gene regulation, development, metabolism, transport and DNA repair.

The identification of many new members of receptor families, cellular components for plant-specific functions, genes of bacterial origin whose functions are now integrated with typical eukaryotic components, independent evolution of several families of transcription factors, and suggestions of as yet uncharacterized metabolic pathways are a few more highlights of this work. The implications of these discoveries are not only relevant for plant biologists, but will also affect agricultural science, evolutionary biology, bioinformatics, combinatorial chemistry, functional and comparative genomics, and molecular medicine.

Overview of sequencing strategy

We used large-insert bacterial artificial chromosome (BAC), phage (P1) and transformation-competent artificial chromosome (TAC) libraries[9–12] as the primary substrates for sequencing. Early stages of genome sequencing used 79 cosmid clones. Physical maps of the genome of accession Columbia were assembled by restriction fragment 'fingerprint' analysis of BAC clones[13], by hybridization[14] or polymerase chain reaction (PCR)[15] of sequence-tagged sites and by hybridization and Southern blotting[16]. The resulting maps were integrated (http://nucleus/cshl.org/arabmaps/) with the genetic map and provided a foundation for assembling sets of contigs into sequence-ready tiling paths. End sequence (http://www.tigr.org/tdb/at/abe/bac_end_search.html) of 47,788 BAC clones was used to extend contigs from BACS anchored by marker content and to integrate contigs.

Ten contigs representing the chromosome arms and centromeric heterochromatin were assembled from 1,569 BAC, TAC, cosmid and P1 clones (average insert size 100 kilobases (kb)). Twenty-two PCR products were amplified directly from genomic DNA and sequenced to link regions not covered by cloned DNA or to optimize the minimal tiling path. Telomere sequence was obtained from specific yeast artificial chromosome (YAC) and phage clones, and from inverse polymerase chain reaction (IPCR) products derived from genomic DNA. Clone fingerprints, together with BAC end sequences, were generally adequate for selection of clones for sequencing over most of the genome. In the centromeric regions, these physical mapping methods were supplemented with genetic mapping to identify contig positions and orientation[17].

Selected clones were sequenced on both strands and assembled using standard techniques. Comparison of independently derived sequence of overlapping regions and independent reassembly sequenced clones revealed accuracy rates between 99.99 and 99.999%. Over half of the sequence differences were between genomic and BAC clone sequence. All available sequenced genetic markers were integrated into sequence assemblies to verify sequence contigs[4–8]. The total length of sequenced regions, which extend from either the telomeres or ribosomal DNA repeats to the 180-base-pair

(bp) centromeric repeats, is 115,409,949 bp (Table 1). Estimates of the unsequenced centromeric and rDNA repeat regions measure roughly 10 megabases (Mb), yielding a genome size of about 125 Mb, in the range of the 50–150 Mb haploid content estimated by different methods[18]. In general, features such as gene density, expression levels and repeat distribution are very consistent across the five chromosomes (Fig. 1), and these are described in detail in reports on individual chromosomes[4–8] and in the analysis of centromere, telomere and rDNA sequences.

We used tRNAscan-SE 1.21 (ref. 19) and manual inspection to identify 589 cytoplasmic transfer RNAs, 27 organelle-derived tRNAs and 13 pseudogenes—more than in any other genome sequenced to date. All 46 tRNA families needed to decode all possible 61 codons were found, defining the completeness of the functional set. Several highly amplified families of tRNAs were found on the same strand[6]; excluding these, each amino acid is decoded by 10–41 tRNAs.

The spliceosomal RNAs (U1, U2, U4, U5, U6) have all been experimentally identified in *Arabidopsis*. The previously identified sequences for all RNAs were found in the genome, except for U5 where the most similar counterpart was 92% identical. Between 10 and 16 copies of each small nuclear RNA (snRNA) were found across all chromosomes, dispersed as singletons or in small groups.

The small nucleolar RNAs (snoRNAs) consist of two subfamilies, the C/D box snoRNAs, which includes 36 *Arabidopsis* genes, and the H/ACA box snoRNAs, for which no members have been identified in *Arabidopsis*. U3 is the most numerous of the C/D box snoRNAs, with eight copies found in the genome. We identified forty-five additional C/D box snoRNAs using software (www.rna.wustl.edu/snoRNAdb/) that detects snoRNAs that guide ribose methylation of ribosomal RNA.

A combination of algorithms, all optimized with parameters based on known *Arabidopsis* gene structures, was used to define gene structure. We used similarities to known protein and expressed sequence tag (EST) sequence to refine gene models. Eighty per cent of the gene structures predicted by the three centres involved were completely consistent, 93% of ESTs matched gene models, and less than 1% of ESTs matched predicted non-coding regions, indicating

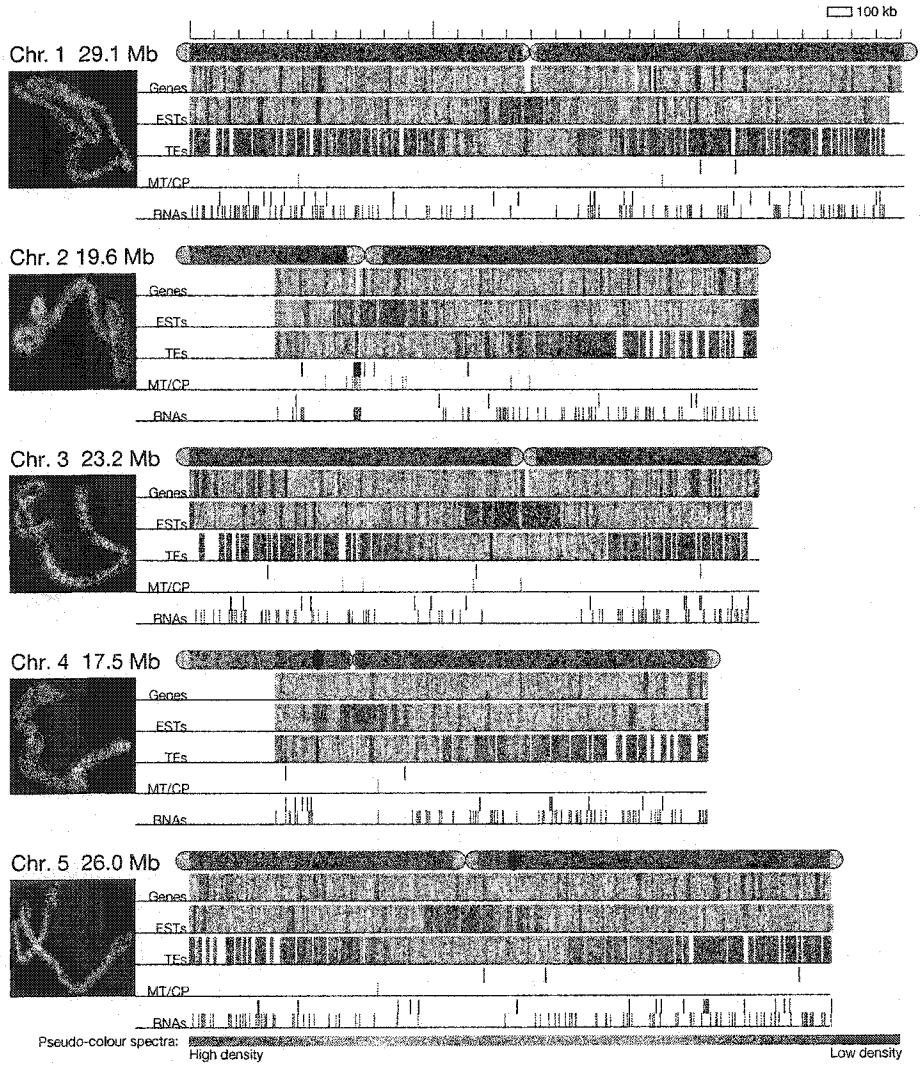


Figure 1 Representation of the *Arabidopsis* chromosomes. Each chromosome is represented as a coloured bar. Sequenced portions are red, telomeric and centromeric regions are light blue, heterochromatic knobs are shown black and the rDNA repeat regions are magenta. The unsequenced telomeres 2N and 4N are depicted with dashed lines. Telomeres are not drawn to scale. Images of DAPI-stained chromosomes were kindly supplied by P. Fransz. The frequency of features was given pseudo-colour assignments, from red (high density) to deep blue (low density). Gene density ('Genes') ranged from 38 per 100 kb to 1 gene per 100 kb; expressed sequence tag matches ('ESTs') ranged from more than 200 per 100 kb to 1 per 100 kb. Transposable element densities ('TEs') ranged from 33 per 100 kb to 1 per 100 kb. Mitochondrial and chloroplast insertions ('MT/CP') were assigned black and green tick marks, respectively. Transfer RNAs and small nucleolar RNAs ('RNAs') were assigned black and red ticks marks, respectively.

that most potential genes were identified. The sensitivity and selectivity of the gene prediction software used in this report has been comprehensively and independently assessed[20].

The 25,498 genes predicted (Table 1) is the largest gene set published to date: *C. elegans*[1] has 19,099 genes and *Drosophila*[2] 13,601 genes. *Arabidopsis* and *C. elegans* have similar gene density, whereas *Drosophila* has a lower gene density; *Arabidopsis* also has a significantly greater extent of tandem gene duplications and segmental duplications, which may account for its larger gene set.

The rDNA repeat regions on chromosomes 2 and 4 were not sequenced because of their known repetitive structure and content. The centromeric regions are not completely sequenced owing to large blocks of monotonic repeats such as 5S rDNA and 180-bp repeats. The sequence continues to be extended further into centromeric and other regions of complex sequence.

Characterization of the coding regions

To assess the similarities and differences of the *Arabidopsis* gene complement compared with other sequenced eukaryotic genomes, we assigned functional categories to the complete set of *Arabidopsis* genes. For chromosome 4 genes and the yeast genome, predicted functions were previously manually assigned[5,21]. All other predicted proteins were automatically assigned to these functional categories[22], assuming that conserved sequences reflect common functional relationships.

The functions of 69% of the genes were classified according to sequence similarity to proteins of known function in all organisms; only 9% of the genes have been characterized experimentally (Fig. 2a). Generally similar proportions of gene products were predicted to be targeted to the secretory pathway and mitochondria in *Arabidopsis* and yeast, and up to 14% of the gene products are

Table 1 Summary statistics of the *Arabidopsis* genome

Feature	Value					
(a) The DNA molecules						
	Chr. 1	Chr. 2	Chr. 3	Chr. 4	Chr. 5	Σ
Length (bp)	29,105,111	19,646,945	23,172,617	17,549,867	25,953,409	115,409,949
Top arm (bp)	14,449,213	3,607,091	13,590,268	3,052,108	11,132,192	
Bottom arm (bp)	14,655,898	16,039,854	9,582,349	14,497,759	14,803,217	
Base composition (%GC)						
Overall	33.4	35.5	35.4	35.5	34.5	
Coding	44.0	44.0	44.3	44.1	44.1	
Non-coding	32.4	32.9	33.0	32.8	32.5	
Number of genes	6,543	4,036	5,220	3,825	5,874	25,498
Gene density (kb per gene)	4.0	4.9	4.5	4.6	4.4	
Average gene length (bp)	2,078	1,949	1,925	2,138	1,974	
Average peptide length (bp)	446	421	424	448	429	
Exons						
Number	35,482	19,631	26,570	20,073	31,226	13,2982
Total length (bp)	8,772,559	5,100,288	6,654,507	5,150,883	7,571,013	33,249,250
Average per gene	5.4	4.9	5.1	5.2	5.3	
Average size (bp)	247	259	250	256	242	
Introns						
Number	28,939	15,595	21,350	16,248	25,352	107,484
Total length (bp)	4,828,766	2,768,430	3,397,531	3,030,649	4,030,045	18,055,421
Average size (bp)	168	177	159	186	159	
Number of genes with ESTs (%)	60.8	56.9	59.8	61.4	61.4	
Number of ESTs	30,522	14,989	20,732	16,605	22,885	105,733
(b) The proteome						
Classification/function						
Total proteins	6,543	4,036	5,220	3,825	5,874	25,498
With INTERPRO	4,194	1,205	2,989	1,545	3,136	13,069
domains	64.1%	29.9%	57.8%	40.4%	53.4%	51.3%
Genes containing at	2,334	1,322	1,615	1,402	1,940	8,613
least one TM domain	35.7%	32.8%	30.9%	36.7%	33.0%	33.8%
Genes containing at	2,513	1,424	1,664	1,304	2,121	9,026
least one SCOP domain	38.4%	35.3%	31.9%	34.1%	36.1%	35.4%
With putative signal peptides						
Secretory pathway	1,242 19.0%	675 16.7%	877 17.0%	659 17.2%	1,014 17.3%	4,467 17.6%
>0.95 specificity	1,146 17.5%	632 15.7%	813 15.7%	632 16.5%	964 16.4%	4,167 16.4%
Chloroplast	866 13.2%	535 13.2%	754 14.6%	532 13.9%	887 15.1%	3,574 14.0%
>0.95 specificity	602 9.2%	290 7.2%	420 8.1%	298 7.8%	475 8.1%	2,085 8.2%
mitochondria	901 13.8%	425 10.5%	554 10.7%	390 10.2%	627 10.7%	2,897 11.4%
>0.95 specificity	113 1.7%	49 1.2%	63 1.2%	59 1.5%	65 1.1%	349 1.4%
Functional classification						
Cellular metabolism	1,188 22.7%	620 23.3%	745 22.8%	588 22.9%	868 21.1%	4,009 22.5%
Transcription	880 16.8%	474 17.8%	566 17.3%	335 13.1%	763 18.6%	3,018 16.9%
Plant defence	640 12.2%	276 10.4%	354 10.8%	295 11.5%	490 11.9%	2,055 11.5%
Signalling	573 11.0%	296 11.1%	356 10.9%	210 8.2%	420 10.2%	1,855 10.4%
Growth	542 10.4%	263 9.9%	357 10.9%	448 17.5%	469 11.4%	2,079 11.7%
Protein fate	520 9.9%	273 10.2%	314 9.6%	264 10.3%	395 9.6%	1,766 9.9%
Intracellular transport	435 8.3%	214 8.9%	269 8.2%	220 8.6%	334 8.1%	1,472 8.3%
Transport	236 4.5%	139 5.2%	155 4.7%	113 4.4%	206 5.0%	849 4.8%
Protein synthesis	216 4.1%	111 4.2%	148 4.5%	90 3.5%	165 4.0%	730 4.1%
Total	5,230	2,666	3,264	2,563	4,110	17,833

The features of *Arabidopsis* chromosomes 1–5 and the complete nuclear genome are listed. Specialized searches used the following programs and databases: INTERPRO[23]; transmembrane (TM) domains by ALOM2 (unpublished); SCOP domain database[121]; functional classification by the PEDANT analysis system[22]. Signal peptide prediction (secretory pathway, targeted to chloroplast or mitochondria) was performed using TargetP[122] and http://www.cbs.dtu.dk/services/TargetP/.

*Default value.

likely to be targeted to the chloroplast (Table 1). The significant proportion of genes with predicted functions involved in metabolism, gene regulation and defence is consistent with previous analyses[5]. Roughly 30% of the 25,498 predicted gene products, (Fig. 2a), comprising both plant-specific proteins and proteins with similarity to genes of unknown function from other organisms, could not be assigned to functional categories.

To compare the functional catagories in more detail, we compared data from the complete genomes of *Escherichia coli*[23], *Synechocystis* sp.[24], *Saccharomyces cerevisiae*[21], *C. elegans*[1] and *Drosophila*[2], and a non-redundant protein set of *Homo sapiens*, with the *Arabidopsis* genome data (Fig. 2b), using a stringent BLASTP threshold value of $E < 10^{-30}$. The proportion of *Arabidopsis* proteins having related counterparts in eukaryotic genomes varies by a factor of 2 to 3 depending on the functional category. Only 8–23% of *Arabidopsis* proteins involved in transcription have related genes in other eukaryotic genomes, reflecting the independent evolution of many plant transcription factors. In contrast, 48–60% of genes involved in protein synthesis have counterparts in the other eukaryotic genomes, reflecting highly conserved gene functions. The relatively high proportion of matches between *Arabidopsis* and bacterial proteins in the categories 'metabolism' and 'energy' reflects both the acquisition of bacterial genes from the ancestor of the plastid and high conservation of sequences across all species. Finally, a comparison between unicellular and multicellular eukaryotes indicates that *Arabidopsis* genes involved in cellular communication and signal transduction have more counterparts in multicellular eukaryotes than in yeast, reflecting the need for sets of genes for communication in multicellular organisms.

Pronounced redundancy in the *Arabidopsis* genome is evident in segmental duplications and tandem arrays, and many other genes with high levels of sequence conservation are also scattered over the genome. Sequence similarity exceeding a BLASTP value $E < 10^{-20}$ and extending over at least 80% of the protein length were used as parameters to identify protein families (Table 2). A total of 11,601 protein types were identified. Thirty-five per cent of the predicted proteins are unique in the genome, and the proportion of proteins belonging to families of more than five members is substantially higher in *Arabidopsis* (37.4%) than in *Drosophila* (12.1%) or

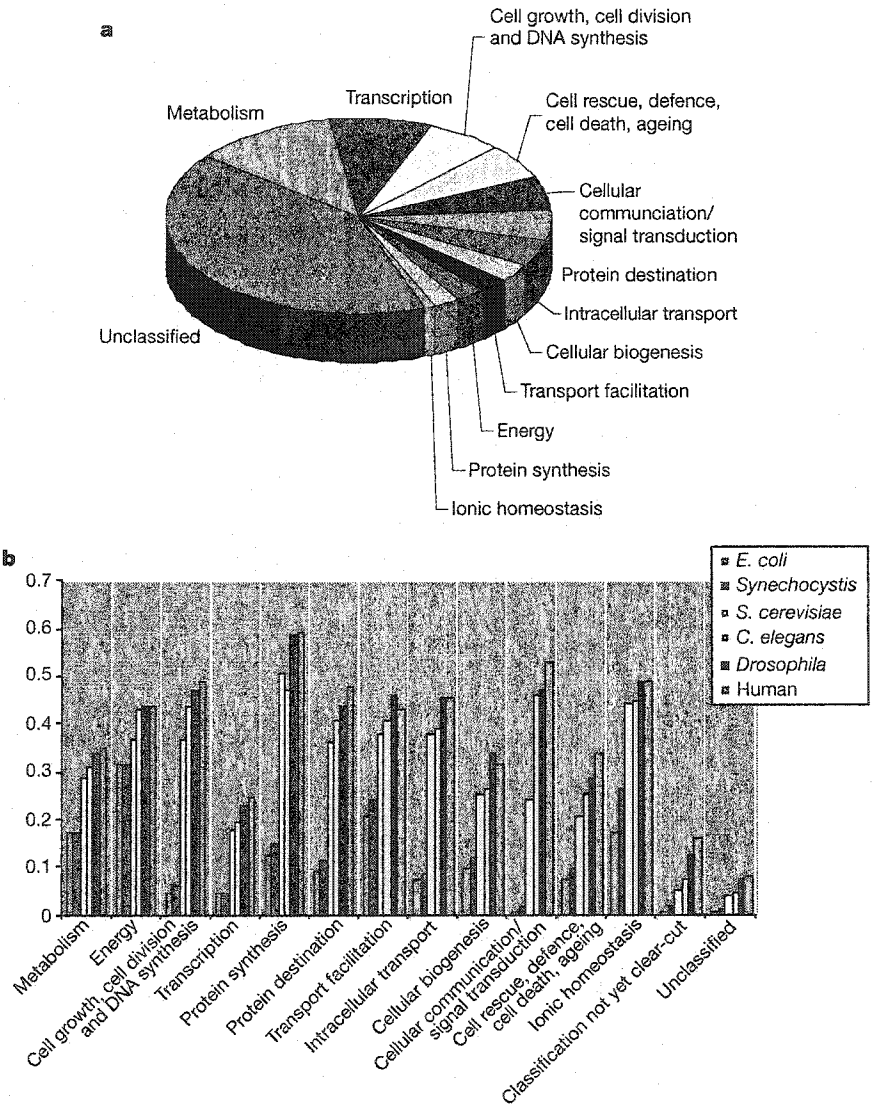


Figure 2 Functional analysis of *Arabidopsis* genes. **a**, Proportion of predicted *Arabidopsis* genes in different functional categories. **b**, Comparison of functional categories between organisms. Subsets of the *Arabidopsis* proteome containing all proteins that fall into a common functional class were assembled. Each subset was searched against the complete set of translations from *Escherichia coli*, *Synechocystis* sp. PCC6803, *Saccharomyces cerevisiae*, *Drosophila*, *C. elegans* and a *Homo sapiens* non-redundant protein database. The percentage of *Arabidopsis* proteins in a particular subset that had a BLASTP match with $E \leq 10^{-30}$ to the respective reference genome is shown. This reflects the measure of sequence conservation of proteins within this particular functional category between *Arabidopsis* and the respective reference genome. *y* axis, 0.1 = 10%.

Table 2 Proportion of genes in different organisms present as either singletons or in paralogous families

	No of singletons and distinct gene families	Unique	Gene families containing				
			2 members	3 members	4 members	5 members	>5 members
H. influenzae	1,587	88.8%	6.8%	2.3%	0.7%	0.0%	1.4%
S. cerevisiae	5,105	71.4%	13.8%	3.5%	2.2%	0.7%	8.4%
D. melanogaster	10,736	72.5%	8.5%	3.4%	1.9%	1.6%	12.1%
C. elegans	14,177	55.2%	12.0%	4.5%	2.7%	1.6%	24.0%
Arabidopsis	11,601	35.0%	12.5%	7.0%	4.4%	3.6%	37.4%

The number of genes in the genomes of *Haemophilus influenzae*, *S. cerevisiae*, *Drosophila*, *C. elegans* and *Arabidopsis* that are present either as singletons or in gene families with two or more members are listed. To be grouped in a gene family, two genes had to show similarity exceeding a BLASTP value $E < 10^{-20}$ and a FASTA alignment over at least 80% of the protein length. In column 1, the number of genes that are unique plus the number of gene families are listed. Columns 2 to 6 give the percentage of genes present as singletons or in gene families of *n* members.

C. elegans (24.0%). The absolute number of *Arabidopsis* gene families and singletons (types) is in the same range as the other multicellular eukaryotes, indicating that a proteome of 11,000–15,000 types is sufficient for a wide diversity of multicellular life. The proportion of gene families with more than two members is considerably more pronounced in *Arabidopsis* than in other eukaryotes (Fig. 3). As segmental duplication is responsible for 6,303 gene duplications (see below), the extent of tandem gene duplications accounts for a significant proportion of the increased family size. These features of the *Arabidopsis*, and presumably other plant genomes, may indicate more relaxed constraints on genome size in plants, or a more prominent role of unequal crossing over to generate new gene copies.

Conserved protein domains revealed more informative differences through INTERPRO[25] analysis of the predicted gene products from *Arabidopsis*, *S. cerevisiae*, *C. elegans* and *Drosophila*. Statistically over-represented domains, and those that are absent from the *Arabidopsis* genome, indicate domains that may have been gained or lost during the evolution of plants (Supplementary Information Table 1). Proteins containing the Pro-Pro-Arg repeat, which is involved in RNA stabilization and RNA processing, are over-represented as compared to yeast, fly and worm; 400 proteins containing this signature were detected in *Arabidopsis* compared with only 10 in total in yeast, *Drosophila* and *C. elegans*. Protein kinases and associated domains, 169 proteins containing a disease resistance protein signature, and the Toll/IL-1R (TIR) domain, a component of pathogen recognition molecules[26], are also relatively abundant. This suggests that pathways transducing signals in response to pathogens and diverse environmental cues are more abundant in plants than in other organisms.

The RING zinc finger domain is relatively over-represented in *Arabidopsis* compared with yeast, *Drosophila* and *C. elegans*, whereas the F-box domain is over-represented as compared with yeast and *Drosophila* only. These domains are involved in targeting proteins to the proteasome[27] and ubiquitinylation[28] pathways of protein degradation, respectively. In plants many processes such as hormone and defence responses, light signalling, and circadian rhythms and pattern formation use F-box function to direct negative regulators

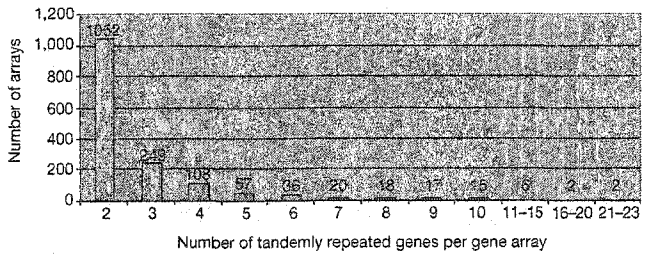


Figure 3 Distribution of tandemly repeated gene arrays in the *Arabidopsis* genome. Tandemly repeated gene arrays were identified using the BLASTP program with a threshold of $E < 10^{-20}$. One unrelated gene among cluster members was tolerated. The histogram gives the number of clusters in the genome containing 2 to *n* similar gene units in tandem.

to the ubiquitin degradation pathway. This mode of regulation appears to be more prevalent in plants and may account for a higher representation of the F box than in *Drosophila* and for the over-representation of the ubiquitin domain in the *Arabidopsis* genome. RING finger domain proteins in general have a role in ubiquitin protein ligases, indicating that proteasome-mediated degradation is a more widespread mode of regulation in plants than in other kingdoms.

Most functions identified by protein domains are conserved in similar proportions in the *Arabidopsis*, *S. cerevisiae*, *Drosophila* and *C. elegans* genomes, pointing to many ubiquitous eukaryotic pathways. These are illustrated by comparing the list of human disease genes[29] to the complete *Arabidopsis* gene set using BLASTP. Out of 289 human disease genes, 139 (48%) had hits in *Arabidopsis* using a BLASTP threshold $E < 10^{-10}$. Sixty-nine (24%) exceeded an $E < 10^{-40}$ threshold, and 26 (9.3%) had scores better than $E < 10^{-100}$ (Table 3). There are at least 17 human disease genes more similar to *Arabidopsis* genes than yeast, *Drosophila* or *C. elegans* genes (Table 3).

This analysis shows that, although numerous families of proteins are shared between all eukaryotes, plants contain roughly 150 unique protein families. These include transcription factors, structural proteins, enzymes and proteins of unknown function. Members of the families of genes common to all eukaryotes have undergone substantial increases or decreases in their size in *Arabidopsis*. Finally, the transfer of a relatively small number of cyanobacteria-related genes from a putative endosymbiotic ancestor of the plastid has added to the diversity of protein structures found in plants.

Genome organization and duplication

The *Arabidopsis* genome sequence provides a complete view of chromosomal organization and clues to its evolutionary history. Gene families organized in tandem arrays of two or more units have been described in *C. elegans*[1] and *Drosophila*[2]. Analysis of the *Arabidopsis* genome revealed 1,528 tandem arrays containing 4,140 individual genes, with arrays ranging up to 23 adjacent members (Fig. 3). Thus 17% of all genes of *Arabidopsis* are arranged in tandem arrays.

Large segmental duplications were identified either by directly aligning chromosomal sequences or by aligning proteins and searching for tracts of conserved gene order. All five chromosomes were aligned to each other in both orientations using MUMmer[30], and the results were filtered to identify all segments at least 1,000 bp in length with at least 50% identity (Supplementary Information Fig. 1). These revealed 24 large duplicated segments of 100 kb or larger, comprising 65.6 Mb or 58% of the genome. The only duplicated segment in the centromeric regions was a 375-kb segment on chromosome 4. Many duplications appear to have undergone further shuffling, such as local inversions after the duplication event.

We used TBLASTX[5] to identify collinear clusters of genes residing in large duplicated chromosomal segments. The duplicated regions encompass 67.9 Mb, 60% of the genome, slightly more than was

found in the DNA-based alignment (Fig. 4), and these data extend earlier findings[4,5,31]. The extent of sequence conservation of the duplicated genes varies greatly, with 6,303 (37%) of the 17,193 genes in the segments classified as highly conserved ($E < 10^{-30}$) and a further 1,705 (10%) showing less significant similarity up to $E < 10^{-5}$. The proportion of homologous genes in each duplicated segment also varies widely, between 20% and 47% for the highly conserved class of genes. In many cases, the number of copies of a gene and its counterpart differ (for example, one copy on one chromosome and multiple copies on the other; see Supplementary Information Fig. 2); this could be due to either tandem duplication or gene loss after the segmental duplication.

What does the duplication in the *Arabidopsis* genome tell us about the ancestry of the species? Polyploidy occurs widely in plants and is proposed to be a key factor in plant evolution[32]. As the majority of the *Arabidopsis* genome is represented in duplicated (but not triplicated) segments, it appears most likely that *Arabidopsis*, like maize, had a tetraploid ancestor[33]. A comparative sequence analysis of *Arabidopsis* and tomato estimated that a duplication occurred ~112 Myr ago to form a tetraploid[34]. The degrees of conservation of the duplicated segments might be due to divergence from an ancestral autotetraploid form, or might reflect differences present in an allotetraploid ancestor. It is also possible, however, that several independent segmental duplication events took place instead of tetraploid formation and stabilization.

The diploid genetics of *Arabidopsis* and the extensive divergence of the duplicated segments have masked its evolutionary history. The determination of *Arabidopsis* gene functions must therefore be pursued with the potential for functional redundancy taken into account. The long period of time over which genome stabilization has occurred has, however, provided ample opportunity for the divergence of the functions of genes that arose from duplications.

Comparative analysis of *Arabidopsis* accessions

Comparing the multiple accessions of *Arabidopsis* allows us to identify commonly occurring changes in genome microstructure. It also enables the development of new molecular markers for genetic mapping. High rates of polymorphism between *Arabidopsis* accessions, including both DNA sequence and copy number of tandem arrays, are prevalent at loci involved in disease resistance[35]. This has been observed for other plant species, and such loci are thought to serve as templates for illegitimate recombination to create new pathogen response specificities[36]. We carried out a comparative analysis between 82 Mb of the genome sequence of *Arabidopsis* accession Columbia (Col-0) and 92.1 Mb of non-redundant low-pass (twofold redundant) sequence data of the genomic DNA of accession Landsberg *erecta* (*Ler*). We identified two classes of differences between the sequences: single nucleotide polymorphisms (SNPs), and insertion–deletions (InDels). As we used high stringency criteria, our results represent a minimum estimate of numbers of polymorphisms between the two genomes.

In total, we detected 25,274 SNPs, representing an average density of 1 SNP per 3.3 kb. Transitions (A/T–G/C) represented 52.1% of the SNPs, and transversions accounted for the remainder: 17.3% for A/T–T/A, 22.7% for A/T–C/G and 7.9% for C/G–G/C. In total, we detected 14,570 InDels at an average spacing of 6.1 kb. They ranged from 2 bp to over 38 kilobase-pairs, although 95% were smaller than 50 bp. Only 10% of the InDels were co-located with simple sequence repeats identified with the program Sputnik. An analysis of 416 relative insertions greater than 250 bp in Col-0 showed that 30% matched transposon-related proteins, indicating that a substantial proportion of the large InDels are the result of transposon insertion or excision. Many InDels contained entire active genes not related to transposons. Half of such genes absent from corresponding positions in the Col-0 sequence were found elsewhere on the genome of *Ler*. This indicates that genes have been transferred to new genomic locations.

Gene structures are often affected by small InDels and SNPs. The positions of SNPs and InDels were mapped relative to 87,427 exons and 70,379 introns annotated in the Col-0 sequence. SNPs were found in exons, introns and intergenic regions at frequencies of 1 SNP per 3.1, 2.2 and 3.5 kb, respectively. The frequencies for InDels were 1 per 9.3, 3.1 and 4.3 kb, respectively. Polymorphisms were detected in 7% of exons, and alter the spliced sequences of 25% of the predicted genes. For InDels in exons, insertion lengths divisible by three are prevalent for small insertions (< 50 bp), indicating that many proteins can withstand small insertions or deletions of amino acids without loss of function.

Our analyses show that sequence polymorphisms between accessions of *Arabidopsis* are common, and that they occur in both coding and non-coding regions. We found evidence for the relocation of genes in the genome, and for changes in the complement of transposable elements. The data presented here are available at http://www.arabidopsis.org/cereon/.

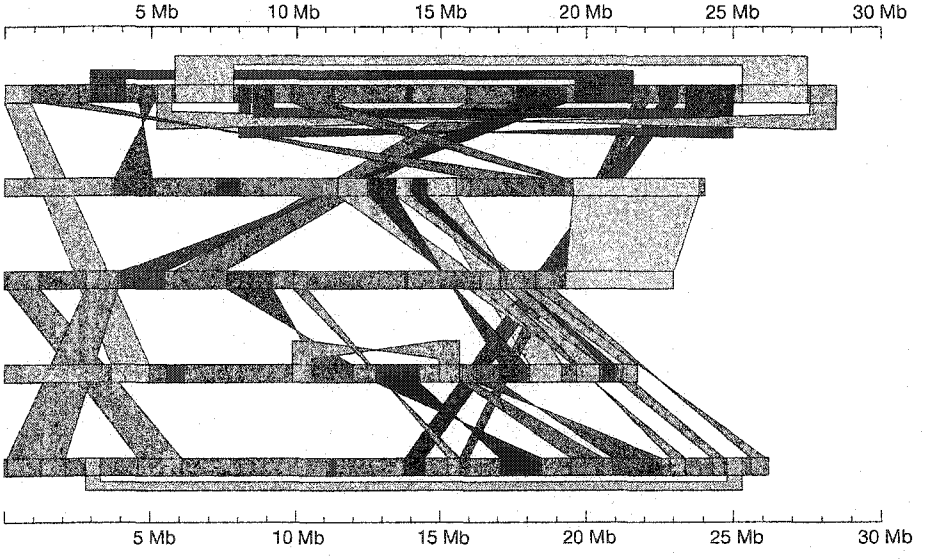


Figure 4 Segmentally duplicated regions in the *Arabidopsis* genome. Individual chromosomes are depicted as horizontal grey bars (with chromosome 1 at the top), centromeres are marked black. Coloured bands connect corresponding duplicated segments. Similarity between the rDNA repeats are excluded. Duplicated segments in reversed orientation are connected with twisted coloured bands. The scale is in megabases.

Comparison of *Arabidopsis* and other plant genera

Comparative genetic mapping can reveal extensive conservation of genome organization between closely related species[37,38]. The comparative analysis of plant genome microstructure reveals much about the evolution of plant genomes and provides unprecedented opportunities for crop improvement by establishing the detailed structures of, and relationships between, the genomes of crops and *Arabidopsis*.

The lineages leading to *Arabidopsis* and *Capsella rubella* (shepherd's purse) diverged between 6.2 and 9.8 Myr ago, and the gene content and genome organization of *C. rubella* is very similar to that of *Arabidopsis*[39], including the large-scale duplications. Alignment of *Arabidopsis* complementary DNA and EST sequences with genomic DNA sequences of *Arabidopsis* and *C. rubella* showed conservation of exon length and intron positions. Coding sequences predicted from these alignments differed from the annotated *Arabidopsis* gene sequences in two out of five cases.

The ancestral lineages of *Arabidopsis* and the *Brassica* (cabbage and mustard) genera diverged 12.2–19.2 Myr ago[40]. *Brassica* genes show a high level of nucleotide conservation with their *Arabidopsis* orthologues, typically more than 85% in coding regions[40]. The structure of *Brassica* genomes resembles that of *Arabidopsis*, but with extensive triplication and rearrangement[41], and extensive divergence of microstructure (Supplementary Information Fig. 3). The divergence between the genomes of *Arabidopsis* and *Brassica oleracea* is in striking contrast to that observed between *Arabidopsis* and *C. rubella*, although the time since divergence is only twofold greater. This accelerated rate of change in triplicated segments of the genome of *B. oleracea* indicates that polyploidy fosters rapid chromosomal evolution.

The *Arabidopsis* and tomato lineages diverged roughly 150 Myr ago, and comparative sequence analysis of segments of their genomes has revealed complex relationships[34]. Four regions of the *Arabidopsis* genome are related to each other and to one region in the tomato genome, suggesting that two rounds of duplication may have occurred in the *Arabidopsis* lineage. The extensive duplication described here supports the proposal that the more recent of these duplications, estimated to have occurred ~112 Myr ago, was the result of a polyploidization event. The lineages of *Arabidopsis* and rice diverged ~200 Myr ago[42]. Three regions of the genome of *Arabidopsis* were related to each other and to one region in the rice genome, providing further evidence for multiple duplication events[43,44].

The frequent occurrence of tandem gene duplications and the apparent deletion of single genes, or small groups of adjacent genes, from duplicated regions suggests that unequal crossing over may be a key mechanism affecting the evolution of plant genome microstructure. However, the segmental inversions and gene translocations in the genomes of both rice and *B. oleracea* that are not found in *Arabidopsis* indicate that additional mechanisms may be involved[40].

Integration of the three genomes in the plant cell

The three genomes in the plant cell—those of the nucleus, the plastids (chloroplasts) and the mitochondria—differ markedly in gene number, organization and stability. Plastid genes are densely packed in an order highly conserved in all plants[45], whereas mitochondrial genes[46] are widely dispersed and subjected to extensive recombination.

Organellar genomes are remnants of independent organisms—plastids are derived from the cyanobacterial lineage and mitochondria from the α-Proteobacteria. The remaining genes in plastids include those that encode subunits of the photosystem and the electron transport chain, whereas the genes in mitochondria encode essential subunits of the respiratory chain. Both organelles contain sets of specific membrane proteins that, together with housekeeping proteins, account for 61% of the genes in the chloroplast and 88 % in the mitochondrion (Table 4). The balances are involved in transcription and translation.

The number of proteins encoded in the nucleus likely to be found

Table 3 *Arabidopsis* genes with similarities to human disease genes

Human disease gene	*E* value	Gene code	*Arabidopsis* hit
Darier–White, SERCA	5.9×10^{-272}	T27I1_16	Putative calcium ATPase
Xeroderma Pigmentosum, D-XPD	7.2×10^{-228}	F15K9_19	Putative DNA repair protein
Xeroderma pigment, B-ERCC3	9.6×10^{-214}	AT5g41360	DNA excision repair cross-complementing protein
Hyperinsulinism, ABCC8	7.1×10^{-188}	F20D22_11	Multidrug resistance protein
Renal tubul. acidosis, ATP6B1	1.0×10^{-182}	AT4g38510	Probable H+-transporting ATPase
HDL deficiency 1, ABCA1	2.4×10^{-181}	At2g41700	Putative ABC transporter
Wilson, ATP7B	7.6×10^{-181}	AT5g44790	ATP-dependent copper transporter
Immunodeficiency, DNA Ligase 1	8.2×10^{-172}	T6D22_10	DNA ligase
Stargardt's, ABCA4	2.8×10^{-168}	At2g41700	Putative ABC transporter
Ataxia telangiectasia, ATM	3.1×10^{-168}	AT3g48190	Ataxia telangiectasia mutated protein AtATM
Niemann–Pick, NPC1	1.2×10^{-166}	F7F22_1	Niemann–Pick C disease protein-like protein
Menkes, ATP7A	1.1×10^{-153}	F2K11_17	ATP-dependent copper transporter, putative
HNPCC*, MLH1	1.5×10^{-150}	AT4g09140	MLH1 protein
Deafness, hereditary, MYO15	2.7×10^{-150}	At2g31900	Putative unconventional myosin
Fam, cardiac myopathy, MYH7	6.5×10^{-147}	T1G11_14	Putative myosin heavy chain
Xeroderma Pigmentosum, F-XPF	1.4×10^{-146}	AT5g41150	Repair endonuclease (gb|AAF01274.1)
G6PD deficiency, G6PD	7.6×10^{-137}	AT5g40760	Glucose-6-phosphate dehydrogenase
Cystic fibrosis, ABCC7	2.3×10^{-135}	AT3g62700	ABC transporter-like protein
Glycerol kinase defic, GK	7.9×10^{-135}	T21F11_21	Putative glycerol kinase
HNPCC, MSH3	6.6×10^{-134}	AT4g25540	Putative DNA mismatch repair protein
HNPCC, PMS2	5.1×10^{-128}	AT4g02460	No title
Zellweger, PEX1	4.1×10^{-125}	AT5g08470	Putative protein
HNPCC, MSH6	9.6×10^{-122}	AT4g02070	G/T DNA mismatch repair enzyme
Bloom, BLM	4.4×10^{-109}	T19D16_15	DNA helicase isoiog
Finnish amyloidosis, GSN	2.2×10^{-107}	AT5g57320	Villin
Chediak–Higashi, CHS1	5.8×10^{-99}	F10O3_11	Putative transport protein
Xeroderma Pigmentosum, G-XPG	7.1×10^{-89}	AT3g28030	Hypothetical protein
Bare lymphocyte, ABCB3	1.3×10^{-84}	AT5g39040	ABC transporter-like protein
Citrullinemia, type I, ASS	3.2×10^{-83}	AT4g24830	Argininosuccinate synthase-like protein
Coffin–Lowry, RPS6KA3	5.2×10^{-81}	AT3g08720	Putative ribosomal-protein S6 kinase (ATPK19)
Keratoderma, KRT9	8.5×10^{-81}	AT3g17050	Unknown protein
Myotonic dystrophy, DM1	1.4×10^{-76}	At2g20470	Putative protein kinase
Bartter's, SLC12A1	1.6×10^{-75}	F26G16_9	Cation-chloride co-transporter, putative
Dents, CLCN5	3.3×10^{-74}	AT5g26240	CLC-d chloride channel protein
Diaphanous 1, DAPH1	1.9×10^{-73}	68069_m00158	Hypothetical protein
AKT2	6.9×10^{-72}	AT3g08730	Putative ribosomal-protein S6 kinase (ATPK6)

in organelles was predicted using default settings on TargetP (Table 1). Many nuclear gene products that are targeted to either (or both) organelles were originally encoded in the organelle genomes and were transferred to the nuclear genome during evolutionary history. A large number also appear to be of eukaryotic origin, with functions such as protein import components, which were probably not required by the free-living ancestors of the endosymbionts.

To identify nuclear genes of possible organellar ancestry, we compared all predicted *Arabidopsis* proteins to all proteins from completed genomes including those from plastids and mitochondria (Supplementary Information Table 2). This search identified proteins encoded by the *Arabidopsis* nuclear genome that are most similar to proteins encoded by other species' organelle genomes (14 mitochondrial and 44 plastid). These represent organelle-to-nuclear gene transfers that have occurred sometime after the divergence of the organelle-containing lineages[47]. There is a great excess of nuclear encoded proteins most similar to proteins from the cyanobacteria *Synechocystis* (Supplementary Information Fig. 4; 806 *Arabidopsis* predicted proteins matching 404 different *Synechocystis* proteins, providing further evidence of a genome duplication). These 806 *Arabidopsis* predicted proteins, and many others of greatly diverse function, are possibly of plastid descent. Through searches against proteins from other cyanobacteria (with incompletely sequenced genomes), we identified 69 additional genes of possibly plastid descent. Only 25% of these putatively plastid-derived proteins displayed a target peptide predicted by TargetP, indicating potential cytoplasmic functions for most of these genes.

The difference between predicted plastid-targeted and predicted plastid-derived genes indicates that there is a probable overestimation by *ab initio* targeting prediction methods and a lack of resolution with respect to destination organelles, the possible extensive divergence of some endosymbiont-derived genes in the nuclear genome, the co-opting of nuclear genes for targeting to organelles, and cytoplasmic functions for cyanobacteria-derived proteins. Clearly more refined tools and extensive experimentation is required to catalogue plastid proteins.

The transfer of genes between genomes still continues (Supplementary Information Table 3). Plastid DNA insertions in the nucleus (17 insertions totalling 11 kb) contain full-length genes encoding proteins or tRNAs, fragments of genes and an intron as well as intergenic regions. Subsequent reshuffling in the nucleus is illustrated by the *atpH* gene, which was originally transferred completely, but is now in two pieces separated by 2 kb. The 13 small mitochondrial DNA insertions total 7 kb in addition to the large insertion close to the centromere of chromosome 2 (ref. 3). The high level of recombination in the mitochondrial genome may account for these events.

Transposable elements

Transposons, which were originally identified in maize by Barbara McClintock, have been found in all eukaryotes and prokaryotes. A subset of transposons replicate through an RNA intermediate (class I), whereas others move directly through a DNA form (class II). Transposons are further classified by similarity either between their mobility genes or between their terminal and/or internal motifs, as well as by the size and sequence of their target site. Internally deleted elements can often be mobilized in *trans* by fully functional elements.

Transposons in *Arabidopsis* account for at least 10% of the genome, or about one-fifth of the intergenic DNA. The *Arabidopsis* genome has a wealth of class I (2,109) and II (2,203) elements, including several new groups (1,209 elements; Supplementary Information Table 4). Mobile histories for many elements were obtained by identifying regions of the genome with significant similarity to 'empty' target sites (RESites) thus providing high-resolution information concerning the termini and target site duplications[48,49]. These regions were readily detected because of the propensity of transposons to integrate into repeats and because of duplications in the genome sequence. In several cases, genes appear to have been included as 'passengers' in transposable units[48]. In some cases, shared sequence similarity, coding capacity and RESites attest to recent activity of transposable elements in the *Arabidopsis* genome. Only about 4% of the complete elements identified correspond to an EST, however, suggesting that most are not transcribed.

Transposable elements found in many other plant genomes are well represented in *Arabidopsis*, including *copia*- and *gypsy*-like long terminal repeat (LTR) retrotransposons, long interspersal nuclear elements (LINEs); short interspersed nuclear elements (SINEs), *hobo*/*Activator*/Tam3 (*hAT*)-like elements, CACTA-like elements and miniature inverted-repeat transposable elements (MITES). Although usually small in size, some larger *Tourist*-like MITEs contain open reading frames (ORFs) with similarity to the transposases of bacterial insertion sequences[48]. *Basho* and many *Mutator*-like elements (MULEs), first discovered in the *Arabidopsis* sequence, represent structurally unique transposons[48–50]. *Basho* elements have a target site preference for mononucleotide 'A' and wide distribution among plants[48,51]. MULEs exhibit a high level of sequence diversity and members of most groups lack long terminal inverted repeats (TIRs). Phylogenetic analysis of the *Arabidopsis* MURA-like transposases suggests that TIR-containing MULEs are more closely related to one another than to MULEs lacking TIRs[49,52].

For many plants with large genomes, class I retrotransposons contribute most of the nucleotide content[53]. In the small *Arabidopsis* genome, class I elements are less abundant and primarily occupy the centromere. In contrast, *Basho* elements and class II transposons such as MITEs and MULEs predominate on the periphery of pericentromeric domains (Fig. 5). In class II transposons, MULEs and CACTA elements are clustered near centromeres and heterochromatic knobs, whereas MITEs and *hAT* elements have a less pronounced bias. The distribution pattern of transposable elements observed in *Arabidopsis* may reflect different types of pericentromeric heterochromatin regions and may be similar to those found in animals.

Numerous centromeric satellite repeats are located between each chromosome arm and have not yet been sequenced, but are represented in part by unanchored BAC contigs (R. Martienssen and M. Marra, unpublished data). End sequence suggests that these domains contain many more class I than class II elements, consistent with the distribution reported here (K. Lemcke and R. Martienssen, unpublished data). We do not know the significance of the apparent paucity of elements in telomeric regions and in the region flanking the rDNA repeats on chromosome 4 (but not on chromosome 2).

Overall, transposon-rich regions are relatively gene-poor and have lower rates of recombination and EST matches, indicating a correlation between low gene expression, high transposon density and low recombination[51]. The role of transposons in genome

Table 4 General features of genes encoded by the three genomes in *Arabidopsis*

	Nucleus/cytoplasm	Plastid	Mitochondria
Genome size	125 Mb	154 kb	367 kb
Genome equivalent/cell	2	560	26
Duplication	60%	17%	10%
Number of protein genes	25,498	79	58
Gene order	Variable, but syntenic	Conserved	Variable
Density (kb per protein gene)	4.5	1.2	6.25
Average coding length	1,900 nt	900 nt	860 nt
Genes with introns	79%	18.4%	12%
Genes/pseudogenes	1/0.03	1/0	1/0.2–0.5
Transposons (% of total genome size)	14%	0%	4%

organization and chromosome structure can now be addressed in a model organism known to undergo DNA methylation and other forms of chromatin modification thought to regulate transposition[52].

rDNA, telomeres and centromeres

Nucleolar organizers (NORs) contain arrays of unit repeats encoding the 18S, 5.8S and 25S ribosomal RNA genes and are transcribed by RNA polymerase I. Together with 5S RNA, which is transcribed by RNA polymerase III, these rRNAs form the structural and catalytic cores of cytoplasmic ribosomes. In *Arabidopsis*, the NORs juxtapose the telomeres of chromosomes 2 and 4, and comprise uninterrupted 18S, 5.8S and 25S units all orientated on the chromosomes in the same direction[54]. In contrast, the 5S rRNA genes are localized to heterogeneous arrays in the centromeric regions of chromosomes 3, 4 and 5 (ref. 55; and Fig. 6). Both NORs are roughly 3.5–4.0 megabase-pairs and comprise ~350–400 highly methylated rRNA gene units, each ~10 kb (ref. 54). The sequence between the euchromatic arms and NORs has been determined. Elsewhere in the genome, only one other 18S, 5.8S, 25S rRNA gene unit was identified in centromere 3. Although minor variations in sequence length and composition occur in the NOR repeats, these variants are highly clustered, supporting a model of sequence maintenance through concerted evolution[55].

Arabidopsis telomeres are composed of CCCTAAA repeats and average ~2–3 kb (ref. 56). For TEL4N (telomere 4 North), consensus repeats are adjacent to the NOR; the remaining telomeres are typically separated from coding sequences by repetitive subtelomeric regions measuring less than 4 kb. Imperfect telomere-like arrays of up to 24 kb are found elsewhere in the genome, particularly

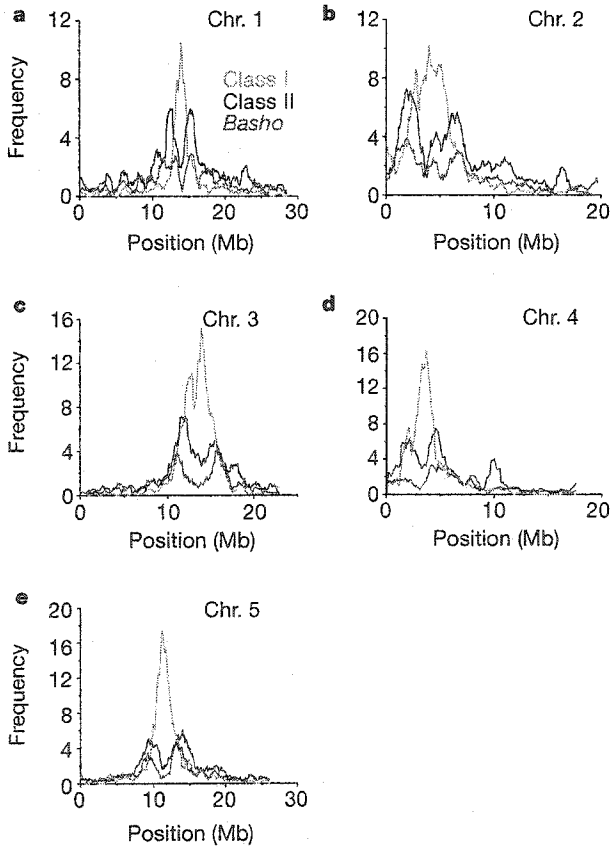


Figure 5 Distribution of class I, II and *Basho* transposons in *Arabidopsis* chromosomes. The frequency of class I retroelements (green), class II DNA transposons (blue) and *Basho* elements (purple) are shown at 100-kb intervals along the five chromosomes (**a**–**e**) of *Arabidopsis*.

near centromeres. These arrays might affect the expression of nearby genes and may have resulted from ancient rearrangements, such as inversions of the chromosome arms.

Centromere DNA mediates chromosome attachment to the meiotic and mitotic spindles and often forms dense heterochromatin. Genetic mapping of the regions that confer centromere function provided the markers necessary to precisely place BAC clones at individual centromeres[17]; 69 clones were targeted for sequencing, resulting in over 5 Mb of DNA sequence from the centromeric regions. The unsequenced regions of centromeres are composed primarily of long, homogeneous arrays that were characterized previously with physical[57] and genetic mapping[17] and contain over 3 Mb of repetitive arrays, including the 180-bp repeats and 5S rDNA[51] (Fig. 6).

Arabidopsis centromeres, like those of many higher eukaryotes, contain numerous repetitive elements including retroelements, transposons, microsatellites and middle repetitive DNA[17]. These repeats are rare in the euchromatic arms and often most abundant in pericentromeric DNA. The repeats, affinity for DNA-binding dyes, dense methylation patterns and inhibition of homologous recombination indicate that the centromeric regions are highly heterochromatic, and such regions are generally viewed as very poor environments for gene expression. Unexpectedly, we found at least 47 expressed genes encoded in the genetically defined centromeres of *Arabidopsis* (http://preuss.bsd.uchicago.edu/arabidopsis.genome.html). In several cases, these genes reside on islands of unique sequence flanked by repetitive arrays, such as 180-bp or 5S rDNA repeats. Among the genes encoded in the centromeres are members of 11 of the 16 functional categories that comprise the proteome. The centromeres are not subject to recombination; consequently, genes residing in these regions probably exhibit unique patterns of molecular evolution.

The function of higher eukaryotic centromeres may be specified by proteins that bind to centromere DNA, by epigenetic modifications, or by secondary or higher order structures. A pairwise comparison of the non-repetitive portions of all five centromeres showed they share limited (1–7%) sequence similarity. Forty-one families of small, conserved centromere sequences (AtCCS, see http://preuss.bsd.uchicago.edu/arabidopsis.genome.html) are enriched in the centromeric and pericentromeric regions and differ from sequences found in the centromeres of other eukaryotes. Molecular and genetic assays will be required to determine whether these conserved motifs nucleate *Arabidopsis* centromere activity. Apart from the AtCCS sequences, most centromere DNA is not shared between chromosomes, complicating efforts to derive clear evolutionary relationships. In contrast, genetic and cytological assays indicate that homologous centromeres are highly conserved among *Arabidopsis* accessions, albeit subject to rearrangements such as inversions to form knobs[5,58,59] and insertions[4]. Further investigation of centromere DNA promises to yield information on the evolutionary forces that act in regions of limited recombination, as well as an improved understanding of the role of DNA sequence patterns in chromosome segregation.

Membrane transport

Transporters in the plasma and intracellular membranes of *Arabidopsis* are responsible for the acquisition, redistribution and compartmentalization of organic nutrients and inorganic ions, as well as for the efflux of toxic compounds and metabolic end products, energy and signal transduction, and turgor generation. Previous genomic analyses of membrane transport systems in *S. cerevisiae* and *C. elegans* led to the identification of over 100 distinct families of membrane transporters[60,61]. We compared membrane transport processes between *Arabidopsis*, animals, fungi and prokaryotes, and identified over 600 predicted membrane transport systems in *Arabidopsis* (http://www-biology.ucsd.edu/~ipaulsen/transport/), a similar number to that of *C. elegans*

(~700 transporters) and over twofold greater than either *S. cerevisiae* or *E. coli* (~300 transporters).

We compared the transporter complement of *Arabidopsis*, *C. elegans* and *S. cerevisiae* in terms of energy coupling mechanisms (Fig. 7a). Unlike animals, which use a sodium ion P-type ATPase pump to generate an electrochemical gradient across the plasma membrane, plants and fungi use a proton P-type ATPase pump to form a large membrane potential (−250 mV)[62]. Consequently, plant secondary transporters are typically coupled to protons rather than to sodium[63]. Compared with *C. elegans, Arabidopsis* has a surprisingly high percentage of primary ATP-dependent transporters (12% and 21% of transporters, respectively), reflecting increased numbers of P-type ATPases involved in metal ion transport and ABC ATPases proposed to be involved in sequestering unusual metabolites and drugs in the vacuole or in other intracellular compartments. These processes may be necessary for pathogen defence and nutrient storage.

About 15% of the transporters in *Arabidopsis* are channel proteins, five times more than in any single-celled organism but half the number in *C. elegans* (Fig. 7b). Almost half of the *Arabidopsis* channel proteins are aquaporins, and *Arabidopsis* has 10-fold more Mfamily major intrinsic protein (MIP) family water channels than any other sequenced organism. This abundance emphasizes the importance of hydraulics in a wide range of plant processes, including sugar and nutrient transport into and out of the vasculature, opening of stomatal apertures, cell elongation and epinastic movements of leaves and stems. Although *Arabidopsis* has a diverse range of metal cation transporters, *C. elegans* has more, many of which function in cell–cell signalling and nerve signal transduction. *Arabidopsis* also possesses transporters for inorganic anions such as phosphate, sulphate, nitrate and chloride, as well as for metal cation channels that serve in signal transduction or cell homeostasis. Compared with other sequenced organisms, *Arabidopsis* has 10-fold more predicted peptide transporters, primarily of the proton-dependent oligopeptide transport (POT) family, emphasizing the importance of peptide transport or indicating that there is broader substrate specificity than previously realized. There are nearly 1,000 *Arabidopsis* genes encoding Ser/Thr protein kinases, suggesting that peptides may have an important role in plant signalling[64].

Virtually no transporters for carboxylates, such as lactate and pyruvate, were identified in the *Arabidopsis* genome. About 12% of the transporters were predicted to be sugar transporters, mostly consisting of paralogues of the MFS family of hexose transporters. Notably, *S. cerevisiae*, *C. elegans* and most prokaryotes use APC family transporters as their principle means of amino-acid transport, but *Arabidopsis* appears to rely primarily on the AAAP family of amino-acid and auxin transporters. More than 10% of the transporters in *Arabidopsis* are homologous to drug efflux pumps; these probably represent transporters involved in the sequestration into vacuoles of xenobiotics, secondary metabolites, and breakdown products of chlorophyll.

Surprisingly, *Arabidopsis* has close homologues of the human ABC TAP transporters of antigenic peptides for presentation to the major histocompatability complex (MHC). In *Arabidopsis*, these transporters may be involved in peptide efflux, or more speculatively, in some form of cell-recognition response. *Arabidopsis* also has 10-fold more members of the multi-drug and toxin extrusion (MATE) family than any other sequenced organism; in bacteria, these transporters function as drug efflux pumps. Curiously, *Arabidopsis* has several homologues of the *Drosophila* RND transporter family Patched protein, which functions in segment polarity, and more than ten homologues of the *Drosophila* ABC family eye pigment transporters. In plants, these are presumably involved in intracellular sequestration of secondary metabolites.

DNA repair and recombination

DNA repair and recombination pathways have many functions in different species such as maintaining genomic integrity, regulating mutation rates, chromosome segregation and recombination, genetic exchange within and between populations, and immune system development. Comparing the *Arabidopsis* genome with other species[65] indicates that *Arabidopsis* has a similar set of DNA repair and recombination (RAR) genes to most other eukaryotes. The pathways represented include photoreactivation, DNA ligation, non-homologous end joining, base excision repair, mismatch excision repair, nucleotide excision repair and many aspects of DNA recombination (Supplementary Information Table 5). The *Arabidopsis* RAR genes include homologues of many DNA repair genes that are defective in different human diseases (for example, hereditary breast cancer and non-polyposis colon cancer, xeroderma pigmentosum and Cockayne's syndrome).

One feature that sets *Arabidopsis* apart from other eukaryotes is the presence of additional homologues of many RAR genes. This is seen for almost every major class of DNA repair, including recombination (four RecA), DNA ligation (four DNA ligase I), photoreactivation (one class II photolyase and five class I photolyase homologues) and nucleotide excision repair (six RPA1, two RPA2, two Rad25, three TFB1 and four Rad23). This is most striking for genes with probable roles in base excision repair. *Arabidopsis* encodes 16 homologues of DNA base glycosylases (enzymes that

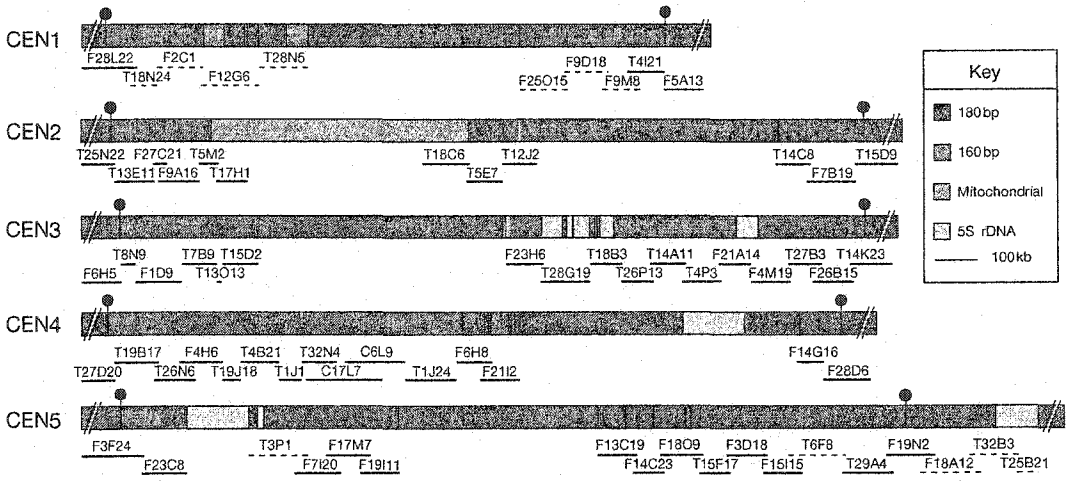


Figure 6 Predicted centromere composition. Genetically defined centromere boundaries are indicated by filled circles; fully and partially assembled BAC sequences are represented by solid and dashed black lines, respectively. Estimates of repeat sizes within the centromeres were derived from consideration of repeat copy number, physical mapping and cytogenetic assays.

recognize abnormal DNA bases and cleave them from the sugar-phosphate backbone)—more than any other species known. This includes several homologues of each of three families of alkylation damage base glycosylases: two of the *S. cerevisiae* MPG; six of the *E. coli* TagI; and two of the *E. coli* AlkA. *Arabidopsis* also encodes three homologues of the apurinic-apyrimidinic (AP) endonuclease Xth. AP endonucleases continue the base excision repair started by glycosylases by cleaving the DNA backbone at abasic sites.

Evolutionary analysis indicates that some of the extra copies of RAR genes in *Arabidopsis* originated through relatively recent gene duplications—because many of the sets of genes are more closely related to each other than to their homologues in any other species. As duplication is frequently accompanied by functional divergence, the duplicate (paralogous) genes may have different repair specificities or may have evolved functions that are outside RAR functions (as is the case for two of the five class I photolyase homologues, which function as blue-light receptors). In most cases, it is not known whether the paralogous gene copies have different functions. The presence of multiple paralogues might also allow functional redundancy or a greater repair or recombination capacity.

The multiplicity of RAR genes in *Arabidopsis* is also partly due to the transfer of genes from the organellar genomes to the nucleus. Repair gene homologues that appear to be of chloroplast origin (Supplementary Information Tables 2 and 5) include the recombination proteins RecA, RecG and SMS, two class I photolyase homologues, Fpg, two MutS2 proteins, and the transcription-repair coupling factor Mfd. Two of these (RecA and Fpg) are involved in RAR functions in the plastid, suggesting that the others may be as well. The finding of an Mfd orthologue of cyanobacterial descent is surprising. In *E. coli*, Mfd couples nucleotide excision repair carried out by UvrABC to transcription, leading to the rapid repair of DNA damage on the transcribed strand of transcribed genes[66] The absence of orthologues of UvrABC in *Arabidopsis* renders the function of Mfd difficult to predict. The presence of Mfd but not UvrABC has been reported for only one other species, a bacterial endosymbiont of the pea aphid.

Other nuclear-encoded *Arabidopsis* DNA repair gene homologues are evolutionarily related to genes from α-Proteobacteria, and thus may be of mitochondrial descent. In particular, the six homologues of the alkyl-base glycosylase TagI appear to be the result of a large expansion in plants after transfer from the mitochondrial genome. Whether any of these TagI homologues function in the repair and maintenance of mitochondrial DNA has not been determined. More detailed phylogenetic analysis may reveal additional *Arabidopsis* RAR genes to be of organellar ancestry.

There are some notable absences of proteins important for RAR in other species, including alkyltransferases, MSH4, RPA3 and many components of TFIIH (TFB2, TFB3, TFB4, CCL1, Kin28). Nevertheless, *Arabidopsis* shows many similarities to the set of DNA repair genes found in other eukaryotes, and therefore offers an experimental system for determining the functions of many of these proteins, in part through characterization of mutants defective in DNA repair[67].

Gene regulation

Eukaryotic gene expression involves many nuclear proteins that modulate chromatin structure, contribute to the basal transcription machinery, or mediate gene regulation in response to developmental, environmental or metabolic cues. As predicted by sequence similarity, more than 3,000 such proteins may be encoded by the *Arabidopsis* genome, suggesting that it has a comparable complexity of gene regulation to other eukaryotes. *Arabidopsis* has an additional level of gene regulation, however, with DNA methylation potentially mediating gene silencing and parental imprinting.

Plants have evolved several variations on chromatin remodelling proteins, such as the family of HD2 histone deacetylases[68]. Although *Arabidopsis* possesses the usual number of SNF2-type chromatin remodelling ATPases, which regulate the expression of nearly all genes, there are significant structural differences between yeast and metazoan SNF2-type genes and their orthologues in *Arabidopsis*. DDM1, a member of the SNF2 superfamily, and MOM1, a gene with similarity to the SNF2 family, are involved in transcriptional gene silencing in *Arabidopsis*. MOM1 has no clear orthologue in fungal or metazoan genomes.

Consistent with its methylated DNA, *Arabidopsis* possesses eight DNA methyltransferases (DMTs). Two of the three types are orthologous to mammalian DMT[69] whereas one, chromomethyltransferase[70], is unique to plants. No DMTs are found in yeast or *C. elegans*, although two DMT-like genes are found in *Drosophila*[71]. *Arabidopsis* also encodes eight proteins with methyl-DNA-binding domains (MBDs). Despite lacking methylated DNA, *Drosophila* encodes four MBD proteins and *C. elegans* has two. These differences in chromatin components are likely to reflect important differences in chromatin-based regulatory control of gene expression in eukaryotes (Supplementary Information Table 6; http://Ag.Arizona.Edu/chromatin/chromatin.html).

The *Arabidopsis* genome encodes transcription machinery for the three nuclear DNA-dependent RNA polymerase systems typical of eukaryotes (Supplementary Information Table 6). Transcription by RNA polymerases II and III appears to involve the same machinery as is used in other eukarotes; however, most transcription factors for RNA polymerase I are not readily identified. Only two polymerase I regulators (other than polymerase subunits and TATA-binding protein) are apparent in *Arabidopsis*, namely homologues of yeast RRN3 and mouse TTF-1. All eukaryotes examined to date have distinct genes for the largest and second largest subunits of polymerase I, II and III. Unexpectedly, *Arabidopsis* has two genes encoding a fourth class of largest subunit and second-largest subunit (Supplementary Information Fig. 5). It will be interesting to determine whether the atypical subunits comprise a polymerase that has a plant-specific function. Four genes encoding single-subunit plastid or mitochondrial RNA polymerases have been identified in *Arabidopsis* (Supplementary Information Table 6). Genes for the bacterial β-, β′- and α-subunits of RNA polymerase are also present, as are homologues of various σ-factors, and these proteins may regulate chloroplast gene expression. Mutations in the *Sde-1* gene, encoding RNA-dependent RNA polymerase (RdRp), lead to defective post-transcriptional gene silencing[72]. We also identified five more closely related RdRp genes.

Our analysis, using both similarity searches and domain matches, has identified 1,709 proteins with significant similarity to known classes of plant transcription factors classified by conserved DNA-binding domains. This analysis used a consistent conservative threshold that probably underestimates the size of families of diverse sequence. This class of protein is the least conserved among all classes of known proteins, showing only 8–23% similarity to transcription factors in other eukaryotes (Fig. 2b). This reduced similarity is due to the absence of certain classes of transcription factors in *Arabidopsis* and large numbers of plant-specific transcription factors. We did not detect any members of several widespread families of transcription factors, such as the REL (Rel-like DNA-binding domain) homology region proteins, nuclear steroid receptors and forkhead-winged helix and POU (Pit-1, Oct- and Unc-8b) domain families of developmental regulators. Conversely, of 29 classes of *Arabidopsis* transcription factors, 16 appear to be unique to plants (Supplementary Information Table 6). Several of these, such as the AP2/EREBP-RAV, NAC and ARF-AUX/IAA families, contain unique DNA-binding domains, whereas others contain plant-specific variants of more widespread domains, such as the DOF and WRKY zinc-finger families and the two-repeat MYB family.

Functional redundancy among members of large families of closely related transcription factors in *Arabidopsis* is a significant potential barrier to their characterization[73]. For example, in the

SHATTERPROOF and SEPALLATA families of MADS box transcription factors, all genes must be defective to produce visible mutant phenotypes[74,75]. These functionally redundant genes are found on the segmental duplications described above. Our analyses, together with the significant sequence similarity found in large families of transcription factors such as the R2R3-repeat MYB and WRKY families, suggest that strategies involving overexpression will be important in determining the functions of members of transcription factor families.

Arabidopsis has two or over three times more transcription factors than identified in *Drosophila*[29] or *C.elegans*[1], respectively. The significantly greater extent of segmental chromosomal and local tandem duplications in the *Arabidopsis* genome generates larger gene families, including transcription factors. The partly overlapping functions defined for a few transcription factors are also likely to be much more widespread, implicating many sequence-related transcription factors in the same cellular processes. Finally, the expanded number of genes involved in metabolism, defence and environmental interaction in *Arabidopsis* (Fig. 2a), which have few counterparts in *Drosophila* and *C. elegans*, all require additional numbers and classes of transcription factors to integrate gene function in response to a vast range of developmental and environmental cues.

Cellular organization

Plant cells differ from animal cells in many features such as plastids, vacuoles, Golgi organization, cytoskeletal arrays, plasmodesmata linking cytoplasms of neighbouring cells, and a rigid polysaccharide-rich extracellular matrix—the cell wall. Because the cell wall maintains the position of a cell relative to its neighbours, both changes in cell shape and organized cell divisions, involving cytoskeleton reorganization and membrane vesicle targeting, have major roles in plant development. Plant cytokinesis is also unique in that the partitioning membrane is formed *de novo* by vesicle fusion. We compared the *Arabidopsis* genome with those of *C. elegans*, *Drosophila* and yeast to glimpse the genetic basis of plant-cell-specific features.

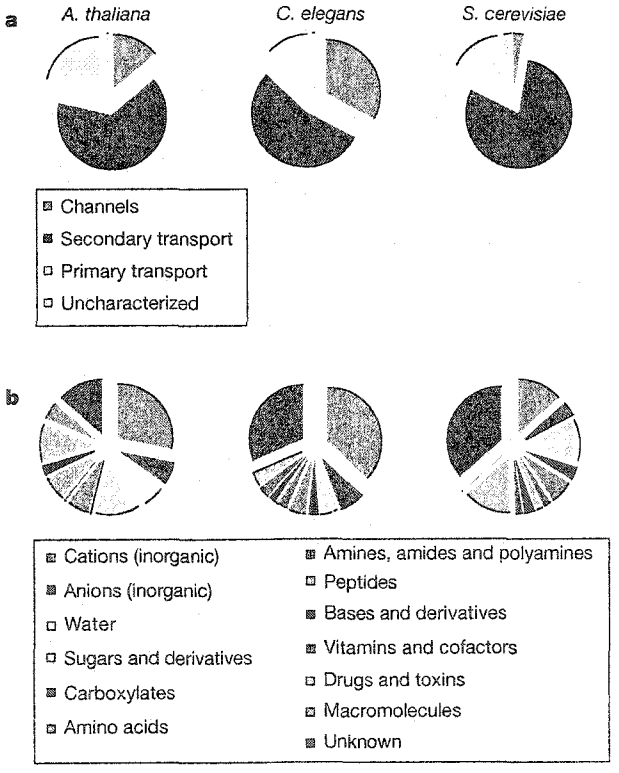


Figure 7 Comparison of the transport capabilities of *Arabidopsis*, *C. elegans* and *S. cerevisiae*. Pie charts show the percentage of transporters in each organism according to bioenergetics (**a**) and substrate specificity (**b**).

The principal components of the plant cytoskeleton are microtubules (MTs) and actin filaments (AFs); intermediate filaments (IFs) have not been described in plants. *Arabidopsis* appears to lack genes for cytokeratin or vimentin, the main components of animal IFs, but has several variants of actin, α- and β-tubulin. The *Arabidopsis* genome also encodes homologues of chaperones that mediate the folding of tubulin and actin polypeptides in yeast and animal cells, such as the prefoldin and cytosolic chaperonin complexes and tubulin-folding cofactors. The dynamic stability of MTs and AFs is influenced by MT-associated proteins and actin-binding proteins, respectively, several of which are encoded by *Arabidopsis* genes. These include the MT-severing ATPase katanin, AF-cross-linking/bundling proteins, such as fimbrins and villins, and AF-disassembling proteins, such as profilin and actin-depolymerizing factor/cofilin. The *Arabidopsis* proteome appears to lack homologues of proteins that, in animal cells, link the actin cytoskeleton across the plasma membrane to the extracellular matrix, such as integrin, talin, spectrin, α-actinin, vitronectin or vinculin. This apparent lack of 'anchorage' proteins is consistent with the different composition of the cell wall and with a prominence of cortical MTs at the expense of cortical AFs in plant cells.

Plant-specific cytoskeletal arrays include interphase cortical MTs mediating cell shape, the preprophase band marking the cortical site of cell division, and the phragmoplast assisting in cytokinesis[76]. Although plant cells lack structural counterparts of the yeast spindle pole body and the animal centrosome, *Arabidopsis* has homologues of core components of the MT-nucleating γ-tubulin ring complex, such as γ-tubulin, Spc97/hGCP2 and Spc98/hGCP3. *Arabidopsis* has numerous motor molecules, both kinesins and dyneins with associated dynactin complex proteins, which are presumably involved in the dynamic organization of MTs and in transporting cargo along MT tracks. There are also myosin motors that may be involved in AF-supported organelle trafficking. Essential features of the eukaryotic cytoskeleton appear to be conserved in *Arabidopsis*.

The *Arabidopsis* genome encodes homologues of proteins involved in vesicle budding, including several ARFs and ARF-related small G-proteins, large but not small ARF GEFs (adenosine ribosylation factor on guanine nucleotide exchange factor), adapter proteins, and coat proteins of the COP and non-COP types. *Arabidopsis* also has homologues of proteins involved in vesicle docking and fusion, including SNAP receptors (SNAREs), *N*-ethylmaleimide-sensitive factor (NSF) and Cdc48-related ATPases, accessory proteins such as Sec1 and soluble NSF attachment protein (SNAP), and Rab-type GTPases. The large number of *Arabidopsis* SNAREs can be grouped by sequence similarity to yeast and animal counterparts involved in specific trafficking pathways, and some have been localized to the trans-Golgi and the pre-vacuolar pathway[77]. *Arabidopsis* also has a receptor for retention of proteins in the endoplasmic reticulum, a cargo receptor for transport to the vacuole and several phragmoplastins related to animal dynamin GTPases. Thus, plant cells appear to use the same basic machinery for vesicle trafficking as yeast and animal cells.

Animal cells possess many functionally diverse small G-proteins of the Ras superfamily involved in signal transduction, AF reorganization, vesicle fusion and other processes. Surprisingly, *Arabidopsis* appears to lack genes for G-proteins of the Ras, Rho, Rac and Cdc42 subfamilies but has many Rab-type G-proteins involved in vesicle fusion and several Rop-type G-proteins, one of which has a role in actin organization of the tip-growing pollen tube[78]. The significance of this divergent amplification of different subfamilies of small G-proteins in plants and animals remains to be determined.

Arabidopsis possesses cyclin-dependent kinases (CDKs), including a plant-specific Cdc2b kinase expressed in a cell-cycle-dependent manner, several cyclin subtypes, including a D-type cyclin that

mediates cytokinin-stimulated cell-cycle progression[79], a retinoblastoma-related protein and components of the ubiquitin-dependent proteolytic pathway of cyclin degradation. In yeast and animal cells, chromosome condensation is mediated by condensins, sister chromatids are held together by cohesins such as Scc1, and metaphase–anaphase transition is triggered by separin/Esp1 endopeptidase proteolysis of Scc1 on APC-mediated degradation of its inhibitor, securin/Psd1. Related proteins are encoded by the *Arabidopsis* genome. Thus, the basic machinery of cell-cycle progression, genome duplication and segregation appears to be conserved in plants. By contrast, entry into M phase, M-phase progression and cytokinesis seem to be modified in plant cells. *Arabidopsis* does not appear to have homologues of Cdc25 phosphatase, which activates Cdc2 kinase at the onset of mitosis, or of polo kinase, which regulates M-phase progression in yeast and animals. Conversely, plant-specific mitogen-actived protein (MAP) kinases appear to be involved in cytokinesis.

Cytokinesis partitions the cytoplasm of the dividing cell. Yeast and animal cells expand the membrane from the surface towards the centre in a cleavage process supported by septins and a contractile ring of actin and type II myosin. By contrast, plant cytokinesis starts in the centre of the division plane and progresses laterally. A transient membrane compartment, the cell plate, is formed *de novo* by fusion of Golgi-derived vesicles trafficking along the phragmoplast MTs[80]. Consistent with the unique mode of plant cytokinesis, *Arabidopsis* appears to lack genes for septins and type II myosin. Conversely, cell-plate formation requires a cytokinesis-specific syntaxin that has no close homologue in yeast and animals. Although syntaxin-mediated membrane fusion occurs in animal cytokinesis and cellularization, the vesicles are delivered to the base of the cleavage furrow. Thus, the plant-specific mechanism of cell division is linked to conserved eukaryotic cell-cycle machinery.

Two main conclusions are suggested by this comparative analysis. First, *Arabidopsis* and eukaryotic cells have common features related to intracellular activities, such as vesicle trafficking, cytoskeleton and cell cycle. Second, evolutionarily divergent features, such as organization of the cytoskeleton and cytokinesis, appear to relate to the plant cell wall.

Development

The regulation of development in *Arabidopsis*, as in animals, involves cell–cell communication, hierarchies of transcription factors, and the regulation of chromatin state; however, there is no reason to suppose that the complex multicellular states of plant and animal development have evolved by elaborating the same general processes during the 1.6 billion years since the last common unicellular ancestor of plants and animals[81,82]. Our genome analyses reflect the long, independent evolution of many processes contributing to development in the two kingdoms.

Plants and animals have converged on similar processes of pattern formation, but have used and expanded different transcription factor families as key causal regulators. For example, segmentation in insects and differentiation along the anterior–posterior and limb axes in mammals both involve the spatially specific activation of a series of homeobox gene family members. The pattern of activation is causal in the later differentiation of body and limb axis regions. In plants the pattern of floral whorls (sepals, petals, stamens, carpels) is also established by the spatially specific activation of members of a family of transcription factors, but in this instance the family is the MADS box family. Plants also have homeobox genes and animals have MADS box genes, implying that each lineage invented separately its mechanism of spatial pattern formation, while converging on actions and interactions of transcription factors as the mechanism. Other examples show even greater divergence of plant and animal developmental control. Examples are the AP2/EREBP and NAC families of transcription factors, which have important roles in flower and meristem development: both families are so far found only in plants (Supplementary Information Table 6).

A similar story can be told for cell–cell communication. Plants do not seem to have receptor tyrosine kinases, but the *Arabidopsis* genome has at least 340 genes for receptor Ser/Thr kinases, belonging to many different families, defined by their putative extracellular domains (Supplementary Information Table 7). Several families have members with known functions in cell–cell communication, such as the CLV1 receptor involved in meristem cell signalling, the S-glycoprotein homologues involved in signalling from pollen to stigma in self-incompatible *Brassica* species, and the BRI1 receptor necessary for brassinosteroid signalling[83]. Animals also have receptor Ser/Thr kinases, such as the transforming growth factor-β (TGF-β) receptors, but these act through SMAD proteins that are absent from *Arabidopsis*. The leucine-rich repeat (LRR) family of *Arabidopsis* receptor kinases shares its extracellular domain with many animal and fungal proteins that do not have associated kinase domains, and there are at least 122 *Arabidopsis* genes that code for LRR proteins without a kinase domain. Other *Arabidopsis* receptor kinase families have extracellular domains that are unfamiliar in animals. Thus, evolution is modular, and the plant and animal lineages have expanded different families of receptor kinases for a similar set of developmental processes.

Several *Arabidopsis* genes of developmental importance appear to be derived from a cyanobacteria-like genome (Supplementary Information Table 2), with no close relationship to any animal or fungal protein. One salient example is the family of ethylene receptors; another gene family of apparent chloroplast origin is the phytochromes—light receptors involved in many developmental decisions (see below). Whereas the land plant phytochromes show clear homology to the cyanobacterial light receptors, which are typical prokaryotic histidine kinases, the plant phytochromes are histidine kinase paralogues with Ser/Thr specificity[84]. Similarly to the ethylene receptors, the proteins that act downstream of plant phytochrome signalling are not found in cyanobacteria, and thus it appears that a bacterial light receptor entered the plant genome through horizontal transfer, altered its enzymatic activity, and became linked to a eukaryotic signal transduction pathway. This infusion of genes from a cyanobacterial endosymbiont shows that plants have a richer heritage of ancestral genes than animals, and unique developmental processes that derive from horizontal gene transfer.

Signal transduction

Being generally sessile organisms, plants have to respond to local environmental conditions by changing their physiology or redirecting their growth. Signals from the environment include light and pathogen attack, temperature, water, nutrients, touch and gravity. In addition to local cellular responses, some stimuli are communicated across the plant body, with plant hormones and peptides acting as secondary messengers. Some hormones, such as auxin, are taken up into the cell, whereas others, such as ethylene and brassinosteroids, and the peptide CLV3, act as ligands for receptor kinases on the plasma membrane. No matter where the signal is perceived by the cell, it is transduced to the nucleus, resulting in altered patterns of gene expression.

Comparative genome analysis between *Arabidopsis*, *C. elegans* and *Drosophila* supports the idea that plants have evolved their own pathways of signal transduction[85]. None of the components of the widely adopted signalling pathways found in vertebrates, flies or worms, such as Wingless/Wnt, Hedgehog, Notch/lin12, JAK/STAT, TGF-β/SMADs, receptor tyrosine kinase/Ras or the nuclear steroid hormone receptors, is found in *Arabidopsis*. By contrast, brassinosteroids are ligands of the BRI1 Ser/Thr kinase, a member of the largest recognizable class of transmembrane sensors encoded by 340 receptor-like kinase (RLK) genes in the *Arabidopsis* genome (Supplementary Information Table 7). With a few notable exceptions, such as CLV1, the types of ligands sensed by RLKs are

completely unknown, providing an enormous future challenge for plant biologists. G-protein-coupled receptors (GPCRs)/ seven-transmembrane proteins are an abundant class of proteins in mammalian genomes, instrumental in signal transduction. INTERPRO detected 27 GPCR-related domains in *Arabidopsis* (Supplementary Information Table 1), although there is no direct experimental evidence for these. *Arabidopsis* contains a family of 18 seven-transmembrane proteins of the mildew resistance (MlO) class, several of which are involved in defence responses. Notably, only single Gα (GPA1) and Gβ (AGB1) subunits are found in *Arabidopsis*, both previously known[86].

Although cyclic GMP has been proposed to be involved in signal transduction in *Arabidopsis*[87], a protein containing a guanylate cyclase domain was not identified in our analyses. Nevertheless, cyclic nucleotide-binding domains were detected in various proteins, indicating that cNMPs may have a role in plant signal transduction. Thus, although cNMP-binding domains appear to have been conserved during evolution, cNMP synthesis in *Arabidopsis* may have evolved independently.

We were unable to identify a protein with significant similarity to known Gγ subunits, but recent biochemical studies suggest that a protein with this functional capacity is likely to be present in plant cells (H. Ma, personal communication). Therefore, there is potential for the formation of only a single heterotrimeric G-protein complex; however, its functional interaction with any of the potential GPCR-related proteins remains to be determined.

Modules of cellular signal pathways from bacteria and animals have been combined and new cascades have been innovated in plants. A pertinent example is the response to the gaseous plant hormone ethylene[88]. Ethylene is perceived and its signal transmitted by a family of receptors related to bacterial-type two-component histidine kinases (HKs). In bacteria, yeast and plants, these proteins sense many extracellular signals and function in a His-to-Asp phosphorelay network[89]. In turn, these proteins physically interact with the genetically downstream protein CTR1, a Raf/MAPKKK-related kinase, revealing the juxtaposition of bacterial-type two-component receptors and animal-type MAP kinase cascades. Unlike animals, however, *Arabidopsis* does not seem to have a Ras protein to activate the MAP kinase cascade. MAP kinases are found in abundance in *Arabidopsis*: we identified ~20, a higher number than in any other eukaryote. As potentially counteracting components, we found ~70 putative PP2C protein phosphatases. Although this group is largely uncharacterized functionally, several members are related to ABI1/ABI2, key negative regulators in the signalling pathway for the plant hormone abscisic acid. Additional components of the His-to-Asp phosphorelay system were also found in *Arabidopsis*, including authentic response regulators (ARRs), pseudoresponse regulators (PRRs) and phosphotransfer intermediate protein (HPt)[90]. We found 11 HKs in the proteome (3 new), 16 RRs (2 new) and 8 PRRs (2 new). The biological roles of most ARRs, PRRs and HPts are largly unknown, but several have been found to have diverse functions in plants, including transcriptional activation in response to the plant hormone cytokinin[91], and as components of the circadian clock[92].

Plants seem to have evolved unique signalling pathways by combining a conserved MAP kinase cascade module with new receptor types. In many cases, however, the ligands are unknown. Conversely, some known signalling molecules, such as auxin, are still in search of a receptor. Auxin signalling may represent yet another plant-specific mode of signalling, with protein degradation through the ubiquitin-proteasome pathway preceding altered gene expression. With many *Arabidopsis* genes encoding components of the ubiquitin-proteasome pathway, elimination of negative regulators may be a more widespread phenomenon in plant signalling.

Recognizing and responding to pathogens

Plants are constantly exposed to pests, parasites and pathogens and have evolved many defences. In mammals, polymorphism for parasite recognition encoded in the MHC genes contributes to resistance. In plants, disease resistance (*R*) genes that confer parasite recognition are also extremely polymorphic. This polymorphism has been proposed to restrict parasites, and its absence may explain the breakdown of resistance in crop monocultures[93]. In contrast to MHC genes, plant resistance genes are found at several loci, and the complete genome sequence enables analysis of their complement and structure. Parasite recognition by resistance genes triggers defence mechanisms through various signalling molecules, such as protein kinases and adapter proteins, ion fluxes, reactive oxygen intermediates and nitric oxide. These halt pathogen colonization through transcriptional activation of defence genes and a form of programmed cell death called the hypersensitive response[94]. The *Arabidopsis* genome contains diverse resistance genes distributed at many loci, along with components of signalling pathways, and many other genes whose role in disease resistance has been inferred from mutant phenotypes.

Most resistance genes encode intracellular proteins with a nucleotide-binding (NB) site typical of small G proteins, and carboxy-terminal LRRs[95]. Their amino termini either carry a TIR domain, or a putative coiled coil (CC). There are 85 TIR–NB–LRR resistance genes at 64 loci, and 36 CC–NB–LRR resistance genes at 30 loci. Some NB–LRR resistance genes express neither obvious TIR nor CC domains at their N termini. This potential class is present seven times, at six loci. There are 15 truncated TIR–NB genes that lack an LRR at 10 loci, often adjacent to full TIR–NB–LRR genes. There are also six CC–NB genes, at five loci. These truncated products may function in resistance. Intriguingly, two TIR–NB–LRR genes carry a WRKY domain, found in transcription factors that are implicated in plant defence, and one of these also encodes a protein kinase domain.

Resistance gene evolution may involve duplication and divergence of linked gene families[36]; however, most (46) resistance genes are singletons; 50 are in pairs, 21 are in 7 clusters of 3 family members, with single clusters of 4, 5, 7, 8 and 9 members, respectively. Of the non-singletons, ~60% of pairs are in direct repeats, and ~40% are in inverted repeats. Resistance genes are unevenly distributed between chromosomes, with 49 on chromosome 1; 2 on chromosome 2; 16 on chromosome 3; 28 on chromosome 4; and 55 on chromosome 5.

In other plant species, resistance genes encode both transmembrane receptors for secreted pathogen products and protein kinases, and some other classes are also found. The *Cf* genes in tomato encode extracellular LRRs with a transmembrane domain and short cytoplasmic domain. Mutation in an *Arabidopsis* homologue, *CLAVATA2*, results in enlarged meristems, but to date no resistance function has been assigned to the 30 *Arabidopsis CLV2* homologues. *CLAVATA1*, a transmembrane LRR kinase, is also required for meristem function. *Xa21*, a rice LRR-kinase, confers *Xanthomonas* resistance, and the *Arabidopsis FLS2* LRR kinase confers recognition of flagellin. It has been proposed that CLV1 and CLV2 function as a heterodimer; perhaps this is also true for Xa21, FLS2 and Cf proteins. There are 174 LRR transmembrane kinases in *Arabidopsis*, with only *FLS2* assigned a role in resistance. A unique resistance gene, beet *Hs1pro-1*, which confers nematode resistance, has two *Arabidopsis* homologues.

The tomato Pto Ser/Thr kinase acts as a resistance protein in conjunction with an NB–LRR protein, so similar kinases might do the same for *Arabidopsis* NB–LRR proteins. There are 860 Ser/Thr kinases in the *Arabidopsis* sequence. Fifteen of these share 50% identity over the Pto-aligned region. The Toll pathway in *Drosophila* and mammals regulates innate immune responses through LRR/TIR domain receptors that recognize bacterial lipopolysaccharides[96]. Pto is highly homologous to *Drosophila* PELLE and mammalian IRAK protein kinases that mediate the TIR pathway.

Additional genes have been defined that are required for resistance by our analysis of the genome sequence. The *ndr1* mutation defines a gene required by the CC–NB–LRR gene *RPS2* and *RPM1*. *NDR1* is 1 of 28 *Arabidopsis* genes that are similar both to each other and to the tobacco *HIN1* gene that is transcriptionally induced early during the hypersensitive response. *EDS1* is a gene required for TIR–NB–LRR function, and like *PAD4*, encodes a protein with a putative lipase motif. *EDS1*, *PAD4* and a third gene comprise the *EDS1/PAD4* family. The *NPR1/NIM1/SAI1* gene is required for systemic acquired resistance, and we found five additional *NPR1* homologues. Recessive mutations at both the barley Mlo and *Arabidopsis LSD1* loci confer broad-spectrum resistance and derepress a cell-death program. There are at least 18 Mlo family members that resemble heterotrimeric GPCRs in *Arabidopsis*, and only two *LSD1* homologues.

One of the earliest responses to pathogen recognition is the production of reactive oxygen intermediates. This involves a specialized respiratory burst oxidase protein that transfers an electron across the plasma membrane to make superoxide. *Arabidopsis* encodes eight apparently functional *gp91* homologues, called *Atrboh* genes. Unlike *gp91*, they all carry an ~300 amino-acid N-terminal extension carrying an EF-hand Ca^{2+}-binding domain. In mammals, activation of the respiratory oxidative burst complex in the neutrophil, which includes gp91, requires the action of Rac proteins. As no Rac or Ras proteins are found in *Arabidopsis*, members of the large rop family of G proteins may carry this out. Similarly, we did not detect any *Arabidopsis* homologues of other mammalian respiratory burst oxidase components (p22, p47, p67, p40).

There are no clear homologues of many mammalian defence and cell-death control genes. Although nitric oxide production is involved in plant defence, there is no obvious homologue of nitric oxide synthase. Also absent are apparent homologues of the REL domain transcription factors involved in innate immunity in both *Drosophila* and mammals. We found no similarity to proteins involved in regulating apoptosis in animal cells, such as classical caspases, bcl2/ced9 and baculovirus p35. There are, however, 36 cysteine proteases. There are also eight homologues of a newly defined metacaspase family[97], two of which, along with LSD1, have a clear GATA-type zinc-finger.

Photomorphogenesis and photosynthesis

Because nearly all plants are sessile and most depend on photosynthesis, they have evolved unique ways of responding to light. Light serves as an energy source, as well as a trigger and modulator of complex developmental pathways, including those regulated by the circadian clock. Light is especially important during seedling emergence, where it stimulates chlorophyll production, leaf development, cotyledon expansion, chloroplast biogenesis and the coordinated induction of many nuclear- and chloroplast-encoded genes, while at the same time inhibiting stem growth. The goal of this process, called photomorphogenesis, is the establishment of a body plan that allows the plant to be an efficient photosynthetic machine under varying light conditions[98]. The signal transduction cascade leading to light-induced responses begins with the activation of photoreceptors. Next, the light signal is transduced via positively and negatively acting nuclear and cytoplasmic proteins, causing activation or derepression of nuclear and chloroplast-encoded photosynthetic genes and enabling the plant to establish optimal photoautotrophic growth. Although genetic and biochemical studies have defined many of the components in this process, the genome sequence provides an opportunity to identify comprehensively *Arabidopsis* genes involved in photomorphogenesis and the establishment of photoautotrophic growth. We identified at least 100 candidate genes involved in light perception and signalling, and 139 nuclear-encoded genes that potentially function in photosynthesis.

The roles have been described of only 35 of the 100 candidate photomorphogenic genes (Supplementary Information Table 8). All of the light photoreceptors had been discovered previously, including five red/far-red absorbing phytochromes (PHYA-E), two blue/ultraviolet-A absorbing cryptochromes (CRY1 and CRY2), one blue-absorbing phototropin (NPH1) and one NPH1-like (or NPL1). In contrast, we uncovered many new proteins similar to the photomorphogenesis regulators COP/DET/FUS, PKS1, PIF3, NDPK2, SPA1, FAR1, GIGANTEA, FIN219, HY5, CCA1, ATHB-2, ZEITLUPE, FKF1, LKP1, NPH3 and RPT2.

Both the phytochromes and NPH1 contain chromophores for light sensing coupled to kinase domains for signal transmission. Phytochromes have an N-terminal chromophore-binding domain, two PAS domains, and a C-terminal Ser/Thr kinase domain[99], whereas NPH1 has two LOV domains (members of the PAS domain superfamily) for flavin mononucleotide binding and a C-terminal Ser/Thr kinase domain[100]. PAS domains potentially sense changes in light, redox potential and oxygen energy levels, as well as mediating protein–protein interactions[99,100]. We searched for uncharacterized proteins with the combination of a kinase domain and either a phytochrome chromophore-binding site or PAS domains. Although we found no new phytochrome-like genes, we did identify four predicted proteins that contain PAS and kinase domains (Supplementary Information Fig. 6). These proteins share 80% amino-acid identity, but, unlike NPH1 and NPL1, have only one PAS domain. The combination of potential signal sensing and transmitting domains makes it tempting to speculate that these proteins may be receptors for light or other signals.

Our screen included searches for components of photosynthetic reaction centres and light-harvesting complexes, enzymes involved in CO_2 fixation and enzymes in pigment biosynthesis. We identified 11 core proteins of photosystem I, including the eukaryotic-specific components PsaG and PsaH[101], and 8 photosystem II proteins, including a single member (psbW) of the photosystem II core. We also found 26 proteins similar to the Chlorophyll-a/b binding proteins (8 Lhca and 18 Lhcb). Of the seven subunits of the cytochrome b_6f complex (PetA–D, PetG, PetL, PetM), only one (*PetC*) was found in the nuclear genome, whereas the remainder are probably encoded in the chloroplast. Similarly, of the nine subunits of the chloroplast ATP synthase complex, three are encoded in the nucleus, including the II- , γ- and δ-subunits; the remaining subunits (I, III, IV, α, β, ϵ) are encoded in the chloroplast[102]. Ten genes were related to the soluble components of the electron transfer chain, including two plastocyanins, five ferredoxins and three ferredoxin/NADP oxidoreductases. Forty genes are predicted to have a role in CO_2 fixation, including all of the enzymes in the Calvin–Benson cycle. For pigment biosynthesis, 16 genes in chlorophyll biosynthesis and 31 genes in carotenoid biosynthesis were found (Supplementary Information Table 8). Our analyses have identified several potential components of the light perception pathway, and have revealed the complex distribution of components of the photosynthetic apparatus between nuclear and plastid genomes.

Metabolism

Arabidopsis is an autotrophic organism that needs only minerals, light, water and air to grow. Consequently, a large proportion of the genome encodes enzymes that support metabolic processes, such as photosynthesis, respiration, intermediary metabolism, mineral acquisition, and the synthesis of lipids, fatty acids, amino acids, nucleotides and cofactors[103]. With respect to these processes, *Arabidopsis* appears to contain a complement of genes similar to those in the photoautotropic cyanobacterium *Synechocystis*[45], but, whereas *Synechocystis* generally has a single gene encoding an enzyme, *Arabidopsis* frequently has many. For example, *Arabidopsis* has at least seven genes for the glycolytic enzyme pyruvate kinase, with an

additional five for pyruvate kinase-like proteins. Whatever the reason for this high level of redundancy, it varies from gene to gene in the same pathway; the 11 enzymes of glycolysis are encoded by up to 51 genes that are present in as few as one or as many as eight copies. Similarly, of the 59 genes encoding proteins involved in glycerolipid metabolism, 39 are represented by more than one gene[104]. Genome duplication and expansion of gene families by tandem duplication have contributed to this diversity.

This high degree of apparent structural redundancy does not necessarily imply functional redundancy. For instance, although there are seven genes for serine hydroxymethyltransferase, a mutation in the gene for the mitochondrial form completely blocks the photorespiratory pathway[105]. Although there are 12 genes for cellulose synthase, mutations in at least 2 of the 12 confer distinct phenotypes because of tissue-specific gene expression[106].

The metabolome of *Arabidopsis* differs from that of cyanobacteria, or of any other organism sequenced to date, by the presence of many genes encoding enzymes for pathways that are unique to vascular plants. In particular, although relatively little is known about the enzymology of cell-wall metabolism, more than 420 genes could be assigned probable roles in pathways responsible for the synthesis and modification of cell-wall polymers. Twelve genes encode cellulose synthase, and 29 other genes encode 6 families of structurally related enzymes thought to synthesize other major polysaccharides[106]. Roughly 52 genes encode polygalacturonases, 20 encode pectate lyases and 79 encode pectin esterases, indicating a massive investment in modifying pectin. Similarly, the presence of 39 β-1,3-glucanases, 20 endoxyloglucan transglycosylases, 50 cellulases and other hydrolases, and 23 expansins reflects the importance of wall remodelling during growth of plant cells. Excluding ascorbate and glutathione peroxidases, there are 69 genes with significant similarity to known peroxidases and 15 laccases (diphenol oxidases). Their presence in such abundance indicates the importance of oxidative processes in the synthesis of lignin, suberin and other cell-wall polymers. The high degree of apparent redundancy in the genes for cell-wall metabolism might reflect differences in substrate specificity by some of the enzymes.

The high degree of apparent redundancy in the genes for cell wall metabolism might reflect differences in substrate specificity by some of the enzymes. It is already known that cell types have different wall compositions, which may require that the relevant enzymes be subject to cell-type-specific transcriptional regulation. Of the 40 or so cell types that plants make, almost all can be identified by unique features of their cell wall[107]. A large number of genes involved in wall metabolism have yet to be defined. Although more than 60 genes for glycosyltransferases can be found in the genome sequence, most of these are probably involved in protein glycosylation or metabolite catabolism and do not seem to be adequate to account for the polysaccharide complexity of the wall. For instance, at least 21 enzymes are required just to produce the linkages of the pectic polysaccharide RGII, and none of these enzymes has been identified at present. Thus, if these and related enzymes involved in the synthesis of other cell-wall polymers are also represented by multiple genes, a substantial number of the genes of currently unknown function may be involved in cell-wall metabolism.

Higher plants collectively synthesize more than 100,000 secondary metabolites. Because flowering plants are thought to have similar numbers of genes, it is apparent that a great deal of enzyme creation took place during the evolution of higher plants. An important factor in the rapid evolution of metabolic complexity is the large family of cytochrome P450s that are evident in *Arabidopsis* (Supplementary Information Table 1). These enzymes represent a superfamily of haem-containing proteins, most of which catalyse NADPH- and O_2-dependent hydroxylation reactions. Plant P450s participate in myriad biochemical pathways including those devoted to the synthesis of plant products, such as phenylpropanoids, alkaloids, terpenoids, lipids, cyanogenic glycosides and glucosinolates, and plant growth regulators, such as gibberellins, jasmonic acid and brassinosteroids. Whereas *Arabidopsis* has ~286 P450 genes, *Drosophila* has 94, *C. elegans* has 73 and yeast has only 3. This low number in yeast indicates that there are few reactions of basic metabolism that are catalysed by P450s. It seems likely that many animal P450s are involved in detoxification of compounds from food plant sources. The role of endogenous enzymes is poorly understood; only a few dozen P450 enzymes from plants have been characterized to any extent. The discrepancy between the number of known P450-catalysed reactions and the number of genes suggests that *Arabidopsis* produces a relatively large number of metabolites that have yet to be identified.

In addition to the large number of cytochrome P450s, *Arabidopsis* has many other genes that suggest the existence of pathways or processes that are not currently known. For instance, the presence of 19 genes with similarity to anthranilate *N*-hydroxycinnamoyl/benzoyl transferase is currently inexplicable. This enzyme is involved in the synthesis of dianthramide phytoalexins in Caryophyllaceae and Gramineae. No phytoalexins of this class have been described in *Arabidopsis* as yet. Similarly, the presence of 12 genes with sequence similarity to the berberine bridge enzyme, ((*S*)-reticuline:oxygen oxidoreductase (methylene-bridge-forming); EC 1.5.3.9), and 13 genes with similarity to tropinone reductase, suggests that *Arabidopsis* may have the ability to produce alkaloids. In other plants, the berberine bridge enzyme transforms reticuline into scoulerine, a biosynthetic precursor to a multitude of species-specific protopine, protoberberine and benzophenanthridine alkaloids. The discovery of these and many other intriguing genes in the *Arabidopsis* genome has created a wealth of new opportunities to understand the metabolic and structural diversity of higher plants.

Concluding remarks

The twentieth century began with the rediscovery of Mendel's rules of inheritance in pea[108], and it ends with the elucidation of the complete genetic complement of a model plant, *Arabidopsis*. The analysis of the completed sequence of a flowering plant reported here provides insights into the genetic basis of the similarities and differences of diverse multicellular organisms. It also creates the potential for direct and efficient access to a much deeper understanding of plant development and environmental responses, and permits the structure and dynamics of plant genomes to be assessed and understood.

Arabidopsis, *C. elegans* and *Drosophila* have a similar range of 11,000–15,000 different types of proteins, suggesting this is the minimal complexity required by extremely diverse multicellular eukaryotes to execute development and respond to their environment. We account for the larger number of gene copies in *Arabidopsis* compared with these other sequenced eukaryotes with two possible explanations. First, independent amplification of individual genes has generated tandem and dispersed gene families to a greater extent in *Arabidopsis*, and unequal crossing over may be the predominant mechanism involved. Second, ancestral duplication of the entire genome and subsequent rearrangements have resulted in segmental duplications. The pattern of these duplications suggests an ancient polyploidy event, and mutant analysis indicates that at least some of the many duplicate genes are functionally redundant. Their occurrence in a functionally diploid genetic model came as a surprise, and is reminiscent of the situation in maize, an ancient segmental allotetraploid. The remarkable degree of genome plasticity revealed in the large-scale duplications may be needed to provide new functions, as alternative promoters and alternative splicing appear to be less widely used in plants than they are in animals. Apart from duplicated segments, the overall chromosome structure of *Arabidopsis* closely resembles that of *Drosophila*; transposons and other repetitive sequences are concentrated in the heterochromatic regions surrounding the centromere.

whereas the euchromatic arms are largely devoid of repetitive sequences. Conversely, most protein-coding genes reside in the euchromatin, although a number of expressed genes have been identified in centromeric regions. Finally, *Arabidopsis* is the first methylated eukaryotic genome to be sequenced, and will be invaluable in the study of epigenetic inheritance and gene regulation.

Unlike most animals, plants generally do not move, they can perpetuate indefinitely, they reproduce through an extended haploid phase, and they synthesize all their metabolites. Our comparison of *Arabidopsis*, bacterial, fungal and animal genomes starts to define the genetic basis for these differences between plants and other life forms. Basic intracellular processes, such as translation or vesicle trafficking, appear to be conserved across kingdoms, reflecting a common eukaryotic heritage. More elaborate intercellular processes, including physiology and development, use different sets of components. For example, membrane channels, transporters and signalling components are very different in plants and animals, and the large number of transcription factors unique to plants contrasts with the conservation of many chromatin proteins across the three eukaryotic kingdoms. Unexpected differences between seemingly similar processes include the absence of intracellular regulators of cell division (Cdc25) and apoptosis (Bcl-2). On the other hand, DNA repair appears more highly conserved between plants and mammals than within the animal kingdom, perhaps reflecting common factors such as DNA methylation. Our analysis also shows that many genes of the endosymbiotic ancestor of the plastid have been transferred to the nucleus, and the products of this rich prokaryotic heritage contribute to diverse functions such as photoautotrophic growth and signalling.

The sequence reported here changes the fundamental nature of plant genetic analysis. Forward genetics is greatly simplified as mutations are more conveniently isolated molecularly, but at the same time extensive gene duplications mean that functional redundancy must be taken into account. At a biochemical level, the specificity conferred by nucleotide sequence, and the completeness of the survey allow complex mixtures of RNA and protein to be resolved into their individual components using micro-arrays and mass spectrometry. This specificity can also be used in the parallel analysis of genome-wide polymorphisms and quantitative traits in natural populations[109]. Looking ahead, the challenge of determining the function of the large set of predicted genes, many of which are plant-specific, is now a clear priority, and multinational programs have been initiated to accomplish this goal using site-selected mutagenesis among the the necessary tools[110]. Finally, productive paths of crop improvement, based on enhanced knowledge of *Arabidopsis* gene function, will help meet the challenge of sustaining our food supply in the coming years.

Note added in proof: at the time of publication 17 centromeric BACs and 5 sequence gaps in chromosome arms are being sequenced. □

Methods

The three centres used similar annotation approaches involving in silico gene-finding methods, comparison to EST and protein databases, and manual reconciliation of that data. Gene finding involved three steps: (1) analysis of BAC sequences using a computational gene finder; (2) alignment of the sequence to the protein and EST databases; (3) assignment of functions to each of the genes. Genscan[111], GeneMark.HMM[112], Xgrail[113] Genefinder (P. Green, unpublished software) and GlimmerA[114] were used to analyse BAC sequences. All of these systems were specially trained for *Arabidopsis* genes. Splice sites were predicted using NetGene2[115], Splice Predictor[116] and GeneSplicer (M. Pertea and S. Salzberg, unpublished software). For the second step, BACs were aligned to ESTs and to the *Arabidopsis* gene index[117] using programs such as DDS/GAP2[118] or BLASTN[119]. Segmental duplications were analysed and displayed using a modified version of DIALIGN2 (ref. 120).

Received 20 October; accepted 15 November 2000.

1. The *C. elegans* Sequencing Consortium. Sequence and analysis of the genome of *C. elegans*. *Science* **282,** 2012–2018 (1998).
2. Adams, M. D. The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
3. Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. D. & Koornneef, M. *Arabidopsis thaliana*: a model plant for genome analysis. *Science* **282,** 662–665 (1998).
4. Lin, X. *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402,** 761–768 (1999).
5. Mayer, K. *et al.* Sequence and analysis of chromsome 4 of the plant *Arabidopsis thaliana*. *Nature* **402,** 769–777 (1999).
6. Theologis, A. *et al.* Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature* **408,** 816–820 (2000).
7. Salanoubat, M. *et al.* Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature* **408,** 820–822 (2000).
8. Tabata, S. *et al.* Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature* **408,** 820–822 (2000).
9. Choi, S. D., Creelman, R., Mullet, J. & Wing, R. A. Construction and characterisation of a bacterial artificial chromosome library from *Arabidopsis thaliana*. *Weeds World* **2,** 17–20 (1995).
10. Mozo, T., Fischer, S., Shizuya, H. & Altmann, T. Construction and characterization of the IGF *Arabidopsis* BAC library. *Mol. Gen. Genet.* **258,** 562–570 (1998).
11. Lui, Y. -G., Mitsukawa, N., Vazquez-Tello, A. & Whittier, R. F. Generation of a high-quality P1 library of *Arabidopsis* suitable for chromosome walking. *Plant J.* **7,** 351–358 (1995).
12. Lui, Y. -G. *et al.* Complementation of plant mutants with large genomic DNA fragments by a transformation-competent artificial chromosome vector accelerates positional cloning. *Proc. Natl Acad. Sci. USA* **96,** 6535–6540 (1999).
13. Marra, M. *et al.* A map or sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genet.* **22,** 265–270 (1999).
14. Mozo, T. *et al.* A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nature Genet.* **22,** 271–275 (1999).
15. Sato, S. *et al.* Structural analysis of *Arabidopsis thaliana* chromosome 5. I. Sequence features of the 1. 6 Mb regions covered by twenty physically assigned P1 clones. *DNA Res.* **4,** 215–230 (1997).
16. Bent, E., Johnson, S. & Bancroft, I. BAC representation of two low-copy regions of the genome of *Arabidopsis thaliana*. *Plant J.* **13,** 849–855 (1998).
17. Copenhaver, G. P. *et al.* Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* **286,** 2468–2474 (1999).
18. Meyerowitz, E. M. & Somerville, C. R. *Arabidopsis* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1994)
19. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
20. Pavy, N. *et al.* Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *BioInformatics* **15,** 887–900 (1999).
21. Mewes, H. W. *et al.* Overview of the yeast genome. *Nature* **387** (Suppl.) 7–65 (1997).
22. Frishman, D. *et al.* Functional and structural genomics using PEDANT. *BioInformatics* (in the press).
23. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1462 (1997).
24. Kotani, H. & Tabata, S. Lessons from the sequencing of the genome of a unicellular cyanobacterium, *Synechocystis* SP. PCC6803. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49,** 151–171 (1998).
25. Apweiler, R. *et al. INTERPRO* (http://www. ebi. ac. uk/interpro/). Collaborative Computer Project 11 Newsletter no. 10 (Cambridge, 2000).
26. Bent, A. F. *et al. RPS2* of *Arabidopsis thaliana* a leucine-rich repeat class of plant disease resistance genes. *Science* **265,** 1856–1860 (1994).
27. Skowyra, D. *et al.* F box proteins are receptors that recruit phsphorylated substrates to the SCF ubiquitin-ligase complex. *Cell* **91,** 209–219 (1997).
28. Joazeiro, C. A. P. & Weissman, A. M. RING finger proteins: mediators of ubiquitin ligase activity. *Cell* **102,** 549–552 (2000).
29. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).
30. Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* **27,** 2369–2376 (1999).
31. Blanc, G. *et al.* Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* **12,** 1093–1102 (2000).
32. Wendel, J. F. Genome evolution in polyploids. *Plant Mol. Biol.* **42,** 225–249 (2000).
33. Gaut, B. S. & Doebley, J. F. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl Acad. Sci. USA* **94,** 6809–6814 (1997).
34. Ku, H. -M., Vision, T., Liu, J. & Tanksley, S. D. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl Acad. Sci. USA* **97,** 9121–9126 (2000).
35. Noel, L. *et al.* Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of *Arabidopsis*. *Plant Cell* **11,** 2099–2111 (1999).
36. Ellis, J., Dodds, P. & Pryor, T. Structure, function, and evolution of plant disease resistance genes. *Trends Plant Sci.* **3,** 278–284 (2000).
37. Tanksley, S. D. *et al.* High density molecular linkage maps of the tomato and potato genomes. *Genetics* **132,** 1141–1160 (1992).
38. Moore, G., Devos, K. M., Wang, Z. & Gale, M. D. Grasses, line up and form a circle. *Curr. Biol.* **5,** 737–739 (1995).
39. Acarkan, A., Rossberg, M., Koch, M. & Schmidt, R. Comparative genome analysis reveals extensive conservation of genome organisation for *Arabidopsis thaliana* and *Capsella rubella*. *Plant J.* **23,** 55–62 (2000).
40. Cavell, A., Lydiate, D., Parkin, I., Dean, C. & Trick, M. A 30 centimorgan segment of *Arabidopsis thaliana* chromosome 4 has six collinear homologues within the *Brassica napus* genome. *Genome* **41,** 62–69 (1998).
41. O'Neill, C. & Bancroft, I. Comparative physical mapping of segments of the genome of *Brassica oleracea* var *alboglabra* that are homologous to sequenced regions of the chromosomes 4 and 5 of *Arabidopsis thaliana*. *Plant J.* **23,** 233–243 (2000).
42. Wolfe, K. H., Gouy, M., Yang, Y. -W., Sharp, P. M. & Li, W. -H. Date of the monocot-dicot divergence estimated from the chloroplast DNA sequence data. *Proc. Natl Acad. Sci. USA* **86,** 6201–6205 (1989).
43. van Dodeweerd, A. -M. *et al.* Identification and analysis of homologous segments of the genomes of rice and *Arabidopsis thaliana*. *Genome* **42,** 887–892 (1999)
44. Mayer, K. Sequence level analysis of homologous segments of the genomes of rice and *Arabidopsis thaliana*. *Genome Res.* (submitted).
45. Sato, S. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Research* **6,** 283–290 (1999).
46. Unseld, M., Marienfeld, J., Brandt, P. & Brennicke, A. The mitochondrial genome in *Arabidopsis*

thaliana contains 57 genes in 366,924 nucleotides. *Nature Genet.* **15,** 57–61 (1997).
47. Palmer, J. D. *et al.* Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. *Proc. Natl Acad. Sci. USA* **97,** 6960–6966 (2000).
48. Le, Q.-H. *et al.* Transposon diversity in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **97,** 7376–7381 (2000).
49. Yu, Z., Wright, S. & Bureau, T. *Mutator*-like elements (MULEs) in *Arabidopsis thaliana*: Structure, diversity and evolution. *Genetics* (in the press).
50. Feschotte, C. & Mouches, C. Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Mol. Biol. Evol.* **17,** 730–737 (2000).
51. Martienssen, R. Transposons, DNA methylation and gene control. *Trends Genet.* **14,** 263–264 (1998).
52. Singer, T., Yordan, C. & Martienssen, R. Robertson's *Mutator* transposons in *Arabidopsis* are regulated by the chromatin-remodeling gene *Decrease in DNA Methylation (DDM1)*. *Genes Dev.* (in the press).
53. SanMiguel, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274,** 765–768 (1996).
54. Copenhaver, G. P. & Pikaard, C. S. Two-dimensional RFLP analyses reveal megabase-sized clusters of rRNA gene variants in *Arabidopsis thaliana*, suggesting local spreading of variants as the mode for gene homogenization during concerted evolution. *Plant J.* **9,** 273–282 (1996).
55. Fransz, P. *et al.* Cytogenetics for the model system *Arabidopsis thaliana*. *Plant J.* **13,** 867–876 (1998).
56. Richards, E. J. & Ausubel, F. M. Isolation of a higher eukarotic telomere from *Arabidopsis thaliana*. *Cell* **53,** 127–136 (1988).
57. Round, E. K., Flowers, S. K. & Richards, E. J. *Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure. *Genome Res.* **7,** 1045–1053 (1997).
58. The CSHL/WUGSC/PEB *Arabidopsis* Sequencing Consortium. The complete sequence of a heterochromatic island from a higher eukaryote. *Cell* **100,** 377–386 (2000).
59. Fransz, P. F. *et al.* Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: Structural organization of heterochromatic knob and centromere region. *Cell* **100,** 367–376 (2000).
60. Paulsen, I. T., Nguyen, L., Sliwinski, M. K., Rabus, R. & Saier, M. H. Jr Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.* **301,** 75–101 (2000).
61. Paulsen, I. T., Sliwinski, M. K., Nelissen, B., Goffeau, A. & Saier, M. H. Jr Unified inventory of established and putative transporters encoded within the complete genome of *Saccharomyces cerevisiae*. *FEBS Lett.* **430,** 116–125 (1998).
62. Hirsch, R. E., Lewis, B. D, Spalding, E. P. & Sussman, M. R. A role for the AKT1 potassium channel in plant nutrition. *Science* **280,** 918–921 (1998).
63. Slayman, C. L. & Slayman, C. W. Depolarization of the plasma membrane of Neurospora during active transport of glucose: evidence for a proton-dependent cotransport system. *Proc. Natl Acad. Sci. USA* **71,** 1035–1939 (1974).
64. Ryan, C. A. & Pearce, G. Systemin: a polypeptide signal for plant defensive genes. *Annu. Rev. Cell. Dev. Biol.* **14,** 1–17 (1998).
65. Eisen, J. A. & Hanawalt, P. C. A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat. Res.* **435,** 171–213 (1999).
66. Selby, C. P. & Sancar, A. Structure and function of transcription-repair coupling factor. Structural domains and binding properties. *J. Biol. Chem.* **270,** 4882–4889 (1995).
67. Britt, A. B. Molecular genetics of DNA repair in higher plants. *Trends Plant Sci.* **4,** 20–25 (1999).
68. Dangl, M. Response to Aravind, L. & Koonin, E. V. Second Family of Histone Deacetylases. *Science* **280,** 1167 (1998).
69. Cao, X. *et al.* Conserved plant genes with similarity to mammalian de novo DNA methyltransferases. *Proc. Natl Acad. Sci. USA* **97,** 4979–4984 (2000).
70. Henikoff, S. & Comai, L. A DNA methyltransferase homologue with a chromodomain exists in multiple polymorphic forms in *Arabidopsis*. *Genetics* **149,** 307–318 (1998).
71. Hung, M. -S. *et al. Drosophila* proteins related to vertebrate DNA (5-cytosine) methyltransferases. *Proc Natl Acad. Sci. USA* **96,** 11940–11945 (1999).
72. Dalmay, T., Hamilton, A. J., Rudd, S., Angell, S. & Baulcombe, D. C. An RNA-dependent-RNA polymerase in *Arabidopsis* is required for post transcriptional gene silencing mediated by a transgene but not by a virus—the truth. *Cell* **101,** 543–553 (2000).
73. Riechmann, J. L. & Ratcliffe, O. J. A genomic perspective on plant transcription factors. *Curr. Opin. Plant Biol.* **3,** 423–434 (2000).
74. Liljegren, S. J. *et al.* SHATTERPROOF MADS-box genes control seed dispersal in *Arabidopsis*. *Nature* **404,** 766–770 (2000).
75. Pelaz, S. *et al.* B and C floral organ identity functions require SEPALLATA MADS-box genes. *Nature* **405,** 200–203 (2000).
76. Canaday, J., Stoppin-Mellet, V., Mutterer, J., Lambert, A. M. & Schmit, A. C. Higher plant cells: gamma-tubulin and microtubule nucleation in the absence of centrosomes. *Microsc. Res. Technol.* **49,** 487–495 (2000).
77. Bassham, D. C. & Raikhel, N. V. Unique features of the plant vacuolar sorting machinery. *Curr. Opin. Cell Biol.* **12,** 491–495 (2000).
78. Zheng, Z. L. & Yang, Z. The Rrop GTPase switch turns on polar growth in pollen. *Trends Plant Sci.* **5,** 298-303 (2000).
79. den Boer, B. G. & Murray, J. A. Triggering the cell cycle in plants. *Trends Cell Biol.* **10,** 245–250 (2000).
80. Heese, M., Mayer, U. & Jurgens, G. Cytokinesis in flowering plants: cellular process and developmental integration. *Curr. Opin. Plant Biol.* **1,** 486–491 (1998).
81. Meyerowitz, E. M. Plants, animals, and the logic of development. *Trends Genet.* **15,** M65–M68 (1999).
82. Wang, D. Y. C. *et al.* Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B Bio.* **266,** 63–171 (1999).
83. Torii, K. Receptor kinase activation and signal transduction in plants: an emerging picture. *Curr. Opin. Plant Biol.* **3,** 362–367 (2000).
84. Yeh, K. C. & Lagarias, J. C. Eukaryotic phytochromes: Light-regulated serine/threonine protein kinases with histidine kinase ancestry. *Proc. Natl Acad. Sci. USA* **95,** 13976–13981 (1998).
85. McCarty, D. R. & Chory, J. Conservation and innovation in plant signaling pathways. *Cell* **103,** 201–211 (2000).
86. Weiss, C. A., Garnaat, C., Mukai, K., Hu, Y. & Ma, H. Molecular cloning of cDNAs from maize and *Arabidopsis* encoding a G protein beta subunit. *Proc. Natl Acad. Sci. USA* **91,** 9554–9558 (1994).
87. Bowler, C. *et al.* Cyclic GMP and calcium mediate phytochrome phototransduction. *Cell* **77,** 73–81 (1994).
88. Stepanova, A. & Ecker, J. R. Ethylene signaling: from mutants to molecules. *Curr. Opin. Plant Biol.* **3,** 353–360 (2000).
89. Urao, T., Yamaguchi-Shinozaki, K. & Shinozaki, K. Two-component systems in plant signal transduction. *Trends Plant Sci.* **5,** 67–74 (2000).
90. Makino, S. *et al.* Genes encoding pseudo-response regulators: Insight into His-to-Asp phosphorelay and circadian rhythm in *Arabidopsis thaliana*. *Plant Cell Physiol.* **41,** 791–803 (2000).
91. D'Agostino, I. B. & Kieber, J. J. Phosphorelay signal transduction: the emerging family of plant response regulators. *Trends Biol. Sci.* **24,** 452–456 (1999).
92. Strayer, C. *et al.* Cloning of the *Arabidopsis* clock gene TOC1, an autoregulatory response regulator homologue. *Science* **289,** 768–771 (2000).
93. Stahl, E. A. & Bishop, J. G. Plant-Pathogen arms races at the molecular level. *Curr. Opin. Plant Biol.* **3,** 299–304 (2000).
94. McDowell, J. M. & Dangl, J. L. Signal transduction in the plant innate immune response. *Trends Biochem. Sci.* **25,** 79–82 (2000).
95. Van der Biezen, E. A. & Jones, J. D. Plant disease-resistance proteins and the gene-for-gene concept. *Trends Biochem Sci.* **23,** 454–456 (1998).
96. Belvin, M. P. & Anderson, K. V. A conserved signaling pathway: the *Drosophila* toll-dorsal pathway. *Annu. Rev. Cell. Dev. Biol.* **12,** 393–416 (1996).
97. Uren, A. G. *et al.* Identification of paracaspases and metacaspases: Two ancient families of caspase-like proteins, one of which plays a key role in MALT lymphoma. *Mol. Cell* **6,** 961–967 (2000).
98. Fankhauser, C. & Chory, J. Light control of plant development. *Annu. Rev. Cell. Dev. Biol.* **13,** 203–229 (1997).
99. Briggs, W. R. & Huala, E. Blue-light photoreceptors in higher plants. *Annu. Rev. Cell. Dev. Biol.* **15,** 33–62 (1999).
100. Christie, J. M., Salomon, M., Nozue, K., Wada, M. & Briggs, W. R. LOV (light, oxygen, or voltage) domains of the blue-light photoreceptor phototropin (nph1): binding sites for the chromophore flavin mononucleotide. *Proc. Natl Acad. Sci. USA* **96,** 8779–8783 (1999).
101. Golbeck, J. H. Structure and function of photosystem I. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **43,** 293–324 (1992).
102. Maier, R. M., Neckermann, K., Igloi, G. L. & Kossel, H. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *J. Mol. Biol.* **251,** 614–28 (1995).
103. Buchanan, B. B., Gruissem, W. & Jones, R. L. in *Biochemistry and Molecular Biology of Plants* 1367 (Am. Soc. Plant Physiol., Rockville, Maryland, 2000).
104. Mekhedov, S., Martínez de Ilárduya, O. & Ohlrogge, J. Toward a functional catalog of the plant genome. A survey of genes for lipid biosynthesis. *Plant Physiol.* **122,** 389–401 (2000).
105. Somerville, C. R., & Ogren, W. L. Photorespiration deficient mutants of *Arabidopsis thaliana* lacking mitochrondrial serine transhydroxymethylase activity. *Plant Physiol.* **67,** 666–671 (1981).
106. Richmond, T., & Somerville, C. R. The cellulose synthase superfamily. *Plant Physiol* **124,** 495–499 (1999).
107. Carpita, N. Vergara C: A recipe for cellulose. *Science* **279,** 672–673 (1998).
108. De Vries, H. Sur la loi de disjonction des hybrides. *C. R. Acad. Sci. Paris* **130,** 845–847 (1900).
109. Alonso-Blanco, C. & Koornneef, M. Naturally occurring variation in *Arabidopsis*: an underexploited resource for plant genetics. *Trends Plant Sci.* **5,** 1360–1385 (1999).
110. Chory, J. Functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. *Plant Physiology* **123,** 423–425 (2000).
111. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
112. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26,** 1107–1115 (1998).
113. Uberbacher, E. C. & Mural, R. J. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl Acad. Sci. USA* **88,** 11261–11265 (1991).
114. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59,** 24–31 (1999).
115. Hebsgaard, S. M. *et al.* Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Res.* **24,** 3439–3452 (1996).
116. Brendel, V. & Kleffe, J. Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. *Nucleic Acids Res.* **26,** 4748–4757 (1998).
117. Quackenbush, J., Liang, F., Holt, I., Pertea, G. & Upton, J. The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* **28,** 141–145 (2000).
118. Huang, X., Adams, M. D., Zhou, H. & Kerlavage, A. R. A tool for analyzing and annotating genomic sequences. *Genomics* **46,** 37–45 (1997).
119. Altschul, S. F. *et al.* Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
120. Morgenstern, B. DIALIGN2: improvement of the segment-to-segment approach to multiple sequence alignment. *BioInformatics* **15,** 211–218 (1999).
121. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540 (1995).
122. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300,** 1005–1016 (2000).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

This work was supported by the National Science Foundation (NSF) Cooperative Agreements (funded by the NSF, the US Department of Agriculture (USDA) and the US Department of Energy (DOE)), the Kazusa DNA Research Institute Foundation, and by the European Commission. Additional support from the USDA, Ministère de la Recherche, GSF-Forschungszentrum f. Umwelt u. Gesundheit, BMBF (Bundesministerium f. Bildung, Forschung und Technologie), the BBSRC (Biotechnology and Biological